

**RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH**

**INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS**

**PROCEEDINGS
OF THE SECOND
INTERNATIONAL CONFERENCE
ON BIOINFORMATICS
OF GENOME REGULATION
AND STRUCTURE**

Volume 2

**BGRS'2000
Novosibirsk, Russia
August 7-11, 2000**

ICG, Novosibirsk, 2000

International Program Committee

Nikolay Kolchanov, Institute of Cytology and Genetics, Novosibirsk, Russia (Chairman of the Conference)
G. Christian Overton, Center for Bioinformatics, University of Pennsylvania, USA (Co-Chairman of the Conference)
 Ralf Hofestadt, University Magdeburg, Germany (Co-Chairman of the Conference)
 Patrizio Arrigo, Institute of Electronic Circuits, CNR, Italy
 Martin Bishop, Human Genome Mapping Project Resource Centre, UK
 Philip Bourne, SDSC, San-Diego, USA
 Philipp Bucher, Swiss Institute for Experimental Cancer Research, Switzerland
 Chris Burge, MIT Center for Cancer Research, Cambridge, MA, USA
 Julio Collado-Vides, National University of Mexico, Mexico
 Jim Fickett, SmithKline Beecham Pharmaceuticals, USA
 Mikhail Gelfand, Institute of Protein Research, RAS, Moscow, Russia
 Charlie Hodgman, GlaxoWellcome Research Medicine Center, UK
 Minoru Kanehisa, Kyoto University, Kyoto, Japan
 Kotoko Nakata, National Institute of Health Sciences, Tokyo, Japan
 Leonid Kalinichenko, Institute of Problems of Informatics RAN, Moscow, Russia
 Lev Kisselev, Engelhardt Institute of Molecular Biology, Moscow, Russia
 Luhua Lai, Institute of Physical Chemistry, Peking University, Beijing, China
 Hwa A. Lim, D'Trends, Inc, USA
 Gerhard Michal, Tutzingen, Germany
 George Michaels, Genomics Lead Development, Monsanto Co.USA
 Luciano Milanese, ITBA, Milan, Italy
 Andrey Mironov, State Center for Applied Genetics, Moscow
 Ken Nishikawa, Center for Information Biology, National Institute of Genetics, Japan
 Manuel Peitsch, Glaxo Wellcome Experimental Research SA, Geneva, Switzerland
 Mikhail Ponomarenko, Institute of Cytology and Genetics, Novosibirsk
 Vadim Ratner, Institute of Cytology and Genetics, Novosibirsk, Russia
 John Reinitz, Mt. Sinai Med. School, USA
 Aida Romashchenko, Institute of Cytology and Genetics, Novosibirsk, Russia
 Akinori Sarai, RIKEN Tsukuba Life Science Center, Tsukuba, Japan
 Victor Solovyev, The Sanger Centre, Cambridge, UK
 Masaru Tomita, Bioinformatics Laboratory of Keio University, Japan
 Eduard Trifonov, Weizmann Institute of Science, Rehovot, Israel
 Vladimir Tumanian, Engelhardt Institute of Molecular Biology, Moscow, Russia
 Edgar Wingender, GBF, Braunschweig, Germany
 Michael Zhang, Cold Spring Harbor Laboratory, Cold Spring Harbor, USA

Local Organizing Committee

Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk
 Dmitry Afonnikov, Institute of Cytology and Genetics, Novosibirsk
 Vasily Areschenko, Siberian Branch of RAS, Novosibirsk
 Elena Borovskikh, Institute of Cytology and Genetics, Novosibirsk
 Dmitry Grigorovich, Institute of Cytology and Genetics, Novosibirsk
 Nadya Omelanchuk, Institute of Cytology and Genetics, Novosibirsk
 Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk
 Anatoly Kushnir, Institute of Cytology and Genetics, Novosibirsk
 Anatoly Kurbatov, Institute of Archeology and Ethnography, Novosibirsk
 Sergey Lavryushev, Institute of Cytology and Genetics, Novosibirsk
 Yuri Orlov, Institute of Cytology and Genetics, Novosibirsk
 Galina Orlova, Institute of Cytology and Genetics, Novosibirsk

INTRODUCTION

Two volumes of Proceedings of the International Conference BGRS-2000 encouning about 180 abstracts are aimed to direct an attention to the actual problems in bioinformatics of genome regulation and structure. The Conference BGRS-2000 organized by the Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences will be held in Novosibirsk, Russia, in August 7-11, 2000. This Conference will be the second in the series: the First International Conference on Bioinformatics of Genome Regulation and Structure – BGRS-98 was held in Novosibirsk in August 1998.

The question may arise: Why the Conferences BGRS attract their attention directly to the problems dealing with genome regulation and structure? The answer could be as follows: the structure and regulation of genome are the counterparts of life at molecular level; that is why understanding of fundamental principles of regulatory genomic machinery is impossible unless their structural organization is known, and *vice versa*.

During two years that have passed from the first BGRS Conference, the experimental genome study including applications to direct sequencing and mapping became of ever-growing scale. The huge bulk of experimental data on nucleotide sequences of complete bacterial genomes, the sequencing of which became a routine procedure in molecular biology, are being accumulated. Besides, complete genome of *Drosophila* is being deciphered, and human genome sequencing is drawing towards completion.

The ever-growing impact in genome studying is produced by novel experimental techniques. In particular, the EST technique is widely used in studying gene structure and gene expression patterns. Besides, microarray methods aimed at extracting unique and complete information on genome functioning and enabling to study simultaneously the expression patterns of dozen thousands of genes including those obtained at a single cell level, become implemented massively. In addition, single nucleotide polymorphism (SNP) technique provides a huge bulk of experimental data for studying regularities in mutation-assisted genome variability. Large-scale proteomic initiatives in the near future will lead to accumulation of large massifs of information on structure-functional organization of proteins.

The huge volume of experimental data that has been acquired on genome structure, functioning and gene expression regulation demonstrate the blistering growth. Development of informational-computational technologies of novel generation is a challenging problem of bioinformatics. Bioinformatics has entered that very phase of development, when decisions of the challenging problems determine the realization of large-scale experimental research projects directed to studying genome structure, function, and evolution.

By analyzing the papers submitted for publication in the two-volume issues of the BGRS-2000, the Organizing Committee came to a conclusion that participants of the Conference have concentrated their attention at consideration of the hottest items in bioinformatics listed below:

(1) Development of the novel generation of databases providing more complex, deep, and comprehensive description of (i) genome structure, function, and evolution, (ii) regulatory genome sequences, (iii) regulatory proteins, (iv) genetic networks, (v) signal transduction pathways and genetically controlled metabolic pathways.

(2) Development of computer technologies for automated knowledge discovery and data mining in the databases: ultra-rapid experimental methods developed for extracting molecular-biological data should correspond to similarly advanced technologies designed for automated treatment of these data, these technologies enabling to get the maximum of reliable and significant knowledge about genome function, regulation, and structure out of computer databases.

(3) Development of rigorous scientific methods for analysis of gene structure, discovery and modeling. Along with traditional approaches based on recognition of potential regulatory sites and coding regions, and their combinations, the more resolving power in solving this problem is demonstrated by the approaches based on comparative genomics, including both data search throughout databases and comparison of extended genome regions and even complete genomes (in case of bacteria).

(4) Development and improvement of methods in comparative genomics that became one of the most high-powered and perspective directions in modern bioinformatics. The efficient algorithms developed within the frames of comparative genomics appear to be more reliable tool acquired for gene recognition and gene reconstruction throughout *de novo* sequenced genome DNA, for recognition of regulatory elements controlling genome functions and gene expression regulation. Besides, comparative genomics will go a long way towards revealing fundamental principles of genome organization and regularities in genome molecular evolution.

(5) Further mastering of approaches designed within the frames of comparative genomics strongly depends upon comprehension of fundamental regularities in genome organization and evolution. That is why computer analysis and modeling of genome mutability, together with studying of fundamental laws of evolution of genomes, coding gene regions and regulatory genomic sequences become the matter of especial importance. Accumulation of knowledge in this field will certainly help in searching for objective methods in annotating and finding of genes and regulatory signals in genomic sequences.

(6) Development of novel generation of mathematical algorithms implemented for analysis of regulatory genome sequences (RGS) and for accounting of real complexity of RGS. These algorithms are characterized by a large variety of parameters significant for gene functioning, by blockwise structure, and hierarchy in RGS organization. On the background of these algorithms, the fine accuracy methods are being developed for

recognition and prediction of quantitative values of regulatory genomic sequences activity of various types, which provide implementation of numerous genome functions regulating basic stages of gene expression.

(7) Revealing of fundamental regularities in structure-functional organization of RGS controlling basic types of molecular-genetical processes (i.e., replication, transcription, splicing, polyadenylation/processing, translation, etc). Besides revealing the regularities in structure-functional organization of RGS that are valuable for increasing the accuracy of their recognition, this analysis allows to obtain a fundamental knowledge on molecular mechanisms of RGS functioning, thus enabling to solve one of the main problems in bioinformatics of genome regulation and structure.

(8) Development of methods aimed at prediction and recognition of structure-functional organization of proteins encoded by the genes detected within *de novo* sequenced genome sequences. The lack of unified technological production line processing from the coding gene regions in the sequenced genomes to prediction of structure-functional organization of proteins encoded by these genes serves as the stopping brakes for implementation of large-scale genome projects. It should be stressed that during the solving of the task, a large attention should be paid to detecting fundamental principles of protein organization and evolution. The most important is the studying of aspects of protein function and structure related to genome regulation. During the recent years, the tendency manifested itself in convergence and intersection of the lines in bioinformatics of genome regulation and structure and in computer-assisted proteomics. This observation is clearly approved in Proceedings of BGRS-2000.

(9) Large-scale genome analysis. The other day computer analysis was restricted to studying of local context regularities in genome structure. Currently, due to widespread sequencing of complete genomes and their extra-extended fragments, a possibility first appeared to analyze large-scale context dependencies in genome DNA organization. To tackle this problem, it is necessary to develop operative methods aimed at analysis of extra-extended genome sequences.

(10) Description in databases and modeling of genetical networks, which control the processes of basic metabolism, cell division and differentiation, organ- and tissue morphogenesis, growth and development of an organism; support of homeostasis of molecular, biochemical, and physiological parameters of organisms, etc. Systemic investigation of mechanisms related to genome functioning and gene expression regulation at the level of gene networks and signal transduction pathways should be provided. On the grounds of these very processes, the key problem in bioinformatics, that is, recognition of phenotypical characteristics of an organism on the basis of information encoded in their genomes will be solved in future.

(11) Development of efficient technologies for integration of informational and software resources on the structure and regulation of genomes and designing on this basis of super-large computer systems implemented for analysis and modeling of intricate molecular-genetic systems and processes.

(12) Analysis of fundamental regularities in (i) genome functioning, organization, and evolution, (ii) the mechanisms governing the coding of genetical information, (iii) molecular bases of realization of genetical language, principles of organization, functioning, and evolution of genetical networks and molecular-genetic systems.

All the questions listed above will be suggested to consideration of participants of BGRS'2000 at 7 sections and presented in a form of plenary lectures, oral communications, posters, Internet computer demonstrations and round table discussions.

BGRS'2000 will bring together the experts in Bioinformatics to discuss the progress in the field of bioinformatics of genome regulation and structure achieved at the end of 20th century, the basic approaches devoted (i) to data description and analysis; (ii) modeling of complex molecular-genetical systems; (iii) to revealing of fundamental principles of genome organization and evolution and of mechanisms of genetical information coding; (iv) to evaluation and marking off the future trends in this field.

The researchers working in the fields of experimental biology and interested in application of Bioinformatics methods in their work are also the participants of the Conference. With this respect, the Conference is expected to be a stimulating event not only giving a future development of bioinformatics as it is, but also establishing new links between Bioinformatics and experimental research.

By working out the BGRS2000 schedule, the Organizing Committee has tried to keep the balance between technical (applied) and fundamental aspects in bioinformatics. This principle has a clear reflection in contents of Proceedings of the Conference. Herein, we have tried to follow the well-known and far-back principle: «nothing is more practical than the good theory».

Professor Nikolay Kolchanov,
Co-Chairman of the Conference,
Head of Laboratory of Theoretical Genetics,
Vice-Director of the Institute of Cytology and Genetics,
Novosibirsk, Russia



G. Christian Overton

15.02.1948 – 1.06.2000

The death of Dr. Overton, premature and unexpected, is a body blow for all Bioinformatics community and for us, his friends and colleagues in Russia. Chris always paid a great attention to strengthening of the international cooperation in the Bioinformatics research. That is why, he left behind along with brilliant scientific results an example of real international collaboration in science. As an example of international cooperation initiated by Chris may serve the collaboration with Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences. From 1995, this laboratory and the Center for Bioinformatics at University of Pennsylvania together made a common research despite of distance between Philadelphia and Novosibirsk, and we are very grateful to Chris for his support, assistance and understanding friendly provided.

Dr. G. Christian Overton was the founding Director of the Center for Bioinformatics at Penn, established in 1997 as an interdisciplinary venture between the Schools of Medicine, Arts and Sciences, and Engineering and Applied Science. He was also an Associate Professor in the Department of Genetics, and held a secondary appointment in the Department of Computer and Information Science in the School of Engineering and Applied Science. Dr. Overton received his Bachelor of Science degree in Mathematics and Physics from the University of New Mexico in 1971, his Ph.D. in Biophysics from the Johns Hopkins University in 1978, and his M.S.E. in Computer and Information Science from the University of Pennsylvania in 1986. After receiving his M.S.E, he returned to the University of Pennsylvania in 1991 as an Associate Professor. In addition to his research, Dr. Overton was an Editor for the Journal of Computational Biology, Bioinformatics, and Gene/Gene-COMBIS as well as the Member of the Board of Directors for the International Society for Computational Biology.

Dr. Overton's brilliant skills in biophysics and bioinformatics, his deep understanding of the challenges in biology, medicine, and computer science enabled him to organize many outstanding research projects, which bridge the gap between experimental biology and computer science aimed to experimental data treatment. Dr. Overton is internationally recognized as a pioneer in genomic research and application of computational approaches for solving biological problems. He focused on problems associated with database integration, genome annotation, gene recognition, and detection of regulatory elements governing the expression of many genes that comprise the human genome.

Chris was one of the Co-Organizers of the BGRS2000 Conference. Due to his activity, the Organizing Committee managed to put together the effort of those interested in the basic approaches and trends in bioinformatics. Chris will be remembered for his love of science, his charm, good nature and his great intelligence. For people here in Novosibirsk who met him and knew him well he will be remembered as a faithful and good friend and as a researcher devoted to science.

Professor Nikolay Kolchanov,
Head of Laboratory of Theoretical Genetics,
Vice-Director of the Institute of Cytology and Genetics,
Novosibirsk, Russia

CONTENTS

CONTENTS	6
SECTION 3. BIOINFORMATICS OF GENOME STRUCTURE AND EVOLUTION	13
AUTOMATED COMPARATIVE ANALYSIS OF REGULATORY PATTERNS: SUGAR METABOLISM AND TRANSPORT SYSTEMS IN GAMMA PURPLE BACTERIA.....	14
MIRONOV A.A., *GELFAND M.S	
PRO-FRAME: SIMILARITY-BASED GENE RECOGNITION IN EUKARYOTIC DNA SEQUENCES WITH ERRORS	16
MIRONOV A.A., *GELFAND M.S	
GENOMEEXPLORER: SOFTWARE FOR ANALYSIS OF COMPLETE BACTERIAL GENOMES	18
MIRONOV A.A., VINOKUROVA N.P., *GELFAND M.S	
COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: CATABOLITE REPRESSION IN GAMMA-PROTEOBACTERIA	20
NOVICHKOVA E.S., NOVICHKOV P.S., MIRONOV A.A., *GELFAND M.S.	
COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: SOS REPAIR IN BACILLUS SUBTILIS.....	24
¹ PERMINA E.A., ² MIRONOV A.A., ^{2*} GELFAND M.S.	
HrcA REGULATES NOT ONLY CHAPERONIN GENES	26
*GELFAND M.S., MIRONOV A.A.	
SOFTWARE FOR ORTHOLOGY ANALYSIS IN COMPLETE BACTERIAL GENOMES	28
BAITALYUK M.V., NOVICHKOV P.S., *GELFAND M.S., MIRONOV A.A.	
FUNCTIONAL GENOMICS: AN ALGORITHMIC PERSPECTIVE	29
CALIFANO A.	
COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: TRANSCRIPTION REGULATORY SITES IN ARCHAEA	34
^{1*} GELFAND M.S., ² KOONIN E.V., ¹ MIRONOV A.A.	
InterPro AS A NEW TOOL FOR WHOLE GENOME ANALYSIS. A COMPARITIVE ANALYSIS OF MYCOBACTERIUM TUBERCULOSIS, BACILLUS SUBTILIS AND ESCHERICHIA COLI AS A CASE STUDY	37
*MULDER N.J., FLEISCHMANN W., APWEILER R.	
NO MYSTERY OF ORFans IN GENOMICS - GENERATION OF ORFans IN THE ANTISENSE OF CODING SEQUENCES	40
MACKIEWICZ P., KOWALCZUK M., GIERLIK A., SZCZEPANIK D., NOWICKA A., DUDEK M.R., *CEBRAT S.	
Pro-Gen: PREDICTION OF THE EXON-INTRON STRUCTURE BY COMPARISON OF GENOMIC SEQUENCES.....	44
¹ NOVICHKOV P.S., ^{1,2*} GELFAND M.S., ^{1,2} MIRONOV A.A.	
COMPARATIVE GENOMICS: HOMOLOGY BASED GENE IDENTIFICATION AND GENE STRUCTURE VALIDATION.....	46
* ¹ WIEHE T, ¹ GEBAUER-JUNG S., ² ABRIL J. AND ² GUIGÒ R.	
MOUSE AQUAPORIN 4 GENE: PREDICTION OF A NEW EXON AND EXPERIMENTAL CONFIRMATION.	48
*BONDAR A.A., ALIKINA T.YU., ZELENIN S.M.	

COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: MULTIDRUG RESISTANCE SYSTEMS IN GAMMA-PROTEOBACTERIA	51
RODIONOV D.A., *GELFAND M.S., MIRONOV A.A. RAKHMANINOVA A.B.	
REGULATION OF DAHP-SYNTASES IN GAMMA-PROTEOBACTERIA: FEEDBACK INHIBITION AND REPRESSION OF TRANSCRIPTION	53
¹ PANINA E.M., ² MIRONOV A.A., ^{2*} GELFAND M.S.	
COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: ATTENUATORS OF AROMATIC AMINO ACID OPERONS OF GAMMA-PROTEOBACTERIA	57
¹ VITRESCHAK A.G., ^{2*} GELFAND M.S.	
AVOIDANCE OF PALINDROMES IN PROCARYOTIC GENOMES AND RESTRICTION-MODIFICATION SYSTEMS	59
¹ PANINA E.M., ^{2*} GELFAND M.S.	
ONE APPROACH FOR ANNOTATION AND CONFIRMATION OF DISCOVERIES OF ALTERNATIVE SPLICE EVENTS USING RECONSTRUCTED EXTENDED UNSPLICED TRANSCRIPTS FROM GENES ON CHROMOSOME 22	61
* ^{1,2} BABENKO V., ¹ VAN HEUSDEN P., ¹ HIDE W.	
FAST SEARCH OF ALL TANDEM REPETITIONS IN NUCLEOTIDE SEQUENCES.....	63
¹ GIRAUD M., ² KOLPAKOV R., ^{3*} KUCHEROV G.	
RECONSTRUCTION OF THE OPEN READING FRAMES BY USING EST MULTIPLE ALIGNMENT AND DYNAMIC PROGRAMMING	64
*VISHNEVSKY O.V., KATOKHIN A.V. BABENKO V.N.	
ON RELATIONSHIPS BETWEEN GENE EXPRESSION EFFICIENCY AND NUCLEOTIDE CONTENT OF THE PROTEIN-CODING SEQUENCES	67
*LIKHOUSHVAI V.A. MATUSHKIN YU.G.	
DETERMINING MARKOV MODEL OF GENETICAL TEXTS BY STOCHASTIC COMPLEXITY ESTIMATION.....	71
*ORLOV YU.L., ¹ POTAPOV V.N.	
NEW ALGORITHMS FOR LARGE-SCALE EST CLUSTERING.....	74
*PTITSYN A., HIDE W.	
PERIODIC PATTERNS IN SEQUENCE ORGANIZATION OF REPLICATION ORIGIN OF <i>ESCHERICHIA COLI</i> K-12 CHROMOSOME	76
*KRAVATSKAYA G.I., ESIPOVA N.G.	
ESTMAP: A PROGRAM FOR ESTs MAPPING ON A GENOMIC SEQUENCE	79
*MILANESI L., *ROGOZIN I.B.	
COMPLEXITY MEASURES OF SYMBOLIC SEQUENCES AND THEIR APPLICATION TO DNA ANALYSIS.....	81
^{1,2} CHUZHANOVA N., ¹ KRAWCZAK M., ² GUSEV V.D., ² NEMYTIKOVA L.A., ^{1*} COOPER D.N.	
ANALYSIS OF MUTATIONAL HOTSPOTS IN HUMAN DISEASE GENES AND MUTATIONAL SPECTRA	84
*ROGOZIN I.B., ¹ BERIKOV V.B., GLAZKO G.V.	
COMPARATIVE ANALYSIS OF FUNCTIONAL SITE MOTIFS OF MGE <i>COPIA</i>-GROUP RELATIVE TO THEIR POSSIBLE MOLECULAR FUNCTIONS.....	87
*AMIKISHIEV V.G., RATNER V.A.	

S/MARs AND SOME ELEMENTS FROM DIFFERENT REPETITIVE FAMILIES ARE COLOCALIZED IN HUMAN GENOME	91
*GLAZKO G.V., KOCHETOV A.V., ROGOZIN I.B.	
STUDYING CORRELATIONS OF COMPUTATIONALLY PREDICTED ORIGINS OF REPLICATION AND BASE SKEWS IN THE <i>SACCHAROMYCES CEREVISIAE</i> GENOME	95
KORBEL J.O., ASSMUS H., KIELBASA SZ.M., *HERZEL H.	
BASIO: A SOFTWARE SYSTEM FOR SEGMENTATION OF BIOLOGICAL SEQUENCES INTO DOMAINS WITH HOMOGENOUS COMPOSITION	98
¹ *RAMENSKY V.E., ¹ MAKEEV V.JU., ² ROYTBERG M.A., ¹ TUMANYAN V.G.	
CAN GENETIC ALGORITHMS ASSIST IN GENOMIC RESEARCH?	100
WESTON P.S.	
A DATABASE OF GENETIC TEXTS WITH LATENT PERIODICITY (LPD)	102
*CHALEY M.B., KOROTKOV E.V.	
DNA SEQUENCE ASSEMBLY ALGORITHMS BASED ON CLUSTERING APPROACHES	105
ELLOUMI M.	
THE EVOLUTION OF REGULATORY FAMILIES IN ARCHEA AND EUBACTERIA: A COMMON ORIGIN OF TRANSCRIPTIONAL REPRESSORS	106
ERNESTO PEREZ-RUEDA, *J. COLLADO-VIDES	
USING LOCUS-SPECIFIC DATABASES OF HUMAN MUTATIONS FOR ESTIMATING PERNUCLEOTIDE RATE OF SPONTANEOUS MUTATION	107
KONDRASHOV A.S.	
COMPUTER MODELING OF EVOLUTION OF THE GENETIC DIVERSITY OF INTERACTING POLYGENIC SYSTEMS AND PATTERNS OF MOBILE GENETIC ELEMENTS IN THE COURSE OF SELECTION FOR THE QUANTITATIVE CHARACTER	108
^{1,2} *RATNER V.A. ¹ YUDANIN A.YA., ² EGOROVA A.V.	
PATTERNS OF MOBILE GENETIC ELEMENTS (MGEs) GENOMIC LOCALIZATION: INDUCTION OF TRANSPOSITIONS BY STRESS FACTORS, RESPONSE TO SELECTION AND POSSIBLE EVOLUTIONARY CONSEQUENCES	111
^{1,2} *RATNER V.A., ^{1,2} VASILYEVA L.A., ¹ BUBENSHCHIKOVA E.V., ^{1,2} ANTONENKO O.V.	
EVOLUTION OF THE CODE AND THE EARLIEST PROTEINS. RECONSTRUCTION FROM PRESENT-DAY SEQUENCES	113
TRIFONOV E.N.	
EVOLUTION OF PLANT REGULATORY SEQUENCES	115
*GOEBEL U., WIEHE T., MITCHELL-OLDS, T.	
A NEW VERSION OF SYNAP COMPUTER PROGRAM FOR LOGICAL MODELING OF PHYLOGENY	117
* ¹ BAIKOV K.S., ² ZVEREV A.A.	
SECTION 4. BIOINFORMATICS OF DNA, RNA, AND PROTEIN STRUCTURE. STRUCTURAL GENOMICS	120
STRUCTURE AND FORMAT OF THE EnPDB DATABASE ACCUMULATING SPATIAL STRUCTURES OF DNA, RNA AND PROTEINS	121
GRIGOROVICH D.A., *IVANISENKO V.A., KOLCHANOV N.A.	
MODEL OF PCR KINETICS	124
TITOV I.I.	

INFORMATION SYSTEM 'HIV VACCINE DEVELOPMENT'	127
*BELOVA O.E., BAZHAN S.I.	
RNA-POLYMERASE – PROMOTER RECOGNITION. SPECIFIC FEATURES OF ELECTROSTATIC POTENTIAL OF “EARLY” T4 PHAGE DNA PROMOTERS	130
¹ *DZHEL'YADIN T.R., ¹ SOROKIN A.A., ¹ IVANOVA N.N., ¹ SIVOZHELEZOV V.S., ¹ KAMZOLOVA S.G., ² POLOZOV R.V.	
STRUCTURE-BASED TARGET PREDICTION OF TRANSCRIPTION FACTORS.....	133
* ¹ SARAI A., ¹ SELVARAJ S., ¹ PRABAKARAN P., ² KONO H.	
REGIONS OF POTENTIAL INTERACTIONS IN RNA MOLECULES AS FOUND BY COMPUTER SEARCH	135
SHABALINA S.A.	
STRUCTURAL FEATURES OF mRNA 5'UTRs OF EUKARYOTIC GENES EXPRESSED AT HIGH AND LOW LEVELS	137
*VOROBIEV D.G., TITOV I.I., KOCHETOV A.V., KOLCHANOV N.A.	
MASS ANALYSIS OF RNA SECONDARY STRUCTURES USING A GENETIC ALGORITHM	140
*TITOV I.I., VOROBIEV D.G., KOLCHANOV N.A.	
TRANSTERM - A DATABASE OF RNA COMPONENTS AND MOTIFS	144
*JACOBS G.H., STOCKWELL P.A., BROWN C.M.	
CRASP: SOFTWARE PACKAGE FOR ANALYSIS OF PHYSICOCHEMICAL PARAMETERS OF ALIGNED SEQUENCES OF PROTEIN FAMILIES	147
AFONNIKOV D.A.	
NOVEL FUNCTIONAL FEATURES OF DNA-BINDING DOMAIN OF THE “HOMEODOMAIN” CLASS REVEALED BY ANALYSIS OF CORRELATIONS OF AMINO ACID SUBSTITUTIONS IN ITS POSITIONS	151
AFONNIKOV D.A.	
ANALYSIS OF STRUCTURAL MOTIFS IN PROTEINS.....	154
¹ JIANGHONG A.N., ² WAKO H., ¹ *SARAI A.	
‘IN SILICO’ ANALYSIS OF POINT MUTATION EFFECTS ON THE CODING REGION OF HUMAN BETA GLOBIN GENE.....	157
* ¹ ARRIGO P., ² IVALDI G., ³ CARDO P.P.	
DESIGN AND IMPLEMENTATION OF THERMODYNAMIC DATABASE FOR PROTEIN-NUCLEIC ACID INTERACTIONS	159
PRABAKARAN P., AN J., GROMIHA M.M., SELVARAJ S., UEDAIRA H., ¹ KONO H. AND *SARAI A.	
TOWARDS A STRUCTURAL BASIS OF HUMAN NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS.....	161
* ^{1,2,3} SUNYAEV S., ³ RAMENSKY V., ^{1,2} BORK P.	
THE DATABASE ASPD ON EXPERIMENTS WITH APPLICATION OF PHAGE DISPLAY TECHNIQUE	162
*VALUEV V. P., AFONNIKOV D.A., PETRENKO O., BEYLINA A.G., LOKHOVA I.V., GRIGOROVICH D.A., FOKIN O.N., IVANISENKO V.A.	
3-DIMENSIONAL PROTEIN STRUCTURAL CLASS RECOGNITION	166
*VALUEV V.P.	
BLOCKWISE EVOLUTION OF HEMOSTASIS AND COMPLEMENT FUNCTIONAL SYSTEMS	169
*ANANKO G.G.	

PDBSite: A DATABASE ON BIOLOGICALLY ACTIVE SITES AND THEIR SPATIAL SURROUNDINGS IN PROTEINS WITH KNOWN TERTIARY STRUCTURE.....	173
*IVANISENKO V.A., GRIGOROVICH D.A., KOLCHANOV N.A.	
RECEPTOR DATABASE (RDB) AS AN ANALYTICAL TOOL FOR THE DRUG DESIGN.....	177
*NAKATA K., TAKAI T., NAKANO T. AND KAMINUMA T.	
PROTEIN PRIMARY SEQUENCES AS MARKOV CHAINS	180
¹ *MITRA CHANCHAL K., ² SEN ARUSHARKA	
FROM GENOMES TO PROTEIN SPACE.....	182
*PEITSCH MANUEL C., SCHWEDE TORSTEN, DIEMAND ALEXANDER. AND GUEX NICOLAS	
DATABASE OF PATTERNS PROF_PAT, USED TO DETECT LOCAL SIMILARITIES	183
¹ *BACHINSKY A.G., ² GRIGOROVICH D.A., ¹ NAUMOCHKIN A.N., ¹ NIZOLENKO L.PH., ¹ YARIGIN A.A.	
ESTIMATION OF THE ENTROPY CHANGE UPON H-BOND FORMATION IN PROTEINS	187
RAKHMANINOVA A.B., *MIRONOV A.A.	
THE SEARCH OF REGIONS IN HIV-1 PROTEINS THAT HAVE LOCAL SIMILARITIES WITH HUMAN PROTEINS	190
*BAZHAN S.I., BACHINSKY A.G., MAKSYUTOV A.Z.	
L-ZIP MOTIF AS A PROBABLE DIMERIZATION MOTIF OF LTB4 RECEPTOR	193
LUKASHEV V.A., LUKASHOVA V.V., ROLA-PLESZCZYNSKI M., *STANKOVA J.	
INVESTIGATION OF THE AMINO ACID SEQUENCES OF MYCOBACTERIUM TUBERCULOSIS COMPLETE GENOME WITH PROTEIN FAMILY PATTERNS BANK PROF_PAT 1.3	196
*NIZOLENKO L.PH., KOZHINA E.M., YARIGIN A.A., BACHINSKY A.G.	
EFFECT OF HUMAN NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS UPON A PROTEIN STRUCTURE	199
^{1,2,3} SUNYAEV S., ³ *RAMENSKY V., ^{1,2} BORK P.	
ANALYSIS OF HEPATITIS C VIRUS PROTEINS USING SEQUENCE AND PUBLISHED DATA	200
SOBOLEV B.N., POROIKOV V.V., MATVEEV I.V., OLENINA L.V., KOLESANOVA E.F., *ARCHAKOV A.I.....	
AN APPROACH TO STRUCTURAL ALIGNMENT WITH GENETIC ALGORITHM	203
*PARK S.-J., YAMAMURA M.	
THEORETICAL MODEL OF INTERACTION: PLATELETE ACTIVATING FACTOR RECEPTOR (PAFR) AND TYROSINE	206
KINASE TYK2	206
¹ *LUKASHOVA V.V., LUKASHEV V.A., ² LUKASHEV V.V., ¹ ROLA-PLESZCZYNSKI M., ¹ STANKOVA J.	
STABILITY OF PARTIAL CORRELATION COEFFICIENT ESTIMATES FOR RESIDUE CHARACTERISTICS AT DIFFERENT POSITIONS OF AMINO ACID SEQUENCES	209
D.A. AFONNIKOV.....	
CONTEXT DEPENDENCIES IN AMINO ACID SEQUENCIES OF PROTEIN DOMAINS	213
*ORLOV YU. L., IVANISENKO V.A., ¹ POTAPOV V.N.	
HIERARCHICAL FEATURE DECOMPOSITION IN FUNCTIONAL DOMAINS	218
MURRAY D., HONIG B.H., *CALIFANO A. ¹	
SECTION 5	222
BIOINFORMATICS OF GENOME STRUCTURE AND EVOLUTION	222

MONO- AND BIVARIATE FLUORIMETRIC FLOW SORTING OF HUMAN CHROMOSOMES: QUANTITATIVE DATA ANALYSIS	223
*KRAVATSKY YU.V., POLETAEV A.I.	223
EDUCATION ON THE BASIS OF THE GENEEXPRESS SYSTEM: BUSINESS GAME "REGULATORY SIGNALS"	226
*PONOMARENKO M.P., PONOMARENKO J.V., LAVRYUSHEV S.V., VOROBIEV D.G., [§] MININA A.V., [§] IVASHIN S.A., [§] MIKHAILOV YU.I.	226
COURSE "INTRODUCTION TO BIOINFORMATICS"	231
*VALUEV V.P., AFONNIKOV D.A.	231
BIOINFORMATICS: NOVEL PROFILE OF VOCATIONAL EDUCATION.....	234
¹ *VALISHEV A.I., ² KOLCHANOV N.A., ² PODKOLODNY N.L., ¹ MELNIKOV V.N., ¹ ALSYNBAYEVA L.G., ¹ YAROSLAVTSEVA R.G., ³ HAANS W.J.A.	234
MINK ENTERITIS VIRUS VP2 GENE FRAGMENTS ANALYSIS.....	236
*TKACHEV S.E.	236
AUTHOR INDEX.....	239
KEYWORDS INDEX	241



SECTION 3.
BIOINFORMATICS OF GENOME
STRUCTURE AND EVOLUTION

AUTOMATED COMPARATIVE ANALYSIS OF REGULATORY PATTERNS: SUGAR METABOLISM AND TRANSPORT SYSTEMS IN GAMMA PURPLE BACTERIA

Mironov A.A., *Gelfand M.S.

State Scientific Center for Biotechnology NII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Resume

Comparative analysis of bacterial genomes is a powerful tool in analysis of regulatory patterns. However, until now it has been used mostly for analysis of regulons that has been studied in experiment, at least partially [Mironov A.A. et al., 2000, Gelfand M.S. et al., 2000] and other papers in this volume.

However, the same approaches can be applied to analysis of completely regulons for which no regulatory sites are available. At that, it is assumed that the set of co-regulated genes and the regulatory signal itself is conserved in related genomes.

Analysis of the purine regulon in genomes of the *Pyrococcus* species allowed us to determine the regulatory signal and to find a new purine transporter [Gelfand M.S. et al., 2000]. However, that study considered a regulon that could be narrowly defined by functional annotation of proteins and analysis of the metabolic map. Here we explore possibilities for a large scale analysis, using as a test case genes implicated in sugar metabolism and transport in gamma purple bacteria.

We were able to derive several known signals of transcription factors for sugar regulons using a formalized protocol. Several new patterns are now studied manually in order to determine their possible function.

Data

Complete genome sequences of *Escherichia coli* and *Haemophilus influenzae* were extracted from GenBank [Benson D.A. et al., 1999]. Partially sequenced genome of *Vibrio cholerae* was extracted from the TIGR WWW site (TIGR).

Methods

Rows of orthologous genes relevant for sugar metabolism, transport, and regulation were compiled using COGs [Tatusov R.L. et al., 1997] and GenomeExplorer [Mironov A.A. et al., 2000]. Approximately 300 rows were constructed (10% of the *E. coli* genome).

For each gene the upstream region was defined as the segment (−200)–(+50) for *E. coli* and *H. influenzae* and (−300)–(+100) for *V. influenzae*, where candidate genes were defined as maximal open reading frames longer than 100 codons. Upstream regions of potentially co-transcribed genes were grouped together and considered as a single regulatory region. Then all (groups of) regulatory regions were combined into one training sample.

Candidate signals were constructed using the iterative procedure of profile generation described in [Gelfand M.S. et al., 2000]. The minimum number of sites per profile was 3 (“seed sites”), but it could be increased by addition of other high-scoring sites. The profile length was chosen 20. Only palindromic profiles were considered.

At the filtration step only profiles involving at least a pair of orthologous genes were retained. It was required that the seed sites and the sites from the orthologous genes did not overlap with protein-coding regions in *E. coli* and *H. influenzae* (gene starts in *V. cholerae* are not reliable). Finally, the candidate profiles were ordered by decrease of the total information content. The profiles with the highest information content were analyzed manually. In particular, they were compared with known consensi of transcription factor binding sites taken from the literature and the database DPlnteract [Robison K. et al., 1998].

Results and Discussion

Fifteen most informative profiles are listed in Table 1. Half of them represent well known regulatory signals. Using these profiles as a starting point for manual analysis, we were able to identify the corresponding regulons in *H. influenzae* and *V. cholerae*. However, another half of the profiles are new signals. They could be false positives, or represent new regulatory patterns. We are currently studying these signals in order to determine their functional relevance.

Note that the main part of this study was done automatically without any external intervention. The analyzed set of genes constitutes about 10% of the *E. coli* genome. Thus the results of this pilot study demonstrate that purely automated analysis of regulation by comparison of bacterial genomes is feasible.

Table 1. Fifteen strongest candidate signals upstream of sugar metabolism and transport genes. I: total information content. I/L: information content per position.

	predicted consensus	I	I/L	regulator	known consensus
1	aCAGCGAAACGTTTCGCTGa	25.06	1.3	RbsR	tyAkCGAAACGTTTCGmTra
2	CAAAAATCTGCAGATTTTTG	23.77	1.2		
3	AATCGGGAACGTTCCCGATT	23.53	1.2	TreR	waakGGGAACGTTCCCmttw
4	AATGTTTCGTTAACGAACATT	23.25	1.2	GlpR	watGtTCgttaacGAaCatw
5	TAGTGTAACGTTTACACTA	22.68	1.1	GalS, GalR	gTGtAAAnCGnTTaCAc
6	TGAATCAGGATCCTGATTCA	22.58	1.1		
7	TTGTGAACAATTGTTCAAA	22.57	1.1		
8	TATTGTTGCATGCAACAATA	22.42	1.1		
9	TTGTGAAACCGGTTTCACAA	22.37	1.1	AsgG?	
10	AGATTCACATATGTGAATCT	22.00	1.1		
11	ATATGTTACGCGTAACATAT	21.87	1.1	GntR	wnAwGTTACssGTAACwTnw
12	GCGCTGAAaGCaTTCAGCGC	21.80	1.1	FruR	rcTGAAAtCGaTTCAGy
13	AAATTTTAAGCTTAAAATTT	21.78	1.1		
14	TTAACGCTGTACAGCGTTAA	21.76	1.1		
15	TTTGTGATCTAGATCACAAA	21.66	1.1	CRP	ttTGtGAtnnnnaTCACAaa

Acknowledgements

We are grateful to M. Galperin for the help in compiling the sample of sugar metabolism and transport genes and to E. Koonin for useful discussions. This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

Preliminary sequence data for *V. cholerae* were obtained from The Institute for Genomic Research WWW site.

References

1. Benson, D.A. et al. (1999) *Nucleic Acids Res.*, 27, 12-17.
2. Gelfand, M.S., Koonin, E.V., Mironov, A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative-genomic approach, *Nucleic Acids Res.*, 28, 695-705.
3. Gelfand, M.S., Koonin, E.V., Mironov, A.A. (2000) Comparative approach to analysis of regulation in complete genomes: transcription regulatory sites in Archaea. (This volume).
4. Mironov, A.A., Vinokurova, N.P. and Gelfand, M.S. (2000) GenomeExplorer: software for analysis of complete bacterial genomes. (This volume).
5. Robison, K., McGuire, A.M. and Church, G.A. (1998) *J. Mol. Biol.*, 284, 241-254.
6. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science*, 278, 631-637.
7. TIGR. <http://www.tigr.org>

PRO-FRAME: SIMILARITY-BASED GENE RECOGNITION IN EUKARYOTIC DNA SEQUENCES WITH ERRORS

*Mironov A.A., *Gelfand M.S.*

State Scientific Center for Biotechnology NII Genetika, Moscow, Russia

Anchorgen, Inc., Santa Monica, USA

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: gene recognition, spliced alignment algorithm, sequencing errors, eukaryotes

Resume

Performance of existing algorithms for similarity-based gene recognition in eukaryotes drops when the genomic DNA has been sequenced with errors. A modification of the spliced alignment algorithm allows for gene recognition in sequences with errors, in particular frameshifts. It tolerates up to 5% of sequencing errors without considerable drop of prediction reliability when a sufficiently close homologous protein is available (normalized similarity score 50% or higher).

Availability:

<http://www.anchorgen.com>

Analysis of sequence similarity is a powerful tool for gene recognition. It is employed in a number of database search programs, most notably BLASTX (Gish and States, 1993), and programs for exact prediction of exon-intron structure, in particular, Procrustes (Gelfand et al., 1996; Mironov et al., 1998), INFO (Hultner et al., 1994; Laub et al., 1998), GeneWise (Birney et al., 1997). The common idea behind these algorithms is that among numerous possible exon chains, an algorithm chooses the chain having the highest similarity to a related protein (target). This is done by modified dynamic programming treating introns as a special case of gaps (GeneWise) or by spliced alignment (Procrustes).

Testing of similarity-based gene recognition programs demonstrated that given sufficiently close relatives, they produce highly reliable predictions. In particular, the correlation between predicted and real human genes is 96-99% when homologous vertebrate genes are available (Mironov et al., 1998; Laub et al., 1998). However, the quality of gene predictions when the genomic DNA contains sequencing errors is much lower (Bursset and Guigo, 1996). One possibility to avoid this problem is to use DNA spliced alignment instead of aligning translated candidate exons with proteins (Sze and Pevzner, 1997). However, it is well known that protein alignments are much more sensitive to distant similarities than nucleotide alignments. Thus it is indicative that there exist numerous protein-DNA alignment algorithms accounting for frameshifts (Posfai and Roberts, 1992; Birney et al., 1996; Guan and Uberbacher, 1996; Zhang et al., 1997; Pearson et al., 1997). However, none of them handles introns.

We have implemented a modified version of the spliced alignment algorithm performing gene recognition in the presence of frameshift errors. The algorithm treats introns as non-penalized gaps that may start only at dinucleotide GT and end at dinucleotide AG. Frameshifts and in-frame stop codons in the genomic sequence are allowed, but heavily penalized. There is an option for acceleration of dynamic programming, using k-tuple alignment technique due to M. Roytberg (Nazipova et al., 1995). Since sequencing errors can destroy invariant dinucleotides at splicing sites, the program has a post-processing step. At this step the program identifies local drops of similarity at exon termini, and observing a sharp drop, moves the exon-intron boundary even if there are no suitable dinucleotides.

Results of testing the algorithm on a sample of human genes and related proteins from (Mironov et al., 1998) are given in Fig. 1. The performance at different error levels is estimated using the standard correlation coefficient measure (Bursset and Guigo, 1996; Mironov et al., 1998). For comparison we present also the correlation coefficient demonstrated by the original Procrustes algorithm. Since the performance depends on the similarity between the gene and a target, the figure features plots of the correlation coefficient at different similarity levels. The similarity measure is the score of the alignment of the actual and target proteins divided by the halfsum of the scores of (trivial) alignments of the actual protein and the target protein with themselves. Such normalization accounts for varying protein length and amino acid composition. Sequencing errors were modeled as random nucleotide substitutions (80%), insertions (10%) and deletions (10%).

It is noteworthy that in the absence of sequencing errors Pro-Frame performs almost as well as Procrustes when the target protein is close to the analyzed gene, but the performance drops for distant relatives. This agrees with our observations about importance of the statistical filtering procedure implemented in Procrustes

(Mironov et al., 1998). On the other hand, up to 3% rate of sequencing errors does not considerably influence the reliability of predictions, and further, up to 6% of errors are easily tolerated if the target protein is sufficiently close to the analyzed gene.

The above results demonstrate that Pro-Frame may be a useful tool for analysis of preliminary sequencing data, e.g. phase II output of major sequencing projects.

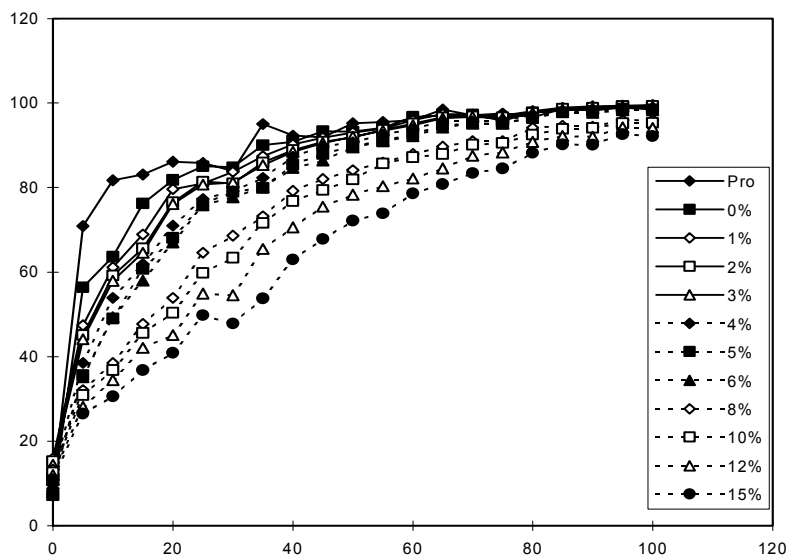


Figure 1. Testing of Pro-Frame on a sample of human genes. Horizontal axis: similarity between actual genes and related proteins. Vertical axis: correlation coefficient. Each curve corresponds to a specific level of sequencing errors (the percent of positions with errors is given in the legend on the right). "Pro" corresponds to the original Procrustes algorithm (Mironov et al., 1998).

Acknowledgements

We are grateful to Drs. V. Bafna, J.W. Fickett, P. Pevzner and M.A. Roytberg for useful discussions. This work was partially supported by Anchorgen, Inc. (<http://www.anchorgen.com>).

References

1. Birney, E., Thompson, J.D. and Gibson, T.J. (1996) PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.*, **24**, 2730-2739.
2. Birney, E. and Durbin, R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *Proc. 5th Int. Conf. on Intelligent Systems for Molecular Biology*, pp. 56-64 (AAAI Press, Menlo Park, CA).
3. Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353-367.
4. Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci.*, **93**, 9061-9066.
5. Gish, W. and States, D.J. (1993) Identification of protein-coding regions by database similarity search. *Nature Genet.*, **3**, 266-272.
6. Guan, X. and Uberbacher, E.C. (1996) Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci.*, **12**, 31-40.
7. Hultner, M., Smith, D.W. and Wills C. (1994) Similarity landscapes: A way to detect many structural and sequence motifs in both introns and exons. *J. Mol. Evol.*, **38**, 188-203.
8. Laub, M.T. and Smith, D.W. (1998) Finding intron/exon splice junctions using INFO, INterruption Finder and Organizer. *J. Comput. Biol.*, **5**, 307-321.
9. Mironov, A.A., Roytberg, M.A., Pevzner, P.A. and Gelfand, M.S. (1998) Performance-guarantee gene predictions via spliced alignment. *Genomics*, **51**, 332-339.
10. Nazipova, N.N., Shabalina, S.A., Ogurtsov, A.Yu., Kondrashov, A.S., Roytberg, M.A., Buryakov, G.V., Vernoslov, S.E. (1995) SAMSON: a software package for the biopolymer primary structure analysis. *Comput. Appl. Biosci.*, **11**, 423-426.
11. Pearson, W.R., Wood, T., Zhang, Z., Miller, W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24-36.
12. Posfai, J. and Roberts, R.J. (1992) Finding errors in DNA sequences. *Proc. Natl. Acad. Sci. USA*, **89**, 4698-4702.
13. Sze, S.-H. and Pevzner P.A. (1997) Las Vegas algorithms for gene recognition: Suboptimal and error-tolerant spliced alignment. *J. Comput. Biol.*, **4**, 297-310.
14. Zhang Z, Pearson WR, Miller W (1997) Aligning a DNA sequence with a protein sequence. *J. Comput. Biol.*, **4**, 339-349.

GENOMEEXPLORER: SOFTWARE FOR ANALYSIS OF COMPLETE BACTERIAL GENOMES

*Mironov A.A., Vinokurova N.P., *Gelfand M.S.*

State Scientific Center for Biotechnology NII Genetika, Moscow, Russia

Anchorgen, Inc., Santa Monica, USA

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: regulatory signal, protein similarity, profile search, bacterial genome

Resume

Analysis of regulatory signals in complete bacterial genomes is a time-consuming process involving multiple protein similarity searches, profile searches in DNA, analysis of orthology relationships etc. Although the standard software tools can be applied for the individual tasks, differences in the input and output format, lack of an interface, and the use of non-compatible hardware make the use of these tools inconvenient. *GenomeExplorer* is a user-friendly program for analysis of complete bacterial genomes that can be used to analyze gene functions and predict regulatory patterns (see accompanying papers in this volume).

Availability:

<http://www.anchorgen.com>

Contact:

mgelfand@anchorgen.com

GenomeExplorer is a program for analysis of complete bacterial genomes. It combines tools for analysis of protein similarity and DNA pattern and profile search with a user-friendly interface allowing for easy integration with external programs and Internet servers. We have successfully used it for analysis of regulatory patterns in various bacterial genomes (Gelfand and Mironov, 1998; Gelfand and Mironov, 1999; Gelfand et al., 2000; Mironov et al., 1999; Panina et al., 2000).

GenomeExplorer runs under *Windows* and retains the general layout and defaults of this operating system. Simultaneous use of several copies of *GenomeExplorer* allows one to perform comparative analysis of multiple genomes. The use of the clipboard allows for easy import and export of data to *Word* and *Excel* as well as the use of external Internet tools.

The program has two main windows. The annotation window shows the data about genes, as described in GenBank or EMBL feature tables. The map window provides graphical information about gene location. The program allows the user to perform a variety of textual searches. Options for the protein similarity analysis include simple homology search using the Smith-Waterman algorithm, search for paralogous genes (in the same genome) and search for orthologues, defined as bidirectional best hits. Results of the similarity search are output to a special window in convenient graphic form. They can also be output to a file or to the system clipboard. There is also an option for DNA-protein similarity search (an analog of TBLASTN). The DNA analysis tools include procedures for profile search, construction of profiles given training sets of experimentally determined sites, and iterative site search using re-computation of profiles from user-defined sets of candidate sites (an analog of PSI-BLAST).

A typical scenario of application of *GenomeExplorer* is the following. Given a set of experimentally determined regulatory sites in an "old" genome, one needs to find analogous sites in a "new" genome and possibly find new sites in both genomes.

Comparative analysis makes sense only if the new genome contains an orthologue of the responsible transcription factor. The similarity should be sufficiently high and evenly distributed along the alignment.

Then one finds in the new genome orthologues of the old regulon. The site search procedure with a lower threshold is used to find candidate sites in upstream regions of these genes. The obtained sites can be used to construct a new profile, either independently, or in addition to the old sites.

To find new regulon members, one should perform site search independently in the two genomes. The sets of genes having candidate sites in upstream regions are compared, and pairs of orthologous genes are selected as candidate regulon members.

If only one experimentally determined gene is available, one should use the iterative profile construction procedure. To find candidate regulon members at each iteration one can use both functional and comparative considerations.

Thus the main underlying assumption is that occurrence of candidate sites upstream of orthologous genes are independent events, providing for effective filtering of false positives. Thus specific care should be exercised for analysis of closely related genomes. Sometimes it is difficult to resolve the orthology relationships in large multigene families, such as ABC transporters or transcription factors. In this case positional analysis can be applied to distinguish orthologues and paralogues. Gene starts in computer-annotated genomes are often unreliable: a warning sign is low similarity at N-termini of aligned proteins. Finally, one has to take into account changes in operon structure. To do that, it is sufficient to analyze sets of potentially co-regulated genes, that is, genes located on the same strand at a close distance. GenomeExplorer has tools that allow one to account for all above possibilities.

For a more detailed description of GenomeExplorer see (Mironov et al., 2000).

Acknowledgements

We are grateful to E.V.Koonin, Yu.I.Kozlov, A.B.Rakhmaninova, D.A.Rodionov, M.A.Roytberg for useful discussions. Development of *GenomeExplorer* is supported by Anchorgen, Inc. (<http://www.anchorgen.com>).

References

1. Gelfand, M.S. and Mironov, A.A. (1998) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *1st Conf. BGRS-98*, vol. 1, pp. 147-149.
2. Gelfand, M.S., Mironov, A.A., Jomantas, J., Kozlov, Yu.I. and Perumov, D.A. (1999) A conserved RNA structure element involved in the regulation of bacterial riboflavin biosynthesis genes. *Trends Genet.*, **15**, 439-442.
3. Gelfand, M.S., Koonin, E.V., Mironov, A.A. (2000) Prediction of transcription regulatory sites in *Archaea* by a comparative-genomic approach, *Nucleic Acids Res.*, **28**, 695-705.
4. Gelfand, M.S., Koonin, E.V. and Mironov, A.A. (2000) Comparative approach to analysis of regulation in complete genomes: transcription regulatory sites in *Archaea*. (This volume).
5. Mironov, A.A., Vinokurova, N.P. and Gelfand M.S. (2000) Software for analysis of bacterial genomes. *Mol. Biol.*, **34**, no. 2, in press.
6. Mironov, A.A., Koonin, E.V., Roytberg, M.A. and Gelfand, M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **27**, 2981-2989.
7. Panina, E.M., Mironov, A.A. and Gelfand, M.S. (2000) Regulation of bacterial DAPH-synthases: retroinhibition and repression of transcription. (This volume).

COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: CATABOLITE REPRESSION IN GAMMA-PROTEOBACTERIA

Novichkova E.S., Novichkov P.S., Mironov A.A., *Gelfand M.S.

State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: regulatory signal, catabolite repression, profile search, gamma-proteobacteria

Resume

Catabolite repression in gamma-proteobacteria is mediated by transcription factor CRP binding to operators with a loose consensus `wwwTGtGAtyyrgwTCACtTwt`. The CRP-regulon is one of the largest and best studied regulons in *Escherichia coli*, both by experimental and computational methods (Studnicka, 1987; Berg and von Hippel, 1988; Stormo and Hartzell, 1989; Ebright, 1993; Perez-Rueda et al., 1998).

We have attempted to describe the CRP-regulons of other gamma-proteobacteria (*Haemophilus influenzae* and *Vibrio cholerae*) and to find new members of this regulon using the methods of comparative genomics (Gelfand and Mironov, 1998; Mironov et al., 1999). These particular genomes were chosen because they are close (in particular, contain orthologues of CRP), but still sufficiently diverse to assume absence of non-functional sequence conservation in non-coding regions. In addition, the *E. coli* and *H. influenzae* genomes are complete (Blattner et al., 1997; Fleischmann et al., 1995), and the *V. cholerae* genome is almost complete (TIGR).

The training set for profile construction consisted of 49 operators corresponding to 37 *E. coli* genes taken from the DPLInteract database (Robison et al., 1998) (marked "T" in Table 1). The minimum observed score on the training set was 3.0, and it should be noted that almost 50% of all *E. coli* genes have a candidate site of such strength in the upstream region. Thus only candidate sites scoring above 3.5 were considered significant (marked by boldface in Table 1). Even with this more restrictive threshold overprediction is about 95%, and thus only filtering of candidate sites by genomic comparison gives any hope of making a reliable prediction. All analyses were done using GenomeExplorer (Mironov et al., 2000).

Table 1 lists all genes having upstream candidate sites in *E. coli* and at least one of the two other genomes. The following observations can be made. The regulon seems to be well conserved, although the degree of conservation seems to be less than that of more specific regulons (Gelfand and Mironov, 1998). Surprisingly, we could identify a large number of additional members of the regulon. Most of them encode enzymes, transport systems, and transcriptional regulators of the sugar metabolism. Indeed, exhaustive search of the literature allowed us to find indication of CRP regulation of some of the predicted CRP-regulon members. There are several additional genes involved in catabolism of amino acids and thus also likely to belong to the CRP-regulon, in particular, a second L-asparaginase gene. Several genes probably are false positives, and among the latter, there are several known or likely members of the FNR regulon. Indeed, FNR and CRP are closely homologous transcriptional regulators with rather similar recognition sites (Bell et al., 1989; Sawers et al., 1995). Finally, the predicted regulon contains some genes with unknown or only generally known function, and we feel that the prediction procedure has been demonstrated to be sufficiently specific to expect that a considerable portion of these genes indeed are regulated by CRP.

Further analysis will be directed towards more detailed analysis of sugar regulons, analysis of the operon structure for the identified genes; positional analysis aimed at prediction of the activation/repression regulation mode (Ebright, 1993; Perez-Rueda et al., 1998); resolving the overlap between the FNR and CRP regulons (Bell et al., 1989; Sawers et al., 1995), and analysis of the nucleotide regulons, in particular, the CytR regulon that is a subset of the CRP regulon (Valentin-Hansen et al., 1996).

Acknowledgements

We are grateful to A.B.Rakhmaninova, D.Rodionov and V.P.Veiko for useful discussions.

This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

Preliminary sequence data for *V. cholerae* were obtained from The Institute for Genomic Research WWW site.

Table 1. Results of analysis. Column 1: *E. coli* gene; divergently transcribed genes are separated by slash "/" (both genes are given independently); genes are listed in alphabetic order within the major functional categories, with unknown genes listed at the end of the category. Columns 2-4: EC: score of the *E. coli* candidate site; HI: score of the *H. influenzae* candidate site; VC: score of the *V. cholerae* candidate site; boldface: scores exceeding 3.5; blank cell: no orthologous gene. Column 5: boldface: regulators; italics: transport systems. Column 6: "T": gene from the training set; "P": gene subsequently found to be CRP-regulated; "F": FNR-regulated genes; "+": genes likely to be CRP-regulated based on their function; "-": genes unlikely to be CRP-regulated based on their function.

1	2	3	4	5	6
GENE	EC	HI	VC	FUNCTION	
<i>crp</i>	3,6	3,1	4,2	cyclic AMP receptor protein	T
				<u>SUGAR METABOLISM</u>	
<i>aldB</i>	3,0		3,0	aldehyde dehydrogenase b	T
<i>araB</i>	3,9			L-ribulokinase	T
<i>deoC</i>	4,7		3,3	deoxyribose-phosphate aldolase	T
<i>fucA</i>	4,2	3,6		fucose-1-phosphate aldolase	+
<i>fruB</i>	4,3	3,2	3,9	<i>pts system, fructose-specific IIA/FPR component</i>	+
<i>galE</i>	3,8	3,3	3,0	UDP-glucose 4-epimerase	T
<i>galS</i>	3,8	4,4	3,2	mgl repressor and galactose ultrainduction factor	P
<i>galU</i>	3,8	3,2	3,7	glucose-1-phosphate uridylyltransferase	+
<i>gapA</i>	4,9	4,5	3,4	glyceraldehyde 3-phosphate dehydrogenase A	+
<i>glpA/glpT</i>	4,4	5,1	3,9	anaerobic glycerol-3-phosphate dehydrogenase, subunit A	T
<i>glpD/glpE</i>	4,2		4,6	aerobic glycerol-3-phosphate dehydrogenase	T
<i>glpE/glpD</i>	4,2			gene of glp regulon	T
<i>glpT/glpA</i>	4,4	5,1	3,7	<i>glycerol-3-phosphatase transporter</i>	T
<i>glpF</i>	3,7	3,0	3,6	facilitates glycerol diffusion	T
<i>gnd/yifB</i>	3,5	4,0	3,0	6-phosphogluconate dehydrogenase, decarboxylating	+
<i>gntV</i>	4,4		4,4	gluconate kinase	+
<i>lacZ</i>	4,5		3,6	beta-galactosidase	T
<i>malK/malE</i>	4,0		3,7	<i>cytoplasmic membrane protein for maltose uptake</i>	T
<i>malE/malK</i>	4,0		4,0	<i>periplasmic maltose-binding protein</i>	T
<i>malP</i>	4,2	3,1	3,8	maltodextrin phosphorylase	+
<i>malS/udp</i>	3,9		3,6	alpha-amylase	+
<i>malT</i>	4,1		4,3	positive regulatory gene for mal regulon	T
<i>meiR</i>	3,3			regulatory gene	T
<i>mglB</i>	4,3	3,6	4,4	<i>d-galactose-binding periplasmic protein</i>	P
<i>mtlA</i>	4,2		4,2	<i>mannitol-specific enzyme II of phosphotransferase system</i>	T
<i>nagB</i>	4,1	3,2	2,9	glucosamine-6-phosphate isomerase	T
<i>nagE</i>	4,1		4,0	<i>pts system, N-acetylglucosamine-specific IIBC component</i>	T
<i>pckA</i>	4,3	4,5	3,3	phosphoenolpyruvate carboxykinase	P
<i>pdhR</i>	3,9		3,8	pyruvate dehydrogenase complex repressor	+
<i>ptsG</i>	3,8		3,9	<i>PTS system, glucose-specific IIBC component</i>	+
<i>ptsH</i>	4,0			<i>phosphocarrier protein Hpr</i>	T
<i>rbsD/infA</i>	3,7	4,3		<i>high affinity ribose transport protein RbsD</i>	+
<i>rhaS/rhaB</i>	4,4			L-rhamnose operon regulatory protein rhas	T
<i>rhaB/rhaS</i>	4,4			rhamnulokinase	T
<i>srlA_1</i>	4,4			<i>PTS system, glucitol/sorbitol-specific IIBC component</i>	T
<i>treB</i>	3,5		3,6	<i>phosphotransferase system trehalose permease</i>	+
<i>uxuA</i>	3,6	3,4		D-mannonate hydrolase	T
<i>uxuR</i>	3,8	4,9		uxu operon regulator	+
<i>xylA</i>	4,1	4,3		D-xylose isomerase	+
<i>xylR</i>	3,6	3,5		xylose operon regulatory protein	+
<i>xylF</i>	4,1	4,3		<i>xylose binding protein transport system</i>	+
<i>b0820/glyA</i>	4,3		3,6	<i>hypothetical ABC transporter ATP-binding protein YbiT</i>	+?
<i>b3575</i>	3,9	4,2		putative dehydrogenase	?
				<u>AMINO ACID CATABOLISM</u>	
<i>ansB</i>	3,5	4,7		L-asparaginase	T,F

1	2	3	4	5	6
GENE	EC	HI	VC	FUNCTION	
<i>aspA</i>	4,1	4,7	3,4	aspartate ammonia-lyase (aspartase)	P
<i>dadA</i>	4,3		2,8	D-amino acid dehydrogenase	T
<i>dsdC</i>	4,2		3,6	D-serine deaminase activator	+
<i>glyA/b0820</i>	3,8	3,2	4,0	serine hydroxymethyltransferase	+
<i>ilvB</i>	3,8			acetohydroxy acid synthase I, small subunit	T
<i>oppA</i>	3,6	3,8	3,9	<i>periplasmic oligopeptide-binding protein</i>	+
<i>ppiA</i>	3,8		2,8	peptidyl-prolyl cis-trans isomerase a	T
<i>sdaC</i>	4,0	4,6	3,9	<i>putative serine transporter</i>	P
<i>serC/yieS</i>	4,4	3,5	3,8	phosphoserine aminotransferase	+
<i>tdcA</i>	4,3			tdcABC operon transcriptional activator	T
<i>tnaL</i>	4,9			tna operon leader peptide	T
<i>ybiK</i>	3,5		3,7	putative L-asparaginase precursor	+
<i>b1729</i>	4,0	3,0	3,7	<i>H/Na-glutamate symport (glutamate-aspartate carrier)</i>	+
				NUCLEOTIDE METABOLISM	
<i>cdd</i>	4,8	4,7	4,2	cytidine deaminase	T
<i>cpdB</i>	3,6	3,6	3,1	2',3'-cyclic-nucleotide 2'-phosphodiesterase	+
<i>cyaA</i>	4,0	5,1	4,7	<i>cyaA</i>	T
<i>cytR</i>	4,0		4,1	transcriptional repressor cytR	T
<i>folE</i>	3,8	3,7	2,6	GTP cyclohydrolase I	+?
<i>guaB</i>	4,0	3,5		inosine-5'-monophosphate dehydrogenase	+
<i>hpt</i>	4,8	3,5	3,1	hypoxanthine phosphoribosyltransferase	+?
<i>nrdD</i>	4,2	4,7	3,2	anaerobic ribonucleoside-triphosphate reductase	+
<i>nupG</i>	5,0			<i>nucleoside permease NupG</i>	T
<i>tsx</i>	4,0			<i>nucleoside-specific channel-forming protein Tsx</i>	T
<i>udp/malS</i>	4,7	3,5	3,9	uridine phosphorylase	P
<i>ung</i>	3,8	4,4	2,6	uracil-DNA glycosylase	+?
<i>ushA</i>	3,5		3,6	UDP-sugar hydrolase precursor	+
				OTHER, UNRELATED, and UNKNOWN PROTEINS	
<i>ackA</i>	3,9	3,6	3,3	acetate kinase	F?
<i>deaD</i>	4,5	3,2	4,1	presumed ATP-dependent RNA helicase	-
<i>fadL</i>	4,4	3,6	2,7	<i>long-chain fatty acid transport protein precursor</i>	-?
<i>fdhD</i>	4,2	5,0		nitrate inducible formate dehydrogenase activity	F?
<i>focA</i>	3,6	3,5	3,0	<i>probable formate transporter (formate channel)</i>	F
<i>fur</i>	3,8	3,7	3,3	ferric uptake regulation protein	T
<i>fusA</i>	3,8	3,4	3,8	protein chain elongation factor EF-G	-
<i>hemC</i>	4,0		4,7	porphobilinogen deaminase	-?
<i>infA/rbsD</i>	3,8	3,7		initiation factor IF-1	-
<i>moeA</i>	3,5	3,7	2,6	molybdopterin biosynthesis MoeA protein	-?
<i>ndh</i>	3,8	3,9	3,6	NADH dehydrogenase	F
<i>ompA</i>	3,5	3,0		<i>outer membrane protein A</i>	T
<i>ompR</i>	3,5			positive regulatory gene for ompC and ompF	T
<i>sodA</i>	3,8	3,7	2,8	manganese superoxide dismutase	-?
<i>tesB/b1873</i>	3,9	2,8	3,9	acyl-coA thioesterase II	-
<i>tgt</i>	3,8	3,8	3,2	queuine tRNA-ribosyltransferase; tRNA-guanine transglycosylase	-
<i>topA</i>	3,7	3,4	4,2	DNA topoisomerase I, omega protein I	-
<i>trxB</i>	3,5	3,2	3,9	thioredoxin reductase	F?
<i>yaiD</i>	3,6	3,7	3,3	orf	?
<i>yciD</i>	4,6		3,8	<i>outer membrane protein w precursor</i>	?
<i>yejM</i>	3,7	3,9		putative sulfatase	-?
<i>yfiD</i>	3,8	4,4		putative formate acetyltransferase	F?
<i>yhdG</i>	4,6	3,9	3,6	putative dehydrogenase	+?
<i>yhjA</i>	4,1		3,7	probable cytochrome c peroxidase	F?
<i>yiaJ</i>	3,9	4,2		hypothetical transcriptional regulator	?

1	2	3	4	5	6
GENE	EC	HI	VC	FUNCTION	
<i>yjeS/serC</i>	3,8		3,5	hypothetical 43.1 KD protein	?
<i>yifB/gnd</i>	3,8	4,7	3,3	putative 2-component regulator	?
<i>yjfS</i>	4,1		4,1	putative transport protein SgaT	?
<i>yjil</i>	4,7	4,6		orf	?
<i>b0426</i>	4,0	3,6	2,7	orf	?
<i>b0618/19</i>	4,4	3,7	3,6	citrate lyase syntetase	F?
<i>b0619/18</i>	4,4		3,8	histidin kinase	-?
<i>b1036</i>	4,4		4,6	hypothetical 18.8 KD protein	?
<i>b1873/tesB</i>	3,5		3,5	probable cytochrome c-type protein	F?
<i>b2463</i>	3,9	4,0	3,5	putative multimodular enzyme	?
<i>b2736</i>	4,6	4,3		putative dehydrogenase	+?
<i>b2777</i>	3,6		3,8	<i>putative transport protein</i>	?
<i>b2794</i>	4,5	3,6	3,2	orf	?
<i>b2900</i>	3,6	3,5		orf	?
<i>b3865</i>	4,0	3,9		orf	?
<i>b2899</i>	3,5		3,8	hypothetical 23.8 KD protein	?

References

- Bell, A.I., Gaston, K.L., Cole, J. and Busby, S.J.W. (1989) *Nucleic Acids Res.*, 17, 3865-3874.
- Berg, O.G. and von Hippel, P.H. (1988) *J. Mol. Biol.*, 200, 709-723.
- Blattner, F.R. et al. (1997) *Science*, 277, 1453-1462.
- Ebright, R.H. (1993) *Mol. Microbiol.*, 8, 797-802.
- Fleischmann, R.D. et al. (1995). *Science*, 269, 496-512.
- Gelfand, M.S. and Mironov, A.A. (1998) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. 1st Conf. BGRS-98, vol. 1, pp. 147-149.
- Mironov, A.A., Koonin, E.V., Roytberg, M.A. and Gelfand, M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, 27, 2981-2989.
- Mironov, A.A., Vinokurova, N.P. and Gelfand, M.S. (2000) *GenomeExplorer*: software for analysis of complete bacterial genomes. (This volume).
- Perez-Rueda, E., Gralla, J.D. and Collado-Vides, J. (1998) *J. Mol. Biol.*, 165-170.
- Robison, K., McGuire, A.M. and Church, G.A. (1998) *J. Mol. Biol.*, 284, 241-254.
- Sawers, G., Laiser, M., Sirko, A. and Freundlich, M. (1995) *Mol. Microbiol.*, 23, 835-845.
- Stormo, G.D. and Hartzell, G.W., III (1989) *Proc. Natl. Acad. Sci. USA*, 86, 1183-1187.
- Studnicka, G. (1987) *Gene*, 58, 45-57.
- TIGR. <http://www.tigr.org>
- Valentin-Hansen, P., Sogaard-Andersen, L. and Pedersen, H. (1996) *Mol. Microbiol.*, 20, 461-466.

COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: SOS REPAIR IN BACILLUS SUBTILIS

¹Permina E.A., ²Mironov A.A., ^{2*}Gelfand M.S.

¹Moscow State University, Department of Biology,

²State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: regulatory signal, repressor, SOS repair, comparative analysis, *Bacillus subtilis*

Resume

SOS-repair in Gram-positive bacteria is regulated by the DinR repressor which is homologous to the LexA repressor of *Escherichia coli* (Winterling et al., 1998). The consensus of the DinR recognition site, called the Cheo box, is cGAACnnnnATTcG.

The SOS regulon of *E. coli* is well characterized (Walker, 1996). In particular, it includes the *umuDC* operon involved in UV-mutagenesis. This operon has numerous plasmid homologues that also are subject to LexA regulation. The C-terminal domain of UmuD is homologous to the C-terminal domain of LexA, whereas the N-terminal domain of UmuC is homologous to the N-terminal domain of another *E. coli* protein, DinP, that has a weak candidate LexA binding site.

The known members of the *Bacillus subtilis* SOS regulon include *recA*, *ruvAB*, *dinB*, *dinC* operons. We have used the known DinR binding sites to construct a recognition profile for Cheo boxes and then applied the comparative approach to find new members of the *B. subtilis* SOS regulon.

A number of new candidate DinR binding sites have been identified. This list includes sites upstream of *yolD*, *yozL* and *yqjW* (Table 1). All these genes have two candidate Cheo boxes at a distance 28-30 nucleotide between the corresponding positions. Thus both sites are located approximately on the same side of the DNA helix.

Table 1. Predicted Cheo boxes of *Bacillus subtilis*. Right column: position relative to the gene start.

gene	1st site	spacer	2nd site
<i>yolD</i>	AcATcGAACtTtTGTTcTgg	10	AtgaAGAACgTtTGTTcTgT
<i>yozL</i>	AcATcGAACtTtTGTTcTga	10	AtAagGAACgTtTGTTcTgT
<i>yqjW</i>	AAAacGAACATactTTCGCa	8	AAAGcGAACATaaGTTcTtT

Gene *yqjW* is orthologous to *umuC*. Not surprisingly, there are no candidate Cheo boxes upstream of *yqjH*, which is the *B. subtilis* orthologue of *dinP*.

The situation with *yolD* and *yozL* is even more interesting. These two genes reside in prophage regions (Kunst et al., 1997). The genes around these sites form collinear chains (Fig. 1). There is a frameshift breaking a single gene into parts *yozK* and *yobH*. There are no clues as to whether this is a sequencing error or an inactive gene. Based on the small intergenic distances, it is likely that there are operons *yolD-uvrX* and *yozL-yozK-yobH*. The corresponding copies of the site pairs are considerably more similar than non-corresponding copies (Table 2). Thus these loci arose from a relatively recent duplication or integration of a phage.

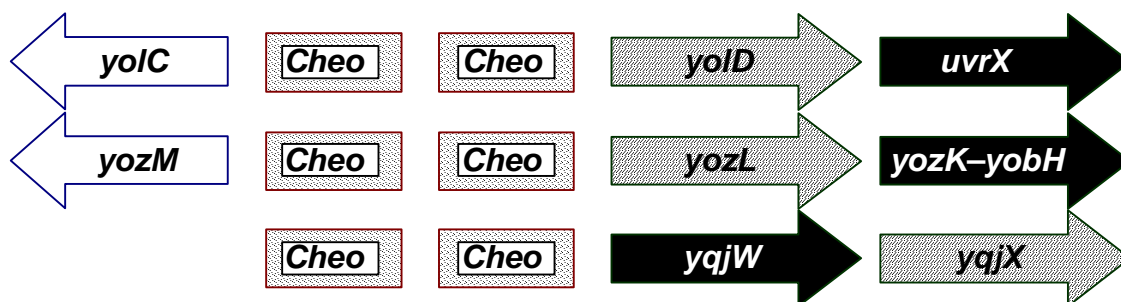


Figure 1. Genome of *Bacillus subtilis*: *yolDuvrX*, *yozLYobH* and *yqjWX* loci.

Table 2. Number of mismatches between prophage sites.

	<i>yoID-1</i>	<i>yoID-2</i>
<i>yoZL-1</i>	1	6
<i>yoZL-2</i>	5	2

Finally, we have noted that the upstream genes of the two prophage operons, *yoID* and *yoZL*, are homologous to the downstream gene of the chromosomal operon, *yqjX* (Fig. 2), whereas the downstream genes of the prophage operons, *uvrX* and *yoZKyobH* belong to the *umuC* family. Based on the concept of functional relatedness of co-localized genes (Overbeek et al., 1999) we can suspect that the proteins encoded by these small genes are involved in UV-mutagenesis.

```

YozL      MML-EQLIQLKQDLIDGSKVEKPSLDDKQIDEMDILVSEALEFNKELKFK
YoID      MMLPEHLTQLKQDLIDVSKIEKPSLDDQIEEMDILVSEALEFNKELQFK
YqjX      MFLPEHKQSLLEKRLKQKLQKPIIDDPDKLEEMNQTLCAAMEFAQDITVS
          *:* * : . * : . * : : * : * * . : : * : : . * : * * : : . .

YozL      LFNKGFVENVTGRV-HYINFEQQKLHVKDQNDNTVYINMNNIIRVIYND
YoID      LFNHNGFVENVTGRV-HYINFEQQKLHVKDQNDNTVYINMNNIIGVTYND
YqjX      CFQDGEIVCCTGKICRYEEFEKAVWIKGDE-DQLYKCLKLDQVLDIVL--
          * : . * : * * : : * : * * : * : * : : : : : : : :

```

Figure 2. Alignment of *yqjX*, *yoID* and *yoZL*

Thus we have demonstrated that the UV-mutagenesis system in *B. subtilis*, like its Gram-negative counterpart, is associated with mobile genomic elements. It is tempting to predict that the proteins YoID/YozL/YqjX are involved in activation of the respective UmuC-family proteins, similarly to UmuD activating UmuC in *E. coli*.

This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Kunst, F. et al. (1997) *Nature*, 390, 249-256.
2. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) *Proc. Natl. Acad. Sci.*, **96**, 2896-2901.
3. Walker, G.C. (1996) *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology. F.C.Neidhardt, ed. Wash. DC: ASM Press. Vol. 1, ch. 89, pp. 1400-1416.
4. Winterling, K.W., Chafin, D., Hayes, J.J., Sun, J., Levine, A.A., Yasbin, R.E. and Woodgate, R. (1998) *J. Bacteriol.*, 180, 2201-2211.

HrcA REGULATES NOT ONLY CHAPERONIN GENES

***Gelfand M.S., Mironov A.A.**

State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: comparative analysis, gene regulation, heat shock response, bacteria

Resume

The heat shock response in various bacteria is regulated by at least twelve diverse systems (Narberhaus, 1999). One of these systems is negative regulation by HrcA repressor binding to the CIRCE element (an inverted repeat of the form TTAGCACTC–N₉–GAGTGCTAA) (Zuber and Schumann, 1994). An unusual feature of this system is its conservation across very diverse taxa: Gram-positive and Gram-negative bacteria, cyanobacteria, chlamydias, and spirochaetes (Segal and Ron, 1996; Hecker et al., 1996). In all cases studied so far CIRCE element occurred upstream of operons containing chaperonin genes *groES*, *groEL*, *grpE*, *dnaK*, and *dnaJ* in various combinations (Hecker et al., 1996). The CIRCE-regulated operons often, but not always, include the regulator itself, although e.g. in *Caulobacter crescentus* the *hrcA* gene is regulated via a RpoH promoter rather than CIRCE element (Roberts et al., 1996).

We have analyzed distribution of the CIRCE element in completely sequenced bacterial genomes. All chaperonin operons and all genes with upstream CIRCE elements are shown in Fig. 1. The composition of the chaperonin operons and the CIRCE regulon is very diverse. In particular, autoregulation of the *hrcA* gene is not an obligatory feature. A specific feature of the mycoplasma genomes is the fact that there the CIRCE regulon includes not only chaperonins, but also genes for heat-shock proteases homologous to *lon* and *clpB* from *Bacillus subtilis*. In other bacteria orthologues of these proteases are regulated by other heat shock systems.

We have also analyzed selected incomplete genomes. A surprising observation is that in *Bordetella pertussis* a perfect CIRCE element is found upstream the gene highly homologous, and likely orthologous, to the *rpoH* gene from *Escherichia coli* encoding the heat shock sigma factor.

Thus the CIRCE regulon is much more diverse than previously thought. It is likely that careful comparative analysis of the heat shock systems will uncover more regulatory cascades and loops.

Acknowledgements

We are grateful to E. Koonin and Y. Kogan for useful discussions. This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Hecker, M., Schumann, W. and Volker, U. (1996) Mol. Microbiol., 19, 417-428.
2. Narberhaus, F. (1999) Mol. Microbiol., 31, 1-8.
3. Roberts, R.C., Toochinda, C., Avedissian, M., Baldini, R.L., Gomes, S.L. and Shapiro, L. (1996) J. Bacteriol., 178, 1829-1841.
4. Segal, G. and Ron, E.Z. (1996) FEMS Microbiol. Lett., 138, 1-10.
5. Zuber, U. and Schumann, W. (1994) J. Bacteriol., 176, 1359-1363.

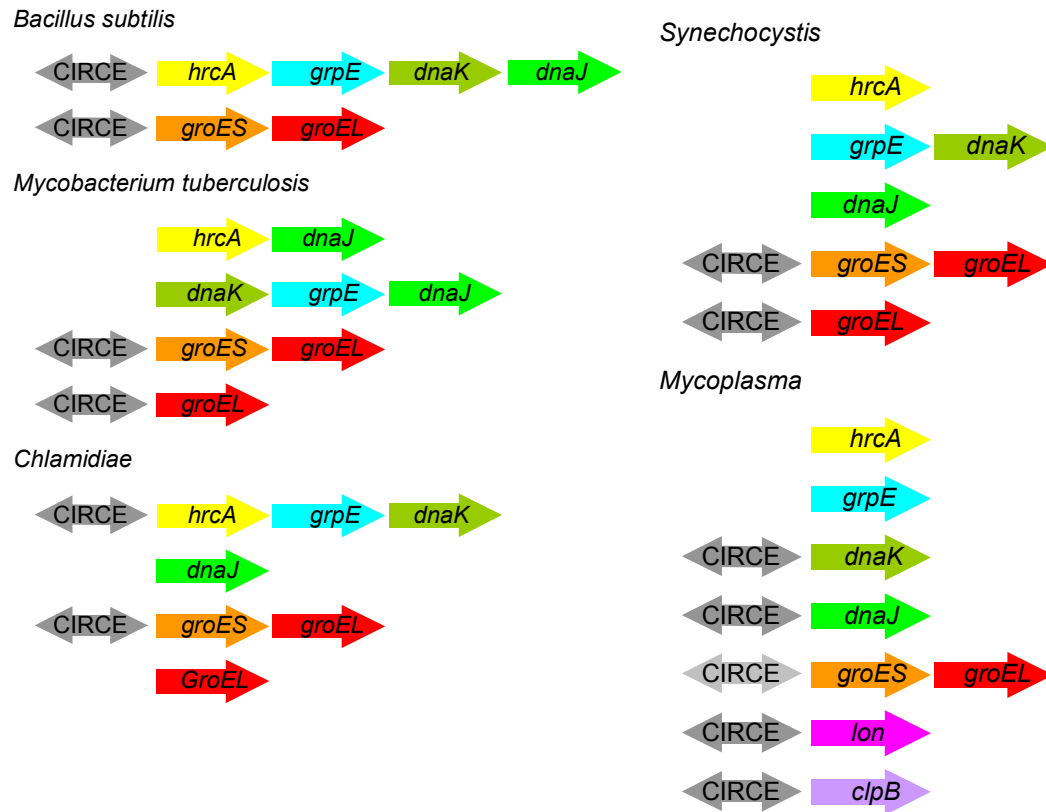


Figure 1. CIRCE regulons and chaperonin operons. Colored arrows: genes; grey arrows: CIRCE element.

SOFTWARE FOR ORTHOLOGY ANALYSIS IN COMPLETE BACTERIAL GENOMES

*Baitalyuk M. V., Novichkov P. S., *Gelfand M. S., Mironov A. A.*

State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: orthologous genes, comparative analysis, bacterial genomes

Resume

Many problems of comparative genomics involve analysis of orthologous genes. Rigorous determination of orthology relationships requires construction of a large number of phylogenetic trees. A reasonable approximation is to consider bidirectional best hits (BETs) as done, for example, in the COG system (Tatusov et al., 1997). Two genes g from genome A and h from genome B form a BET if the similarity between these genes $s(g,h)$ exceeds the similarity for any other choice of either member of the pair: $s(g,h) > s(g,y)$ and $s(g,h) > s(x,h)$ for any $x \neq g$ from A and $y \neq h$ from B .

GenomeExplorer (Mironov et al., 2000) has convenient tools for manual analysis of orthologues. However, large scale analyses require tools for automated processing of gene complements. Orthologator is a set of programs for identification of BET pairs in bacterial genomes. It uses BLAST (Altschul et al., 1990) for identification of homologues.

Analysis of a pair of genomes involves the following steps:

Task	Input	Output	Time (PC)
1. Proteomes from genome annotation	genomes in the GenBank or EMBL formats	proteome in the FASTA format retaining gene names, positions and other necessary data	few seconds
2. BLAST pre-processor	proteome in the FASTA format	BLAST index files	few seconds
3. BLAST	proteome in the FASTA format, index files	homologs for each protein in the other genomes ordered by decrease of similarity	twice about 40 min.
4. Identification of BETs	lists of homologues for each gene (from both genomes)	list of BETs	about 30 min.

The two intermediate steps are performed by the standard software available at the BLAST ftp site (BLAST). Pre- and post-processing is done by our customized software. Complete processing of two genomes takes about 2 hours.

Further development will be directed towards the use of ORTHOLOGATOR in large scale analysis of regulatory patterns, creation of a database of orthologues in complete bacterial genomes, correction of annotation errors (wrong gene starts), and development of more sensitive tools for analysis of orthologous domains, gene fusions, and paralogous families. It will be used in large scale analyses of bacterial regulation (e.g. Mironov and Gelfand, 2000).

Acknowledgements

This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Altschul, S.F., Gish, W., Miller, A., Myers, E.W., Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403-410.
2. BLAST. <ftp://ncbi.nlm.nih.gov/blast/>
3. Mironov, A.A., and Gelfand, M.S. (2000) Automated comparative analysis of regulatory patterns: sugar metabolism and transport systems in gamma purple bacteria. (This volume).
4. Mironov, A.A., Vinokurova, N.P. and Gelfand, M.S. (2000) GenomeExplorer: software for analysis of complete bacterial genomes. (This volume).
5. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) *Science*, **278**, 631-637.

FUNCTIONAL GENOMICS: AN ALGORITHMIC PERSPECTIVE

Califano A.

IBM Computational Biology Center, T.J. Watson Research Center, Yorktown Heights, NY, USA
e-mail: acal@us.ibm.com

Keywords: genomics, pattern analysis, gene expression, microarray data

Resume

Several biological mechanisms manifest themselves through a rich variety of genomic patterns. Protein motifs, TATA boxes, and gene expression clusters are but a few of the more obvious examples.

Unfortunately the identification of biologically relevant patterns, especially of subtle, less obvious ones, is still a challenging task. This paper discusses advances in the deterministic and exhaustive identification of patterns and in their classification based on statistical significance criteria. Several biologically relevant applications of pattern discovery coupled with a statistical analysis framework will be discussed.

First, we will show how pattern analysis can be used to identify, with high probability, regions of protein sequences that are responsible for their functional or structural properties. This work has been validated by exhaustive analysis and comparison of the PROSITE database and of the GPCR superfamily. These patterns can then be used to seed sensitive and accurate functional and structural sequence annotation algorithms.

Similarly, we will discuss how gene expression patterns can be used for the accurate and sensitive prediction of cell phenotype from microarray data. Given gene expression measurements for a set of phenotype cells, e.g. a specific cancer morphology, and for a set of control cells, pattern discovery can rapidly identify significant gene expression clusters, without having to explicitly explore the exponential search space of all possible combinations of genes and experiments. Statistically significant clusters can then be used to build complex multivariate classifiers. Results on the analysis of several morphological and drug related phenotypes in 60 human cancer cell lines will be discussed.

Introduction

Whenever Nature finds a "recipe" to grant differential fitness to an organism, chances are that such recipe will be conserved through evolution. At the molecular level, this means that biological sequences, belonging sometimes to widely distant species, will likely share common motifs, or highly conserved, ungapped regions of a protein or DNA sequence [1]. Similarly, when genes are expressed in a coregulated fashion within a cell, gene expression patterns arise that can be measured by a variety of microarray devices [2].

As a result, the identification of sparse patterns in biological databases is becoming a very transited venue within the bioinformatics community. Pattern databases such as PROSITE [3] are examples of this trend. And pattern and association discovery algorithms are becoming increasingly relevant in computational biology [4].

In recent years a number of interesting algorithms have emerged. These are divided in two main categories: statistical algorithms, such as the Gibbs Sampler [5] and MEME [6], which use heuristics to improve efficiency, and deterministic algorithms, such as SPLASH [7], Pratt [8], and Teiresias [9]. A statistical framework has also been proposed to help determine the statistical significance of discovered patterns [10].

Based on several benchmark criteria [7], SPLASH has emerged as an extremely efficient and versatile algorithm that can discover patterns and associations defined through a distance or similarity metric on either discrete or continuous data. SPLASH has been applied to the discovery of functionally relevant motifs in protein families [11], as well as to orphan annotation, and to phenotype prediction from gene expression microarray data [12]. In this paper we give a summary of several applications of SPLASH to biologically relevant problems.

Functional analysis of protein families

Rapid advancement in sequencing technology and exponential growth in genomic databases are spurring the development of techniques for the identification of sequence motifs and sequence classification. This is commonly accomplished by defining sequence signatures that distinguish all members of the respective family from the complete sequence database and allow the classification of new proteins into these families [13]. Definition of the sequence signatures can range from simple consensus patterns in sequences, often called sequence motifs, to more elaborate and rigorous descriptors, termed position specific scoring matrices or profiles, to Hidden Markov Models [14]. Currently, there are several well curated and established compilations of sequence motifs, such as PROSITE [3], PRINTS [15], PFAM [16], and BLOCKS [17]. The latter is a comprehensive non-redundant database of blocks derived from several databases of sequence motifs and profiles.

Among these databases, PROSITE is especially relevant because of the high biological significance of the reported patterns. The PROSITE database version 15.0 contains extensively annotated collections of 1352 motifs grouped into 1014 protein families. Each PROSITE entry stems from a set of protein sequences grouped by an expert, using biological information which is provided as documentation. For almost all entries, PROSITE provides a sequence motif that characterizes the functionally relevant residues of a protein family. These are obtained by selecting regions of sequences that have a documented functional significance and by performing multiple sequence alignment over these selected regions to identify consensus patterns. Beyond the inherent utility of the simple consensus patterns, PROSITE also serves as a very useful database of seed locations to guide the automated development of more complex descriptors, as in BLOCKS. Similar seed entries are created semi-manually for other curated databases as well, such as PRINTS and PFAM. Efforts are underway to integrate these into a single resource called InterPro (www.ebi.ac.uk/interpro). Curation of these databases is a labor-intensive task, which is increasingly challenged by the rapid explosion of data in genomic repositories. In [11], it has been shown that SPLASH can automatically reproduce the great majority of results in the PROSITE database, often improving on the selectivity and specificity of the motifs. These results support the use of automatic pattern discovery, coupled with a statistical analysis framework, for the automatic identification of protein regions that carry important functional or structural roles. We have analyzed each one of the 974 protein families associated with one or more PROSITE motifs by characterizing three important parameters: 1) the sensitivity and specificity of the automatically discovered patterns 2) the correlation between the statistical significance of a pattern and its performance, in terms of sensitivity and specificity 3) the overlap between automatically discovered patterns and PROSITE patterns. We emphasize that the entire process is automated and performed identically for all PROSITE families, with no effort to tune parameters for specific families and no utilization of PROSITE patterns. The tabulated analysis is available at www.research.ibm.com/spat.

Figure 1 plots a scatter graph of the values of Δn_{fn} and Δn_{fp} across all 974 families. These are respectively the difference in false negatives (selectivity) and false positives (specificity) between SPLASH and PROSITE.

Negative values better SPLASH performance. The histogram of the scatter plot for both Δn_{fn} and Δn_{fp} is also reported. This shows that most patterns are accumulated in a few bins around the center of the plot. The associated table shows that for 76.3% of the families (thick rectangle or lower left quadrant) SPLASH patterns (or S-Pattern) perform at least as well as the corresponding PROSITE patterns (or P-Patterns). For 28% of the families, the S-patterns strictly outperform the P-patterns. If the ranking is done based on the sum of false positives and false negatives, The number of S-patterns that perform as well or better than P-patterns increases to 80.6%.

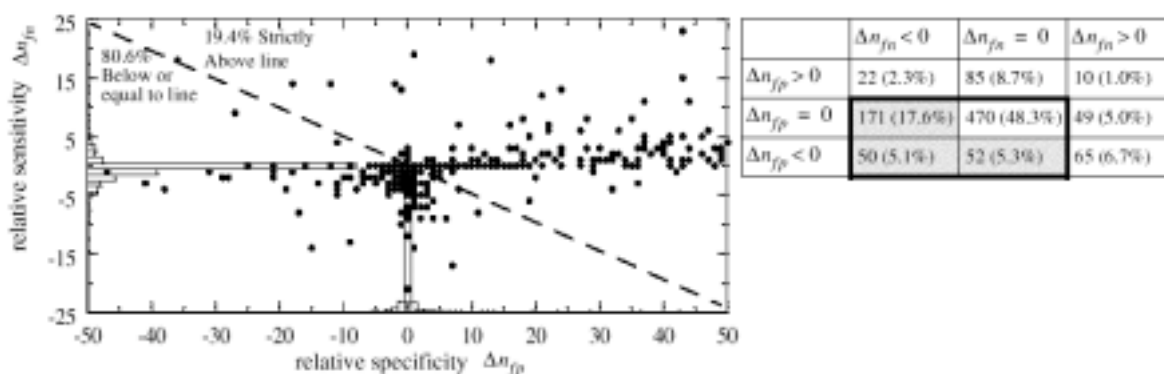


Figure 1. Scatter plot of Δn_{fn} vs. Δn_{fp} for all 974 PROSITE families. The separate histograms across the Δn_{fn} and Δn_{fp} axis are on a scale 0 to 1800. The center of the plot is at 900. 580 patterns are in the (0,0) bin. 74% of the patterns are in the lower left quadrant.

Fig. 2 plots how many of the top scoring patterns on the x axis (in percent) have a ratio of overlap larger than a given percent (on the y axis). The three lines correspond to (a) All families with at least 20 elements, (b) all families with at least 60 elements, and (c) all families with at least 120 elements. About 72% of the S-patterns, for families with 20 or more elements, overlap at least 50% with their corresponding PROSITE patterns. That is, they tend to identify the same region of the protein sequence. This ratio increases to about 77% for families with at least 120 elements. The relatively small improvement hints that about 20 sequences may be sufficient to identify biologically relevant regions from purely statistical criteria. Overall, Figs. 1 and 2 show a remarkable relationship between patterns that are identified based on purely statistical criteria and those in PROSITE, which are assumed to have a significant biological role. This suggests that patterns generated by our methodology would be useful as seeds for further refinement with PSSM or profile HMM.

Gene expression analysis

Recent advances in DNA microarray technology [18, 2] are for the first time offering us exhaustive snapshots of some of the cell's most intimate genetic mechanisms, and creating a unique opportunity to improve our knowledge of the cellular machinery. Previous work on DNA microarrays has concentrated primarily on the identification of coregulated genes [19, 20, 21, 22, 23, 24] to decipher the underlying structure of genetic networks and/or the molecular classification of diseased tissues [25, 26, 27].

A few data analysis schemes [26, 27, 28, 29, 30, 31] have been proposed aimed at extracting useful information from microarray data. Two general strategies have been followed, based on either supervised [27, 31] or unsupervised [26, 28, 29, 30] learning algorithms. In these approaches, the quantitative expression of n genes in k samples are considered as either n vectors in k -dimensional space or as k vectors in n -dimensional space, and various metrics are used. Supervised learning approaches are designed to assess whether cells belong to a class characterized by a known phenotype or not, based on their gene expression profiles. This is also known as the cell phenotype prediction problem. Typically, these algorithms are first trained on two example sets: a *positive example set* or *phenotype set*, with data for cells characterized by a predefined phenotype, and a *negative example set* or *control set*, with data for cells that do not exhibit that phenotype. In [27], a vector of "marker genes," or signature, is used for classification. Marker genes are selected based on the individual discriminative power of their statistics. In [31] it is shown that Support Vector Machine (SVM) based algorithms outperform other standard learning algorithms.

In [12] we have introduced a novel supervised learning algorithm for cell phenotype prediction which differs on several counts. First, rather than relying on a unique best-fit model, which optimally discriminates between the phenotype and control set, the new algorithm uses multiple, optimally discriminative models. It is shown that this improves the analysis of complex phenotypes when these are mixtures of multiple, simpler sub-phenotypes.

Also, our analysis selects genes based on their collective discriminative power, rather than on their individual one. This significantly increases the signature dimensionality and, as shown by the results, its discriminative power. Our aim, therefore, is to find gene vectors whose expression is tightly clustered in a subset of the phenotype set but not otherwise clustered over any significant subset of the control set. Given a microarray with N_g genes and N_e experiments, there are 2^{N_g} potential vectors and 2^{N_e} potential subspaces.

The approach has four steps. First, we transform the gene expression axis, on a gene by gene basis, using a non-linear metric. This metric is such that the distance between two expression values in the phenotype set is equal to the integral of the gene expression probability density estimated from the control set. In other words, expression values that are highly probable are spread apart while unlikely values are compressed together.

We then use the SPLASH [7] deterministic pattern discovery algorithm to find any sub-matrix of the transformed $N_g \times N_e$ gene expression matrix, such that the expression values in each column are tightly bound.

Each discovered pattern is defined by a subset of the genes and by the subset of the phenotype set over which these genes are tightly clustered. Worst case analysis shows that the number of such gene expression patterns could be exponential. In practice, however, their number tends to be relatively small. For the analysis reported in this paper, the complete set of expression patterns is found in just seconds on a standard workstation.

Patterns that are not statistically significant are discarded, using an analytical model, which is shown to be in extremely good agreement with experimental results. Finally, an optimal, mostly orthogonal pattern subset is chosen using a greedy set covering algorithm. For typical cases, this set is small, consisting of one to three signatures. Each signature in the set is used to build an independent "sub-phenotype" multivariate probability density model. A "control" probability density model is also built for each gene from the samples in the control set. A standard classification scheme based on the ratio of the two probability densities is then used.

We have applied our supervised learning scheme to the classification of 60 human cancer cell lines [32], from data obtained with Affymetrix HU6800 GeneChips. Cell lines have been analyzed according to a variety of phenotypes, such as cancer morphology, a mutations in the p53 oncogene, or the efficiency of a given anti-cancer compound. Complex phenotypes, such as the p53 related one, are likely to be mixtures of simpler unknown sub-phenotypes at the molecular level, each one characterized by a possibly independent gene marker vector. We show that these complex cases perform poorly with methods that rely on a single model, as truly there is no single model that describes the entire set.

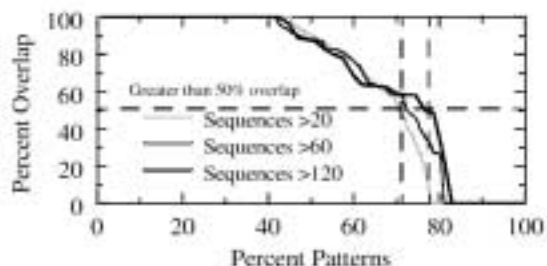


Figure 2. Cumulative of patterns with overlap better than y .

Systematic comparative analysis of false positives (FP%) and false negatives (FN%) rates has been performed using a standard leave-one-out cross-validation scheme. Results for melanoma, a p53 mutation, and the efficiency of Chlorambucil are shown in Fig. 3, for our method (PD) that of [27] (Gene By Gene), and that of [31] (SVM). The sum of false positive and false negatives is shown on the y-axis. The classifier threshold is shown on the x-axis. For simple phenotypes, such as a specific cancer morphology, a unique model is sufficient.

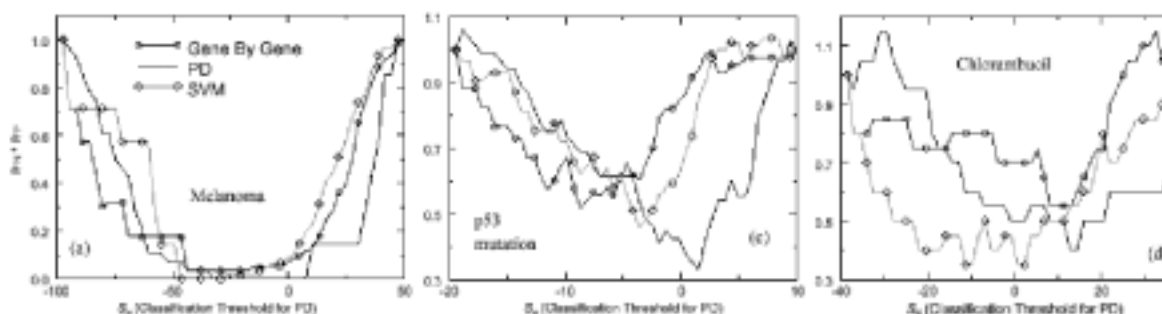


Figure 3. Comparative analysis of phenotype prediction algorithm performance.

In that case, performance is similar for all methods. In more complex cases such as with the p53 related phenotype, where multiple models clearly emerge, the new technique outperforms the others.

References

- Bailey T.L. and Gribskov M., "Methods and statistics for combining motif match scores." *J.Comp.Bio.* **5**, 211-221, (1998)
- Brown, P.O., and Botstein, D. "Exploring the new world of the genome with DNA microarrays", *Nature Genetics Suppl.*, **21**, Jan 99, 33-37 (1999).
- Hoffman K., Bucher P., Falquet L., and Bairoch A. "The PROSITE database, its status in 1999" *Nucleic Acids Research* **27**, 215-219 (1999).
- Brazma A. et al.: "Approaches to the Automatic Discovery of Patterns in Biosequences" *J.Comp.Bio.* 5(2):279-305, (1998)
- Neuwald, Liu, & Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats" *Protein Science* 4:1618-1632, (1995)
- Bailey T.L. and Elkan C. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers" in *Proc. of 2nd ISMB Conf.*, 28-36, AAAI Press, Menlo Park (1994)
- Califano A. "SPLASH: structural pattern localization analysis by sequential histograms", *Bioinformatics*, 16(4), 341-357, (2000).
- Jonassen I., Collins J.F., Higgins D.G. "Finding flexible patterns in unaligned protein sequences" *Protein Science* 4:1587-1595, (1995)
- Rigoutsos I. and Floratos A. "Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm," *Bioinformatics* 14(1):56-67, (1998)
- Stolovitzky G. and Califano A. "Pattern Statistics in Biological Datasets," to be communicated to *J.Comp.Bio.*, available at www.research.ibm.com/topics/popups/deep/math/html/statistics.pdf (1999)
- Hart R., Royyuru A.K., Stolovitzky G., and Califano A. "Systematic and Automated Discovery of Patterns in PROSITE Families", in *Proceedings of Fourth Annual International Conference on Computational Molecular Biology*, Tokyo, 2000.
- Califano A., Stolovitzky G. and Tu Y., "Analysis of Gene Expression Microarrays for Phenotype Classification", in *Proceedings of the 8th Symposium on Intelligent Systems for Molecular Biology*, San Diego, 2000.
- Bork P., Koonin E. V. "Protein sequence motifs" *Curr. Opin. Struct. Biol.* 6: 366-376 (1996).
- Durbin R., Eddy S., Krogh A. and Mitchison G. "Biological sequence analysis: probabilistic models of protein and nucleic acids". Cambridge University Press, 1998.
- Attwood T.K., Flower D.R., Lewis A.P., Mabey J.E., Morgan S.R., Scordis P., Selley J.N. and Wright W. "PRINTS prepares for the new millenium" *Nucleic Acids Research* 27: 220-225 (1999).
- Bateman A., Birney E., Durbin R., Eddy S.R., Finn R.D. and Sonnhammer E.L.L.. "Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins" *Nucleic Acids Research* 27: 260-262 (1999).
- Henikoff S., Henikoff J.G. and Pietrovski S. "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations" *Bioinformatics* 15: 471-479 (1999).
- Lockhart D. J., Dong H. et al. "Expression monitoring by hybridization to high-density oligonucleotide arrays" *Nat Biotechnol* 14(13): 1675-80, (1996).
- DeRisi J., Penland L. et al. "Use of a cDNA microarray to analyse gene expression patterns in human cancer" *Nat Genet* 14(4): 457-60, (1996).
- Wodicka L., H. Dong, et al. "Genome-wide expression monitoring in *Saccharomyces cerevisiae*" *Nat Biotechnol* 15(13): 1359-67, (1997).

21. Cho R.J., Campbell M.J. et al. "A genome-wide transcriptional analysis of the mitotic cell cycle" *Mol Cell* 2(1): 65-73, (1998).
22. Chu S., DeRisi J. et al. "The transcriptional program of sporulation in budding yeast" *Science* 282(5389): 699-705, (1998).
23. Iyer V.R., Eisen M.B. et al. "The transcriptional program in the response of human fibroblasts to serum" *Science* 283(5398): 83-7, (1999).
24. DeRisi J. L., Iyer V.R. et al. "Exploring the metabolic and genetic control of gene expression on a genomic scale" *Science* 278(5338): 680-6, (1997).
25. Perou Ch., Jeffrey S.S. et al. "Distinctive gene expression patterns in human mammary epithelial cells and breast cancers" *Proc Natl Acad Sci U S A* 96(16): 9212-7, (1999).
26. Alon U., Barkai N., Notterman D.A., Grish K., Ybarra S., Mack D. and Levine A.J. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays" *Proc Natl Acad Sci U S A* 96(12):6745-50, (1999).
27. Golub T.R., Slonim D.K. et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring" *Science*, 286, 531-7, (1999).
28. Eisen M.B., Spellman P.T. et al. "Cluster analysis and display of genome-wide expression patterns" *Proc Natl Acad Sci U S A* 95(25): 14863-8, (1998).
29. Tamayo P., Slonim D. et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." *Proc Natl Acad Sci U S A* 96(6): 2907-12.
30. Ben-Dor A and Yakhini Z. "Clustering Gene Expression Patterns" *Proc. of the 3rd International Conference on Computational Molecular Biology*, April 1999, 33-42, (1999).
31. Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet C., Ares M. and Haussler D. "Support Vec-tor Machine Classification of Microarray Gene Expression Data", University of California Technical Report USCC-CRL- 99-09. (1999). (Available at: <http://www.cse.ucsc.edu/research/compbio/genex>).
32. Weinstein J.N., Myers T.G. et al. "An information-intensive approach to the molecular pharmacology of cancer" *Science* 275(5298): 343-9, (1997).

COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: TRANSCRIPTION REGULATORY SITES IN ARCHAEA

¹*Gelfand M.S., ²Koonin E.V., ¹Mironov A.A.

¹State Scientific Center GosNII Genetika, Moscow, 113545, Russia,

²National Center for Biotechnology Information, Bethesda, USA

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: transcription regulation, regulons, profile search, comparative analysis, Archaea

Introduction

Very little is known about regulation of transcription in Archaea. The basal transcription machinery of the archaea is closely related to that of eukaryotes (Bell and Jackson, 1998). On the other hand, archaeal genomes encode a large number of proteins containing helix-turn-helix motifs and resembling eubacterial transcriptional regulators (Aravind and Koonin, 1999). Only a few sets of co-regulated genes have been characterized experimentally, in particular, heat-shock genes in *Haloferax volcanii* (Thompson et al., 1999), nitrogen-fixation genes in *Methanococcus maripaludis* (Cohen-Kupiec et al., 1997), gas-vehicle genes in *Halobacterium salinarium* (Pfeifer et al., 1997), and bacteriorhodopsin genes in *Halobacterium* sp. (Baliga and DasSarma, 1999).

We have applied the recently developed comparative approach (Gelfand and Mironov, 1998; Mironov et al., 1999; Gelfand et al., 1999) to analysis of several archaeal regulons. It is based on the assumption that regulons (sets of co-regulated genes) are conserved in related genomes. Thus reliable predictions can be made even with weakly specific rules: true sites consistently occur upstream of orthologous genes, whereas false positives are scattered at random.

We have used several variants of the basic technique, based on availability of experimentally determined regulatory sites. At that, a set of experimentally determined heat shock promoters was used to derive the recognition rule for the heat shock regulon, comparison of upstream gene regions from various archaea was used to construct the nitrogen fixation box profile, and comparison of upstream regions of genes encoding purine metabolism enzymes in the three *Pyrococcus* species was used for construction of the purine box profile.

Data and Methods

Genomes of *Methanococcus jannaschii* (Bult et al., 1996), *Methanobacterium thermoautotrophicum* (Smith et al., 1997), *Archaeoglobus fulgidus* (Klenl et al., 1997), *Pyrococcus horikoshii* (Kawarabayashi et al., 1998), were downloaded from GenBank (Benson et al., 1999). Unannotated genome of *Pyrococcus furiosus* was downloaded from Utah Genome Center. Unannotated genome of *Pyrococcus abyssi* was downloaded from Genoscope.

Positional nucleotide weights were defined as previously described (Mironov et al., 1999):

$$W(b,k) = \log(N(b,k)+0.5) - 0.25 \sum_{i=A,C,G,T} \log(N(b,i)+0.5)$$

where $N(b,k)$ is the count of nucleotide b at position k , the base of the logarithm is chosen so that the random deviation of the score

$$Z(b_1 \dots b_L) = \sum_{k=1 \dots L} W(b_k, k)$$

on random oligomers equals 1.

A simple iterative procedure is performed in order to construct a profile from a set of upstream gene fragments. Weak palindromes are selected in each region. Each palindrome is compared to all palindromes, and the palindromes most similar to the initial one (at most one from each region) are used to make a profile. These profiles are used to scan the set of palindromes again, and the procedure is iterated until convergence. Thus a set of profiles is constructed. The quality of a profile is defined as its information content (Schneider et al., 1986)

$$I = \sum_{k=1 \dots L} \sum_{i=A,C,G,T} f(i,k) \log(f(i,k) / 0.25).$$

The best profile is used as the recognition rule.

Identification of orthologues and site search were performed using the program GenomeExplorer (<http://www.anchor.gen.com>).

Results and Discussion

Heat shock regulons in *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus* and *Pyrococcus horikoshii* were analyzed using a training set of published heat shock promoters (Thompson and Daniels, 1998). The HSP profile was derived with consensus CCGAAAAGTTTATATAGAA. The following genes were consistently preceded by candidate HSP sites: HSP60-class chaperones (thermosome components), small heat shock proteins (hsp20), and AAA+ superfamily ATPases, including a family of 26S protease regulatory subunit proteins (Table 1). However, the most interesting fact is likely heat shock regulation of the archaeal homologues of the H4 histone. There are two families of these proteins. All of the genes encoding these proteins with a single exception (an extrachromosomal gene from *M. jannaschii*) contain candidate HSP in upstream regions. Euryarchaeal histones can mediate topological changes in DNA during thermal stress as proposed for the bacterial histone-like HU proteins (Ogata et al., 1997; Mizushima et al., 1997) and the histone-like protein Sso7d from the crenarchaeon *Sulfolobus solfataricus* (Lopez-Garcia et al., 1998).

Table 1. Heat shock regulons in archaea. Boldface: genes with candidate heat shock promoters. Boldface italics: genes in operons with candidate heat shock promoters.

	<i>A. fulgidus</i>	<i>M. thermo-autotrophicum</i>	<i>M. jannaschii</i>	<i>P. furiosus</i>
HSP60	AF2238 AF1451	MTH794 MTH218	MJ0999	PH0017
HSP20	AF1971 AF1296	MTH1366 MTH859	MJ0285	PH1842
AAA+ ATPase	AF1297 AF2098	MTH1639	MJ1156	PH1840
ATP-dependent 26 protease subunit	AF1976	MTH728	MJ1176	PH0201
histone H4 homologue, subfamily 1	AF0337	MTH821	MJ0932 MJ1258	PHs046
histone H4 homologue, subfamily 2	AF1493	MTH254 MTH1696	MJ0168 MJECL29	PHs051

The initial profile of the nitrogen fixation box (NIF) was derived from comparison of upstream regions of *nifH* and *glnA* genes of *Methanococcus* and *Methanobacterium* species from GenBank excluding the complete genomes. The consensus of the NIF box is TCGGAAATATATTTCCGA. The profile was then used to scan the genomes of *M. thermoautotrophicum* and *M. jannaschii* (*A. fulgidus* and *Pyrococcus* spp. do not fix nitrogen). Candidate sites were found upstream of *glnA* genes and *nif* operons, as well as in regulatory regions of *glnK* (glutamine synthetase regulator) and *amtB* (ammonium transporter) genes (Fig. 1). The latter two genes have been recently found to be transcriptionally linked in a variety of bacterial and archaeal genomes (Thomas et al., 2000).

M. thermoautotrophicum



M. jannaschii

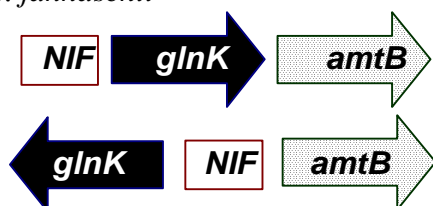


Figure 1. *glnK-amtB* loci of *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii*.

No experimental data are available about regulation of purine metabolic pathway genes in archaea. Thus a different approach was applied to analysis of purine regulon. The initial set of genes in the three *Pyrococcus* genomes was identified by analysis of the metabolic map. The procedure of profile construction was applied to the three samples independently. The resulting profile (PUR box) was weak, but the consensus TTTAACATATATATGTTAAA was the same in all three cases. When the profile was used to scan the genomes, we noted that the upstream regions of genes encoding the purine metabolism enzymes contained two candidate PUR boxes with a spacer of fixed length 22-23 base pairs. The rule requiring presence of two boxes was absolutely specific: thus defined signal appeared upstream of purine metabolism genes only. One additional gene having this signal, PH1162 in *P. horikoshii* and its orthologues in the other two *Pyrococcus*

genomes, encodes a protein orthologous to purine permeases from eubacteria, identified in (Gelfand and Mironov, 1998; Mironov et al., 1999).

These three examples demonstrate that the comparative approach allows one not only to transfer information about known regulons to new genomes, but also to identify new members of regulons and even to find regulatory patterns in the absence of any experimental information. For more details see (Gelfand et al., 2000).

Acknowledgements

This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Aravind, L. and Koonin, E.V. (1999) *Nucleic Acids Res.*, **27**, 4658-4670.
2. Baliga, N.S. and DasSarma, S. (1999) *Mol. Microbiol.*, **181**, 2513-2518.
3. Bell, S.D. and Jackson, S.P. (1998) *Trends Microbiol.*, **6**, 222-228.
4. Benson, D.A. et al. (1999) *Nucleic Acids Res.*, **27**, 12-17.
5. Bult, C.J. et al. (1996) *Science*, **273**, 1058-1073.
6. Cohen-Kupiec, R., Blank, C. and Leigh, J.A. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 1316-1320.
7. Gelfand, M.S. and Mironov, A.A. (1998) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *1st Conf. BGRS-98*, vol. **1**, pp. 147-149.
8. Gelfand, M.S., Mironov, A.A., Jomantas, J., Kozlov, Yu.I. and Perumov, D.A. (1999) A conserved RNA structure element involved in the regulation of bacterial riboflavin biosynthesis genes. *Trends Genet.*, **15**, 439-442.
9. Gelfand, M.S., Koonin, E.V., Mironov, A.A. (2000) Prediction of transcription regulatory sites in *Archaea* by a comparative-genomic approach, *Nucleic Acids Res.*, **28**, 695-705.
10. Genoscope (National Center for Sequencing, France). <http://www.genoscope.cns.fr>
11. Kawarabayashi, Y. et al. (1998) *DNA Res.*, **5**, 145-155.
12. Klenk, H.P. et al. (1997) *Nature*, **390**, 364-370.
13. Lopez-Garcia, P., Knapp, S., Ladenstein, R. and Forterre, P. (1998) *Nucleic Acids Res.*, **26**, 2322-2328.
14. Mironov, A.A., Koonin, E.V., Roytberg, M.A. and Gelfand, M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **27**, 2981-2989.
15. Mizushima, T., Kataoka, K., Ogata, Y., Inoue, R. and Sekimizu, K. (1997) *Mol. Microbiol.*, **23**, 381-386.
16. Ogata, Y., Inoue, R., Mizushima, T., Kano, Y., Miki, T. and Sekimizu, K. (1997) *Biochim. Biophys. Acta*, **1353**, 298-306.
17. Pfeifer, F., Kruger, K., Roder, R., Mayr, A., Ziesche, S. and Offner, S. (1997) *Arch. Microbiol.*, **167**, 259-268.
18. Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.*, **188**, 415-431.
19. Smith, D.R. et al. (1997) *J. Bacteriol.*, **179**, 7135-7155.
20. Thomas, G., Coutts, G. and Merrick, M. (2000) *Trends Genet.*, **16**, 11-14.
21. Thompson, D.K. and Daniels, C.J. (1998) *Mol. Microbiol.*, **27**, 541-551.
22. Thompson, D.K., Palmer, J.R. and Daniels, C.J. (1999) *Mol. Microbiol.*, **33**, 1081-1092.
23. Utah Genome Center (University of Utah, USA). <http://www.genome.utah.edu>

InterPro AS A NEW TOOL FOR WHOLE GENOME ANALYSIS. A COMPARITIVE ANALYSIS OF *MYCOBACTERIUM TUBERCULOSIS*, *BACILLUS SUBTILIS* AND *ESCHERICHIA COLI* AS A CASE STUDY

***Mulder N.J., Fleischmann W., Apweiler R.**

EMBL Outstation – European Bioinformatics Institute, Cambridge, United Kingdom

e-mail: Mulder@ebi.ac.uk

*Corresponding author

Keywords: database, protein, domain, function, family, repeat, computer tool

Resume

Motivation:

Several pattern-recognition methods have evolved to address different protein sequence analysis problems, resulting in rather different and mostly independent databases. InterPro was developed as a new integrated documentation resource for protein families, domains and functional sites, to rationalise the complementary efforts of the PROSITE, PRINTS, Pfam and ProDom database projects. With the rapid emergence of uncharacterised sequences from genome sequencing projects, the emphasis has moved towards automatic annotation of sequences and the study of whole genomes rather than single genes. InterPro has applications in computational functional classification of newly determined sequences lacking biochemical characterisation, and in comparative genome analysis.

Results:

The first release of InterPro was built from Pfam 5.0, PRINTS 25.0 and PROSITE 16.0. and contains nearly 3000 entries, representing families, domains, repeats and PTMs. Overall, InterPro entries match more than 300,000 sequences in SWISS-PROT and TrEMBL. This new resource provides an integrated view of the pattern databases, and provides an intuitive interface for text- and sequence-based searches.

InterPro was used for whole proteome analysis of the pathogenic microorganism, *Mycobacterium tuberculosis*, and comparison with the predicted protein coding sequences of the complete genomes of *Bacillus subtilis* and *Escherichia coli*. At present, 64.5% of the non-redundant proteome set of *M. tuberculosis* SWISS-PROT and TrEMBL proteins lack biochemical classification, and are labelled as hypothetical. 55.6% of all the *M. tuberculosis* proteins in the set matched InterPro entries, and these could be classified according to function. The comparative genome analysis with *B. subtilis* and *E. coli* provided information on the most common protein families and domains, the most highly represented families, and the representation of different regulatory protein families in each organism.

Availability:

The database is accessible for text- and sequence-based searches at <http://www.ebi.ac.uk/interpro/>. The InterPro flatfile may be retrieved from the EBI anonymous-ftp server <ftp://ftp.ebi.ac.uk/pub/databases/interpro>.

Introduction

Pattern databases have become vital tools for identifying distant relationships in novel sequences and hence for inferring protein function. Currently, the most commonly used pattern databases include PROSITE (Hofmann et al., 1999); Pfam (Bateman et al., 2000); and PRINTS (Attwood et al., 2000). These methods provide tools for identifying sequence relationships and inferring protein function, however, they have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods. In an attempt to address some of these issues, we have developed InterPro, and found applications for the database in developing automatic methods for the functional classification and annotation of raw sequence data from genome sequencing projects, and for whole proteome analysis. The availability of complete genome sequences provides a new approach for the study of different organisms and analysing, for example, the relationship between pathogenic and non-pathogenic organisms, which may be useful in determining virulence and disease-related proteins and thus identifying potential drug targets.

Source databases and methods

The first release of InterPro (Release 1.0, March 2000) was built from Pfam 5.0 (2,008 domains), PRINTS 25.0 (1,260 fingerprints) and PROSITE 16.0 (1,370 families). Flat-files submitted by each of the groups were systematically merged and dismantled. Where relevant, family annotations were amalgamated, and all method-

specific annotation separated out. Different types of parent-child relationship were evident both between entries in the same database, and between entries in different databases, leading us to recognise 'sub-types' and 'sub-strings'. All recognisably distinct entities were assigned unique accession numbers (which take the form IPRxxxxxx). An InterPro entry contains the list of member database signatures, HMMs, profiles or fingerprints associated with the entry, an abstract describing the domain, repeat, family or PTM, and links to a tabular or graphical view of the matches to the SWISS-PROT and TrEMBL protein sequence databases.

A comparative analysis of the predicted protein coding sequences of three complete prokaryotic genomes, *M. tuberculosis*, *B. subtilis* and *E. coli* was performed by running a non-redundant set of proteins against the InterPro database. A manual inspection of the results of the InterPro runs was done to calculate general statistics of protein families.

Implementation and results

We will illustrate the use of InterPro in whole proteome analysis of *M. tuberculosis* as shown in the graph in Figure 1.

An application of InterPro in the comparative genome analysis of *M. tuberculosis*, *B. subtilis* and *E. coli* is shown in Table 1 and Figure 2.

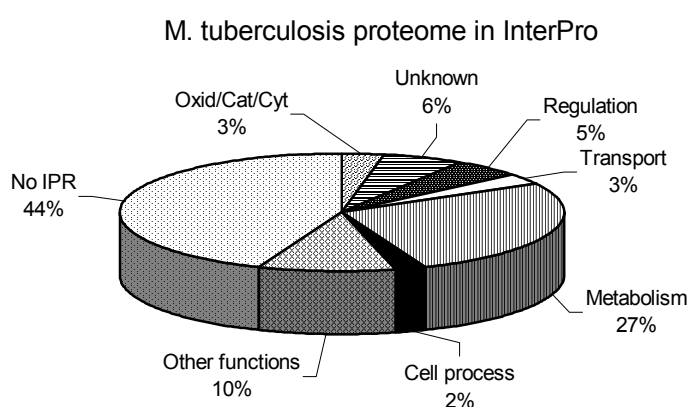


Figure 1. A pie graph representing the coverage of InterPro protein functions in the *M. tuberculosis* proteome.

Table 1. The 10 biggest InterPro families for *M. tuberculosis*, a comparative view with *B. subtilis* and *E. coli*.

InterPro Acc. No.	InterPro Entry Name	<i>M. tub</i> proteins	<i>B. subtilis</i> proteins	<i>E. coli</i> proteins
IPR000084	PE family	86	0	0
IPR000030	PPE family	66	0	0
IPR000379	Esterase/lipase/thioesterase	65	36	23
IPR000051	SAM (and other nucleotide) binding motif	53	27	30
IPR002198	Short-chain dehydrogenase/reductase family	52	33	18
IPR001617	ABC transporters family	42	81	78
IPR001647	Bacterial regulatory proteins, TetR family	42	19	11
IPR000873	AMP-binding domain	41	24	9
IPR001051	ATP-binding transport protein, P-loop motif	40	85	81
IPR000205	NAD binding site	34	22	30

The potential use of InterPro for classification of hypothetical proteins, and an in depth analysis of the relative representation of specific regulatory proteins in the three organisms based on InterPro families will be presented in additional figures.

Discussion

We have developed the InterPro database, an integrated resource of protein domains and functional sites. By uniting the databases, we have capitalised on their individual strengths, producing a single entity that is far greater than the sum of its parts. InterPro can streamline the analysis of newly determined sequences for the individual user, and makes a significant contribution in the demanding task of automatic annotation of predicted proteins from genome sequencing projects. It has been used here for the comparative genome analysis of the complete proteomes of *M. tuberculosis*, *B. subtilis* and *E. coli*, and provided information on the proteome structure of these organisms. The InterPro analysis has also facilitated the potential functional classification of

many hypothetical proteins generated from genome sequencing projects. InterPro has also proven its usefulness for whole proteome analysis of *Drosophila melanogaster* (Rubin *et al.*, 2000).

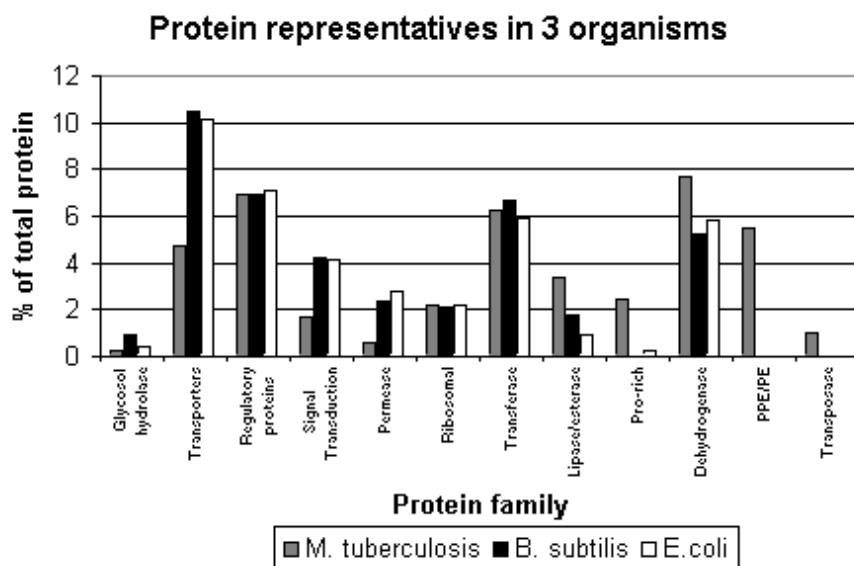


Figure 2. Graph of the relative representation of specific protein families in *M. tuberculosis*, *B. subtilis* and *E. coli* based on an InterPro analysis.

Acknowledgements

The InterPro Consortium: R.Apweiler 1, T.K.Attwood 4, A.Bairoch 2, A.Bateman 5, E.Birney 1, M.Biswas 1, P.Bucher 3, L.Cerutti 5, M.D.R.Croning 1,4, R.Durbin 5, W.Fleischmann 1, H.Hermjakob 1, N.Hulo 2, D.Kahn 6, A.Kanapin 1, Y.Karavidopoulou 1, R.Lopez 1, B.Marx 1, N.J.Mulder 1, T.M.Oinn 1, C.J.A.Sigrist 2, E.Zdobnov 1. (1 EMBL Outstation – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 2 Swiss Institute for Bioinformatics, Geneva, Switzerland; 3 Swiss Institute for Experimental Cancer Research, Lausanne, Switzerland; 4 School of Biological Sciences, The University of Manchester, Manchester, UK; 5 The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; 6 CNRS/INRA, Toulouse, France)

The InterPro project is supported by grant number BIO4-CT98-0052 of the European Commission. TKA is a Royal Society University Research Fellow.

References

1. Attwood, T.K., Croning, M.D.R., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N. and Wright, W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, 28, 225-227.
2. Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, 28, 45-48.
3. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L.L. (2000) The Pfam Protein Families Database. *Nucleic Acids Res.*, 28, 263-266.
4. Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, 28, 267-269.
5. Etzold, T., Ulyanov, A. and Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, 266, 114-128.
6. Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, 27, 215-219.
7. Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vossball, L.B., Zhang, J., Zhao, Q., Zheng, X.H., Zhong, F., Zhong, W., Gibbs, R., Venter, J.C., Adams, M.D., Lewis, S. (2000) Comparative genomics of the eukaryotes. *Science*, 287, 2204-2215.

NO MYSTERY OF ORFans IN GENOMICS - GENERATION OF ORFans IN THE ANTISENSE OF CODING SEQUENCES

*Mackiewicz P., Kowalczyk M., Gierlik A., Szczepanik D., Nowicka A., Dudek M.R., *Cebrat S.*

Institute of Microbiology, Wrocław University, Poland

e-mail: cebrat@angband.microb.uni.wroc.pl

*Corresponding author

Keywords: ORFan, *Saccharomyces cerevisiae*, gene number, coding probability, DNA asymmetry, DNA walk, random walk, long range correlation, antisense

Resume

Motivation:

Despite the growing number of known sequences coding for proteins or even completely sequenced genomes, the fraction of Open Reading Frames (ORFs) without known function or homology to other known coding sequences (so-called ORFans) is not diminishing. This phenomenon is known as Mystery of ORFans. There have been many attempts to explain this paradox but only one is in fact reasonable: a large fraction of ORFans do not code for proteins. Therefore, another problem arises: how these long, noncoding ORFs have been generated.

Introduction

Analyses of several completely sequenced genomes have revealed that many ORFs longer than 100 codons have no assigned functions or homologues. They make about one third of all ORFs in every genome. During sequencing of genomes the fraction of these ORFs (ORFans) grew much quicker than the fraction of homologues, which is a paradox because the more known genes, the higher fraction of homologues and the lower fraction of orphans should be found among newly sequenced ORFs. This paradox was called the "mystery of orphans" [Dujon, 1996; Casari et al., 1996]. After researching updated databases for homologues, ORFans still exist in the number much higher than expected [Fischer & Eisenberg, 1999].

Because of problems with finding homologues for ORFans and classifying them to known protein families, many authors consider ORFans fast evolving proteins or sequences unique to an organism or to a closely related group of organisms. Assuming this, ORFans should form new unknown protein superfamilies of unique function and structure. If it was true, almost every ORFan should define a new superfamily and the number of protein superfamilies ought to be several times larger than earlier estimations [Fischer & Eisenberg, 1999].

We have approximated the total number of coding ORFs longer than 100 codons in the most intensively studied genome, yeast *Saccharomyces cerevisiae*. Based on the analysis of asymmetry between coding and non-coding strands, we have found no more than 4700-4800 coding ORFs in this genome [Cebrat et al., 1997; Cebrat et al., 1998a; Kowalczyk et al., 1999]. It is about 1000 less than 5800-6000 which is the total number of ORFs annotated in data bases [Goffeau et al., 1996; Mewes et al., 1997]. The result indicates that about 1000 ORFs considered ORFans in the yeast genome data bases should be eliminated as non-coding.

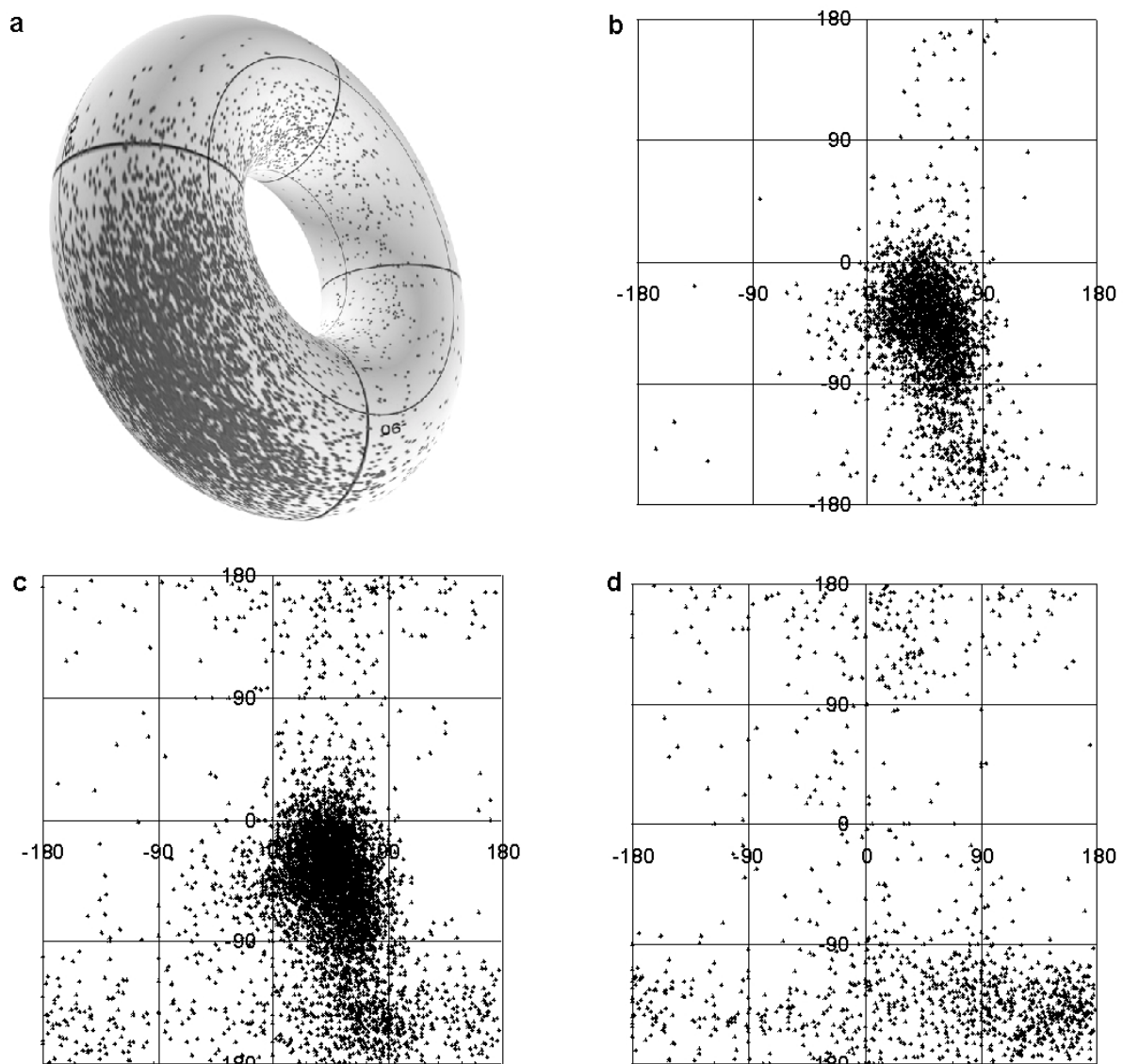
Thus, we suggest to use the Okham razor to solve the problem of ORFans - we just claim that the overwhelming fraction of ORFans do not code for proteins.

It is difficult to accept such high number of non-coding long ORFs in the yeast genome if we assume after Senapathy (1986), Sharp and Cowe (1991) that there is a small chance of occurring of long ORFs in a random sequence of the same size as the real yeast genome. Nevertheless, Cebrat and Dudek, (1996) and Cebrat et al., (1998b) have shown that the genetic code and coding sequences have specific properties of generating long ORFs, especially in the antisense strand.

In this paper we have shown that many ORFans in the yeast genome do not code for proteins and have been generated in protein coding sequences, mostly in their antisense strand.

Methods

We have parameterised coding sequence composition counting $\arctan([G-C]/[A-T])$ for the first and the second codon positions separately. Each sequence is represented by a point on the surface of torus, with the values of these two parameters as co-ordinates. The distributions of different sets of ORFs from the yeast genome are presented in Fig. 1. For details of the method see our papers [Cebrat et al., 1997; Cebrat et al., 1998a] and our web page (<http://smorfland.microb.uni.wroc.pl>).



Figur 1. Distribution of ORFs of the yeast genome on the torus projection. a - distribution of all ORFs on the torus, b - distribution of ORFs with known phenotype on the torus projection, c - distribution of all ORFs annotated in MIPS data base, d - distribution of baby ORFs (generated inside coding sequences, see text for a more detailed explanation).

Results

ORFs with known phenotypes form a compact set of points on the torus surface - about 98% of sequences are situated on about 11% of the torus projection (Fig. 1b). When all ORFs annotated in the yeast data bases are plotted using this method, they form a much more dispersed set of points, but still not evenly dispersed (Fig. 1c). Our previous studies [Cebrat & Dudek, 1996; Cebrat et al., 1998b] have shown that coding sequences preferentially generate noncoding overlapping ORFs (called *baby*) in the antisense in phases 3/3 and 2/2 (numbers indicate the positions in codons which overlap). In fact there are many (about 1500) such generated ORFs in the yeast genome. Their distribution is shown in Fig 1d.

Comparing all plots on Fig. 1 it is visible that many ORFs annotated in the yeast data bases are located in regions occupied by *baby* ORFs. On the other hand, these regions are very poor in known genes. It suggests that some ORFs (possessing properties of *baby* ORFs) have been generated in the same way as *baby* and probably do not code for proteins. Many of these ORFs may have arisen by ancient duplications of coding sequences nesting noncoding ORFs. Duplicated sequences accumulated mutations which eventually eliminated the proper reading frames of the original genes, leaving generated *baby* ORFs.

To prove that, we have translated the antisense of 2840 ORFs (without determined functions or distinct homologues, grouped in MIPS data base in classes 3-6). Then, we have searched protein databases for homologues using FASTA search program. We have found significant homologues for 757 ORFs with E value < 0.01 and for 603 ORFs with E value < 0.001 [Mackiewicz et al., 1999].

Table 1. Fractions of ORFs for which antisense homologues were found, depending on the distance to centre of distribution of ORFs with known phenotypes; A_d - distance to the centre, N - number of homologues found, N_f - number of homologues for which the generating phase was properly predicted.

A_d	Number of sequences	N	%	N_f	% _f
1-2	886	111	12	5	4
2-3	351	65	18	38	58
>3	587	298	51	217	73

We have grouped analysed ORFs according to their distance from the centre of distribution of known genes on the torus projection. This distance is anti-correlated with the ORFs' coding probability. For half of ORFs with low coding probability we have found homologues for their antisense (Table 1). For about 70 % of these ORFs we have predicted properly (based on our base content parameters) frame in which they had been generated.

Almost 80% of generated ORFs arose in antisense, 50% of which in the sixth phase - overlapping 3/3 and 28% in the fourth phase - overlapping 2/2, which is in agreement with our previous observations on generating overlapping ORFs [Cebrat et al., 1998b].

Discussion

One would argue that the set of about 3000 ORFs with recognised phenotypes (Fig. 1b) is not representative for all coding sequences in the yeast genome - for unknown reasons. Thus, one has to accept an implication of his argument that ORFs coding for unknown protein superfamilies have very specific properties - they resemble the antisense of coding sequences at least in their nucleotide composition of specific positions in codons, since they are dispersed non-evenly on the torus surface, grouping preferentially in the regions where generated ORFs are grouped. Defenders of the larger number of coding ORFs in the yeast genomes could argue that, still for unknown reasons, perhaps structural constraints of DNA molecule, ORFans (their double strand structure) have to possess the overall nucleotide structure of normal, known coding sequences and the only difference between them and the already known genes is in the phase they are coding in and which strand is coding.

If we agree with such arguments another question would rise: should we expect homologues between the antisense of known genes and presumed product of ORFans? It is hard to assume that it would be very easy to adopt the antisense information for producing functional proteins. In fact we have found a few such homologues between known coding sequences, but they are very rare cases [Cebrat et al., 1998b]. If there are no phylogenetic relations between them, there should be no ORFans homologous to antisense of coding sequences. As we have proved, it is not the case, a lot of ORFans have homologues in the antisense of ORFs with known phenotypes.

If anybody still defended the position of the large number of coding ORFans in the genomes, they would have to accept another, perhaps very plausible hypothesis that duplication and exploiting the antisense for a new function is a very common way of new gene evolution. But in the view of the data shown above, one has to accept the implication: such "inverted genes" have very specific functions, because they preferentially escape the traditional methods of finding the gene phenotypes - if not, we could find genes with known phenotypes in both classes, "normal phase" and "inverted". Furthermore, they diverge much faster, evolving into huge number of genes coding for new protein superfamilies.

Our question is: why to not cut off all these beings with the Okham razor and accept the thesis that a lot of ORFs in the genomes have been generated inside coding sequences and by the common recombination events were translocated into other genome regions where they can accumulate mutations very fast but they are still visible as the "antisense pseudogenes"? In this thesis, there are no assumptions (*beings*) like these:

Two third of ORFs in the yeast genome with known phenotypes make a very unrepresentative set of all genes in this genome.

Protein coding ORFans for unknown reasons have structural properties of ORFs generated spontaneously in the highest frequency in the antisense of known coding sequences.

Many (coding!) ORFans developed a product function by simple reading their information in the antisense of another coding sequence.

The rate of divergence of antisense ORFans is much faster than that of normal genes.

The generation and evolution of ORFans is unidirectional - "normal genes" are primordial.

References

1. G. Casari, A. de Druvar, C. Sander and R. Shneider, "Bioinformatics and the discovery of gene function" *Trends Genet.* **12**, 244 (1996).
2. S. Cebrat and M.R. Dudek, "Generation of overlapping open reading frames" *Trends Genet.* **12**, 12 (1996).
3. S. Cebrat, M.R. Dudek and P. Mackiewicz, "Sequence asymmetry as a parameter indicating coding sequence in *Saccharomyces cerevisiae* genome" *Theory in BioSciences* **117**, 78 (1998a).
4. S. Cebrat, M.R. Dudek, P. Mackiewicz, M. Kowalczyk and M. Fita, "Asymmetry of Coding versus Noncoding Strand in Coding Sequences of Different Genomes" *Microb. & Comp. Genom.* **2**, 259 (1997).
5. S. Cebrat, P. Mackiewicz and M.R. Dudek, "The role of the genetic code in generating new coding sequences inside existing genes" *Biosystems* **42**, 165 (1998b).
6. B. Dujon, "The yeast genome project, what did we learn" *Trends Genet.* **12**, 263 (1996).
7. D. Fischer and D. Eisenberg "Finding families for genomic ORFans" *Bioinformatics* **15**, 759 (1999).
8. A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston et al., "Life with 6000 genes" *Science* **274**, 546 (1996).
9. M. Kowalczyk, P. Mackiewicz, A. Gierlik, M.R. Dudek and S. Cebrat, "Total Number of Coding Open Reading Frames in the Yeast Genome" *Yeast* **15**, 1031 (1999).
10. P. Mackiewicz, M. Kowalczyk, A. Gierlik, M.R. Dudek and S. Cebrat, "Origin and properties of noncoding ORFs in the yeast genome" *Nucleic Acids Res.* **27**, 3503 (1999).
11. H.-W. Mewes, K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S.G. Oliver, F. Pfeiffer and A. Zollner, "Overview of the yeast genome" *Nature* **387**, 7 (1997).
12. P. Senapathy "Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications". *Proc. Natl. Acad. Sci. USA* **83**, 2133 (1986).
13. P.M. Sharp and E. Cowe, "Synonymous codon usage in *Saccharomyces cerevisiae*" *Yeast* **7**, 657 (1991).

Pro-Gen: PREDICTION OF THE EXON-INTRON STRUCTURE BY COMPARISON OF GENOMIC SEQUENCES

¹Novichkov P.S., ^{1,2}*Gelfand M.S., ^{1,2}Mironov A.A.

¹State Scientific Center for Biotechnology NII Genetika, Moscow, Russia

²Anchorgen, Inc., Santa Monica, USA

e-mail: misha@imb.imb.ac.ru, mgelfand@anchorgen.com

*Corresponding author

Keywords: eukaryotic genes, exon-intron structure, prediction

Resume

An algorithm for prediction of the exon-intron structure of higher eukaryotic genes is suggested. The algorithm is based on comparison of genomic sequences of homologous genes from different species. It uses the fact that protein-coding sequences evolve slower than non-coding regions. Unlike existing comparison methods, the suggested algorithm compares not nucleotide, but amino acid sequences, which increases its sensitivity. Conservation of the exon-intron structures of the compared genes is not assumed.

The algorithm is implemented in the program Pro-Gen. Testing of the program has demonstrated that it can be successfully applied to prediction of genes from human and model genomes (mouse, *Xenopus*, *Drosophila*). The algorithm overcomes deficiencies of the existing methods, both statistical (insufficient reliability) and similarity-based (inapplicability to completely new genes).

Availability:

<http://www.anchorgen.com>

One of the most reliable methods of gene recognition in DNA sequences of higher eukaryotes is comparison with already known proteins [Mironov et al., 1998]. However, this approach does not allow for prediction of genes having no known homologues. These can, in principle, be found by statistical algorithms, but the latter are insufficiently reliable [Bursset & Guigo, 1996]. Large-scale sequencing of eukaryotic genomes, and in particular, syntenic regions, allows one to perform gene recognition via comparison of genomic sequences. This approach is based on the fact that protein-coding regions evolve at a slower rate than non-coding regions (introns and intergenic spacers). Recently it has been successfully applied to analysis human-mouse and nematode *C. elegans* - *C. briggsae* gene pairs (Ansari-Lari et al., 1998; Thacker et al., 1999). However, these studies relied on comparison of nucleotide sequences analyzed using dot-matrices (Thacker et al., 1999) or local similarity analysis (Ansari-Lari et al., 1998).

We have developed a modified version of the spliced alignment algorithm that finds most similar chains of protein-coding exons in a pair of genomic sequences containing homologous genes [Novichkov et al., 2000]. The algorithm identifies candidate exons as regions of strong amino acid similarity and then constructs the highest scoring chains of candidate exons in the two genomic sequences using dynamic programming. Not only similar regions corresponding to complete exons, but also partial candidate exons are considered. Thus the program can successfully predict genes whose exon-intron structure in the two genomes is different. Early use of the translated sequences allows the program to successfully predict genes even in the cases when the nucleotide identity is low, and also decreases the influence of conservation of non-coding (e.g. regulatory) regions.

The program was tested on a sample of human-mouse, human-*Xenopus* and human-*Drosophila* gene pairs. The vertebrate gene pairs were taken from the HOVERGEN database (HOVERGEN). The human-*Drosophila* gene pairs were taken from the Berkeley *Drosophila* Project sample (GASP). Only protein pairs with identity exceeding 50% were considered. This is motivated by the observation that lower identity usually corresponds to homologous domains rather than orthologous genes, and thus the protein length in the case of lower identity are very different. Note that domain organization of orthologous genes of higher eukaryotes is usually conserved [Mushegian et al., 1997].

The results are given in Table 1. In all three samples the number of exact predictions exceeds 50%, fair predictions (with specificity or sensitivity in at least one gene worse than 90%) are rare, whereas bad predictions (specificity or sensitivity less than 80%) comprise less than 5% of predictions. We were able to identify a number of obvious annotation errors, mostly incorrectly identified start codons.

The main type of errors for the human-mouse sample are overpredictions caused by conservation of non-coding regions. This leads either to generation of spurious exons, or extension of exons. Some of these cases

may in fact be due to alternative splicing. In the human-*Xenopus* and human-*Drosophila* samples the program tends to underpredict: weakly conserved exons are not identified by the initial local similarity search.

Another source of errors is due to very short initial and terminal exons (that can be as short as 3 nucleotides). This leads both to overprediction (if stronger extensions are found) and underprediction (if neither the correct exons, nor positive-scoring alternatives can be found).

Table 1. Testing results. Human-mouse, human-*Xenopus* and human-*Drosophila* gene pairs are considered. The quality is ascribed according to the worse of the two predictions (e.g. if specificity of the predicted mouse gene is 85%, sensitivity of the mouse gene is 100%, and both specificity and sensitivity of the predicted human gene is 95% the prediction is counted as "fair").

	mouse	<i>Xenopus</i>	<i>Drosophila</i>
number of pairs	73	14	23
exact predictions	39	8	12
good predictions (specificity and sensitivity >90%)	25	4	10
fair predictions (specificity and sensitivity >80%)	6	1	1
bad predictions	3	1	0
spliced alignment score > correct alignment score	22	5	9
spliced alignment score < correct alignment score	12	1	2

One possible difficulty not considered in this study is analysis of multigene genomic fragments. In the human-mouse case, where the order of orthologous genes is conserved in syntenic regions, this can lead to creation of chimeric exon chains due to merging of adjacent genes. However, this should not be a problem for analysis of more distant genomes.

Further research will be directed towards improvement of the local similarity search procedure, increasing its sensitivity. We plan to perform large scale analysis of the *Drosophila* genome when it becomes available, comparing it with the nematode genome and available portions of the human genome.

Acknowledgements

We are grateful to Drs. J.W. Fickett and M.A. Roytberg for useful discussions. This work was partially supported by Anchorgen, Inc. (<http://www.anchorgen.com>).

References

1. M.A. Ansari-Lari, J.C. Oeltjen, S. Schwartz, Z. Zhang, D.M. Muzny, J. Lu, J.H. Gorrell, A.C. Chinault, J.W. Belmont, W. Miller, and R.A. Gibbs, *Genome Res.*, **8**, 29 (1998).
2. M. Burset and R. Guigo, "Evaluation of gene structure prediction programs" *Genomics*, **34**, 353 (1996).
3. GASP. <http://www.fruitfly.org/sequence/{human-datasets,drosophila-datasets}.html>
4. HOVERGEN. <http://pbil.univ-lyon1.fr/databases/hovergen.html>
5. A.A. Mironov, M.A. Roytberg, P.A. Pevzner and M.S. Gelfand, "Performance-guarantee gene predictions via spliced alignment" *Genomics*, **51**, 332 (1998).
6. Novichkov, P.S., Gelfand, M.S. and Mironov, A.A. (2000) "Prediction of the exon-intron structure by comparison of genomic sequences." *Molecular Biology*, **34**, 200-206.
7. Thacker, C., Marra, M.A., Jones, A., Baillie, D.L., and Rose, A.M.,(1999) *Genome Res.*, **9**, 348-359. "Functional genomics in *Caenorhabditis elegans*: An approach involving comparisons of sequences from related nematodes."
8. Mushegian, A.R., Bassett, D.E., Jr., Boguski, M.S., Bork, P. and Koonin, E.V. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 5831-5836. "Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs"

COMPARATIVE GENOMICS: HOMOLOGY BASED GENE IDENTIFICATION AND GENE STRUCTURE VALIDATION

**¹Wiehe T, ¹Gebauer-Jung S., ²Abril J. and ²Guigò R.*

¹Department of Genetics and Evolution, Max Planck Institute for Chemical Ecology, Germany

²Departament d'Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Spain

e-mail: twiehe@ice.mpg.de

*Corresponding author

Keywords: Homology, gene prediction, alignment, evolution, phylogenetic footprint

Resume

Motivation:

In the era of large scale genomic sequencing reliable methods for gene and gene structure prediction are highly demanded. So far, no single tool performs with the desired accuracy, in particular so, when large genomic sequences are to be analyzed. As a remedy, investigators and annotators may rely upon an array of diverse and independent prediction programs. Another problem is that most existing programs are extremely species-specific. Their applicability is often limited to species for which they have been trained. Gene prediction by homology addresses both problems: it is a method for gene identification independent from existing ones, and it is not limited to a particular species.

Methods and Results:

We have devised a program, SGP, to identify the structures of protein coding genes. It works simultaneously in both of two aligned homologous genomic sequences. We evaluated the prediction accuracy on several test sets from vertebrate and angiosperm species. We compared the prediction results with gene structures as they are annotated in GenBank entries. Our results indicate that gene identification by homology is very reliable when based on pairs of evolutionarily closely related species, irrespective of their origin, plant or animal. Closely related in the above sense are, for example, pairs of mammalian or dicotyledon sequences. Furthermore, we show how SGP can be applied for the validation of gene structure annotation in GenBank entries.

Availability:

Our program is written in C. It is available on request from the authors. The installation of an interactive GUI at <http://www.soft.ice.mpg.de> is under construction.

Contact: twiehe@ice.mpg.de

Introduction

Genome projects have entered a stage in which megabases of homologous genomic sequences from a spectrum of diverse species will soon be available and which provide the raw material for comparative genome analyses. To address questions of genome and gene structure evolution computer tools are needed to predict and compare - simultaneously in more than one genome - coding as well as regulatory sequences. Here, we deal with the first task: predict and compare coding sequences.

Methods and Algorithms

Given a pairwise alignment, as produced by, say, a dynamic programming method such as sim96 (<http://globin.cse.psu.edu>), a post-processing step yields a so-called lean alignment. This ensures that any given site is aligned at most once with the partner sequence. Furthermore, two lists of potential exons, so called pre-candidates, are generated. The only properties of pre-candidates are to agree with the standard splice site consensus and to sense-translate in at least one frame. In a filtering step pre-candidates are compared with the alignment. Only pairs which are compatible with the alignment are retained, yielding a list of candidates. In a scoring step individual scores are assigned to the candidates. We have tested different scoring schemata. The choice of the scoring scheme is a trade-off between versatility of the program (species independent scoring) and additional accuracy (use of species-specific substitution scores). Finally, based on their scores, exon candidates are selected and assembled [Guigò, 1998]. Multiple genes on forward and reverse strands are simultaneously predicted on both sequences (see Fig. 1).

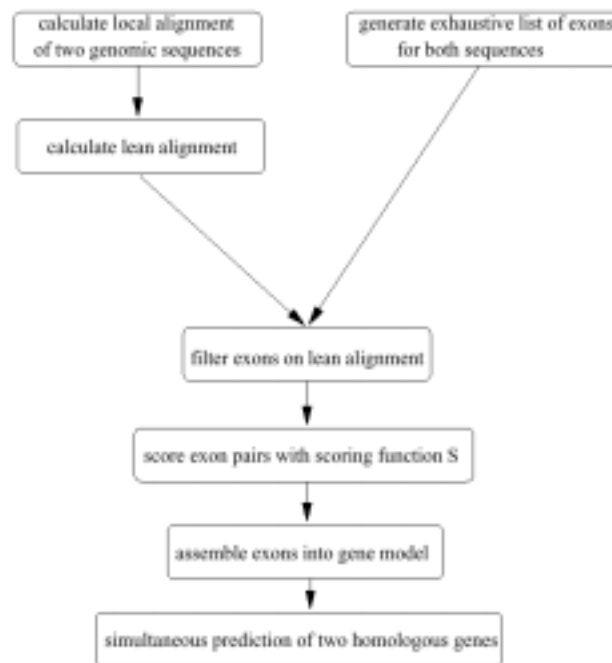


Figure 1.

Results

We have applied the program to several test sets. The largest one contains human/rodent sequence pairs (see, for example, <http://www.sanger.ac.uk/Software/Alfresco/mmhs.shtml>). Furthermore, we compared mouse/rat, human/rabbit, human/chicken, human/xenopus, maize/rice and several pairs from the family of Brassicaceae: *Arabidopsis thaliana*, *Arabidopsis drummondii*, *Arabidopsis lyrata* and *Brassica oleracea*. The test sets consist mostly of single gene sequences since large scale ($>0.1\text{Mb}$), fully annotated and homologous genomic sequences are currently still rare. We evaluated the performance of the program in terms of standard measures of prediction accuracy [Burset & Guigó, 1996]. We obtain best results when a species specific substitution matrix for splice sites is available and integrated into the scoring scheme. On nucleotide level, accuracy (AC value) is above 96%, on exon level above 90%. The most remarkable result is the extremely low percentage of false positive predictions.

Discussion

Homology based gene identification does not only serve as a prediction tool, but is also a means to validate gene annotations. In our test sets we encountered several examples where the annotated gene structure appears highly questionable when compared with a homologue. In particular, improbable reading frame shifts or splice junctions can be detected and highlighted. In connection with automated tools for comparative analysis of untranslated and/or regulatory regions [Abstract by Goebel *et al.*, this conference], syntenic gene prediction should be valuable not only for functional but also evolutionary genomics.

Acknowledgments

This work has been supported by Proyecto del Plan Nacional de I+D, BIO98-0443-C02-01, Beca de Formación de Personal Investigador, FP95-3881 7943 from the Ministerio de Educación y Ciencia (Spain). We would like to thank Laurent Duret and Matthias Platzer for valuable comments and discussions.

References

1. M. Burset, and R. Guigó, "Evaluation of gene structure prediction programs" *Genomics* 34, 353 (1996).
2. R. Guigó, "Assembling genes from predicted exons in linear time with dynamic programming" *J. Comp. Biol.* **5**, 681 (1998).

MOUSE AQUAPORIN 4 GENE: PREDICTION OF A NEW EXON AND EXPERIMENTAL CONFIRMATION.

**Bondar A.A., Alikina T.Yu., Zelenin S.M.*

Novosibirsk Institute of Bioorganic Chemistry, SB RAS, Russia

e-mail: alex_bondar@mail.ru

*Corresponding author

Keywords: aquaporin 4, AQP4, isoforms, mRNA, gene, exon localization

Resume

Aquaporin 4 (AQP4) is abundantly expressed in the brain, and also in lung and kidney. This water channel has been suggested to play a role in the regulation of brain water homeostasis. AQP4 is unique among other aquaporins in that its water transport function is not inhibited by mercurials. Its original name is mercurial insensitive water channel (MIWC) [1]. It has been suggested that AQP4 expression disturbances connected with a variety of serious disorders such as brain edema, ataxia et al. [Ma et al., 1996; Turtzo et al., 1997; Neely et al., 1999; Zelenin et al., 2000]. There are at least two isoforms of the AQP4 protein, M1 and M23. The M1 form contains 22 more amino acids in the N-terminus than the M23 form. It has recently been demonstrated that AQP4 forms heterotetramers from M1 and M23 isoforms and that the 22-amino acid sequence at the N-terminus of M1 does not influence water permeability but may contribute to membrane trafficking or assembly of arrays [Neely et al., 1999].

M1 and M23 AQP4 isoforms of mouse [Ma et al., 1996; Turtzo et al., 1997; Neely et al., 1999], human and rat has been shown to be encoded at least by two different mRNA. The M23 isoform has been reported to be encoded by four exon segments 1-4. The mRNA encoding the M1 protein contain an additional nucleotide sequence corresponding to an exon segment 0, upstream of partially spliced exon 1 in the mouse AQP4 gene. There are however conflicting data [Ma et al., 1996; Turtzo et al., 1997] concerning the mRNA sequence of AQP4 and the number of distinct AQP4 isoforms. AQP4 genomic nucleotide sequence is known mostly in the exons 1, 2, 3 and 4. Ma T et al. [Ma et al., 1996] have demonstrated at least three mouse AQP4 mRNA named MIWC1 (corresponding M23), MIWC2 (corresponding M1) and MIWC3. According to Ma T et al. [1] MIWC3 mRNA encodes the third form of mouse AQP4. Turtzo L.C. et al. [Turtzo et al., 1997] could not find MIWC3 and have cloned mouse AQP4 mRNA with open reading frames from Met-1 and Met-23. AQP4 mRNAs cloned have sizes not more than 1700 bp, although about 5500 bp long RNA transcripts have been revealed by Northern-blot analysis [Ma et al., 1996; Turtzo et al., 1997].

Here we report that mouse AQP4 gene contains an additional exon, named by us as X ("extra - exon"), that encodes mRNA (exon X, partially spliced 1, 2, 3, 4) distinct from M23 mRNA (exon 1-4) but encoding the same protein M23 AQP4.

By screening of a mouse genomic library with a 1.6 kb PCR fragment, corresponding to the exon 1-intron-exon 2 fragment of the mouse AQP4 gene (mAQP4), 2 clones with 17000 and 17500 bp inserts were obtained. By use of PCR and restriction analysis we found that this contig of clones covered approximately 26500 bp genomic fragment and included all known AQP4 exons and also 8500 bp upstream of the 5'-end of exon 0. Selected lambda clones DNA was partially digested by restriction endonucleases, separated by agarose gel electrophoresis and hybridized with a specific probe. Then exon-intron and restriction map of the mouse AQP4 gene has been constructed for XbaI, SacI, HindIII and BamHI (Fig.1). DNA inserts from lambda clones were subcloned into pBlueScript II vectors as a set of clones fully overlapping exon-intron and about 8500 bp of the 5'-upstream regions.

By introducing unidirected deletions of predictable size (300-600 bp) with Exo III/Mung Bean nucleases and following sequencing we revealed nucleotide sequence of the both strands of the 4964 bp mAQP4 gene fragment covering exon 0-exon 1 region (Fig.1, GenBank - AF219992).

Genomic DNA sequence obtained was compared with all known mouse AQP4 mRNA (M1[Turtzo et al., 1997]/MIWC2, MIWC3, MIWC1[Ma et al., 1996]) sequence data registered in GenBank.

It appears that nucleotide sequence of the mAQP4 gene fragment contains three instead of two, corresponding to exon 0 and exon 1, highly homologous to mAQP4 mRNAs regions.

This third homology cluster of the mAQP4 gene was located 582 nucleotide downstream from exon 0. It was identical to the 5'-end of the MIWC3 mRNA [Ma et al., 1996] (Fig.1) but contained 206 bp long deletion in the MIWC3 mRNA nucleotide sequence in the middle of the homology cluster. To confirm either two or one new exon ("X") interspaced between exon 0 and 1 exists we prepared total RNA from mouse brain, made RT-PCR analysis, following direct sequencing of PCR products and analysis of revealed AQP4 cDNA structures. By use

of all combinations of sense and antisense primers corresponding to exons 0, X, 1 (Fig.1) and 4 (not shown), we identified three mAQP4 mRNAs. By comparing of mAQP4 genomic structure with mAQP4 mRNA nucleotide sequences obtained we revealed that correspondent regions of mRNAs are identical to the exon 0, X and 1 in a pattern showed in Fig.1, B). Exon X identical sequence was located at the 5'-end of the AQP4 mRNA that is distinct from all known but similar to MIWC3 mRNA. We did not find in this new AQP4 mRNA any deletion as in the 5'-end of MIWC3 mRNA when compared to exon X (Fig.1). We found also multiple stop codons in all three reading frames in the region correspondent to deletion in MIWC3 mRNA. An open reading frame (ORF) analysis did reveal that ATG initiating codon of the first optimal context for initiation of translation encodes Met-23 of aquaporin 4 protein. We named this new mRNA as AQP4 M23X mRNA. It is differ from M23 mRNA by 5'-end but encodes the same aquaporin protein.

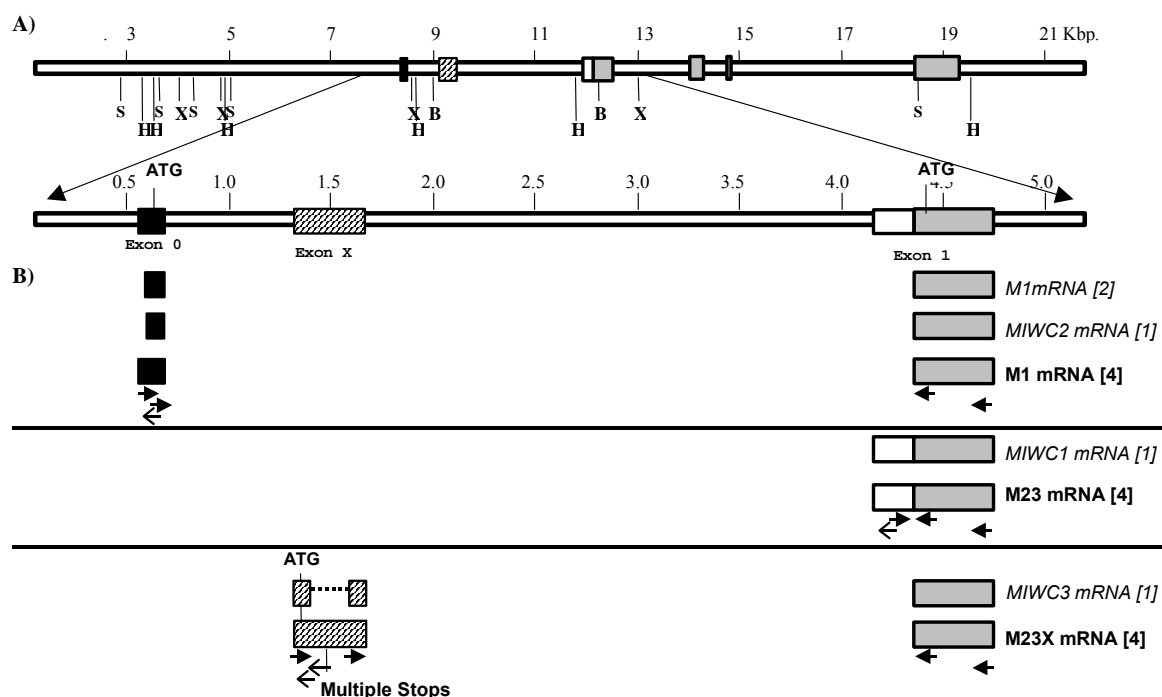


Figure 1. Exon-intron structure of the mouse aquaporin 4 gene.

A) mAQP4 gene map. Endonuclease restriction sites are denoted as S for SacI, X for XbaI, B for BamHI and H for HindIII. Size scale of the cloned mAQP4 gene is in kilobase pairs (Kbp). Exons 1, X, 0 was located according genomic nucleotide sequence analysis and comparing with mRNAs. Exons 2, 3 and 4 was mapped by use of PCR and restriction analysis. Mouse AQP4 gene fragment sequenced and analyzed here is indicated by arrows and shown below. AQP4 gene exons are depicted by blocks with different pattern according to their homology with AQP4 mRNAs. ATG means initiation translation codons of the M1, MIWC3 and M23 AQP4 isoforms.

B) Schematic presentation of a homologous clusters of the mAQP4 genomic fragment and mAQP4 mRNA nucleotide sequences. Homologous regions of mRNAs are depicted by similar pattern blocks along the gene. Names of mouse AQP4 mRNAs are in the author's nomenclature. MIWC3 mRNA deletion region is depicted by dotted line. *M1 mRNA* [Zelenin et al., 2000], *M23 mRNA* [Zelenin et al., 2000] and *M23X mRNA* [Zelenin et al., 2000] denotes mRNA nucleotide sequences obtained in the present study. Filled arrows represent location of the AQP4 gene exon-specific oligonucleotide primers used for RT-PCR analysis of the mouse brain total RNA. Open arrows represent location of the AQP4 gene exon-specific oligonucleotide primers used for the primer extension. "Multiple stops" means multiple stop codons in the M23X mRNA nucleotide sequence located in frame of MIWC3 and two others translation frames.

The sequence of AQP4 M1 mRNA we obtained was similar to that one of the M1 mRNA reported by Turtzo et al. [Turtzo et al., 1997] and MIWC2 reported by Ma et al. [Ma et al., 1996] and has ORF starting from AQP4 Met-1. We found two silent nucleotide substitutions and one changing Gly-4 to Arg-4 that is known as polymorphism [Turtzo et al., 1997]. However, in contrast to Turtzo et al., we found 105 additional nucleotides at the 5'-end of the exon 0 by primer extension combined with RT-PCR products sequencing as described above. We did not confirm also the presence of the first 32 nucleotides at the 5' end of the M1 mRNA reported by Turtzo et al. [Turtzo et al., 1997]. These 32 nucleotides are not present also in the structure of AQP4 genomic fragment.

The sequence of AQP4 M23 mRNA we obtained was similar to the AQP4 mRNA mMIWC1 reported by Ma et al. [Ma et al., 1996].

By comparing with genomic structure it was found that splicing site for the M23X and MIWC3 mRNA between exons 0/1, and one for the M1 and MIWC2 mRNA between exons X/1 are identical and located inside of exon 1.

We conclude that mouse AQP4 gene contains additional new exon X interspaced between exon 0 and 1. At least three (M1, M23X, M23) AQP4 mRNAs are expressed in mouse brain and encode M1 and M23 AQP4

proteins. Of great interest are the mechanisms of different mouse AQP4 mRNAs expression regulation that appears to be controlled by distinct, probably independently functioning promoters.

This work was supported in part by the RFBR grant no.98-04-49369 and by the INTAS grant no.11404.

References

1. T.Ma, B. Yang, A.S. Verkman, "Gene structure, cDNA cloning, and expression of a mouse mercurial-insensitive water channel" *Genomics* **33**, 382 (1996).
2. L.C. Turtzo, M.D. Lee, M. Lu, B.L. Smith, N.G. Copeland, D.J. Gilbert, N.A. Jenkins, P. Agre, "Cloning and chromosomal localization of mouse aquaporin 4: exclusion of a candidate mutant phenotype, ataxia" *Genomics* **41**, 267 (1997).
3. J.D. Neely, B.M. Christensen, S. Nielsen and P. Agre, "Heterotetrameric composition of aquaporin-4 water channels" *Biochemistry* **38**, 11156 (1999).
4. S. Zelenin, E. Gunnarson, T. Alikina, A. Bondar, and A. Aperia, "Identification of a new form of AQP4 mRNA that is developmentally expressed in brain" *Pediatric Research* (2000, in press).

COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: MULTIDRUG RESISTANCE SYSTEMS IN GAMMA-PROTEOBACTERIA

Rodionov D.A., *Gelfand M.S., Mironov A.A. Rakhmaninova A.B.

State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: complete genome, phylogenetic fingerprinting, multidrug systems

Resume

Results:

The comparative approach is a powerful tool of analysis of gene regulation in completely sequenced bacterial genomes (Gelfand & Mironov, 1998; Mironov et al., 1999; the papers in this volume; for a review see Gelfand, 1999). However, in the above studies it was applied to considerably diverged genomes. At that, occurrences of candidate sites upstream of orthologous genes can be considered as statistically independent events, and thus consistency in distribution of regulatory sites is a powerful sign of their functional relevance.

A different variant of the comparative technique can be applied for analysis of closely related genomes. In this case the gene upstream regions are less diverged and often can be aligned without ambiguities. However, the regulatory sites diverge slower than the non-coding regions in general, and thus are visible as strongly conserved islands. This approach, extensively applied for analysis of eukaryotic regulation under the name "phylogenetic fingerprinting" (reviewed in Duret & Bucher, 1997; Gelfand, 1999) has been recently shown to be applicable for analysis of prokaryotic sequences as well (Stojanovic et al., 1999). Here we apply it to analysis of regulation of multidrug resistance transport systems in enteric bacteria *Escherichia coli*, *Salmonella typhimurium*, *Klebsiella pneumoniae* and *Yersinia pestis*.

The results are shown in Fig. 1. The simplest case is that of MarA. This transcriptional activator interacts with several footprinted sites (Martin & Rosner, 1995; Martin et al., 1999) that have been used for construction of the recognition profile. We were able to find candidate MarA binding sites in regulatory regions of the *mdlAB* operon and *envR-acrEF* locus encoded multidrug transport systems. Comparison of the non-coding regions in the *marC-marRAB* loci from different genomes allowed us to identify a new candidate regulatory site (MarX box). Similarly, comparison of the non-coding regions in the *acrR-acrAB* and *emrRAB* loci lead to identification of the candidate EmrR signal.

More instances of these signals were found upstream of known and hypothetical multidrug transport genes *emrKY*, *b2409-b2410* and *slyA-b1644/5* as well as upstream of porin genes (Fig. 1).

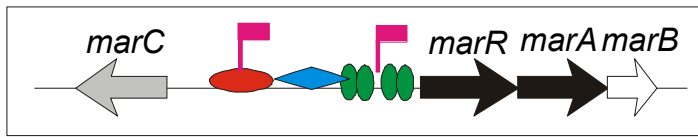
Thus the multidrug systems in enterobacteria seem to be regulated by several diverse regulatory systems. Many operons are regulated by several regulators: a global one (MarA, MarX, EmrR) and a local one (MarR, AcrR, EnvR, b0447, b2409, SlyA).

This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476). Preliminary sequence data were obtained from The Institute for Genomic Research WWW site.

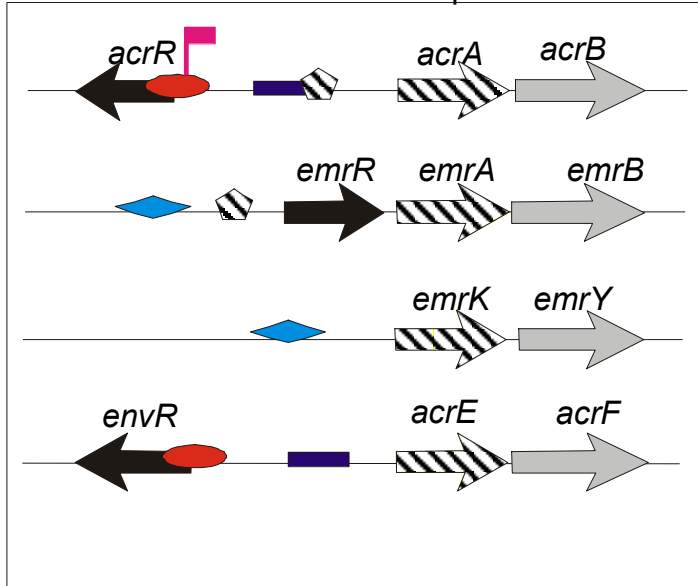
References

1. L. Duret and P. Bucher, *Curr. Opin. Struct. Biol.*, **7**, 399 (1997).
2. M.S. Gelfand, Recognition of regulatory sites by genomic comparison. *Res. Microbiol.*, **150**, 755 (1999).
3. R. Martin and J. Rosner, *Proc. Natl. Acad. Sci. USA*, **92**, 5456 (1995).
4. R. Martin et al. *Mol. Microbiol.*, **34**, 431 (1999).
5. N. Stojanovic, L. Florea, C. Riemer, D. Gumuchio, J. Slightom, M. Goodman, W. Miller, and R. Harrison, *Nucleic Acids Res.*, **27**, 3899, (1999).

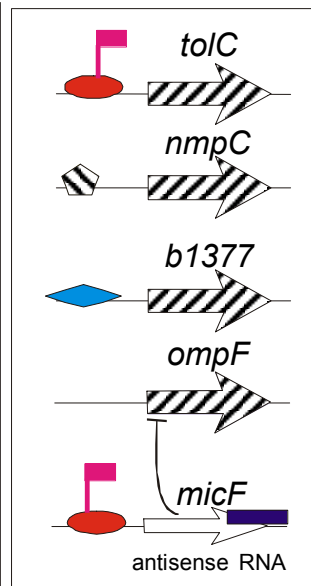
The regulatory locus *mar*



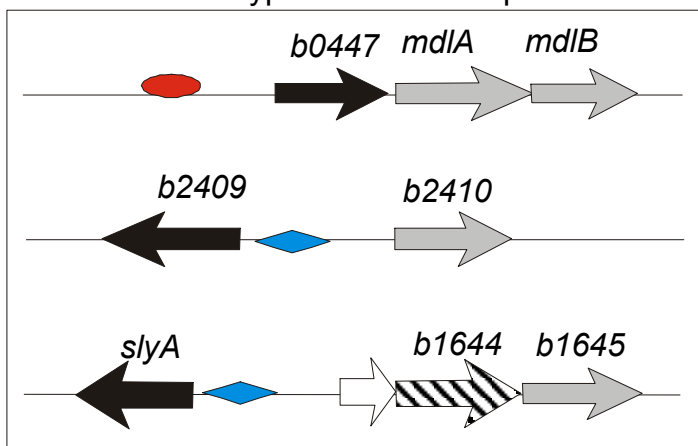
The known MDT operons



Porins (*E. coli*)



The hypothetical MDT operons



Operators

- known sites
- MarA-box
- MarR-box
- MarX-box
- EmrR-box
- AcrR-box

Genes

- transmembrane MDT proteins
- periplasmic MDT proteins
- transcriptional regulators
- porins

Figure 1. Multidrug resistance regulatory network.

REGULATION OF DAHP-SYNTASES IN GAMMA-PROTEOBACTERIA: FEEDBACK INHIBITION AND REPRESSION OF TRANSCRIPTION

¹Panina E.M., ²Mironov A.A., ^{2*}Gelfand M.S.

¹Moscow State University, Russia

²State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: DAHP-synthase, regulation, comparative genomics

Resume

Introduction

DAHP-synthase performs conversion of phosphoenolpyruvate and erythrose 4-phosphate to 3-deoxy-D-arabino-heptulosonate-7-phosphate. This is the first step of the common pathway leading to aromatic amino acids (Srinivasan & Sprinson, 1959). *Escherichia coli* and related bacteria contain three DAHP-synthase isoenzymes AroF, AroG and AroH feedback inhibited by tyrosine, phenylalanine and tryptophan respectively (Smith, 1967; Camakaris & Pittard, 1974; Doy, 1967). In *E. coli* and *Salmonella typhimurium* transcription of *aroF* and *aroG* is regulated by repressor TyrR (Grove & Gunsalus, 1987), and transcription of *aroH*, by TrpR (Doy & Brown, 1965).

Here we have applied the comparative genomics approach (Gelfand & Mironov, 1998; Mironov et al., 1999) to analysis of transcriptional regulation of DAHP-synthase genes in other proteobacteria from the gamma-subdivision. In addition, we describe evolution of the feedback inhibition site in the protein molecule.

Data and Methods

Complete genome sequences of *Escherichia coli* (EC), and *Haemophilus influenzae* (HI) were extracted from GenBank (Benson et al., 1999). Partially sequenced genomes of *Salmonella typhimurium* (ST), *Klebsiella pneumoniae* (KP), *Yersinia pestis* (YP), *Vibrio cholerae* (VC), *Actinobacillus actinomycetemcomitans* (AA), *Pseudomonas aeruginosa* (PA), *Shewanella putrefaciens* (SP) were extracted from GenBank (Benson et al., 1999) and from the TIGR WWW site (TIGR).

Profiles for recognition of TrpR and TyrR binding sites (TRP and TYR boxes respectively) were taken from (Mironov et al., 1999). Multiple protein alignment was done using CLUSTAL (Thompson et al., 1997). Analysis of the protein 3D structure was done using RASMOL (<http://www.umass.edu/microbio/rasmol/>). Phylogenetic trees were constructed using PHYLIP (<http://evolution.genetics.washington.edu/phylip.html>) and plotted with GeneTree (Page, 1998; <http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html>). Genomic analyses (protein similarity search, analysis of orthology, DNA profile search) were done using GenomeExplorer (Mironov et al., 2000; <http://www.anchorgen.com>).

Results and discussion

Transcriptional regulation of DAHP-synthase genes

Table 1 features candidate TYR and TRP boxes upstream of the DAHP-synthase genes. It has been observed that all enterobacterial genomes harbor three differently regulated DAHP-synthase isoenzymes (Pittard & Gibson, 1970). In fact, it has been suggested that appearance of the most recently diverged DAHP-synthase-PHE (*aroG*) coincided with the enteric lineage (Ahmed et al., 1988). However, there are representatives of all three lineages in the genome of *V. cholerae* and *S. putrefaciens*, and thus duplication leading to *aroG* has occurred before divergence of *Enterobacteriaceae*, *Vibrionaceae* and *Alteromonadaceae*. However, it has been suggested that *Buchnera aphidicola* has only one DAHP-synthase (Kolibachuk et al., 1995) belonging to the enteric AroH subfamily (Fig. 1).

Absence of *aroG* in *K. pneumoniae* is likely to be due to incompleteness of its genome. There is a very weak candidate TRP box upstream of *aroH* gene of *Y. pestis* and *Erwinia herbicola* and no such box before this gene in *B. aphidicola* and *V. cholerae* (Fig. 2), although both phylogenetic analysis (Fig. 1) and analysis of the feedback inhibition site (below) show that these genes encode proteins regulated by tryptophan. It appears that TrpR regulation of *aroH* existed in the ancestor of enteric lineage, but was lost in *Y. pestis* and *E. herbicola* due to point mutations in the TRP box. The *aroH* upstream region of *B. aphidicola* cannot be aligned with the *aroH* upstream regions of other enterobacteria. Thus the TrpR regulation appeared after divergence of *Buchnera* and

the remaining clades. However, one cannot completely rule out the recombination-caused loss of the TRP box region in these tryptophan-producing endosymbionts of aphids.

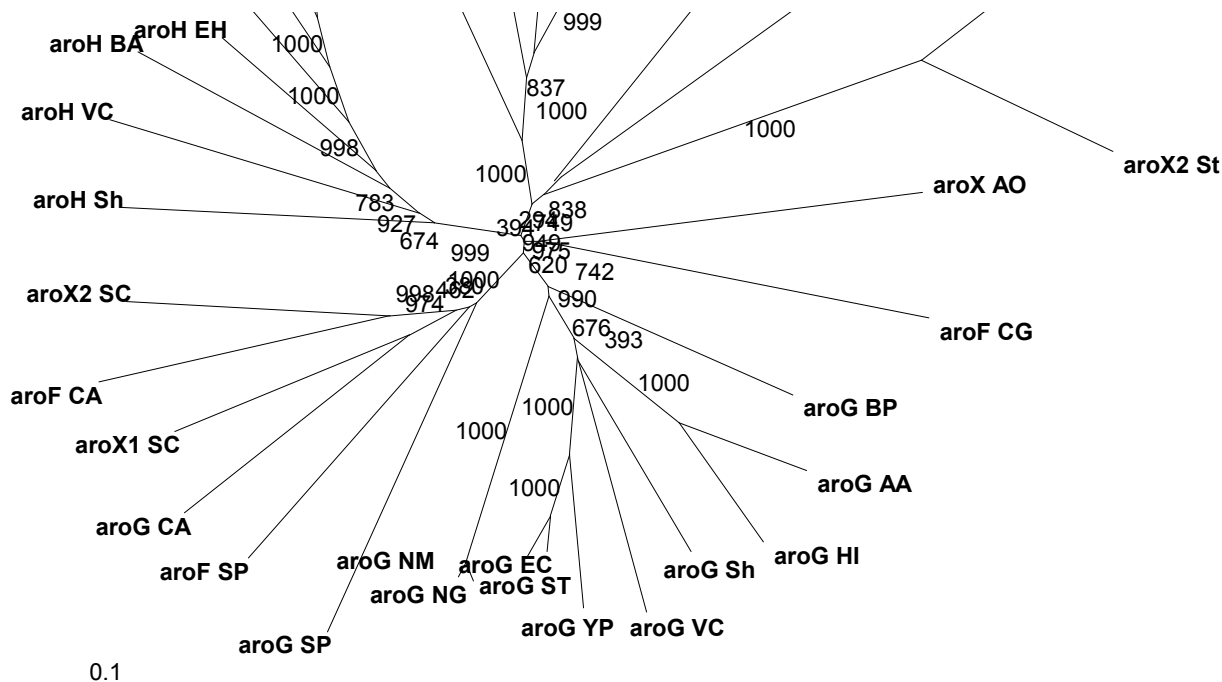


Figure 1. Evolutionary tree of DAHP-synthases. Numbers at internal nodes: bootstrap values. Additional notation: AM: *Amycolatopsis methanolica*, AO: *Amycolatopsis orientalis*, BA: *Buchnera aphidicola*, BP: *Bordetella pertussis*, CA: *Candida albicans*, CG: *Corynebacterium glutamicum*, DR: *Deinococcus radiodurans*, EH: *Erwinia herbicola*, NG: *Neisseria gonorrhoeae*, NM: *Neisseria meningitidis*, PA: *Pseudomonas aeruginosa*, SC: *Saccharomyces cerevisiae*, Sh: *Shewanella putrefaciens*, SP: *Schizosaccharomyces pombe*, St: *Streptococcus pneumoniae*. Numbers on internal nodes: bootstrap values (1000 replications).

```

EC  gggc-cttagtcgcccgaATGACTAGAGAACTAGTGCATtagcttatttttttgttatcatg
ST  gggg-cttaaccgcaggATGACTAGTAAACTAGTTTAAtggattgggttttt-gttatcatg
KP  gggg-ctgaatcgtagaATGACTAATACACTAGTACATtattaaaaatttg-tttatgctg
YP  aggg-atgagtcgctcggATGTTTAACTAAATATTTTCATgagtgattatttgcggtttata
EH  tgggcctgactcgccgcATGACTGATGGGTAACCGCGGctgaactgtaaagcgg-cgctg
BA  acttctgtcactgtatttttttagtataactattaaacttatcctccctaataattacaatt

```

```

EC  ctaaccaccgcccggaggtgtgacacacctcgacttgaaatcagcagcgattggtttatcg
ST  ccgttctgctcgcgaggtgtgacacgcctcgacttgaaatcagcagggattggtttatcg
KP  tcacccgtggcctatacaggatcatgcccctgctcttgcaataagcagggattggtttatcg
YP  tcctctaaatttgataaaattacagtgatgacggttgaattcatgtgattttgggtttatg
EH  ccgtcgcttttgatttcattattaaattccg-gggtgaattcgagactaatggtttatg
BA  tttatattgattaatttttaaaatttattag-aaatgagtttatgactagtttagtttatcg

```

```

EC  tgatgcgcatcacttcccggcagtcctgcccgtagaagcaacaaatttctgagacttgtaatg
ST  tgatgcccacttcccggtagtcctgcccgtgaaacaacaaatttctgagacttgtaatg
KP  tgatgcgcatcacttcccggcattttcgcgcggaagaacagatttttgagacccc-aatg
YP  tgatgcccacttcccggatttctgcccgcagagaagaacagatttagagagttttcatg
EH  tgaccaccatcacttcccggatttc-accgctgcgaagaaccgttttcgagaaacc-catg
BA  tgattt-tatcacatcttgctattcttttcaggataaaaaattcatttagagaacaa-aatg

```

Figure 2. Tentative alignment of *aroH* upstream regions of enteric bacteria. EH: *Erwinia herbicola*. BA: *Buchnera aphidicola*. Notation: **BOLDFACE ITALIC**, TRP box (*E. coli*); **BOLDFACE**, predicted TRP box; **CAPITALS**, region homologous to the TRP box of *E. coli*; underlined: 5' UTR of the *E. coli aroH* gene.

Unique DAHP-synthases from *H. influenzae* and *A. actinomycetemcomitans* belong to the AroG lineage (Fig. 1) and have a typical phenylalanine inhibition site (below). However, in both genomes *aroG* genes are preceded by strong TRP boxes, but have no candidate TYR boxes (Mironov et al., 1999).

Table 1. Candidate TYR and TRP boxes. "Pos." column: position relative to the start codon. Italics: diverged "box shadows" of *Y. pestis* (see text).

gene	aroH			aroG			aroF		
	Box	Score	Pos.	Box	Score	Pos.	Box	Score	Pos.
<i>E. coli</i>	TRP	5.60	-166	TYR	4.52	-91	TYR TYR TYR	5.28 4.83 4.08	-165 -113 -90
<i>S. typhimurium</i>	TRP	5.74	-165	TYR	4.37	-91	TYR TYR TYR	4.46 4.98 4.18	-164 -113 -90
<i>K. pneumoniae</i>	TRP	5.52	-165	no gene			TYR TYR TYR	4.97 5.14 4.11	-164 -111 -88
<i>Y. pestis</i>	TRP	2.73	-166	TYR	3.74	-94	TYR TYR TYR	2.91 4.59 4.53	-169 -119 -97
<i>V. cholerae</i>	none			TYR	5.36	-189	TYR TYR TYR	5.13 3.82 4.92	-124 -101 -29
<i>H. influenzae</i>	no gene			TRP	6.20	-81	no gene		
<i>A. actinomycet.</i>	no gene			TRP	5.15	-105	no gene		
<i>S. putrefaciens</i>	none?			none?			none?		
<i>P. aeruginosa</i>	no gene			none?			none?		

Finally, the absence of both TRP and TYR candidate boxes in *S. putrefaciens* and *P. aeruginosa* is not surprising, since transcriptional regulation of aromatic amino acid metabolism in these bacteria seems to be quite different from that in other gamma-proteobacteria (Panina, unpublished observations).

Allosteric control of DAHP synthase activity by feedback inhibition

Experiments (Ray et al., 1988; Weaver and Herrmann, 1990) have shown that mutations of Val and Gly at positions 147, 149 in AroH and Pro at position 148 in AroF are sufficient to confer feedback resistance to tryptophan and tyrosine respectively. Fig. 3 shows alignment of this site together with a second, spatially close region. Together these two regions form the amino acid binding pocket.

	****	*****
AroH_EC	LDMVTGQFIAD	HMFLSPDKDGQMTIYQT
AroH_KP	LDMVTGQFIAD	HMFLSPDKQGQMTIYQT
AroH_YP	LNMVTGQYIAD	HMFLSPDKTGQMTIYQT
AroH_EH	LDMVIGQFIAD	HMFLSPDKLGQMTIYQT
AroH_BA	LDMVIGQFIAD	HLFFAPNKDGQMTINHT
AroH_VC	LDMITGQYIAD	HYFYSPDKNGRMTVYRT
AroH_SP	LDMVNGQYIAD	HIFYSPDKDGAMSVYRT
AroG_EC	LDMITPQYLAD	HCFLSVTKWGHSAIVNT
AroG_ST	LDMITPQYLAD	HCFLSVTKWGHSAIVNT
AroG_YP	LDMITPQYLAD	HCFLSVTKWGHSAIVNT
AroG_VC	LDMITPQYVAD	HHFLSVTKFGHSAIVET
AroG_SP	LDMITPQYVAD	HHFLSVTKFGHSAIVST
AroG_AA	LDMITPQYLAD	HHFLSVTKFGHSAIVST
AroG_HI	LDMITPQYLAD	HYFLSVTKFGHSAIVST
AroG_NG	LDMITPQYYAD	HHFLSVTKAGHSAIVHT
AroG_NM	LDMITPQYYAD	HHFLSVTKAGHSAIVHT
AroG_BP	LDMITPQYIAD	HHFLSVTKGGHSAIVST
AroF_EC	LDPNSPQYLGD	HRFVGINQAGQVALLQT
AroF_ST	LDPNSPQYLGD	HRFVGINQAGQVALLQT
AroF_KP	LDPNSPQYLGD	SRFVGINQAGQVCLLQT
AroF_YP	LDPNSPQYLGD	HRFMGINQSGQVCLLQT
AroF_VC	LDPISPQYLAD	HRFMGINREGQVALLTT
AroF_SP	LDPISPQYISE	HRFMGINQQGQVALLQT
AroF_PA	LDPISPQYLQD	HRFLGINQQGGVSVITV
AroG_CG	LEPNSPQYYAD	HFFFGTSDDGALSVVET
AroG_PA	LQPLAAGYFDD	HRHFGLDPHGH PALIET
AroF_DR	LDPFAPQYLFD	HAFFTIDEDGRAIVHT

Figure 3. Alignment of the feedback inhibition site in bacterial DAHP-synthases. Notation: see the legend to Fig. 2.

Analysis of the specific patterns allows us to predict phenylalanine inhibition for the DAHP-synthase from *Bordetella pertussis*.

The first site in the DAHP-synthase from *Corinebacterium glutamicum* coincides with the pattern in tyrosine-inhibited isoenzymes from gamma-proteobacteria, although the second site is less conserved. However, this fact and position of this protein in the phylogenetic tree makes it likely that this is AroF rather than AroG, as annotated in GenBank.

The second site of AroH in *Buchnera aphidicola* has lost some residues absolutely conserved in other tryptophan-inhibited DAPH-synthetases. This can be explained by selection for tryptophan production in these bacteria causing possible loss of feedback inhibition (cf. above).

Finally, the experimental evidence about inhibition of DAHP-synthases in *Pseudomonas aeruginosa* is contradictory. One of them (AroF) is probably inhibited by tyrosine, whereas the other, annotated as AroG, is said to be tryptophan-dependent (Whitaker et al., 1982) or phenylalanine-dependent (Maksimova et al., 1991). Analyses of the feedback inhibition site or the phylogenetic tree cannot resolve this problem.

Acknowledgements

We are grateful to Dr. V. Alyoshin for discussions. This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program «Human Genome», and INTAS (99-1476). Preliminary sequence data were obtained from The Institute for Genomic Research WWW site.

References

1. S. Ahmed, B. Rightmire, and R.A. Jensen, *Mol. Biol. Evol.*, **5**, 282 (1988).
2. D.A. Benson, M.S. Boguski, D.J. Lipman, J. Ostell, B.F. Ouellette, B.A. Rapp and D.L. Wheeler, *Nucleic Acids Res.*, **27**, 12 (1999).
3. J. Camakaris, and J. Pittard, *J. Bacteriol.*, **120**, 590 (1974).
4. C.H. Doy, *Biochem. Biophys. Res. Commun.*, **26**, 187 (1967).
5. C.H. Doy and K.D. Brown, *Biochim. Biophys. Acta*, **104**, 377 (1965).
6. M.S. Gelfand, and A.A. Mironov, "Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes" 1st Conf. BGRS-98, **1**, 147 (1998).
7. C.L. Grove, and R.P. Gunsalus, *J. Bacteriol.*, **169**, 2158 (1987).
8. N.P. Maksimova, I.N. Olekhovich and Iu.K. Fomichev, *Genetika*, **27**, 217 (1991).
9. A.A. Mironov, E.V. Koonin, M.A. Roytberg and M.S. Gelfand "Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes" *Nucleic Acids Res.*, **27**, 2981 (1999).
10. A.A. Mironov, N.P. Vinokurova and M.S. Gelfand, "GenomeExplorer: software for analysis of complete bacterial genomes" (This volume) (2000).
11. R. D. M. Page, *Bioinformatics*, **14**, 819 (1998).
12. J. Pittard, and F. Gibson, *Curr. Top. Cell. Regul.*, **2**, 29 (1970).
13. J.M. Ray, C. Yanofsky and R. Bauerle, *J. Bacteriol.*, **170**, 5500 (1988).
14. O.H. Smith, *J. Biol. Chem.*, **237**, 3566 (1967).
15. P.R. Srinivasan, and D.B. Sprinson, *J. Biol. Chem.*, **234**, 716 (1959).
16. J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins, *Nucleic Acids Res.*, **25**, 4876 (1997).
17. TIGR. <http://www.tigr.org>
18. L.M. Weaver and K.M. Herrmann, *J. Bacteriol.*, **172**, 6581 (1990).
19. R.J. Whitaker, M.J. Fiske and R.A. Jensen, *J. Biol. Chem.*, **257**, 12789 (1982).

COMPARATIVE APPROACH TO ANALYSIS OF REGULATION IN COMPLETE GENOMES: ATTENUATORS OF AROMATIC AMINO ACID OPERONS OF GAMMA-PROTEOBACTERIA

¹Vitreschak A. G., ²Gelfand M. S.

¹Institute for Problems of Information Transmission RAS, Moscow, Russia

²State Scientific Center GosNII Genetika, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: attenuators, operons, RNA secondary structure

Resume

Comparative analysis is a powerful method of prediction of the RNA secondary structure. It has been used for prediction of both regulatory and structural RNAs [Waterman, 1989; Gelfand et al., 1999]. A somewhat different approach is to predict gene regulation by analysis of RNA patterns [Dandekar and Sibbald, 1990; Lisacek et al., 1994; Knoop et al., 1994; Biolloud et al., 1996; Vitreschak et al., 1998, 1999]. We have used to analyze attenuators of transcription of the aromatic amino acid operons from the genomes of gamma proteobacteria (*Salmonella typhi*, *Yersinia pestis*, *Vibrio cholerae*, *Haemophilus influenzae*, *Actinobacillus actinomycetemcomitans* and several other species for which only isolated fragments of the genomes are available). Biosynthetic operons *trp* and *phe*, aminoacyl-tRNA synthetase operon *pheST* and catabolic operon *tnaAB* have been considered.

Regulation of the *trp* operon is similar in *E. coli*, *S. typhi*, *Y. pestis* and *V. cholerae*. However, the RNA sequence is diverged, and only the high level features, namely the alternative secondary structures and the leader peptide enriched by the tryptophan codons are conserved. In *Pasteurellaceae* the situation is more complicated. In these genomes the *trp* operon is split into two parts, *trpEDC* and *trpBA*. Candidate attenuator structures can be constructed upstream of both operons. It is noteworthy that candidate binding sites for the tryptophan repressor TrpR also can be found upstream of both operons [Mironov et al., 1999]. A candidate attenuator structure also can be found upstream of the *trpBA* operon of a distant bacterium *Chlamydia trachomatis*. It is likely that this operon has been horizontally transferred from some gamma proteobacterium [Stephens et al., 1998]. Indeed, it is absent from the genome of a closely related *C. pneumoniae* [Kalman et al., 1999], whereas in *C. trachomatis* it is transcribed divergently from the repressor gene *trpR*. Gamma proteobacteria is the only taxonomic group known to contain the *trpR* gene.

Comparison of candidate attenuators for the *phe* operon from *E. coli*, *S. typhi*, *Erwinia herbicola*, *Y. pestis*, *V. cholerae*, *H. influenzae*, *A. actinomycetemcomitans*, and *Xanthomonas campestris* showed some specific features of taxonomic groups. Thus, in enteric bacteria and *V. cholerae* the distance between the pause stem-loop and the terminator is several nucleotides, whereas in *Pasteurellaceae* (*H. influenzae* and *A. actinomycetemcomitans*) it exceeds 25 nucleotides. In *Y. pestis* the *phe* attenuator is disrupted by an insertion of IS200 on the complementary strand. At that, the terminal stem-loops of IS200 substitute for disrupted elements of the attenuator retaining the alternative structure at both parts. The distal candidate attenuator contains all the structural elements, but it is very distant from the *pheA* gene, so that transcription from the original promoter regulated by the distal attenuator yields an untranslated leader of more than 800 nucleotides. The proximal structure contains all the RNA elements, but no candidate leader peptide (no open reading frame).

Candidate *pheST* attenuators were constructed for *S. typhi*, *Y. pestis*, and *H. influenzae*. Candidate *tnaAB* attenuators were constructed for *E. coli*, *Enterobacter aerogenes*, *Proteus vulgaris*, and *H. influenzae*.

Acknowledgements

We are grateful to A.A. Mironov and E. Panina for useful comments. This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Biolloud, B., Kontic, M. and Viari, A. (1996), *Nucleic Acids Res.*, **24**, 11395-1403.
2. Dandekar, T. and Sibbald, P.R. (1990) *Nucleic Acids Res.*, **18**, 4719-4725.
3. Gelfand, M.S., Mironov, A.A., Jomantas, J., Kozlov, Yu.I. and Perumov, D.A. (1999) A conserved RNA structure element involved in the regulation of bacterial riboflavin biosynthesis genes. *Trends Genet.*, **15**, 439-442.
4. Kalman, S. et al. (1999) *Nature Genetics*, **21**, 385-389.
5. Knoop, U., Kloska, S. and Brennicke, A. (1994) *J. Mol. Biol.*, **242**, 389-396.

6. Lisacek, F., Diaz, Y. and Michel, F. (1994) *J. Mol. Biol.*, **235**, 1206-1217.
7. Mironov, A.A., Koonin, E.V., Roytberg, M.A. and Gelfand, M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **27**, 2981-2989.
8. Stephens, R.S., et al. (1998) *Science*, **282**, 754-759.
9. Vitreschak, A., Bansal, AK. and Gelfand, M.S. (1998) Conserved RNA structures regulate initiation of translation of *Escherichia coli* and *Haemophilus influenzae* ribosomal protein operons. *1st Conf. BGRS-98*, vol. **1**, p. 229.
10. Vitreschak, A., Bansal, AK., Titov, I.I. and Gelfand, M.S. (1999) Computer analysis of regulatory patterns in complete bacterial genomes. Translation initiation of ribosomal protein operons. *Biofizika*, **44**, 601-610 (in Russian).
11. Waterman, M.S. (1989) *Mathematical Methods for DNA sequences*, Boca Raton: CRC Press, ch. 8, pp. 185-224.

AVOIDANCE OF PALINDROMES IN PROCARYOTIC GENOMES AND RESTRICTION-MODIFICATION SYSTEMS

¹Panina E.M., ^{2*}Gelfand M.S.

¹Moscow State University, Moscow, Russia

e-mail: misha@imb.imb.ac.ru

*Corresponding author

Keywords: palindromes, restriction-modification systems, procaryotic genomes, complete genome analysis

Resume

Introduction

Avoidance of palindromes in bacterial genomes is a well known phenomenon (e.g. Brendel et al., 1986; Karlin et al., 1992). Recently it was suggested that it is linked to restriction-modification (RM) systems (Gelfand and Koonin, 1997). Indeed, no avoidance of palindromes was observed in the genomes of organelles: mitochondria (derived from gamma-proteobacteria) and chloroplasts (derived from cyanobacteria). This correlates well with the absence of RM-systems in these genomes, in contrast to other gamma-proteobacteria (e.g. *Escherichia coli*) and cyanobacteria (e.g. *Synechocystis* sp.) possessing multiple RM systems and strongly avoiding palindromes. Similarly, genomes of *Chlamidia* have no genes for RM-systems and show very limited avoidance of palindromes, unlike genomes of related *Firmicutes* (e.g. *Bacillus subtilis*).

The number of completely or almost completely sequenced prokaryotic genomes has reached several dozens. Thus the aim of the present study was to check and possibly extend the above observations on this new material.

Data and Methods

Complete genome sequences of eubacteria *Bacillus subtilis*, *Borrelia burgdorferi*, *Chlamidia trachomatis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *M. pneumoniae*, *Rickettsia prowazekii*, *Synechocystis* sp., *Thermotoga maritima*, *Treponema pallidum* and archaea *Archaeoglobus fulgidus*, *Methanococcus jannaschii* and unfinished sequences of eubacteria *Actinobacillus actinomycetemcomitans*, *Bordetella pertussis*, *Campylobacter jejuni*, *Clostridium acetobutylicum*, *Deinococcus radiodurans*, *Enterobacter faecalis*, *Mycobacterium leprae*, *Neisseria gonorrhoeae*, *N. meningitidis*, *Porphyromonas gingivalis*, *Rhodobacter capsulatus*, *Pseudomonas aeruginosa*, *Salmonella typhimurium*, *Streptococcus pneumoniae*, *S. pyogenes*, *Streptomyces coelicolor*, *Vibrio cholerae*, *Yersinia pestis* were taken from GenBank (Benson et al., 1999). Data about RM-systems were taken from REBASE (Roberts and Macelis, 1996).

Let $N(W)$ be the observed count of oligonucleotide $W=w_1...w_m$ in a sequence of length L . The expected count assuming the Markov model of the maximum applicable length $m-2$ is

$$K(W) = K(w_1...w_m) = N(w_1...w_{m-1}) N(w_2...w_m) / N(w_2...w_{m-1}).$$

Contrast $C(W)$ is a measure of difference between the observed and expected count (Brendel et al., 1986; Gelfand and Koonin, 1997):

$$C(W) = (N(W) - K(W)) / (L^{1/2} \sigma)$$

where σ is the standard deviation of the difference $N(W)-K(W)$. According to (Schbath et al., 1995; Gelfand and Koonin, 1997),

$$\sigma^2 = [K(W) / L] [1 - N(w_1...w_{m-1}) / N(w_2...w_{m-1})] [1 - N(w_2...w_m) / N(w_2...w_{m-1})].$$

Results and Discussion

Tetranucleotide palindromes are strongly avoided in *H. influenzae*, *A. actinomycetemcomitans*, *N. gonorrhoeae*, *S. pneumoniae*, *S. pyogenes* and are not avoided in *B. pertussis*, *S. coelicolor*, *P. aeruginosa*, *M. tuberculosis*, *M. leprae*, *M. genitalium*, *M. pneumoniae*. Pentanucleotide palindromes are strongly avoided in *E. coli*, *S. typhimurium*, *Y. pestis*, *V. cholerae*, *H. influenzae*, *A. actinomycetemcomitans*, *B. subtilis*, *H. pylori*, *P. gingivalis* and not avoided in *B. burgdorferi*, *R. prowazekii*, *M. jannaschii*, *T. maritima*, *C. acetobutylicum*. Hexanucleotide palindromes are most strongly avoided in *S. coli*, *S. typhimurium*, *Y. pestis*, *P. gingivalis*, *D. radiodurans*, *R. capsulatus*, considerable avoided in *H. influenzae*, *V. cholerae*, *N. gonorrhoeae*, *P. aeruginosa*, *S. coelicolor*, *M. jannaschii*. They are weakly avoided in *M. tuberculosis*, *M. leprae*, *C. jejuni*, *B. burgdorferi*, *C. acetobutylicum*, *R. prowazekii*, *T. pallidum*, and not avoided in *T. maritima*, *M. pneumoniae*, *M. genitalium*, *E. faecalis*.

Thus there is only weak, if any, avoidance of palindromes in the genomes of obligate intracellular parasites *M. genitalium*, *M. pneumoniae*, *R. prowazekii*, *C. trachomatis*. Neither restriction endonucleases, nor methylases have been found in the genomes of *M. genitalium* and *R. prowazekii*. *M. pneumoniae* has a single restriction endonuclease *hsdS*, and *C. trachomatis* has gene *CT024* encoding an adenine-specific DNA methylase. The latter may explain unusual avoidance of tetranucleotide palindromes in the *C. trachomatis* genome.

However, not only palindromes recognized by the species' own RM-systems are avoided, but rather all palindromes as a specific subset of oligonucleotides. It has been suggested (Gelfand and Koonin, 1997) that it is caused by the fact that RM-systems are a very rapidly evolving system subject to frequent horizontal transfer. Thus avoidance of palindromes may be a relic of all RM-systems encountered by a genome. Some examples directly corroborate this explanation. For instance, CAGCTG is one of the most avoided hexanucleotides in the genomes of enteric bacteria *E. coli*, *S. typhimurium*, *Y. pestis*, *V. cholerae*. There are no known RM-systems with this specificity in these genomes. However, this 6-palindrome is recognized by PvuI system of the closely related bacterium *Proteus vulgaris*. Similar observations can be made about a number of other 6-palindromes, in particular CTGCAG and CCCGGG. On the other hand, GATATC is overrepresented in *E. coli* and other members of *Enterobacteriaceae* despite the fact that it is the recognition site of the EcoRV RM-system from *E. coli*. However, it is avoided in the related bacteria from the families *Vibrionaceae* and *Pasteurellaceae*.

Correlations of palindrome contrasts in related genomes are often stronger than correlations of contrasts of all oligonucleotides. This can be observed for epsilon-proteobacteria *C. jejuni* and *H. pylori*, for gamma-proteobacteria from the family *Enterobacteriaceae*, for *B. subtilis* and *C. acetobutylicum*. This is an additional evidence of the influence of horizontal transfer.

Acknowledgements

We are grateful to Andrey Mironov for useful discussions. This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Benson, D.A. *et al.* (1999) *Nucleic Acids Res.*, **27**, 12-17.
2. Brendel, V., Beckmann, J.S. and Trifonov, E.N. (1986) *J. Biomol. Struct. Dynam.*, **4**, 11-21.
3. Gelfand, M.S. and Koonin, E.V. (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.*, **25**, 2430-2439.
4. Karlin, S., Burge, C. and Campbell, A.M. (1992) *Nucleic Acids Res.*, **20**, 1363-1370.
5. Roberts, R.J. and Macelis, D. (1996) *Nucleic Acids Res.*, **24**, 223-235.
6. Schbath, S., Prum, B. and de Turckheim, E. (1995) *J. Comput. Biol.*, **3**, 417-437.

ONE APPROACH FOR ANNOTATION AND CONFIRMATION OF DISCOVERIES OF ALTERNATIVE SPLICE EVENTS USING RECONSTRUCTED EXTENDED UNSPLICED TRANSCRIPTS FROM GENES ON CHROMOSOME 22

*^{1,2}*Babenko V.*, ¹*van Heusden P.*, ¹*Hide W.*

¹South African National Bioinformatics Institute, University of the Western Cape, South Africa

²Institute of Cytology and Genetics SB-RAS, Novosibirsk, Russia

e-mail: bob@sanbi.ac.za

*Corresponding author

Keywords: human chromosome 22, alternative splicing, splice detection methods, EST

Resume

Results and discussion:

We perform an analysis of alternative splicing using Chromosome 22 as a guide to the genomic counterparts of reconstructed transcripts given following advantages: a priori the gene set is nonredundant, unless gene clusters such as immunoglobins, etc are considered (a); all transcripts have their genomic counterparts (b).

Availability of complete human chromosome sequence provides us with a unique genome model. Our techniques have been developed and aimed at applying the tool using analogous chromosome data, but with the purpose of gaining enough experience that a set of rules can be developed to expand the detection with confidence to likely splice events of non-genome sequence confirmed reconstructed transcripts. We note that current methods for splice detection have a high rate of error due to the lack of genome confirmation and predictive rules.

We created an unspliced mRNA set: we used most proximal and distal 40-bp oligonucleotides for each of 544 mRNAs (we didn't consider the immunoglobulin lambda gene cluster) described as "gene" or "Predicted Gene" in the Sanger genes table (http://www.sanger.ac.uk/cgi-bin/c22_genes_table.pl). We scanned Chromosome 22 [Dunham et al., 1999] for an exact match using a Perl script (<http://www.perl.com>). Genomic (un-spliced) sequences for 467 genes from a known annotated set of 545 loci were successfully located. Further iterations with varying oligonucleotide length allowed us to extract all of the entries. Extraction of genomic sequences included an additional 500bp stretch flanking the start and end of each mRNA. We then employed STACKed ESTs (using STACK-PACK clustering [Miller et al., 1999]) from genes on Chromosome 22 and EST GENOME [Mott, 1997] to parse exon-intron structure of the each gene.

From this set two further sets are compiled, an exon set, comprising all annotated exons, and an intron set, representing the introns. For that we represented the genomic sequences as EMBL – formatted entries with mRNA parsing as the "FT mRNA join" feature. Then we applied the BioPerl (www.bio.perl.org) software suite to produce the intron set. After aligning both sets with STACKdb we report previously undetected alternative splicing on Chromosome 22 by means of CRAW report [Burke et al., 1998]. We used intron-EST hits to select the candidate genes. To check out possible artifacts such as contamination and intron retention [Mironov et al., 1999], [Croft et al., 2000] we checked the consistency of putative novel mRNA by verification of the uninterrupted open reading frame within the coding part of a putative isoform. We also confirm the variation of 3' – 5' UTR by observing additional ESTs within start and end flanks of particular genes. We also plan to use EST redundancy to identify highly and low expressed genes along with UTR length [Kochetov et al., 1998], [Kan et al., 2000].

Investigation of already reported alternative splicing phenomena (short or long isoform) for a 85/88kDa Ca²⁺-independent phospholipase A2 (iPLA2) gene (ACC Numbers: AF102988 (short) and AF102989 (long)) provided an initial experimental testbed. It is known [Larsson Forsell et al., 1997] that exon skipping mechanisms in this gene provides the basis for a short isoform. We observe the same phenomena for EST sequences compared against the mRNA set. In addition we report ESTs demonstrating alternative splicing in the exon 2. We demonstrate isoforms previously unreported.

An alignment of the phospholipase A2 unspliced masked sequence (inverted strand) with human EST sequences reveals both matching EST sequence and also some ESTs that do not match the counterpart mRNA. By use of the unspliced genomic sequence in gene annotation we can improve detection of alternate isoforms of a gene, and rules such as tissue specific isoform variation can be used for consistent detection in non-genome confirmed reconstructed transcripts.

References

1. I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, D.M. Beare, M. Clamp, L.J. Smink, R. Ainscough, J.P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow,
3. K.N. Bates, O. Beasley, C.P. Bird, S. Blakey, A.M. Bridgeman, D. Buck, J. Burgess, W.D. Burrill,
4. K.P. O'Brien, et al., "The DNA sequence of human chromosome 22" *Nature* 402, 489–495(1999).L. Croft, S. Schandorff, F. Clark, K. Burrage, P. Arctander, J.S. Mattick, "ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome" *Nature Genetics* 24, 340-341(2000).
6. A.A. Mironov, J.M. Fickett, M.S. Gelfand, "Frequent alternative splicing of human genes" *Genome Research* 9, 1288-1293(1999).
7. J. Burke, H. Wang, W. Hide, D.B. Davidson, "Alternative gene form discovery and candidate gene selection from the indexing projects" *Genome Research* 8, 2767-290(1998).
8. Z.-Y. Kan, W. Gish, E. Rouchka, J. Glasscock, D. States, "UTR Reconstruction and Analysis Using Genomically Aligned EST Sequences"//The 8th conference on Intelligent Systems in Molecular Biology (ISMB'2000), La Jolla, California. In Press.
9. A.V. Kochetov, I.V. Ischenko, D.G. Vorobiev, A.E. Kel, V.N. Babenko, L.L. Kisselev, N.A. Kolchanov, "Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features" *FEBS Lett.* 440, 351-355(1998).
10. P.K. Larsson Forsell, B.P. Kennedy, H.E. Claesson, "The human calcium-independent phospholipase A2 gene multiple enzymes with distinct properties from a single gene" *Eur. J. Biochem.* 262,575-585(1999).R. Mott, "EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA" *Comput. Appl. Biosci.* 13, 477-478(1997).R.T Miller, A.G. Christoffels, C. Gopalakrishnan, J. Burke, A.A. Ptitsyn, T.R. Broveak, W.A. Hide, "A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base" *Genome Research* 9, 1143-1155(1999).

FAST SEARCH OF ALL TANDEM REPETITIONS IN NUCLEOTIDE SEQUENCES

¹*Giraud M.*, ²*Kolpakov R.*, ^{*3}*Kucherov G.*

¹Ecole Normale Supérieure de Lyon, Lyon, France

²French-Russian Institute for Informatics and Applied Mathematics, Moscow University, Moscow, Russia

³INRIA-Lorraine/LORIA, Villers-lès-Nancy, France

e-mail: kucherov@loria.fr

*Corresponding author

Keywords: maximal repetitions, tandem repeats, nucleotide sequences, algorithm, complexity

Resume

Motivation:

Successive repetitions of a fragment in DNA sequences often bear important information and is characteristic for many genomic structures (telomere regions, tandem repeats and others). From practical viewpoint, satellites and alu-repeats are involved in chromosome analysis and genotyping, and thus are of great interest to genomic researchers. Tools for finding successive repeats are nowadays an obligatory part of integrated systems for analyzing and annotating whole genomes.

Results:

We present an implementation of an efficient (linear-time) algorithm for finding all so-called maximal repetitions of a given sequence. The algorithm was recently proposed by R.Kolpakov and G.Kucherov [1] and is a modification of Main's algorithm [2]. The maximal repetitions can be viewed as an encoding of all exact tandem repeats in the sequence, that is all its fragments repeated contiguously. The algorithm uses advanced word-algorithmic techniques, such as the DAWG (Directed Acyclic Word Graph) data structure, s-factorization, longest common extension functions.

The package mreps has been developed. It includes the program reps implementing the search algorithm, as well as the program graf providing a graphical interface to visualize the repetitive structure of the sequence. The whole package is about 6000 commented lines of C code.

The program has been tested on two prokaryotic (*E.coli* and *B.subtilis*) and one eukaryotic (yeast) genomes. Some interesting long repetitions have been found.

Availability:

<http://www.loria.fr/remag/logiciels.html>

References

1. Kolpakov, R., and Kucherov, G. Finding maximal repetitions in a word in linear time. In: Proceedings of the 1999 Symposium on Foundations of Computer Science, New York (USA) (1999), 596–604.
2. Main, M.G. Detecting leftmost maximal periodicities. *Discrete Applied Mathematics* 25 (1989), 145–153.

RECONSTRUCTION OF THE OPEN READING FRAMES BY USING EST MULTIPLE ALIGNMENT AND DYNAMIC PROGRAMMING

**Vishnevsky O.V., Katokhin A.V. Babenko V.N.*

Institute of Cytology and Genetics of SB RAS, Novosibirsk, Russia

e-mail oleg@bionet.nsc.ru

*Corresponding author

Keywords: EST, ORF reconstruction, errors correction, recognition

Resume

Motivation:

In the recent years, the methods directed to receiving information about gene structure from the EST (Expression Sequence Tags) analysis are being rapidly developed. However, mRNA sequences may be reconstructed on the basis of ESTs with the relatively low accuracy. This may be explained by such facts as numerous mistakes during EST sequencing (nucleotide substitutions, insertions, and deletions) and insufficient accuracy of the software programs reconstructing open reading frames on the basis of EST analysis.

Results:

We have developed the software program ORFScan that reconstructs open reading frame of translation by dynamic programming and multiple alignment of EST sequences without using information on homologues. The program produces the false positive and false negative estimates at the levels of 3% and 1%, respectively.

Introduction

In the recent years, the EST (Expressed Sequence Tags) analysis became one of the most widely used approaches for detecting and deciphering primary gene structure. ESTs are obtained by one-time sequencing of 5' and 3'-ends of clones randomly extracted from cDNA libraries, which are prepared out of mRNA pool extracted from definite cell type. The resulted EST fragments, by means of different computer tools such as Cap2 [3], are compiled into assembles, which are the base for consensus reconstruction. Such approach notably decreases the cost of an experiment and reveals all the variety of mRNAs expressed in definite cell types at different stages of their functioning and stages of individual development of an organism, although it is accompanied by the fall in accuracy. During mRNA reconstruction, the most serious mistakes leading to significant alterations in reconstructed amino acid sequence are the false insertions or deletions of nucleotides shifting translation frame. With this respect, the lowering of mistakes during mRNA reconstruction on the basis of ESTs is a very important problem. Solving of this problem by means of increase in accuracy by multi-stage sequencing needs many expenses. However, in order to find and correct mistakes in coding sequences it is possible to use various computer methods developed for reconstruction of open reading frames (ORF). The examples of such programs are GRAIL [6] built upon neuron networks and dynamic programming; GENSCAN [2] applying hidden Markov model; and GeneFinder [5] based on linear discriminate analysis. However, these programs were developed mainly for the search and reconstruction of the gene structure in extended DNA regions, but they were not intended for the open frame reconstruction on the basis of EST analysis. With this respect, novel tools, e.g. ESTScan [4], are being recently developed that are specially oriented to EST analysis. We have developed a program ORFScan enabling to reconstruct open reading frame of translation by using dynamic programming and multiple alignment of ESTs. As a result of multiple alignment of the EST assemble by the program Cap2 [3], the EST-consensus is formed, which serves as a basis for reconstruction of the open reading frame of translation, and amino acid sequence corresponding to consensus is constructed. The discriminative feature of our program is application of not only resulting consensus for ORF reconstruction, but also the entire process of EST alignment. The program produces the false positive and false negative estimates at the levels of 3% and 1%, respectively.

Materials and Methods

At the first stage of the approach described is compiling of ESTs into assembles, their multiple alignment, and detection of the EST consensus. This stage is produced by the program Cap2 [3]. The derived consensus may contain stop-codons appearing due to erroneous sequencing of ESTs, during formation of the EST assemble, and its alignment.

The task of the following stage is to reveal positions of false nucleotide deletions and insertions in the EST consensus by means of multiple alignment analysis. The mistakes of the types as above cause appearing of

positions with the gaps (denoted as “-”) in the multiple alignment. These positions with gaps can be separated into two types. To the first type are referred those, in which the type of an error could be detected with high significance. In case a gap is present in the most sequences, then false nucleotide insertion is most likely. Under construction of the resulting consensus, these positions with gaps were removed. If a gap is present in a minimal number of sequences, one may suppose an erroneous deletion of a nucleotide during the sequencing. Under construction of the resulting consensus, the nucleotide, which is present in most sequences in a certain position, is inserted into these positions. To the second type of positions are referred those where both the gaps and nucleotides occur in a ratio, which does not permit to detect exactly the type of mistake. In this case, under construction of resulting consensus, such positions are marked by the sign “=”. Let us denote such positions as I/D, by considering them as the place of possible nucleotide insertions or deletions, which may cause the ORF shift, false stop-codon occurrence, and ORF breakage. A position is denoted as I/D if the ratio of sequences G with gaps satisfies to condition $20\% < G < 70\%$.

Then, by means of information on possible nucleotide insertion-deletion, we may reconstruct ORF, corresponding to the EST assemble analyzed, by dynamic programming method. Nucleotide sequence located between two I/D, we call as the interval W_n , $n=1, \dots, N$, where N is the total amount of intervals in consensus. Each of three possible ORFs within the interval W_n we name as the segment $S_{n,m}$ of this interval ($m=1,2,3$). Let the set of segments of consensus be $\mathbf{S} = \{S_{n,m}\}$, $1 \leq n \leq N$, $1 \leq m \leq 3$. Thus, consensus contains $3N$ segments in total.

At the following stage of analysis, the task is to find the route with the maximal length of transitions between the segments of neighboring intervals of consensus. This route should start with AUG and terminate by stop-codon or the ending of a sequence. The search of such route is made by using the method of dynamic programming. The program recurrently analyses the segments of all intervals from the first to the last, by supplementing each segment with the longest route, which passed through preceded segments.

We mean that the route passes from the segment $S_{i,p}$ to the segment $S_{i+1,q}$ with transition of the reading frame if $p \neq q$. Let us introduce the matrix $Q_{i,j}$ ($i \leq (n+1)$, $j \leq 3$), the element of which $q_{i,j}$ equals to the number of transitions with the change of the reading frame between the segments, which occurred along the most continuous route $\langle S_{n,m}, \dots, S_{i,j} \rangle$ in the course of its passing through the j -th segment of the i -th interval.

Let us construct the matrix $|A_{i,j}|_{(n+1) \times 3}$. The element $A_{i,j}$ equals to the length of the most continuous route $\langle S_{n,m}, \dots, S_{i,j} \rangle$, which is acquired during the passing through the j -th segment of the i -th interval. The matrices $|A_{i,j}|$ and $|Q_{i,j}|$ are filled recurrently:

$$A_{i,j} = \max_k (A_{i-1,k} + R_{i,j}) \quad (1)$$

$$k^* = \underset{k=\arg \max(A_{i,j})}{\operatorname{argmin}} (Q_{i,j} / Q_{i,j} = Q_{i-1,k} + \partial_k), \text{ where } \partial_k = \begin{cases} 1, k \neq j \\ 0, k = j \end{cases}$$

$$\text{and } P_{k^*}(S_{i,j}) > \text{BORD} \quad (2)$$

Here $R_{i,j}$ is the distance from beginning of the segment $S_{i,j}$ to the first occurring stop-codon in it, or to the ending of this segment (if it has no stop-codons); ∂_k - is a penalty function for transition between the segments, which belong to different frames; $P_{k^*}(S_{i,j}) = p_{k^*}(S_{i,j})/p_n(S_{i,j})$ - coding potential of $S_{i,j}$. Here $p_{k^*}(S_{i,j})$ is a probability to observe $S_{i,j}$ in the k^* -th frame of the coding region, $p_n(S_{i,j})$ - probability to observe $S_{i,j}$ within mRNA 3'-untranslated region. $p_{k^*}(S_{i,j})$ was calculated according the model of nonhomogeneous Markov chain of the 6th order, $p_n(S_{i,j})$ - homogeneous Markov chain of the 6th order [1]. BORD is a limiting value (0.1).

Following (1), we choose the number of a route k with the maximal length $A_{i,j}$. In case the route with maximal length $A_{i,j}$ occurs several times, condition (2) enables to choose such route k^* , which has the minimal number of transitions $Q_{i,j}$ between the segments of different reading frames. Thus, the matrix $Q_{i,j}$ is constructed at each step simultaneously with the matrix $A_{i,j}$. This way of calculation of the element $A_{i,j}$ provides the maximal elongation of the route supplemented with the minimization of the number of transitions between segments at each stage of the route reconstructing.

To reconstruct the consequence of transitions between the segments for each of three routes, we set the massive of classes $W[r]$, where r is the number of a route ($r=1,2,3$). Each class $W[r]$ carries an information about the number of initial interval for the r -th route and about position of the AUG start-codon in it. Besides, let us organize the class M for storing an information about the current terminated, or ending by stop-codon, ORF with the maximal length (ORFmax), which is formed during the algorithm's processing in the course of passing through the i -th interval. Whence the terminated ORF with the longer length is found at the following stages of the algorithm processing, the information about novel ORF is recorded into the class M. The class M contains an information about the length of the ORFmax, its starting and ending positions, and the number of transitions between the segments of different reading frames along ORFmax. This information is sufficient for

reconstruction of the whole route, from the first till the last segment of the ORF with maximal length. The reconstruction is produced by traceback procedure, by means of moving in inverse mode along the matrix A.

In the course of the ORFmax route reconstruction, an insertion of a certain number of symbols "X" is made into positions of transitions between the segments of different reading frames. Under transition from the segment $S_{i,j}$ to $S_{i-1,j}$ segment, the number of inserted symbols X equals to $|j-j^*|$. On the basis of nucleotide sequence obtained as described above, the corresponding amino acid sequence is formed. To the triplets containing the symbol "X", corresponds the symbol of undetermined amino acid "?".

Results and Discussion

For evaluation of the ORF reconstruction accuracy, an analysis was performed as follows:

1. From the DOTS (Database of Transcribed Sequences) (http://www.cbil.upenn.edu/DOTS_ANNOT/annot?page=home), accumulating partial sequences of murine and human genes, an assemble of ESTs aligned by the Cap2 program [3] was extracted.
2. By means of BLASTN program (<http://www.ncbi.nlm.nih.gov/BLAST/>), in GenBank database, the real nucleotide sequence is found, which is at most homologous to the sequence of consensus constructed.
3. Then the amino acid sequence R2 is extracted from the GeneBank, this sequence corresponding to the real nucleotide sequence detected.
4. By applying the program ORFScan described here, the ORF with the maximal length is detected and amino acid sequence R1 encoded by this ORF is constructed.
5. By pairwise alignment, a comparison of amino acids R1 and R2 is made; then the false positives and false negatives are estimated. False positives appear when amino acids occurring in the real sequence R1 are absent in reconstructed sequence R2. The value of false positive estimate α_1 equals to the ratio of the number of amino acids, which are absent in the sequence R2, to the total length of the sequence R1.

False negatives appear when amino acids, which are present in reconstructed sequence R2, lack in the real sequence R1. The value of false negative estimates α_2 is equaling to the ratio of number of amino acids lacking in the sequence R1 to the total length of the sequence R1. Averaged results of accuracy evaluation of ORF reconstruction by the ORFscan program are given in Table 1.

A comparison of the quality of the ORFScan program processing with the other Internet-accessible programs (GENSCAN (<http://ccr-081.mit.edu/GENSCAN.html>), GeneFinder (<http://genomic.sanger.ac.uk/gf/gf.shtml>), GRAIL (<http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm>) and ESTScan (<http://www.ch.embnet.org/software/ESTScan.html>)) is performed. The values α_1 and α_2 for each of these programs were obtained similarly in accordance with the scheme described above.

Table 1. Efficiency of some software packages application to ORF reconstruction.

	False positives, α_1	False negatives, α_2
ORFScan	0.03	0.01
ESTScan	0.13	0.08
GenScan	0.08	0.01
Grail	0.27	0.01
GeneFinder	0.09	0.01

As can be seen from the Table, our program ORFscan has demonstrated the minimal false positive estimate (0.03), whereas the GRAIL is characterized by the maximal false positives. The comparison of several computer packages has revealed that increase of false positives in GRAIL is mainly produced by the fact that this program loses the inner parts of the coding sequences due to their false cutting. This mistake is also typical to the other packages (except ESTscan and ORFscan), but in a less extent. Besides, for the programs oriented to gene structure reconstruction in genome sequences (GenScan, Grail, and GeneFinder), some increase in false positives is typical, which occurs due to earlier termination of the coding sequence in comparison with the ORFscan program, which is originally designed for analysis of ESTs. However, late termination in the program ESTscan causes increase in false negatives number.

Acknowledgement

The work is supported by Integration Project of SB RAS No 65. The authors are grateful to G. Orlova for translation of the paper into English.

References

1. Borodovsky M., Sprizhitskii Yu., Golovanov E. and Alexandrov A., Statistical patterns in the primary structures of functional regions in E.coli., *Molekulyarnaya Biologiya*, 20, 1390-1398.
2. Burge C., S. Karlin. Prediction of Complete Gene Structures in Human Genomic DNA. *J.Mol.Biol.*, 268, 1997, 78-94.
3. Huang X. An improved sequence assembly program. *Genomics*, Apr 1;33(1), 1996, 21-31.
4. Iseli C., Jongeneel C.V. and Bucher P., ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *ISMB* 7, 1999, 138-148.
5. Solovyev V., Salamov A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Ismb*. 5, 1997, 294-302
6. Uberbacher E.C. and Mural R.J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA*. 88, 1991, 11261-11265.

ON RELATIONSHIPS BETWEEN GENE EXPRESSION EFFICIENCY AND NUCLEOTIDE CONTENT OF THE PROTEIN-CODING SEQUENCES

**Likhoshvai V.A. Matushkin Yu.G.*

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

e-mail: likho@bionet.nsc.ru

*Corresponding author

Keywords: efficiency, gene expression, local complementarity, codon frequencies, mathematical model, computer analysis

Resume

Motivation:

Measures based only on the frequencies of codon occurrence (analogous to codon adaptation index) do not completely reflect gene expression efficiency.

Results:

In the present work, the more generalized measure is suggested (elongation efficiency index), which takes into account both codon occurrence frequencies and the level of the local mRNA complementarity. An adequate recognition of highly expressed genes according to their protein-coding sequences was performed in 23 unicellular organisms, namely, in 22 procaryotes and 1 eukaryota (yeast).

Availability:

Application of the measure developed enables adequate recognition of highly expressed genes according to their protein-coding regions.

Introduction

Expression efficiency is one of the most general characteristics of protein-coding genes. In some organisms, protein synthesis efficiency is well correlated to codon occurrence frequencies in the coding genes (Sharp & Li, 1986; Sharp & Devine, 1989; Shields et al., 1988; Shields & Sharp, 1987). This phenomenon may be explained by accumulation in the genes, coding effectively synthesized proteins, of the most quickly "readable" codons. The rate of codon reading is related to concentrations of corresponding tRNA fractions (Ikemura, 1985; Yamao et al., 1991). An analysis of dynamical models of mRNA translation taking into account most general regularities such as matrix pattern of protein synthesis, genetic code degeneracy, the presence of tRNA-adaptors, develops the hypothesis given above and reveals some regularities of evolutionary drive of the gene codon content (Bulmer, 1987; Shields, 1990; Bagnoli & Lio, 1995). For example, it is proved that codons divergence in evolution by the rates of reading them by a ribosome and increase in the number of the "quickly readable" codons within gene content with the growth of expression efficiency is a clear consequence of co-evolution in codon content and tRNA concentrations (Likhoshvai, 1992). This points to the principle possibility to predict gene expression efficiency by the frequencies of occurring in them codons. For some organisms, this possibility is really true (Li & Luo, 1996).

In the present work, by the example of 23 organisms with the known nucleotide content, we demonstrate that gene nucleotide content is optimized in such a way that average period for one peptide bond formation is inversely proportional to gene expression efficiency. As a result, we have developed a measure for evaluation of nucleotide content quality within protein coding gene regions (elongation efficiency index), which accounts both codon content features and peculiarities of local mRNA secondary structure and which allows to estimate relative gene expression efficiency.

Methods and algorithms

The genes of the following organisms were analysed: *Aquifex aeolicus*, *Archaeoglobus fulgidus*, *Bacillus subtilis*, *Borrelia burgdorferi*, *Chlamydia trachomatis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Methanobacterium thermoautotrophicum*, *Methanococcus jannagchii*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Pyrococcus horikoshii*, *Rickettsia prowazekii*, *Treponema pallidum*, *Synechocystis* и *Saccharomyces cerevisiae*. Nucleotide sequences of genomes and their marking-off were extracted from the из банка EMBL gene bank, issue 56, provided by the Russian program "Human genome".

Designations

N – total number of genes in the sample analyzed; i – gene number in some pre-determined numeration (i = 1, N); n_i – total number of codons in the i-th gene; j – codon number in the i-th gene, starting from codon next to the initiator one; C – total number of codons in genetic code; $\delta(i, j)$ – number of codon in some pre-fixed numeration of all codons of the genetic code, which is located in j-th position of the i-th gene ($1 \leq \delta(i, j) \leq C$); \mathfrak{R} – the set of genes selected by a definite parameter from the set of analyzed genes (the representative set of genes); Ω – the set of gene numbers from \mathfrak{R} . If not specified, we consider that the set of genes \mathfrak{R} is determined.

Codon efficiency index (CEI)

The quality of nucleotide content was estimated by the value of codon efficiency index:

$$CEI(i) = \left(\sum_{j=1}^{n_i} \alpha_{\delta(i,j)} \right) / n_i, \text{ where } \alpha_{\delta(i,j)} \text{ is the frequency of codon } \delta(i,j) \text{ occurrence in } \mathfrak{R}.$$

Elongation efficiency index (EEI)

The quality of the i-th gene nucleotide content was evaluated by the value of elongation efficiency index EEI(i), which has two components. The first is related to codon content of a sequence and evaluates the average time of tRNA exposition in the ribosomal A-site (Likhoshvai, 1992); and the second – nucleotide content of a sequence viewed as their ability to form “hairpins”, i.e., secondary structures, which retard ribosome movement.

$$EEI(i) = \left[\left(\sum_{j=1}^{n_i} \beta_{\delta(i,j)} \right) / n_i + t_{\max} \cdot \frac{r_1 LCI(i)^{\log_2 \left(\frac{r_2(1-r_1)}{r_1(1-r_2)} \right)}}{1 - r_1 + r_1 LCI(i)^{\log_2 \left(\frac{r_2(1-r_1)}{r_1(1-r_2)} \right)}} \right]^{-1}, \text{ where } 0 \leq t_{\max}, 0 \leq r_1 \leq r_2 \leq 1, \text{ are the given}$$

parameters, hereinafter, the calculations are made under $t_{\max}=1$, $r_1=0.05$, $r_2=0.5$. These parameters were found experimentally for two organisms, *E.coli* and *M.genitalium*. They proved to be suitable for the rest 21 organisms.

The value of EEI(i) in the first approximation, with the accuracy of proportionality coefficient, is equaling to the averaged constant of elongation rate for codons of the i-th gene.

$$\beta_{\delta(i,j)} = \frac{\sum_{m=1, C} \sqrt{\sum_{i' \in \Omega} \sum_{k: \delta(i', k)=m} 1}}{\sqrt{\sum_{i' \in \Omega} \sum_{k: \delta(i', k)=\delta(i, j)} 1}}, \text{ reverse value } 1/\beta_{\delta(i,j)} \text{ in the simplest case corresponds to the optimal relative}$$

concentration of t-RNA, which is complementary to the j-th codon of genetic code.

$$LCI(i) = \frac{\sum_{m=1}^{3n_i - s_{\max} - l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[\sum_{l=l_{\min}}^{l_{\max}} \zeta \left(\text{con}(m, m+s-1), \overline{\text{con}(m+s+l-1, 2m+2s+l-2)} \right) \right] \right\}}{3n_i - s_{\max} - l_{\max}},$$

where $\text{con}(i, j)$ is a gene context from i-th to j-th nucleotides and $\overline{\text{con}(i, j)}$ is a complementary gene context from the j-th to i-th nucleotides ($i \leq j$), $\zeta(\text{conext1}, \text{conext2})=1$, if the words conext1 and conext2 are identical; otherwise, $\zeta(\text{conext1}, \text{conext2})=0$. The length of the considered reverted repeat is not less than s_{\min} and not higher than s_{\max} . The distance between these inverted regions is not less than l_{\min} and not higher than l_{\max} (in the present paper, we assume $s_{\min}=3$, $s_{\max}=6$, $l_{\min}=3$, $l_{\max}=50$). The LCI(i) value corresponds to the number of complementary nucleotides per one nucleotide of the sequence under analysis. The average time, consumed by ribosome for transpeptidation stage was set equal for all codons and genes and was not taken into account in calculations of EEI(i) index.

We have developed an algorithm for automated ordering of genes, which is by iteration classifies the sequences by the decrease of an index (CEI or EEI). When the gene list with the highest index values becomes constant, the process is interrupted.

To evaluate the relationships between the index values and the gene expression efficiency, we suppose that the genes coding ribosomal proteins are highly expressed in all the organisms analyzed. Then the high values of CEI or EEI indices in these genes may serve as a criterion of coincidence of indices to gene expression efficiency. We consider this criterion significant if probability to obtain the distribution of ribosomal proteins coding genes accidentally was less than 10^{-3} .

Results and discussion

We have determined that if to make the ordering of genes according to increase of the CEI value, then different organisms may principally differ by the pattern of ribosomal genes distribution. For *B.subtilis*, *C.pneumoniae*, *C.trachomatis*, *H.influenzae*, *P.horikosii*, *A.fulgidus*, *P.abbysi*; together with the studied previously *E.coli* and *S.cerevisiae* (Li H. and Luo L., 1996.), we obtained direct correspondence of the CEI value to ribosomal genes expression. However, for *A.aeolicus*, *H.pylori*, *M.thermoautotrophicum*, *M.tuberculosis*, *M.pneumoniae*, *R.prowazekii*, *Synechocystis*, *T.maritima*, *T.pallidum*, this dependancy was either weak or even absent, that is, in these organisms, the genes were randomly distributed within the ordered set. In *B.burgdorferi*, *U.urealyticum*, *M.jannagschii*, *M.genitalium*, *D.radiodurans*, the genes encoding ribosomal proteins are characterized by very low values of indices. Since the index of codon efficiency CEI is primarily a measure of the rate of isoacceptor aminoacyl-tRNA binding to translating A-site and it ignores the stages of transpeptidation and translocation, we make a conclusion that generally evaluation of the gene expression efficiency by its protein-coding region should account all stages of elongation. At the translocation stage, as the factor influencing expression efficiency, may serve a non-uniform ribosome movement along mRNA due to local secondary structures, which may be formed prior the ribosome. To estimate this phenomenon quantitatively, we suggest using the index of local complementarity LCI, whereas as the quantitative measure of nucleotide content of the reading frame, could be used the index of efficiency elongation EEI described in the section "Methods and materials".

Table 1. Quantitative comparison gene expression efficiency prediction based on CEI and EEI values.

Organism	%GC	Number of tRNA precursors	Quality, based on CEI ¹	Quality, based on EEI ²
<i>S.cerevisiae</i>	38	>200	+++++	+++++
<i>E.coli</i>	50.8	86	++++	++++
<i>C.pneumoniae</i>	40.6	38	++++	+++
<i>B.subtilis</i>	43.5	88	+++	+++
<i>H.influenzae</i>	38	56	+++	+++
<i>P.abbysi</i>	44.7	Data are absent	+++	+++
<i>A.fulgidus</i>	48.6	46	++	++
<i>C. trachomatis</i>	41.3	37	++	+
<i>P.horikosii</i>	41.9	46	++	++++
<i>A.aeolicus</i>	39	44	+-	++
<i>H.pylori</i>	37	36	+-	+++
<i>M.pneumoniae</i>	40.0	33	+-	+
<i>M.thermoautotrophicum</i>	50	39	+-	+++
<i>M.tuberculosis</i>	65.6	45	+-	+++
<i>R.prowazekii</i>	29.0	33	+-	+
<i>T.maritima</i>	46.2	46	+-	+++
<i>T.pallidum</i>	53.1	45	+-	++
<i>Synechocystis</i>	47.7	44	+-	+++
<i>B.burgdorferi</i>	28.6	18	-	+++
<i>M.jannagschii</i>	31.4	37	--	++
<i>D.radiodurans</i>	67.0	48	---	+++
<i>M.genitalium</i>	31.7	33	---	+++
<i>U.urealyticum</i>	25.5	30	---	+++

¹-quality of relatedness between the CEI and gene expression efficiency;

²-quality of relatedness between the EEI and gene expression efficiency

We ranged the EEI by increase of values in all annotated genes of 23 organisms and obtained the results presented in Table 1. The main conclusion – in all the organisms considered, the ribosomal proteins coding genes have rather high values of EEI and occupy the top place in the list of ordered genes. As for CEI index, the same result of ordering can be obtained only for some organisms.

The results suggested enable to make some conclusions about possible mechanisms of action of two different factors onto formation of nucleotide content of the protein-coding genome regions in different organisms. The first factor is related to conditions favoring to establishment and supporting in the nucleotide genome content of AT/GC-asymmetry. The second factor is related to evolutionary trend leading to increase in translation efficiency. Both factors are non-specific by nature and may act simultaneously on the same genome regions. However, the whole spectrum of action of these factors on nucleotide content falls between two extremes: a) nucleotide content of the protein coding sequences is selected in such a way that the translocation stage limits elongation; b) nucleotide content of the protein coding sequences is evolves so that the tRNA binding stage becomes limiting for elongation.

For *M.genitalium*, the variant (a) is realized. On the other hand, in genome of this organism due to unknown external reasons, the decreased content of G,C-nucleotides is supported (31,7%) (Table 1). In the low expressed genes, this content is even less (25%, calculations are not shown). Due to above reasoning, translation is strongly inhibited by the local complementarity, this makes a little effect to the cell, because the genes are weakly expressed. However, this organism uses in the process of translation a limited set of tRNAs (Table 1), so many groups of synonymous codons may be recognized by the single tRNA, and mutations of the type purine ↔ pyrimidine at the third codon position are neutral. This enables the cell to increase the G,C-nucleotides content (up to 40%, calculations are not given) and in this way to decrease the level of local complementarity and increase the rate of translation of highly expressed genes.

Another extreme is represented by *E.coli* and *S.cerevisiae*. This option could be realized both under strong selection of genome nucleotide content (*S.cerevisiae*) and under the neutral conditions (*E.coli*) (Table 1). As a crucial factor we consider the appearance of the extended tRNA set, which causes the strong dependency of translation upon codon content of a gene (Table 1). In this case, the factors of selection are combined in such a way that elongation efficiency is mainly determined by the stage of recognition of a codon in the ribosome A-site, whereas the stages of transpeptidation and translocation proceed rather effectively.

In general, several variants of interactions between the impacts of different stages of peptide bond formation into the overall efficiency of elongation stage may be realized. Moreover, it cannot be excluded that for different genes of a single organism, and even for different parts of a gene, drastic variation of the relative impact of different stages into elongation process could happen due to the influence of various and, sometimes, random reasons.

Acknowledgments

The work was supported by the National Program "Human Genome" (Grant №106) and Integration Project of SB RAS No 66. The authors are grateful to G. Orlova for translating the paper into English.

References

1. P.M. Sharp and W.-H. Li, "Codon usage in regulatory genes in Escherichia coli not reflect selection for 'rare' codons" *Nucl.Acids Res.* **14**, №19, 7737 (1986).
2. P.M. Sharp and K.M. Devine "Codon usage and gene expression level in Dictiostelium discoideum: highly expressed genes do 'prefer' optimal codons" *Nucl.Acids Res.* **17**, №13, 5029 (1989).
3. D.C. Shields, P.M. Sharp, D.G. Higgins, F. Wright, "'Silent' sites in Drozophila genes are not Neutral: Evidence of selection among synonymous codons" *Mol.Biol.Evol.* **5(6)**, 704 (1988).
4. D.C. Shields and P.M. Sharp, "Synonymous codons usage in Bacillus subtilis reflects both translational selection and mutational biases" *Nucl.Acids Res.* **15**, №19, 8023 (1987).
5. T. Ikemura, "Codon usage and tRNA content in unisellular and multicellular organisms" *Mol.Biol.Evol.* **2**, 13 (1985).
6. F. Yamao, Y. Andachi, A. Muto, T. Ikemura, S. Osawa, "Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins" *Nucl.Acids Res.*, **19**, №22, 7737 (1991).
7. M. Bulmer "Coevolution of codon usage and transfer RNA abundance" *Nature.* **325**. 728 (1987).
8. D.C. Shields "Switches in species-specific codon preferences: the influence of mutation biases" *J.Mol.Evol.* **31**, 71 (1990).
9. F. Bagnoli and P. Lio, "Selection, mutations and codon usage in a bacterial model" *J.theor.Biol.* **173**, 271 (1995).
10. V.A. Likhoshvai, "Rare codons: Fortuity or regularity?" *In Modelling and Computer Methods in Molecular Biology and Genetics* / eds Ratner A. and Kolchanov N.A. Nova Science Publishers. Inc. USA. 463 (1992).
11. H. Li and L. Luo, "The relation between codon usage, base correlation and gene expression level in Escherichia coli Yeast" *J.theor.Biol.* **181**, 111 (1996).

DETERMINING MARKOV MODEL OF GENETICAL TEXTS BY STOCHASTIC COMPLEXITY ESTIMATION

*Orlov Yu.L., ¹Potapov V.N.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

¹Institute of Mathematics by name of S.L. Sobolev, SB RAS, Novosibirsk, Russia

e-mail: orlov@bionet.nsc.ru

*Corresponding author

Keywords: stochastic complexity, Markov models, genetical texts, functional sites

Resume

Motivation:

For functional annotating of genome sequences, the algorithms are needed that could reveal contextual features of sequences, significant for functional sites recognition.

Results:

A method is suggested for constructing the generating contextual source-tree (a variant of hidden Markov model) of symbolic DNA and protein sequences. As a criterion for a model's ascertainment serves the estimation of stochastic complexity of data within the frames of the model chosen. The regulatory DNA sequences from the "Samples" database (<http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/>) and amino acid protein sequences were taken into analysis.

Availability:

<http://wwwmgs.bionet.nsc.ru/mgs/programs/>.

Introduction

Genetical texts are known to have the Markov property, that is, in real sequences, the probability of a letter occurrence depends upon preceding set of symbols [Durbin R. et al., 1998]. Markov models are intensively used for studying gene structure and for dissection (marking off) of biologically significant regions [Peshkin L. & Gelfand M., 1999]. The tasks appear as to recognize and mark off functional sites, as to ascertain mathematical model for description of these regions.

We suggest to use for studying of DNA sequences the method, which is widely applied in the theory of data coding and data compression [Barron A. et al, 1997]. A statistical model is considered, such that probability of the next in turn symbol occurrence is determined by preceding context. The contexts may be of various length, but neither of them is the ending of another. It is convenient to represent such set of contexts in a graphical form as a binary tree. For a 4-lettered alphabet describing DNA sequences, the tree will be "quaternary", that is, it has 4 branches at each level (see Fig. 1, 2).

The method suggested is based on the algorithm "Context" developed by J. Rissanen in 1983. This algorithm enables to construct, according the sequence, the statistical model (generating source-tree), which generates this sequence, and to calculate the data complexity of this model [Orlov Yu. & Potapov V., 2000].

Methods and algorithms

Let $D=\{A,T,G,C\}$ be an alphabet. By a communication, we denote a sequence (word) $X_1...X_n=X^n \in D^n$. The subsequence of letters $S=X_{m-t}...X_{m-1}$ is called a context with the length t of occurrence of the letter X_m in the word X^n ($t < m \leq n$). By $X^n(S)$, we designate a sequence compiled out of letters occurring in the word X^n of the context S with conservation of the ordering of letters. We consider the DNA sequences as the communications generated by some stationary discrete source. For the Markov model of the k -th order constructed for DNA sequences, we have 4^k generating contexts (or states of a chain), which determine distribution of probabilities of letters' occurrences after these contexts:

$$P(D_i|S_j)=\theta_j^i \text{ and } \sum_{i=1}^4 \theta_j^i = 1, \text{ where } D_i \in \{A,T,G,C\}, S_j \in D^k, j=1,...,4^k.$$

Let us represent the set of states of the Markov source of the k -th order as the leaves of the tree T . Each leave corresponds to the context with the length k (see Fig. 1, $k=3$). If (i) the leaves (suspended vertexes), corresponding to the states of a source, are the brothers (i.e., they are connected with the same vertex) and (ii)

their distributions of probabilities of generating the symbol coincide (the states are equivalent), then can unify these states into a single one (Fig. 2). By unifying the sets of equivalent states, we obtain the minimal tree of states T' . Here, the vertexes at the higher levels correspond to more short contexts. The generating source-tree obtained may occur to be not Markov. However, each source-tree may be supplied to Markov source-tree by reverse procedure.

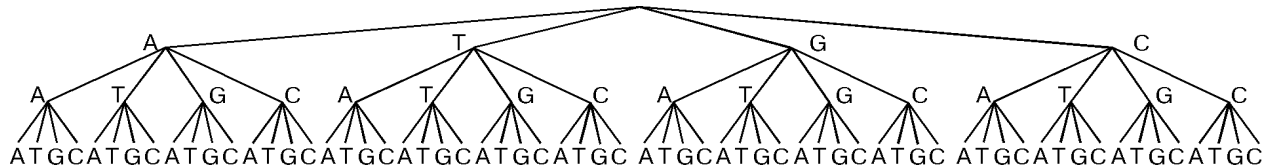


Figure 1. Markov source of the 3-rd order.

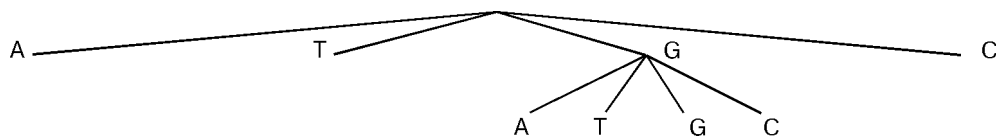


Figure 2. Context source-tree

For example, if a model M has a tree illustrated at Fig. 2, then the probability of the next letter occurrence depends only upon the preceding letter in case this letter is A, G or C, and upon the pair of letters AG, TG, GG, CG, in case the preceding letter is G. The contexts are determined by the letters from the routes directed bottom-up, from suspended vertexes to the route of the tree.

As for Markovian source, the probability of generating the sequence X^n is determined by probabilities of occurrence of the symbols X_i , comprising X^n in corresponding contexts S_i :

$$P(X^n) = P(X_1|S_1) P(X_2|S_2) \dots P(X_n|S_n),$$

where $S_i \in T'$ is a context, in which the letter X_i is contained in the word X^n , T' is a generating tree.

To Bernoulli's model M_0 (stationary source without memory) corresponds a tree, which consists only from the root. To Markovian model of the k -th order M_k corresponds complete tree with the depth k (see Fig. 1). As complexity of the communication X^n (e.g., DNA sequence) relatively the known distribution of probabilities θ , we set the module of the logarithm of the probability of this communication to the base 2,

$$L(X^n) = - \log P(X^n) = \log 1/P(X^n).$$

In particular, for Bernoulli's model, complexity equals to

$$L(X^n, \theta, M_0) = - \sum_{i=1}^4 C_i \log P(D_i), \tag{1}$$

where C_i is the number of the letter D_i occurrence in the sequence X^n ,

$$P(D_i) \text{ is probability of letters } D_i \in \{A, T, G, C\}, \theta = \{P(D_i)\}.$$

If to evaluate the probabilities via the frequencies, by setting $P(D_i) = C_i/n$, then complexity with the accuracy up to normality condition for the length of a sequence n will be equal to Shannon entropy ($-\sum p \log p$), or the sum of products of symbols probabilities to the logarithm of these probabilities. C. Shannon [Shannon C.E., 1948] has demonstrated that the module of the logarithm of the probability of a communication is the length of the shortest (in average) binary code of a communication.

As stochastic data complexity L^0 relatively the model M , we denote the minimal sum of complexities L of communication X_n , under some set parameters of a distribution model θ , and complexity of these parameters description H_n (complexity of a model M):

$$L^0(X^n, M) = L(X^n, \theta(X^n), M) + H_n(M) \tag{2}$$

J. Rissanen has obtained an estimate of complexity $H_n(M_0)$ of description of the model's parameters, this estimate depending upon the sequence length and the alphabet's size [Barron A. et al, 1997]. For a four-lettered alphabet, it equals to

$$H_n(M_0) = (3/2) \log n + 1/2 \log(\pi) - 3/2$$

Following J. Rissanen [Rissanen J., 1983; 1999], we suggest to consider that the model at best corresponds to the data, if the data in this model have minimal stochastic complexity. The limitation for the usage of the contexts with larger length is the growth of complexity H_n in description of added parameters.

An algorithm is the following recurrent procedure. Let T be the tree containing all the vertexes with the depth at least d , i.e., $T=D^d$. (As was shown by our practical analysis of DNA sequences, it is suffice to take $d=10$). For all the contexts S corresponding to suspended vertexes, we set $J(S)=L^0(X(S),M_0)$. For inner vertexes S , we determine

$$J(S)=\min \{L^0(X(S),M_0), \sum_{i=1}^4 J(D_iS)\}$$

If the first element in the brackets is less or equal to the second one, then from the tree T we remote the vertexes, which are the offspring of the vertex S . (The tree vertex offspring are those vertexes, which are connected with this vertex and located downwards). We remote exactly all the offspring. The values $L^0(X(S),M_0)$ are calculated by formulas (1) and (2). The values $J(S)$ are determined recurrently from suspended vertexes to the root R of the tree T , by removing the vertexes corresponding to non-informative contexts. The resulted tree T' corresponds to the required model M corresponding to the sequence X^n . The final complexity $J(R)$ is determined in the root R corresponding to empty context (lack of dependency upon the preceding symbols). As indicated in review by A. Barron (1997), the value $J(R)$ is close to the minimum of stochastic complexity of a sequence in the model among all the models with contextual trees with the depth at most d . The minimal order of Markovian dependency d , from which the construction of a tree begins, does not influence the structure of a tree at the upper levels. By this procedure, theoretically the contexts with maximal length (more than 10 nucleotides prior position analyzed) could be not considered. We failed to meet such dependencies during DNA analysis.

Results and discussion

The graphical representation of the contextual tree structure similar to those illustrated in Fig. 2, is typical for the most considered regulatory sequences, which were extracted from the "Samples" database (<http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/>). By generalizing the method for complexity estimation [Rissanen J., 1999; Orlov Yu. & Potapov V., 2000] for 2-, 3-, and 5-lettered alphabet, we have considered amino acid and nucleotide sequences in different codes: hydrophobic and hydrophilic, positively and negatively charged and neutral amino acids, GC/AT content.

The program developed performs to construct complexity profile in a sliding window in a graphical representation, by analogy to the program by P. Kosarev and V. Babenko (http://wwwmgs.bionet.nsc.ru/mgs/programs/gc_net/). However, our mathematical methodics of stochastic complexity evaluation differs from approach, which was applied by V.Gusev et al. (1987) and which evaluates the complexity of texts measured by the number of operations necessary for generating the text.

Acknowledgements

The authors are grateful to G. Orlova for help in translation of the manuscript into English, to N. Kolchanov and V. Babenko for valuable comments and scientific discussion. The work was supported by Russian Foundation for Basic Research (grant No 99-01-00531) and Integration project SB RAS.

References

1. Barron A., Rissanen J and Yu B. (1997) The minimum description length principle in coding and modelling. *IEEE Trans. Inform. Theory*, **43**, N.5, 669-683.
2. Gusev V.D., Kulichkov V.A., Chupakhina O.M. (1991) Analysis of genome complexity . A measure of complexity and classification of revealed structural peculiarities. *Mol. Biologia (Mosk)*, **25**, 825-834.
3. Durbin R., Eddy S.R., Krogh A. and Mitchison G. (1998) *Biological sequence analysis: probabilistic models of protein and nucleic acids*. Cambridge University Press, 1-347.
4. Orlov Yu.L. and Potapov V.N. (2000) Estimation of stochastic complexity of genetical texts. *Computational technologies (Novosibirsk)*, **5**, spec.issue, 5-15.
5. Peshkin L. and Gelfand M.S. (1999) Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*, **15**, (5), 980-986.
6. Rissanen J. (1983) A universal data compression system. *IEEE Trans.Inform.Theory*, **IT-29**, (5), 656-664.
7. Rissanen J. (1999) Fast universal coding with context models. *IEEE Trans.Inform.Theory*, **45**, (4), 1065-1071.
8. Shannon C.E. (1948) A mathematical theory of communication. *Bell Syst.Tech.J.*, **27**, pt.I., 379-423; pt.II., 623-656.

NEW ALGORITHMS FOR LARGE-SCALE EST CLUSTERING

^{*}Ptitsyn A., Hide W.

South African National Bioinformatics Institute, P/b X17 UWC Bellville 7535

email: ptitsyn@sanbi.ac.za

^{*}Corresponding author

Keywords: EST, clustering, sequence similarity, alternative processing, algorithm development, supercomputing

Motivation:

Fast growing EST databases contain a huge amount of information, but usually of very poor quality. EST clustering is aimed to improve the data quality for further datamining, as well as discovery of differential expression, alternative gene forms and SNPs. This is a challenging task even for the latest available supercomputers. New algorithms are required to improve the quality of the existing databases and cope with the evergrowing amount of data.

Results:

We have developed a new statistical measure of sequence dissimilarity with a linear computation complexity. This statistic is sensitive to local similarities. Based on this new algorithm for sequence comparison, we have developed a number of applications for large-scale EST clustering. The applications were tested on the real data from the human EST library and show a significant improvement in cluster quality and calculation time, compared to the existing systems.

Availability:

The programs will be available within open-source STACKPack package under the general GNU public license. More information can be found at <http://www.sanbi.ac.za>

Introduction

Continuous flow of EST data remains one of the richest sources for discoveries in modern biology. Untranslated 3' and 5' regions are particularly well represented in the EST databases, which make them more attractive for gene expression and regulation studies. The first step in EST datamining is usually associated with EST clustering, the process of grouping of original fragments according to their annotation, similarity to known genomic DNA, or each other. Clustered EST data, accumulated in the databases such as UniGene, STACK and TIGR HGI have proven to be crucial in various research areas from gene discovery to regulation of gene expression (Hillier et al. 1996). STACKpack has been developed for management and analysis of EST clustering at SANBI (Miller et al, 1999). Based on the STACK and STACKpack experience, we have developed a new clustering algorithm, which will be applied to the new generation of STACK or similar databases.

Methods and Algorithms

The key element of the new clustering algorithm is a new statistical measure of sequence similarity. This measure is a further development of the algorithm, proposed for fast sequence comparison by Strelets et al., 1994. We introduce an asymmetric approach, where one of the compared sequences is presented as a hash-table of oligonucleotide words and the other is scanned with a sliding window. The measure has linear computation complexity, yet it's fast enough to compete with sub-linear metrics and so does not require data pre-indexing. The metric is more rapid than the currently more widely used D2 (Torney et al., 1990) measure and demonstrates an improvement in detection of short regions of local similarity. The asymmetric approach helps to reduce both under- and over-prediction of local similarity. Control *in silico* experiments achieve over 96% of correct detection of a pair of fragments with 40bp similarity region. To select a threshold the following procedure was conducted: for each of 100 000 ESTs, randomly picked from the Genbank EST section a random counterpart sequence was generated by shuffling. Each pair of sequences was compared by the new algorithm and a This algorithm automatically ignores low-complexity regions like poly-tracts and short tandem repeats.

Implementation

The algorithm is implemented in two variants: loose clustering with assembly by a third party system and stringent clustering with simultaneous consensus generation and alternative variant apprehension. Each detected match is approved or rejected by following pair-wise alignment. This alignment is built-in for stringent clustering or in case of loose clustering is performed by a third-party software like PHRAP (Green, 1996) on the cluster assembly stage. Stringent clustering application is also able to include mRNA and genomic DNA

information when available. The same sequence comparison algorithm was also used to develop a masking program for repeat and vector fragment masking.

Results and discussion

Current release of STACKPack uses D2_cluster program, which implements a loose clustering approach. Loose clustering application, although showing less impressive performance on the benchmark datasets can be easier introduced as a plug-in into the existing EST clustering system. Stringent clustering application is implemented for testing in MS Windows 32-bit environment and still has to be ported to a Unix-based supercomputer. Its higher performance is mostly explained by the reduction of the data set during the clustering procedure. On-flight consensus generation also makes possible simultaneous detection of branching points and following alternative variants apprehension, as well as utilization of known mRNA or genomic DNA information.

Table 1. Comparison of the clustering software performance. The performance is measured in number of seconds required to complete clustering process. In case of Stringent clustering this also includes cluster assembly.

Clustering program	
D2_cluster	
Loose Cluster	
Stringent Cluster	
Platform	
SGI Origin2000	
(R10000/180MHz)	
	12302
	5660
	-
Pentium II/450	
	17014*
	6021
	2163
Compaq DS20 (EV6/500MHz)	
	5655
	2520
	-
	*data for 400MHz CPU

Most contemporary systems for EST clustering use third-party software for masking repetitive sequences and cloning vector fragments. When different algorithms are used it's always hard to select correct parameters and results in contamination of clusters by incompletely masked repeats. Utilization of the same algorithm for sequence comparison throughout all stages of clustering unifies the parameters and practically excludes potential leaks of unmasked repeats and vectors. This is the main reason for introduction of a new masking program based on the same algorithm as the clustering application.

Clustering applications are tested on a real data set (10.000 ESTs from human eye tissue division). Both variants show dramatic improvement in performance with comparison to D2_cluster in computation time and cluster/singleton ratio. The algorithm for clustering is also suitable for further parallel optimization.

References

1. Green, P. PHRAP 1994-1996. <http://bozeman.mbt.washington.edu/phrap.docs/phrap.html>
2. Hillier, L., N. Clark, T. Dubuque, K. Elliston, M. Hawkins, M. Holman, M. Hultman, T. Kucaba, M. Le, G. Lennon, M. Marra, J. Parsons, L. Rifkin, T. Rohlfing, M. Soares, F. Tan, E. Trevaskis, R. Waterston, A. Williamson, P. Wohldmann, and R. Wilson. 1996. Generation and Analysis of 280,000 Human Expressed Sequence Tags. *Genome Research* 6:807-828.
3. Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* Nov;9(11):1143-55
4. Strelets, V.B., Ptitsyn, A.A., Milanese, L., Lim, H.A. 1994. Data bank homology search algorithm with linear computation complexity. *Comp. Appl. Biosci.*, v.10, n. 3 (1994), pp. 319-322;
5. Torney, D.C., Burks, C., Davison, D., and Sirotkin, K.M. 1990. Computation of d^2 . A measure of sequence dissimilarity. In Bell, G. and Marr, T., eds., *Computers and DNA*, Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, New York.

PERIODIC PATTERNS IN SEQUENCE ORGANIZATION OF REPLICATION ORIGIN OF *ESCHERICHIA COLI* K-12 CHROMOSOME

**Kravatskaya G.I., Esipova N.G.*

V.A. Engelhardt Institute of Molecular Biology, Moscow, Russia

e-mail: GK@imb.imb.ac.ru, nge@imb.imb.ac.ru

*Corresponding author

Keywords: origin of chromosome replication, DNA unwinding, periodicity, matrix Fourier analysis

Resume

Motivation:

The process of initiation of DNA replication is one of the most important and insufficiently known in the cell. This process starts at special DNA sequences called replication origins. *E.coli* chromosomal replication during normal growth initiates bidirectionally at unique sequence (*oriC*). In order to reveal periodic regularities in the primary structure of *oriC* and analyse their role in the process of replication initiation, we applied matrix Fourier analysis.

Results:

oriC Fourier spectra were obtained and compared with that ones of other regions of *E.coli* complete genome. Using matrix Fourier analysis with sliding window technique several sites resembling *oriC* were revealed. Most of them are coincide with the sites of replication initiation in *E.coli* stable DNA replication mutants. The method applied can be useful for analysis of other complete genomes and prediction of the sites of possible replication initiation.

Introduction

The replication of the *E.coli* chromosome is normally initiated at the *oriC* site, the origin of replication. A sequence of 245 base-pairs (*oriC*) in the replication origin of the *E.coli* K-12 chromosome has been shown to provide all the information essential for initiation of bidirectional replication (Asada et al., 1982). It is known that *oriC* sequence is extremely saturated with direct and inverted repeats (Meijer et al., 1979). In this work we attempted to reveal another (periodic) regularities in the *oriC* nucleotide sequence. Discovering of periodicities in the DNA primary structure is important for understanding of regularities of higher order structures formation and stability.

Methods and algorithms

The Fourier transformation of nucleotide sequence is performed as described in (Makeev et al., 1996). The program applied was PERF (Makeev et al., 1996).

Implementation and results

Periodicities in the dispositions of nucleotides and dinucleotides in the origin of chromosome replication *oriC* from *E.coli* were studied by means of matrix Fourier analysis. Peaks corresponding to the periods $T=2$, 17, 93-98 nucleotides are the most high in the Fourier spectrum of *oriC* (Fig.1). Peaks corresponding to the periods $T=3$, 11, 19, 13, 24, 27, 28, 41, 79-81 nucleotides are also prominent, but not so high. The difference between *oriC* Fourier spectrum and that ones of adjacent to *oriC* regions are demonstrated (Fig.2, 3).

We have also demonstrated that *oriC* contains several regions with different periodic organization of nucleotide occurrences. We connected this result with the disposition on *oriC* of binding sites of initiator protein DnaA and regulatory proteins FIS and IHF (Woelker and Messer, 1993).

Matrix Fourier analysis of *E.coli* genome with sliding window (step = 100 nucleotides, the length of the window = 245 nucleotides) technique reveals that only 10 regions of *E.coli* genome have Fourier spectra resembling the Fourier spectrum of *oriC* in the sense of prominent peaks corresponding to the periods of $T=2$, 17, 93-98.

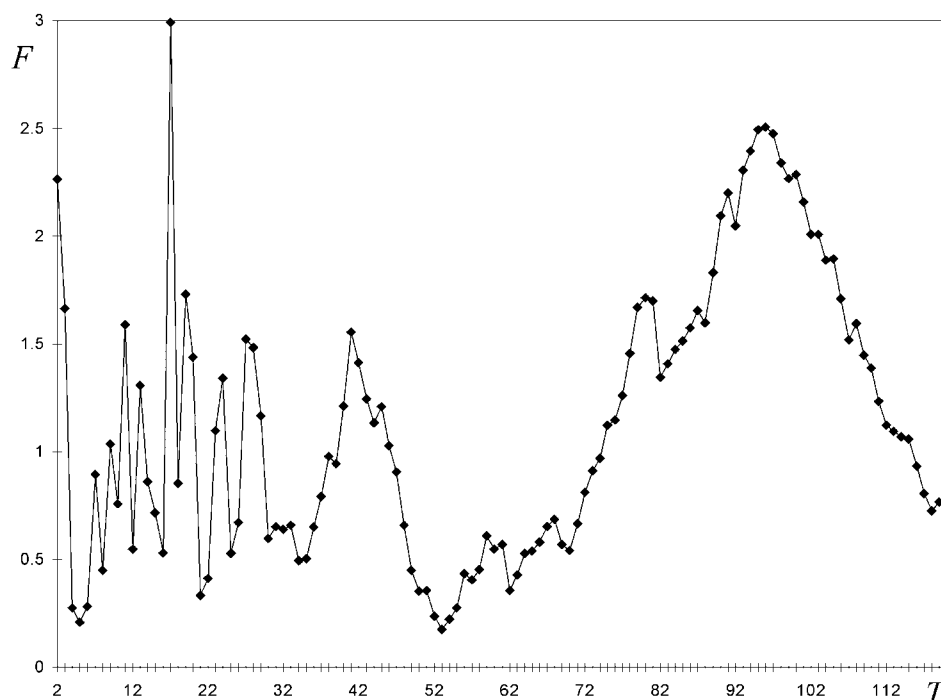


Figure 1. Fourier spectrum of *oriC* in terms of A,C,G,T nucleotide occurrences. T- the length of the period, F- corresponding to the period of T spectral power.

Discussion

The periodicity, corresponding to B-DNA pitch (10.5), is negligible in the Fourier spectrum of *oriC* in contrast to the Fourier spectra of flanking regions (Fig.2). The main periodicities of the *oriC* spectrum are not multiples of the B-DNA sugar-phosphate backbone period, that destabilizes DNA at *oriC* and contributes to the formation of a structure called a *replication bubble*. We suppose that the presence of strong periodicities destabilizing the B-form of DNA also contributes to the spontaneous unwinding (Polaczek, 1998) of DNA in *oriC*.

In stable DNA replication *sdr/rnh* mutants of *E.coli*, initiation of replication occurs in the absence of the normal origin of replication, *oriC* (de Massy, 1984). There are at least four fixed sites or regions of the *sdrA ΔoriC* chromosome from which DNA replication can be initiated in the absence of the *oriC* sequence. Most of the regions revealed by our method are the sites of replication initiation in *E.coli* stable DNA replication (*sdrA/rnh*) mutants. Two of the sites revealed are novel sites. Probably, these sites are also functional but normally repressed. Our results suggests, that the method applied can be useful for analysis of other complete genomes and prediction of the sites of possible replication initiation.

This work was supported by grant N 00-04-48351 from Russian Foundation of Basic Research (RFBR).

References

1. Meijer, M., Beck, E., Hansen, F.G., Bergmans, H. E.N., Messer, W., Meyenburg, K., Schaller, H., "Nucleotide sequence of the origin of replication of the Escherichia coli K-12 chromosome" Proc. Natl. Acad. Sci. USA. **76**, 580-584 (1979).
2. Asada K, Sugimoto K, Oka A, Takanami M, Hirota Y. "Structure of replication origin of the Escherichia coli K-12 chromosome: the presence of spacer sequences in the ori region carrying information for autonomous replication" Nucleic Acids Res. **10**, 3745-54 (1982).
3. Makeev, V.Ju., Tumanyan, V.G. Search of periodicities in primary structure of biopolymers: a general Fourier approach. CABIOS **12**, 49-54 (1996).
4. Makeev, V.Ju., Frank, G.K. , Tumanyan, V.G. "Statistics of periodic patterns in the sequences of human introns." Biophysics **41**, 263-268 (1996).
5. Woelker, B., Messer, W., "The structure of the initiation complex at the replication origin, *oriC*, of *Escherichia coli*" Nucleic Acids Res. **21**, 5025-5033 (1993).
6. Polaczek P, Kwan K, Campbell JL Unwinding of the Escherichia coli origin of replication (*oriC*) can occur in the absence of initiation proteins but is stabilized by DnaA and histone-like proteins IHF or HU. Plasmid **39**, 77-83 (1998).
7. de Massy B, Fayet O, Kogoma T. "Multiple origin usage for DNA replication in *sdrA(rnh)* mutants of Escherichia coli K-12. Initiation in the absence of *oriC*." J Mol. Biol. **178**, 227-36 (1984).

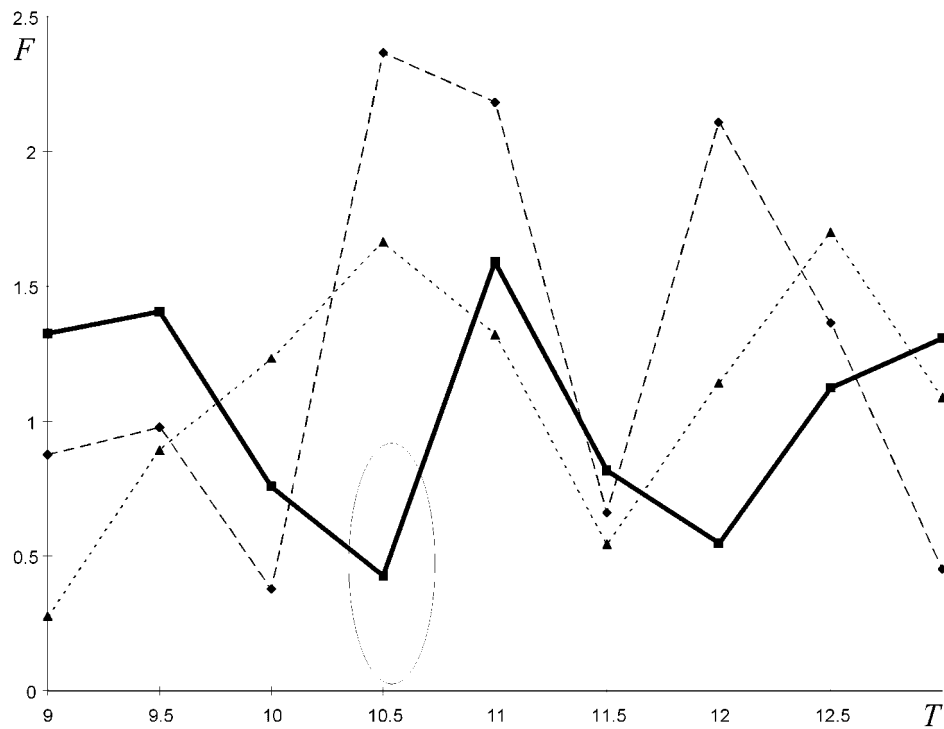


Figure 2. Fragments of the Fourier spectra of *oriC* (solid bold line) and that ones of the regions (dotted line), which flank *oriC* (in terms of A,C,G,T nucleotides occurrences). T- the length of the period, F- corresponding to the period of T spectral power.

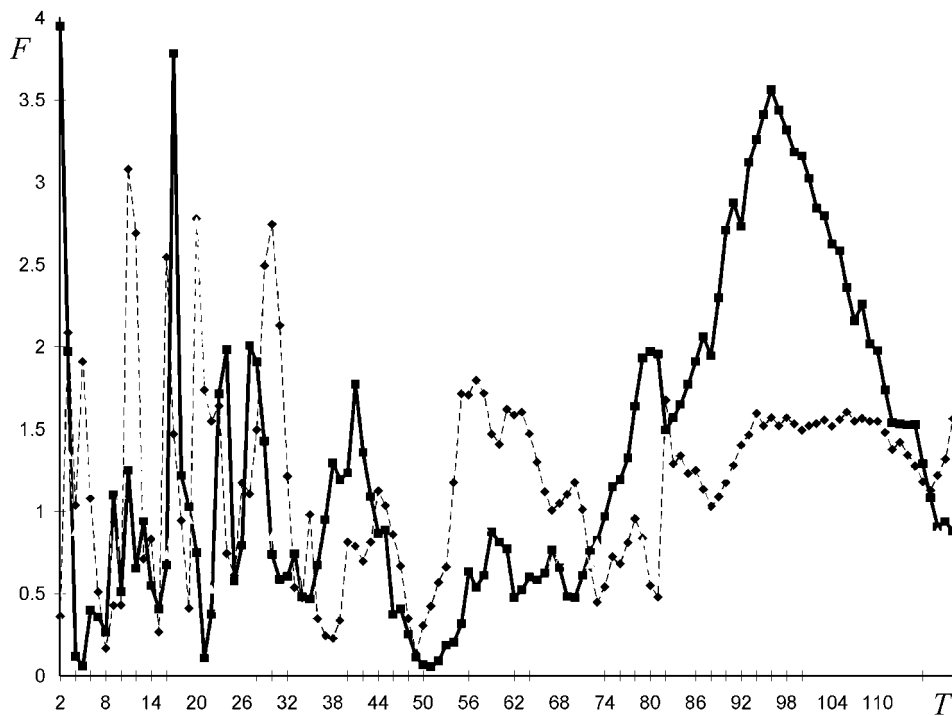


Figure 3. Fourier spectra of *oriC* (solid bold line) and flanking regions (dotted line) in terms of A and T nucleotides occurrences. T- the length of the period, F- corresponding to the period of T spectral power.

ESTMAP: A PROGRAM FOR ESTs MAPPING ON A GENOMIC SEQUENCE

**Milanesi L., *Rogozin I.B.*

Istituto di Tecnologie Biomediche Avanzate CNR, Italy

e-mail: milanesi@itba.mi.cnr.it

Institute of Cytology and Genetics of SB RAS, Russia

e-mail: rogozin@bionet.nsc.ru

*Corresponding author

Keywords: EST, gene prediction, alignment, genomic DNA, database searches, repeated elements

Resume

Motivation:

Prediction of protein-coding genes in newly sequenced DNA becomes very important in large genome sequencing projects. These problems are complicated due to exon-intron of the eukaryotic genes. Currently existing collections of expressed sequence tags (ESTs) are very large and thus very useful for gene mapping.

Results:

BLASTN homology searches against the EST Division of GenBank (dbEST) and Unigene database are used for EST revealing. ESTMAP program extracted "exact" matches with EST sequences (>95% of homology) from BLASTN output file. ESTMAP tries to predict introns in DNA comparing ESTs and a query sequence. Results of ESTMAP are used by a GeneBuilder system for gene structure prediction, and presence of EST matches can significantly improve prediction of gene structure.

Availability:

<http://www.itba.mi.cnr.it/webgene>

Gene identification in the newly-discovered DNA sequences is an important problem in current molecular biology studies. A number of programs have been developed for predicting the protein coding genes (Milanesi and Rogozin, 1998). The most common approach is based on the combination of the potential functional signals with global statistical properties of protein coding regions. Another approach for gene structure prediction is based on the homology detection throughout the databases of nucleotide or amino acid sequences. By using the information available on homologous protein sequences, it is possible to significantly improve the accuracy of gene structure prediction. Currently existing collections of expressed sequence tags (ESTs) are very large and can be very useful for gene mapping. Homology searches against the EST Division of GenBank (dbEST) and Unigene database can be used for this purpose (Boguski et al., 1994).

ESTs (Expressed Sequence Tags) offer a rapid route to gene identification (Adams, et al, 1991, Adams, et al, 1992, Okubo, et al, 1991, Vasmatzis et al, 1998), analysis of expression and regulation data (Vasmatzis et al, 1998), and can highlight multigene family diversity and gene alternative splicing (Wolfberg and Landsman, 1997). EST matches may identify more than half of the known human genes (Hillier et al, 1996). The price of the high-volume and high-throughput nature of the data, however, is that ESTs contain high error rates (Aaronson, et al 1996), do not have a defined protein product, are not well annotated and present only a raw substrate for sequence matching. This low quality sequence data can be significantly improved on, and various statistical and computational approaches are very useful for this purpose (Burke et al., 1998).

A ESTMAP system (<http://www.itba.mi.cnr.it/webgene>) performs a homology search against the EST database and the position of the homologous EST sequences is reported in the output in relation to a query sequence.

The ESTMAP system involves the following procedures:

1. Repeat masking. The repeated elements (for example, the human Alu elements) can be automatically masked in a query sequence before the homology search. Homology searches against the collection of repeated element (Jurka et al., 1992) are used for repeats detection. We implemented a program REPEAT for that purpose. A censored sequence (with 'N's instead of repeated elements) is automatically produced by REPEAT.
2. Homology searches. BLASTN (Altschul et al. 1990) is used for homology searches of the censored query sequence against the EST Division of GenBank (dbEST) and the Unigene database of sequences (www.ncbi.nlm.nih.gov) This step is most time-consuming since these EST datasets are very large.

3. EST mapping. The BLASTN output is used as input information by a EST_GENE program. Information about an EST sequence is used only when the similarity between the EST sequence and the query sequence is greater than 95%. The module EST_GENE is also able to predict the introns in DNA comparing ESTs and a query sequence based on the alignment method suggested by Huang (1994) (a linear-space divide-and-conquer strategy). The GT/AG splicing sites rule is used by EST_GENE, however non-canonical splicing signals (Milanesi and Rogozin, 1998) can also be predicted in cases of unambiguous alignment.
4. Output of results. The graphical visualization of the results is particularly important for the analysis of alternative splicing in a query sequence. By using a Java based graphical interface the user can visualize the EST maps and the sequence pattern of predicted features.

Homology searches are very important for functional mapping, homology with a known functional region can suggest the function of a query sequence. In particular, when the homologous protein sequence is already known and EST matches are detected, then the gene structure can be reconstructed with high accuracy. Information about EST matches is automatically used by the GeneBuilder system (Milanesi et al., 1999).

Acknowledgment

This work was supported by Italian CNR Genetic Engineering Project.

References

1. Aaronson, J.S., B. Eckman, R.A. Blevins, J.A. Borkowski, J. Myerson, S. Imran, and K.O. Elliston. "Toward the Development of a Gene Index to the Human Genome: An Assessment of the Nature of High-throughput EST Sequence Data." *Genome Research* **6**, 829–845 (1996).
2. Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J.C. Venter. «Sequence Identification of 2,375 Human Brain Genes." *Nature* **355**, 632–634 (1992).
3. Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merrill, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McConbie, and J.C. Venter. "Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project." *Science* **252**, 1651–1656 (1991).
4. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ "Basic local alignment search tool.» *J. Mol. Biol.* **215**, 403-410 (1990).
5. Boguski MS, Tolstoshev CM and Bassett DE, Jr "Gene discovery in dbEST." *Science* **265**, 1993-1994 (1994).
6. Hillier, L., N. Clark, T. Dubuque, K. Elliston, M. Hawkins, M. Holman, M. Hultman, T. Kucaba, M. Le, G. Lennon, M. Marra, J. Parsons, L. Rifkin, T. Rohlffing, M. Soares, F. Tan, E. Trevaskis, R. Waterston, A. Williamson, P. Wohldmann, and R. Wilson. "Generation and Analysis of 280,000 Human Expressed Sequence Tags." *Genome Res.* **6**, 807-828 (1996).
7. Huang, X "On global sequence alignment." *Comput. Applic. Biosci.* **10**, 227-235 (1994)
8. Jurka J, Walichiewicz J and Milosavljevic AJ "Prototypic sequences for human repetitive DNA." *J. Mol. Evol.* **35**, 286-291 (1992).
9. Milanesi L., D'Angelo D., Rogozin I.B. "GeneBuilder: interactive *in silico* prediction of genes structure." *Bioinformatics* **15**, 612-621 (1999).
10. Milanesi L., Rogozin I.B. "Prediction of human gene structure." In: *Guide to Human Genome Computing (2nd ed.)* (Ed. M.J.Bishop) Academic Press, Cambridge, 215-259 (1998).
11. Okubo, K., H. Hori, R. Matuba, T. Niiyama, and K. Matsubara "A novel system for large-scale sequencing of cDNA by PCR amplification." *DNA Sequence* **2**, 137-144. (1991)
12. Vasmatazis, G., M. Essand, U. Brinkmann, B. Lee, and I. Pastan "Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis." *Proc. Natl. Acad. Sci. USA* **95**, 300-304 (1998).
13. Wolfberg, T.G. and D. Landsman "A comparison of expressed sequence tags (ESTs) to human genomic sequences." *Nucleic Acids Res.* **25**, 1626-1632 (1997).

COMPLEXITY MEASURES OF SYMBOLIC SEQUENCES AND THEIR APPLICATION TO DNA ANALYSIS

^{1,2}Chuzhanova N., ¹Krawczak M., ²Gusev V.D., ²Nemytikova L.A., ^{1*}Cooper D.N.

¹University of Wales, United Kingdom

e-mail: Nadia.Chuzhanova@cs.cf.ac.uk, krawczak@cardiff.ac.uk, cooperdn@cardiff.ac.uk

²Sobolev Institute of Mathematics of SB RAS, Novosibirsk, Russia

e-mail: gusev@math.nsc.ru, luba@math.nsc.ru

*Corresponding author

Keywords: complexity analysis, promoter shuffling, vertebrate growth hormone genes, evolution, promoter mutation

Resume

Motivation:

Assessing DNA and RNA sequence regularity is motivated by the preferential recruitment of direct, symmetric and palindromic repeats in the regulatory regions of complex genomes. In Gusev (1999) two indicators of sequence regularity termed *complexity measures* have been proposed that generalize the Lempel–Ziv complexity measure (Ziv and Lempel, 1976) by taking into account the occurrence of isomorphic repeats. Here isomorphic repeats are defined as fragments that are identical (or symmetric) modulo some permutation of the alphabet letters. Linear algorithms for computing complexity measures have been proposed. These measures gave rise to a technique termed *complexity analysis* that can be used to explore the regularity and modularity of DNA sequences.

Results:

In this paper, we discuss two applications of complexity analysis. First, we describe the identification of modular components (“blocks”) in the growth hormone (GH) gene promoter sequences of some 22 vertebrate species, from salmon to human. Significant rearrangements of blocks were found to have occurred during evolution. Second, we use the technique to demonstrate that the concomitant change in local DNA sequence complexity is directly related to the likelihood of a regulatory single base-pair substitution coming to clinical attention. Increases in complexity exhibited a higher odds ratio than decreases but only for pyrimidine to purine transversions.

Availability:

A comprehensive definition of blocks in GH gene promoter regions and the DNA context of disease-associated single base-pair substitutions in regulatory regions of human genes are available for inspection via the Internet at <http://www.uwcm.ac.uk/uwcm/mg/ghblock.txt> and <http://www.uwcm.ac.uk/uwcm/mg/regmut/txt>

Introduction

It is well established that the regulatory regions of complex genomes are extremely repetitive. They are rich in direct, symmetric and inverted repeats. Among known measures of regularity, the Lempel–Ziv complexity measure reflects most adequately direct repeats occurring in a given text. However, this measure does not take into account other types of repeats.

In this paper, we give a more general definition of complexity (Gusev, 1999) which takes into account the occurrence of isomorphic repeats. Isomorphic repeats are defined as fragments that are identical (or symmetric) modulo some permutation of the alphabet letters. Application of this technique to DNA sequence has provided some new insights into the evolution of the vertebrate growth hormone gene promoter regions and the sequence dependence of pathological mutations in human gene regulatory regions are reported.

Methods and Algorithms

Let $S=s_1\dots s_N$ be a nucleotide sequence of length N , and denote by $S[i:j]$ the substring of S that starts at position i and ends at position j . For every $1\leq j\leq N$, $S[1:j]$ is called a *prefix* of S . If p is a one-to-one mapping (i.e. a permutation) of the four nucleotides $\{T,C,A,G\}$, then two sequences $S=s_1\dots s_N$ and $Q=q_1\dots q_N$ are called *directly isomorphic* modulo p ($S\equiv^+Q \pmod p$) if $q_k=p(s_k)$ for all $1\leq k\leq N$, and *inversely isomorphic* modulo p ($S\equiv^-Q \pmod p$) if $q_k=p(s_{N-k+1})$ for all $1\leq k\leq N$. We shall consider two specific permutations here, namely $p_1: T\rightarrow T, C\rightarrow C, A\rightarrow A, G\rightarrow G$ (‘identical’), and $p_2: T\rightarrow A, C\rightarrow G, A\rightarrow T, G\rightarrow C$ (‘complementary’).

For every $1 < i \leq N$, let $l(i)$ be the length of the longest prefix of $S[i:N]$ that is directly or inversely isomorphic modulo p_1 or p_2 to a substring of S starting at some position $j < i$. We define $l(i) = 1$ if $s_j \neq s_i$ for all $j < i$ (i.e. when the nucleotide at position i occurs there for the first time) or if $i = 1$. It can readily be shown that exactly one decomposition of S into a list of consecutive substrings $S = S[1:i_1]S[i_1+1:i_2] \dots S[i_{m-1}:N]$ exists such that $i_{k+1} - i_k = l(i_k + 1)$. The number of substrings in this unique decomposition of S , $C(S) = m$ is called *the (scalar) complexity* of S . This parameter indeed represents a suitable measure of regularity since any abundance in S of direct and inverted repeats and/or their inversions thereof serves to reduce C .

The second measure proposed here, termed a *complexity vector*, is designed specifically for a small alphabet such as the alphabet of nucleotides. Each component in a complexity vector corresponds to the complexity computed for only one of the permissible permutations.

The complexity vector serves to characterize a DNA sequence by its complexity values. It highlights which type of regularity is predominant in the sequence that is being analysed. It also helps us to choose the most "significant" permutations in the sense that they give the lowest values of complexity, thereby avoiding "masking" effects.

Scalar complexity measure can be used for the recognition of local structural regularities in DNA sequences by scanning the text using a window of given size. Local regularities appear, in the context of this paper, as fragments with abnormally low complexity. To study the relatedness of two sequences S and Q , sequence Q can be used to define a decomposition of sequence S , or *vice versa*. In this situation, $l(i)$ denotes the length of the longest prefix of $S[i:N]$ that is isomorphic modulo p_1 or p_2 to a substring of Q , irrespective of the starting position in Q . Gusev (1999) have shown that this decomposition can be derived in linear time using some special structural representation.

Fragments from single and pairwise decompositions that occurred for at least two sequences or sequence pairs can be included in a *vocabulary of blocks*. Usually, only exact matches would be considered in the decomposition process. However, when two or more substrings, isomorphic to the respective substrings modulo the same permutation and in the same orientation, are found to be separated by a similar number of nucleotides (± 1) in all instances, they can be merged into a single block.

Implementation and results

Promoter shuffling has occurred during the evolution of the growth hormone gene

Nucleotide sequences, comprising approximately 180 bp each of the *GH1*, *GH2* and *GH* gene promoters of 22 vertebrate species were retrieved from either the EMBL database or GenBank. In addition, the human prolactin (*PRL*) gene promoter sequence was retrieved from EMBL (Accession No. X00368) as an example of a paralogous gene. Attempts to align the 22 vertebrate GH gene promoter sequences by conventional means were only partially successful. Although good nucleotide sequence alignments were obtained within specific mammalian Orders viz. primates, artiodactyls and rodents (Krawczak *et al.*, 1999), attempts to align promoter sequences derived from chicken, bullfrog and five fish species were unsuccessful.

Complexity analysis was used to identify recurring sequence "blocks" in the GH gene promoters of different vertebrates. When the resulting block patterns were compared, significant similarities became apparent between orthologous promoter sequences that would not have been readily identified through the use of conventional sequence alignment procedures. Promoter sequences were found to differ not only in terms of the presence or absence of particular blocks but also with respect to block length, copy number and relative location. In terms of their block patterns, the fish GH gene promoters were the most diverse, yet still recognisably similar to the amphibian (bullfrog) sequence. By contrast, the chicken GH gene promoter was more reminiscent of the mammalian pattern. Even for two species as distant as for example human and salmon, some blocks still appeared to be common to both GH promoter sequences. Furthermore, with increasing evolutionary distance, a continuum of change in block size, number and location became apparent from human via other mammals to fish. The application of complexity analysis also permitted the structural comparison of the promoters of the paralogous human *GH1* and *PRL* genes even although nucleotide sequence alignment had proved impossible.

Analysis of single base-pair substitutions in the regulatory regions of human genes

No general rules have as yet been proposed to account for the functional consequences of regulatory mutations. In an attempt to seek such rules, complexity analysis was performed on the DNA sequence context of 153 different single base-pair substitutions in the regulatory regions of 65 different human genes underlying inherited disease.

For transversions that substitute a purine for a pyrimidine ($Y \rightarrow R$), the odds ratio (OR) of an increase in C causing an inherited disease state (2.69) was estimated to be more than six times higher than for lesions that leave C unchanged (0.41). The OR for decreases of C was intermediate (0.92). Even when multiple testing (4

independent tests) was taken into account, odds ratios within the Y→R category were found to be significantly different from each other ($\chi^2 = 9.891$, 2 d.f., $p = 0.007$, corrected for multiple testing $p = 4 \cdot 0.007 = 0.028$). Odds ratios within the other mutation types were not found to differ significantly from one another, although a trend similar to that observed for Y→R transversions was also apparent for transitions (OR decrease vs. increase: 0.68 vs. 1.28 for R→R; 0.68 vs. 1.07 for Y→Y).

Discussion

Comparative study of vertebrate GH promoter regions by means of complexity analysis has served to extend the concept of "promoter shuffling" (Surguchov, 1991). Such studies have generally failed to detect extensive evolutionary conservation between promoter regions owing to the inability of conventional sequence alignment procedures to cope with the gross rearrangements that become apparent with increasing evolutionary distance (Yowe and Epping, 1995). Similarly, gross rearrangements of orthologous and paralogous promoter regions have largely gone undetected because such analyses have usually been confined to the comparison of relatively similar sequences that may be readily aligned.

A strong requirement for sequence regularity is reflected in the preferential recruitment of direct, symmetrical or palindromic repeats as transcription factor binding sites in human genes (Kel *et al.*, 1995). Complexity analysis as described above provides a means to quantify these regularities. Our analysis has revealed that, for Y→R transversions in the regulatory regions of human genes, concomitant changes in local DNA sequence complexity (C) increase the likelihood that a specific lesion will come to clinical attention.

Acknowledgement

This work was supported by the Russian Foundation for Fundamental Research (grant 00-06-80420).

References

1. Gusev, V.D., Nemytikova, L.A. and Chuzhanova, N.A. "On the complexity measures of genetic sequences." *Bioinformatics* 15, 994-999 (2000).
2. Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender E. and Kolchanov, N.A. "A compilation of composite regulatory elements affecting gene transcription in vertebrates." *Nucleic Acids Res.* 23, 4097-4103 (1995).
3. Krawczak, M., Chuzhanova, N.A. and Cooper, D.N. "Evolution of the proximal promoter region of the mammalian growth hormone gene." *Gene* 237, 143-151 (1999).
4. Surguchov, A. "Migration of promoter elements between genes: a role in transcriptional regulation and evolution." *Biomed. Sci.* 2, 22-28 (1991).
5. Yowe, D.L. and Epping, R.J. "Cloning of the barramundi growth hormone-encoding gene: a comparative analysis of higher and lower vertebrate GH genes." *Gene* 162, 255-259 (1995).
6. Ziv, J. and Lempel, A. "On the complexity of finite sequences." *IEEE Trans. Inf. Theory* IT-22, 75-81 (1976).

ANALYSIS OF MUTATIONAL HOTSPOTS IN HUMAN DISEASE GENES AND MUTATIONAL SPECTRA

*Rogozin I.B., ¹Berikov V.B., Glazko G.V.

Institute of Cytology and Genetics of SB RAS, Novosibirsk, Russia

e-mail: rogozin@bionet.nsc.ru

¹Sobolev Institute of Mathematics, of SB RAS, Novosibirsk, Russia

e-mail: berikov@math.nsc.ru

*Corresponding author

Keywords: mutation, regression analysis, context, hotspot, p53 gene, immunoglobulin gene, sequence

Resume

Motivation:

The study and comparison of mutational spectra is an important problem in molecular biology, because these spectra often reveal important features of the action of various mutagens and the functioning of repair/replication enzymes. As is known, mutability varies significantly along nucleotide sequences: mutations often concentrate at certain positions in a sequence, otherwise termed "hotspots".

Results:

Herein we discussing various aspect of hotspot revealing and analysis. We will illustrate problems of such analysis on examples of the human p53 gene, mammalian immunoglobulin V genes and spontaneous mutations induced in mutT strain of *E.coli*.

With advancements in molecular biology, a large body of experimental data has been acquired on mutations in DNA. Data of this specific kind are otherwise known as "mutational spectra". A mutational spectrum is often a quite informative feature of the functioning of various repair/replication enzymes. It is now believed that mutability varies significantly along nucleotide sequences: mutations, whether induced or spontaneous, occur at higher frequencies at certain positions of a nucleotide sequence (Benzer, 1961). Study of these positions (otherwise termed "hotspots") suggests that there might be an association between the mutations observed and the features of DNA primary structure (otherwise called "context") near the hotspots. Figure 1 exemplifies a mutational spectrum (Fowler and Schaaper, 1997). As can be seen, at some positions mutations occur much more often than at the others. In many cases, the reason for hotspots in a site is certain combinations of neighbouring bases (as reviewed by Boulikas, 1992).

The problem of hotspot context revealing is normally addressed using three methods.

1. Stormo et al. (1986) used multiple linear regression analysis to see how the context affects the mutability of different positions in the *lacI* gene in *E.coli* cells treated with 2-aminopyrine. Data obtained indicate that the bases at positions -2 and -1 relative to the mutating base most strongly affect the frequency of mutations. However, it is assumed within the framework of the method that there is a linear correlation between the frequency of mutations in positions and the context factors; and that the factors are distributed normally. The problem is that in many, if not all, cases, neither assumption fits the context features of real mutational spectra.
2. Rogozin and Kolchanov (1992) employed the heuristics classification approach and the Monte-Carlo procedure to classify and build consensus of hotspots. The procedure of building a consensus (the rule that defines the hotspot context) is based on assessing the nonrandomness of the bases in the neighbourhood of the hotspots. The statistical significance of the constructed set of consensus sequences is estimated using the Monte-Carlo procedure. A statistically significant description of the context of hotspots of somatic mutations (consensus sequences RgYW and TaA, R = A or G, Y = T or C, W = A or T) in immunoglobulin V genes has become possible with this approach. It has been shown that these consensus sequences are characteristic of somatic mutations in various organisms. These consensus sequences and their modified versions are widely used for analysis of hypermutation in immunoglobulin V genes.
3. A new approach for mutational spectra classification based on "regression trees" was suggested (Berikov and Rogozin, 1999). This approach has a number of properties, which makes it very useful for mutational spectrum analysis: this approach allows to utilise mutational spectrum characteristics of heterogeneous nature: qualitative and quantitative; it makes possible to work under condition of high uncertainty (limited data size; absence of *a priori* information about distributions); a regression tree represents hierarchical logical-and-probability model of mutational process. For fast analysis a

modification of dynamic programming method for regression tree design was applied (Berikov and Rogozin, 1999).

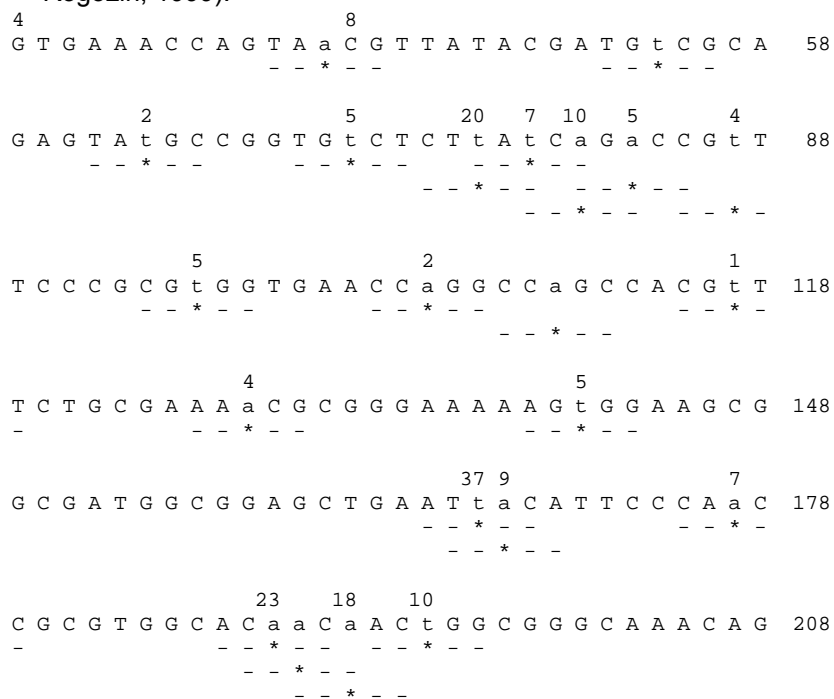


Figure 1. The sequence of the *lacI* gene with spontaneous A:T->C:T mutations induced in mutT strain of *E. coli* (Fowler and Schaaper, 1997). Lower case letters denote the detectable positions. Numbers above the sequence stand for the number of mutations at the position. Available positions are asterisked, available sites are underlined (positions -2 to +2 are shown).

Example of mutational spectrum classification by using a REGRT program (Berikov and Rogozin, 1999) is shown in Fig.2b. One can see that each consensus matches some sites, and mutations should be distributed evenly throughout these sites (all differences between them should be owing to random reasons). Homogeneity was assessed with the X2 statistic or a Monte Carlo test with the X2 statistic (Berikov and Rogozin, 1999). For example, the differences between the number of mutations in the sites corresponding to the aA consensus (20, 37, 23 and 18) can be attributed to chance (Fig.2b).

We will discuss further methodological development of REGRT. New features allow more detailed and complete analysis of mutational spectra.

We will illustrate importance of hotspot analysis on several examples. One of examples is mutational spectra in the human p53 gene. The comparative study of mutational spectra (base substitutions) in the p53 gene from germline cancer-prone families (Li-Fraumeni syndrome) and somatic mutations in tumors of different histogenesis and their derived cell lines was conducted. Previously we revealed that differences in the distribution of hotspots in the p53 gene is influenced by cell growth conditions *in vivo* and *in vitro* (Glazko et al., 1999). However, further analysis revealed that differences in mechanisms of mutagenesis in various spectra may be also an important feature of p53.

a)			b)		
Posi- tion	Site sequence	Number of mutations	Posi- tion	Site sequence	Number of mutations
	- - * - -			- - * - -	
41	T A a C G	4	77	A T a A G	20
54	C G a C A	8	167	G T a A T	37
64	G C a T A	2	189	A C a A C	23
72	A G a C A	5	192	A C a A C	18
77	A T a A G	20		-----	----
79	T G a T A	7	C1	a A	24.5
81	T C a G A	10			
83	A G a C C	5	41	T A a C G	4
87	A A a C G	4	54	C G a C A	8
96	C C a C G	5	64	G C a T A	2
105	C C a G G	2	72	A G a C A	5
110	C C a G C	0	79	T G a T A	7
117	A A a C G	1	81	T C a G A	10
128	A A a C G	4	83	A G a C C	5
141	C C a C T	5	87	A A a C G	4
167	G T a A T	37	96	C C a C G	5
168	T T a C A	9	105	C C a G G	2
177	C A a C C	7	110	C C a G C	0
189	A C a A C	23	117	A A a C G	1
190	C A a C A	0	128	A A a C G	4
192	A C a A C	18	141	C C a C T	5
195	C C a G T	10	168	T T a C A	9
			177	C A a C C	7
			190	C A a C A	0
			195	C C a G T	10
				-----	----
			C2	a B	4.9

Figure 2. Example of regression analysis as applied to the mutational spectrum described in Figure 1. Two consensus sequences (C1 and C2) are revealed (positions -2 to +2 are shown).

Acknowledgment

This work was supported by the Russian Foundation for Basic Research (grant N 99-04-49535).

References

1. Benzer, S. "On the topology of the genetic fine structure." Proc. Natl. Acad. Sci. USA **47**, 403-415 (1961).
2. Berikov V.B., Rogozin I.B. "Regression trees for analysis of mutational spectra in nucleotide sequences." Bioinformatics **15**, 553-562 (1999).
3. Boulinkas, T. "Evolutionary consequences of nonrandom damage and repair of chromatin domains." J. Mol. Evol. **35**, 156-180 (1992).
4. Fowler, R.G. and Schaaper, R.M. "The role of the *mutT* gene of *Escherichia coli* in maintaining replication fidelity." FEMS Microbiol. Rev. **21**, 43-54 (1997).
5. Glazko G.V., Rogozin I.B., Sozinov A.A. "Analysis of mutational spectra of p53 gene in different tumors." Tsitol. Genet. **33**, 15-23 (1999).
6. Rogozin, I.B. and Kolchanov, N.A. "Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis." Biochim. Biophys. Acta, **1171**, 11-18 (1992).
7. Stormo, G.D., Schneider, T.D. and Gold, L. "Quantitative analysis of the relationship between nucleotide sequence and functional activity." Nucleic Acids Res. **14**, 6661-6679 (1986).

COMPARATIVE ANALYSIS OF FUNCTIONAL SITE MOTIFS OF MGE *COPIA*-GROUP RELATIVE TO THEIR POSSIBLE MOLECULAR FUNCTIONS

**Amikishiev V.G., Ratner V.A.*

Novosibirsk State University, Novosibirsk, Russia
Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: ratner@bionet.nsc.ru

*Corresponding author

Keywords: motifs of functional sites, mobile genetics elements, molecular functions

A computer-assisted analysis of DNA sequences of three mobile genetic elements (MGE) referring to the *cop* group of *Drosophila* (i.e., *cop*, *cop*-white, and 1731) was performed. In each case, more than 400 motifs of 75-95 types similar to the known functional sites from the database (storing 277 entries) were detected. It was revealed that the motifs are unevenly distributed along DNA, that is, there are marked "jammings" (condensations) in the possible regulatory zones (LTRs; ORF starts and terminations; and ORF domains). Statistical significance of condensation occurrence was proved in comparison to the random sequences with the same length and nucleotide content as the real MGE. The presence of necessary motifs in condensations related to possible regulatory zones enables to provide the basic molecular functions of MGE: expression of their own ORFs, reproduction (transposition), induction of transpositions, exhibiting of modifying action on the neighboring genes, polygenes, etc. There were detected assumed positions of start and termination of synthesis of full-sized RNA of MGE within the LTR limits. For the *cop*-element, there are the TATA-boxes in positions 197, 256 and 263 bp (within the limits of ILTR) and polyadenylation site in positions 5074, 5087 and 5115 bp (by the example of the rLTR). The regularity of concentration of the sites induced by external signals and located prior and after domain of replicative complex of the reverse transcriptase (RT) enzyme is statistically justified. The distribution of functional site motifs along MGE sequences significantly correlates with the distribution of the nucleotide content (%A+T).

The work was supported by the grants of Russian Foundation for Basic Research (№ 97-04-49232) and the Program of Russian Ministry of Education "Russian Universities – Fundamental Studies" (N 1760).

Designations

On the consequent Figs. 1-3 the abscissa corresponds to the sequence of mobile element DNA in bp; the ordinate of Figs. 1-3 (a) represents the variety of regulatory sites: the **1st group** - the sites of replication and transcription initiation and termination; the **2nd group** - enhancers and silencers of chromosomal, viral, etc., genes; the **3rd group** - the sites recognized by cellular transcription and translation protein factors; and the **4th group** - the sites recognized by protein receptors for inductive signals; the arrows mark the discovered location of regulatory sites on the left-directed (leftward arrows) and right-directed (rightward arrows) DNA strands. Figs. 1-3 (b) express the genetic maps of corresponding MGEs. The other details are in figure legends.

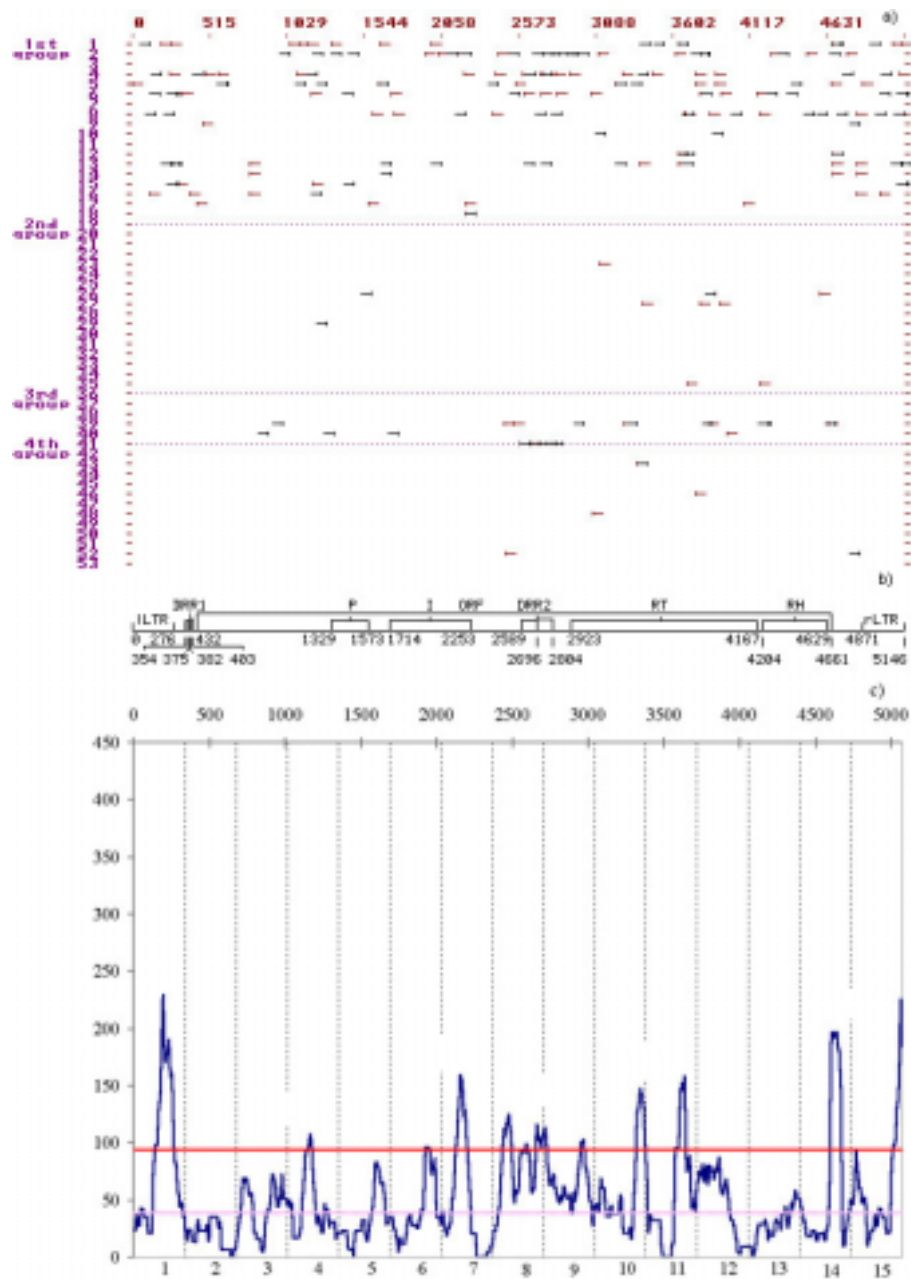


Figure 1. (a) Distribution of the revealed motifs of functional sites along the DNA sequence of *Drosophila* LTR-containing retrotransposon copia-white in the individual samples. (b) Schematic structure of the DNA sequence of the retrotransposon copia-white. Designations. LTR, long terminal repeats; ORF, big open reading frame; DRR1 and DRR2, direct repeats; P is the motif of amino acid sequence of protease domain; RT, of reverse transcriptase domain; RH, of RNase H domain; and I, of integrase domain (c) Consolidated distribution of the revealed motifs of functional sites along the copia-white DNA sequence. On the abscissa the segment numbers of MGE genome are indicated, each size 1/15 length given MGE; on the ordinate there are the total numbers of nucleotides contained in the motifs of functional sites and falling within the scanning window by a size 75 bp. The upper horizontal line - 95%th level of the nonrandomness of the condensations of the motifs of functional sites; the lower line - average by 50 of random sequences of the same lengths and same DNA nucleotide content as copia-white.

Jamming of the motifs correlates with the locations of potential regulatory regions within LTR, DRR2, and in the interval between the ORF and rLTR.

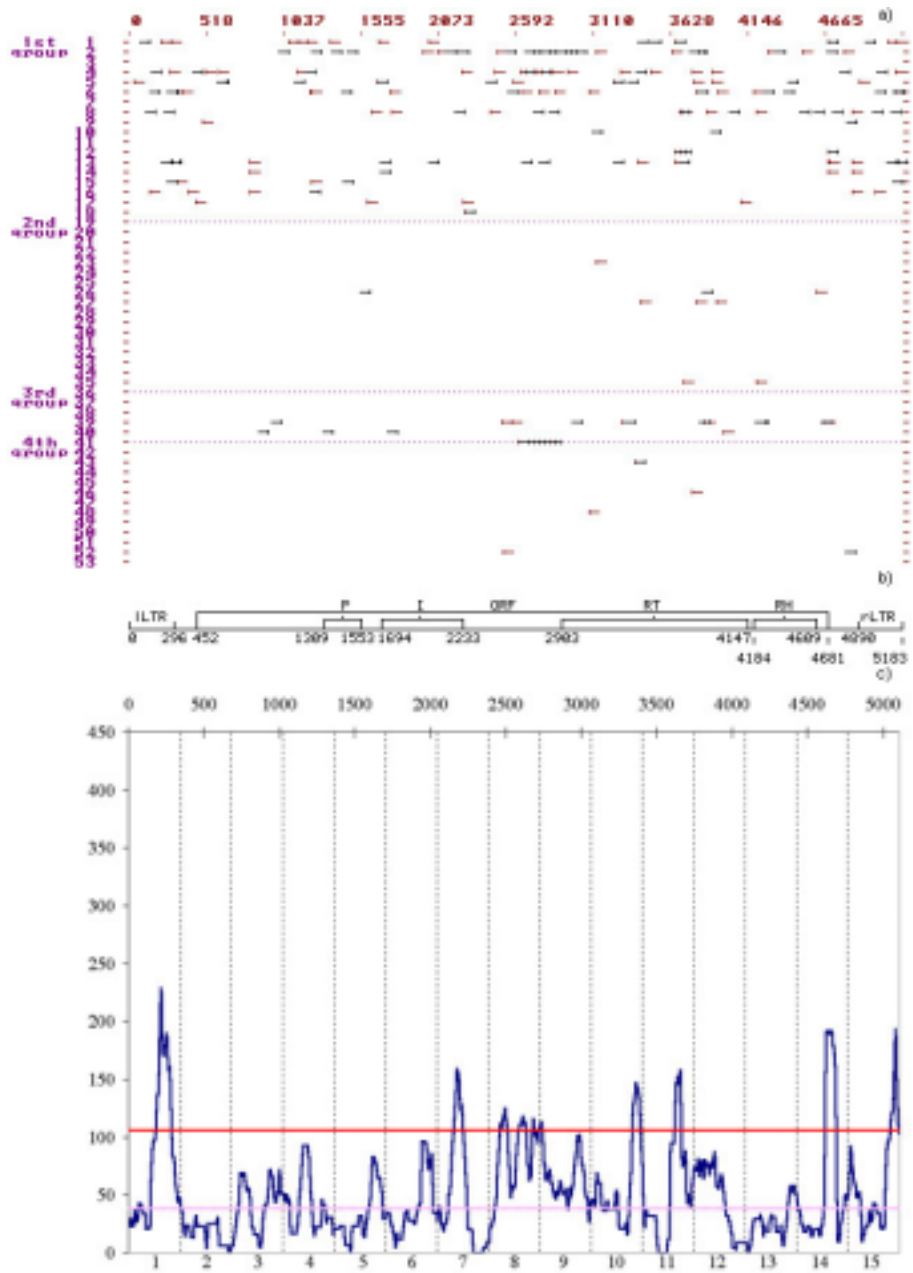


Figure 2. (a) Distribution of the revealed motifs of functional sites along the DNA sequence of *Drosophila* LTR-containing retrotransposon of copia in the individual samples. (b) Schematic structure of the DNA sequence of retrotransposon copia. (c) Consolidated distribution of the revealed motifs of functional sites along the copia DNA sequence. The designations are as in Fig. 1.

Jamming of the motifs correlates with the locations of potential regulatory regions within LTR, DRR2, and in the interval between the ORF and rLTR.

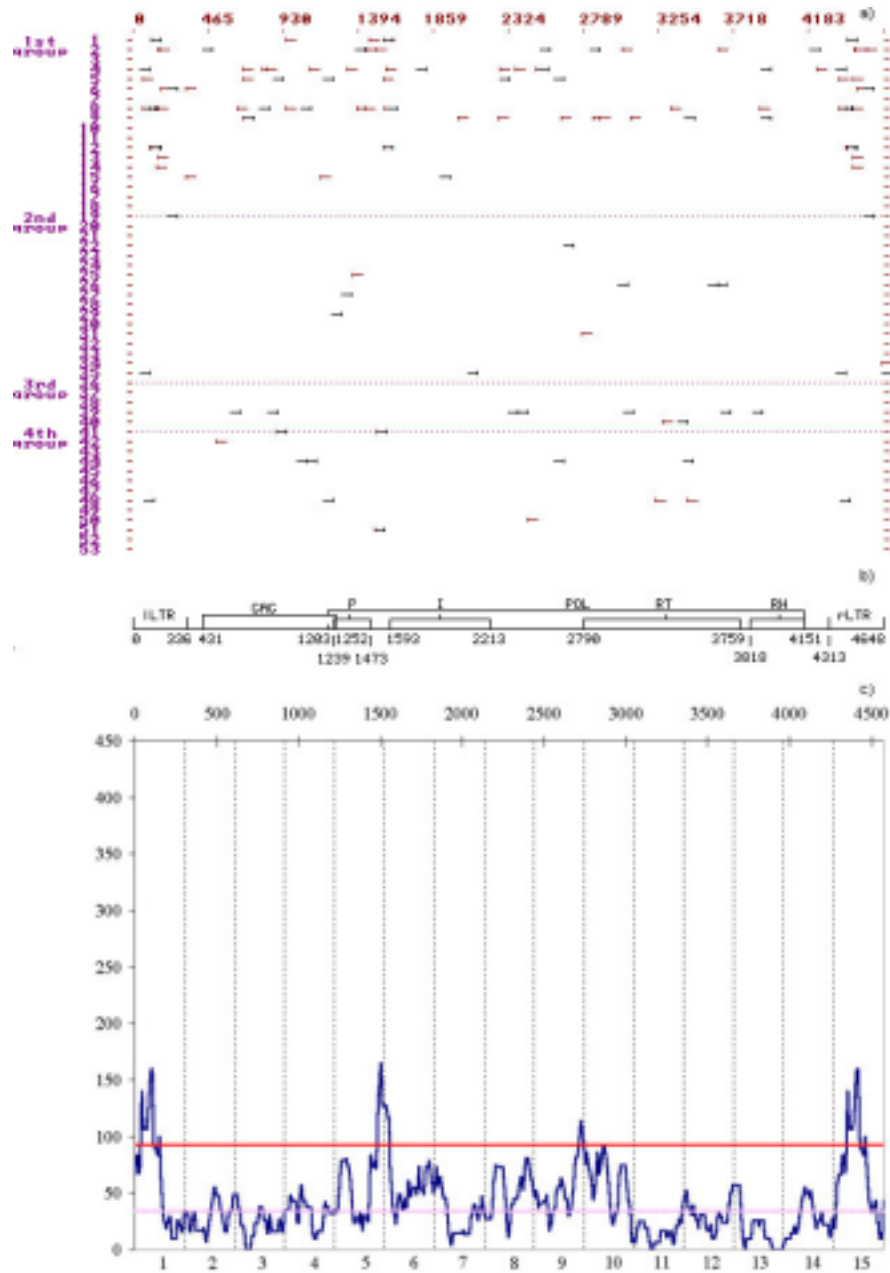


Figure 3. (a) Distribution of the revealed motifs of functional sites along the DNA sequence of *Drosophila* LTR-containing retrotransposon of copia-group 1731 in the individual samples. (b) Schematic structure of the DNA sequence of retrotransposon 1731. Designations: GAG and POL are large open reading frame; (c) Consolidated distribution of the revealed motifs of functional sites along the 1731 DNA sequence. The other designations are as in Fig. 1.

Jamming of the motifs correlates with the potential regulatory regions in the LTR and in the vicinity of the ends (start and beginning) of the ORFs and domains of the ORF2 (POL-gene).

S/MARs AND SOME ELEMENTS FROM DIFFERENT REPETITIVE FAMILIES ARE COLOCALIZED IN HUMAN GENOME

**Glazko G.V., Kochetov A.V., Rogozin I.B.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: glazko@biotech.relc.com

*Corresponding author

Keywords: S/MARs, repetitive DNA families, Repbase Update, colocalization

Resume:

Motivation:

Experimental data suggests that S/MARs and repeats from various repetitive families have been observed together. S/MARs could be described as repetitive family since they have several common features: a high copy number, "hidden similarity" between different elements and they are dispersed throughout whole genome. A general question about any relationship between S/MARs sequence family and different repetitive families can be divided into two ones: (1) are S/MARs sequences the members of some yet well-known repetitive family and (2) does the matrix-associated activity is the intrinsic property of some repetitive family (families)?

Results:

Computer-based prediction of matrix-associated regions in Repbase Update ("human") was carried out. It was shown that: (1) there is not any repetitive family coinciding as whole with S/MARs family; (2) repetitive sequences from a wide variety of different families possess matrix-associated activity. In this case the reveal of matrix-associated activity in the genomic DNA could mark the presence of preexisting integration at this region the member from some repetitive family. Thus the transposition could be putative mechanism to disperse sites associated with nuclear matrix throughout genome.

Introduction

S/MARs sequences could be described as repetitive family

The nuclear matrix is a network of RNA and nonhistone proteins that serves as a scaffold for loops of chromatin, anchored via special DNA sequences called matrix- or scaffold-associated regions (S/MARs) (for a review, see van Driel et al., 1995; Bode et al., 1995). It is generally accepted that S/MARs are implicated in a wide variety of regulatory events thus the investigation of S/MARs' structure-functional organization is crucial to understand the basic principles of genome structure and regulation. One of key points here is the evolutionary origin of S/MARs sequences.

Chromatin loops contain 5-100 kb of DNA (Laemmli et al., 1992). Assuming that this is an inaccurate estimate of S/MARs number in human genome (about $3,3 \cdot 10^9$ bp) constitute the value of 10^5 order. This value is one order less than the number of short interspersed elements ($5-9 \cdot 10^5$ copies for Alu's) and is comparable with the number of long interspersed elements (10^5 copies for L1) (Mighell et al., 1997; Smit et al., 1995). Thus the sequences of matrix-associated elements could be viewed as some specific repetitive family. By definition besides high copy number repetitive family members should possess another key feature: their should be homologous. It is well-known that S/MARs do not have a clearcut consensus sequence. But there are the repetitive sequence joined together in repetitive family without clear similarity on the nucleotide level: in such cases similarity exists on the structural level. For example it is known that long terminal repeats (LTRs) show no overall sequence similarity, but all retrotransposon LTRs share short conservative elements functional in integration and transcription (Smit, 1993).

Most interspersed repeats in eukaryotic genomes have been assigned to one of the following major categories: non-LTR retroposons and LTR-retrotransposons including retroviruses and DNA transposons, which in turn could be joined in different families (MaLRs, MERs) with different level of similarity between family members (Jurka, 1998; Smit&Riggs, 1996; Smit, 1993; Jurka, 1990). Inactive transposons revealed in MER family are significantly (15 to >50%) diverged from other copies of the same element but share key features of any transposons: terminal inverted repeats (TIRs) and target site duplication (Smit&Riggs, 1996). This could serve as another example of "hidden similarity" when the consensus exists on a structural level.

S/MARs sequences possess thus defined similarity. Despite of the lack of long sequence similarity there are at least three classes of motifs, have been observed in all S/MARs and presumably reflect the different possibilities for DNA-protein interactions: (1) sites for direct DNA-protein interaction (as transcription factors binding sites), (2) stress-induced base-unpairing regions and (3) sites directing DNA curvature and kinks. Computer-based

prediction of S/MARs is based on the calculation of such motifs. To summarize, S/MARs can be viewed as repetitive family based upon the presence of some key features of these families: a high copy number, "hidden similarity" between family members, interspersing throughout whole genome. Simultaneously appears the question about possible relationships between S/MARs sequence family and different repetitive families, concerning the origin, evolution and dispersion S/MARs or proto-S/MARs sequences in genome.

Experimental data concerning S/MARs' and repetitive elements colocalization

A general question about any relationship between S/MARs sequence family and different repetitive families can be divided into two ones: (1) are S/MARs sequences the members of some yet well-known repetitive family or (2) does the matrix-associated activity is the intrinsic property of some repetitive family (families).

At least three examples show that first is not allowed. Usually, experimentally obtained S/MARs are searched throughout database of repetitive elements and in some cases a colocalization is observed. For example in the human serpin gene cluster 5 S/MARs were revealed. 83-92% from their whole sequence were constructed from LINE, LTR and SINE fragments (Rollini et al., 1999); in a 16 kb region around the plastocyanin gene of Arabidopsis 3 S/MARs were mapped (van Drunen et al., 1997) and one of them was completely overlapped with transposon ATTIR16T3A (the data not shown); in the co-linear Sh2/A1-homologous regions of rice (30 kb) and sorghum (50 kb) the location of 4 and 7 S/MARs respectively were identified, their majority were colocalized with miniature inverted repeat transposable elements (MITEs) (Avramova et al., 1998). This data suggests that there is no one repetitive family coinciding exactly with S/MARs.

The second question suggests two possible decisions. Experimental one supposes test in vivo the ability bind to the nuclear matrix the thousand members of different repetitive families, that is naturally unreal. Although for some DNA and RNA viruses the regions in vivo associated with nuclear matrix were revealed (Liu et al., 1997; Tan et al., 1998). Computer decision, presented here, assumes the analysis of different repetitive family members to reveal "presumable" S/MARs.

Methods and algorithms

Computer-based prediction of matrix-associated regions in a members of different repetitive DNA families presented in Repbase Update, part "human" (on <http://www.girinst.org>.) was carried out. For this purpose the computer program ChrClass was applied, designed to reveal the DNA fragments, presumably associated with nuclear matrix in vivo, based on the frequencies of the simple nucleotides motifs (Rogozin et al., 2000).

Implementation and results

Sequence data. It is generally accepted that matrix-associated sequences should be at least 300 bp long; thus all SINE elements were excluded from our analysis (they are usually less than 300 bp). The following repetitive families were extracted from Repbase Update, part "human" for computer-based prediction: 1) LINE-elements (1000-7000 bp); 2) MERs-elements (400-7000 bp); 3) LTRs (500-1000 bp) 4) internal part of endogenous retroviruses (without LTRs, 5000-13000 bp); 5) autonomous and non-autonomous transposon elements (400-3000 bp). Partly annotated sequences containing more than >5% non-A,T,G,C nucleotides were also excluded from analysis.

In general we analyzed 37 LINE elements from different subfamilies, 145 LTRs (including MERs, contained «putative LTR» in KW field), 48 autonomous and non-autonomous transposons (including MERs, contained «transposon» in KW field), 36 internal retroviral fragments (including MERs, contained «Internal sequence of retroviral-like element» or «retroelement» in KW field). It was revealed about 10 MERs, annotated as «dispersed repeat» but their were excluded from analysis because of lack of information. Thus, at the end MER-family was divided into subfamilies, consisting from internal sequence of retroviral-like elements, autonomous/nonautonomous transposons and LTRs. This classification is used in Table 1.

Discussion

S/MARs' and repetitive elements colocalization. The data presented in Table 1 suggests that 50-60% of retroviruses and transposons contains fragments, intrinsic structures of which suppose their possible association with nuclear matrix in vivo. About 10% of LTRs as well as 16% of LINE family members might also associate in vivo with nuclear matrix, based on their nucleotide sequence structure. Concerning LINEs it should be noted that only members of ancient inactive subfamilies possessed these properties (Smit et al., 1995).

If different repetitive elements contain fragments with S/MARs sequences similarities then some S/MARs should also contain structural features of these families. All LTRs share 1) terminal 5'TG and 3'CA dinucleotide; 2) RNA polymerase II promoter elements and transcription start site, and 3) a polyadenylation signal and site. S/MARs sequence could possess all these signal even though their AT-richness. The transposons are constructed from ORF for transposase, TIRs and duplicated target sites ("transposon frame") or without ORF in

the case of non-autonomous transposon. The computer analysis of the 16 kb fragment around plastocyanin gene revealed that the frequency of "transposon frame" was two-fold higher in predicted matrix-associated regions than in non matrix-associated ones. This observation as well as results of computer-based prediction in the members of Repbase Update support the hypothesis about presumable relationship of S/MARs origin and some repetitive families. The further investigation should clarify the nature of this relationship.

Table 1. Colocalization of S/MARs and repetitive elements from Repbase Update (bold font marks ID in Repbase; normal marks S/MAR positions in element).

LINE	Retroviruses and retroelements	LTRs	Transposons (AT and non-AT)
L1MA2; MAR:300-1000	MER4I, 6388bp; MARs:300-1100, 1500-4700	LTR11, 684bp; MAR:1-400	TIGGER1, 2418 bp; MAR: 1-400
L1MC3; MAR:1-800	HERVL, 5654 bp; MARs:800-1200,4700-5600	MER11a, 1126bp; MAR:300-1100	HSMAR2, 1301bp; MAR:1-600
L1MC4; MAR:1-800	MER57I, 7537bp; MARs: 3700-4900	MER4D, 1017bp MAR:1-500	CHARLIE1, 2739bp; MAR:300-1800
L1MD2; MAR:1-800	HERVK, 7243bp; MARs:1-1500,1800-4600, 6200-6800	MER41C, 554bp; MAR:1-400	MER80, 508bp;MAR:1-300
L1MC5; MAR: 300-1300	HERV-E, 7813bp; MAR: 6000-6800	MER65A, 445bp; MAR:1-400	GOLEM, 2986 bp; MARs: 400-1200,2100-2400
L1ME4; MAR:1-400	MER65I, MARs: 300-700,1100-3300	MER74, 624 bp; MAR:1-400	ZOMBI, 2806bp; MAR:400-1100;
	HERVK(C4), 5262bp; MAR:900-4800	MER41D, 557bp; MAR:1-500	MER63C, 938bp; MAR:1-300
	MER41I, 3944bp; MAR: 800-3800;	LTR26, 603bp; MAR:1-400	CHESHIRE, 2285bp; MAR:500-2200;
	HERV23, 4823bp; MARs:1-400, 1600-4700	MER89, 559bp; MAR:1-400	MER82, 653bp; MAR1-300;
	MER31I, 4936bp; MAR: 800-3100;	LTR37B, 468bp MAR:1..400	CHARLIE5, 2585bp; MAR:1-800
	HERVK9I, 5937bp; MAR:2300-4000	LTR44, 519bp; MAR:1..400	CHARLIE1B, 518bp; MAR:1..400
	MER66I, 6676bp; MAR: 3200-4800	LTR54, 510bp; MAR:1-500	LOOPER, 1460bp; MAR:300-1300
	HARLEQUIN, 6896bp; MAR:1600-2500, 4800-5100	LTR24B, 576bp; MAR:1-400	CHARLIE2, 2760 bp; MAR:2100-2700
	MER51I, 7816 bp; MARs: 1200-2700, 4800-5500	MER114, 648bp; MAR:1-400	MER45B, 1037bp; MAR: 1-300
	HERVK22I, 6837bp; MARs: 1-900, 2100-3600	LTR66, 610bp; MAR:1-400	GOLEM_B, 1205bp; MAR:1-1000
	MER61I, 5217bp; MARs: 700-2200, 2500-3200	MER45R, 784bp; MAR:1-300	MER97, 1106bp; MAR:300-700
	HERVL68, 3307bp; MAR: 1300-2200	LTR2B, 490bp; MAR:1-400	MER96B, 434bp; MAR:1-400
	MER4BI, 2454bp; MARs:1-1000, 1600-2300		MER100, 1264bp; MAR:1-1100
	HERVK11I, 7953bp; MARs:1300-2800,3100-4700, 7000-7300		MER44B, 719bp; MAR:1-400
	HERVK14I, 5945bp; MAR: 800-1300,2400-4200		RICKSHA_0, 1708bp; MAR:1-600
	HERVK13I, 8116 bp MARs: 1600-2100,4400-4700		
	HERVK14CI, 7417bp; MARs: 2100-4800, 6300-6700		
	PRIMA41, 7756 bp; MARs: 300-700; 3200-3700		

Conclusion

The obtained data suggests that the repetitive sequences from a wide variety of different families may possess matrix-associated activity. In this case regions of matrix-associated activity in the genomic DNA could mark the presence of preexisting integration at this region of some repeated elements. Thus the transposition could be putative mechanism to disperse sites associated with nuclear matrix throughout whole genome.

References

1. Avramova, Z., Tikhonov, A., Chen, M., and Bennetzen, J.L.. (1997) Matrix attachment regions and structural colinearity in the genomes of two grass species. *Nucleic Acids Res.* **26**,761-767.
2. Bode, J., Schlake, T., Rios-Ramirez, M., Mielke, C., Stengert, M., Kay, V., and Khler-Wirth, D. (1995) Scaffold-matrix attached regions: structural properties creating transcriptionally-active loci. *Int. Rev. Cytol.* **162A**, 389-454.
3. Jurka, J. (1990) Novel families of interspersed repetitive elements from the human genome. *Nucl. Acids. Res.* **18**, 137-141.
4. Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. *Cur. Opin. Struct. Biol.* **8**, 333-337.

5. Laemmli, U.K., Kas, E., Poljak, L., and Adachi, Y. (1992) Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Cur. Opin. Gen. Dev.* **2**, 275-285.
6. Liu, J., Bramblett, D., Zhu, Q., Lozano, M., Kobayashi, R., Ross, S.R., and Dudley, J.P. (1997) The matrix-attachment region binding protein SATB1 participate in negative regulation tissue-specific gene expression. *Mol. Cell Biol.* **17**, 5275-5287.
7. Mighell, A.J., Markham, A.F., and Robinson, A.P. (1997) Alu sequence. *FEBS Lett.* **417**, 1-5.
8. Rogozin, I.B., Glazko, G.V., Glazkov, M.V. (2000) Computer prediction of sites associated with various elements of the nuclear matrix. *Briefings in Bioinformatics.* **1**, (in press).
9. Rollini, P., Namciu, S.J., Marsden, M.D., and Fournier, R.E.K. (1999) Identification and characterization of nuclear-matrix attachment regions in the human serpin gene cluster at 14q32.1. *Nucleic Acids Res.* **27**, 3779-3791.
10. Smit, A.F.A.. (1993) Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucl. Acids Res.* **21**, 1863-1872.
11. Smit, A.F.A., G. Toth, A.D. Riggs, and J. Jurka. (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequence. *J. Mol. Biol.* **246**, 401-417.
12. Smit, A.F.A., and A.D. Riggs. (1996) Tiggers and other DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci* **93**, 1443-1448.
13. Tan, S.-H., Bartsch, D., Schwarz, E., and Bernard, H.-U. (1998) Nuclear matrix attachment region of human papillomavirus type 16 point toward conservation of these genomic elements in all genital papillomaviruses. *J. Virol.*, 3610-3622.
14. van Driel, R., Wansink, D.G., van Steensel, B., Grande, M.A., Schul, W., and de Jong, L. (1995) Nuclear domains and the nuclear matrix. *Int. Rev. Cytol.* **162A**, 151-189.
15. van Drunen, C.M., R.W.Oosterling, G.M. Keultjes, P.J. Weisbeek, R. van Driel, and S.C.M. Smeekens. (1997) Analysis of the chromatin domain organization around the plastocyanin gene reveals an MAR-specific sequence element in *Arabidopsis thaliana*. *Nucleic Acids Res.* **25**, 3904-3911.

STUDYING CORRELATIONS OF COMPUTATIONALLY PREDICTED ORIGINS OF REPLICATION AND BASE SKEWS IN THE *SACCHAROMYCES CEREVISIAE* GENOME

Korbel J.O., Assmus H., Kielbasa Sz.M., *Herzel H.

Institute for Theoretical Biology, Humboldt University, Berlin, Germany

e-mail: h.herzel@itb.biologie.hu-berlin.de

*Corresponding author

Keywords: replication, compositional asymmetries, ARS elements, *Saccharomyces cerevisiae*, origin of replication

Resume

Motivation:

Specific compositional bias has been observed in genomes of prokaryotes and large viruses (see, e.g. Lobry, 1996; Mrázek & Karlin, 1998), but previous analyses were not able to find correlations between origins of replication and asymmetries in the *Saccharomyces cerevisiae* genome (Grigoriev, 1998; Gierlik *et al.*, 2000). We present a detailed study of base skews in the genome of *S. cerevisiae* and origins of replication sites predicted by a computer program.

Results:

Origins of replication can be predicted in eubacteria. Correlations of computationally predicted origins of replication and base skews in the *Saccharomyces cerevisiae* genome are studied in detail.

Availability:

The Perl programs used in our analysis are available upon request.

Introduction

When both strands of DNA are analyzed, roughly equal frequencies of the complementary bases C and G or A and T, respectively, are observed. However, asymmetries in the base compositions of DNA single strands have been reported in several organisms. These strand compositional biases have been used to localize origins of replication in bacterial and large viral genomes (Lobry, 1996; Mrázek & Karlin, 1998; Grigoriev, 1998). Skewed base distributions are due to several reasons, such as different mutation rates as a consequence of the asymmetric replication mechanism (Marians, 1992; Kunkel, 1992), a codon preference in connection to a favoured location of genes on the leading strand in bacteria (Zhang & Zhang, 1991), and mutation of the free, non-transcribed strand during transcription (Beletskii & Bhagwat, 1996; Francino *et al.*, 1996).

In contrast to eubacteria, only weak compositional asymmetries have been detected in eukaryotes (Grigoriev, 1998). While eubacteria usually possess a single circular chromosome and only one origin of replication, eukaryotes contain linear chromosomes and several origins. Yeast origins of replications are present in excess (Herrick & Bensimon, 1999) and are more or less randomly chosen during a replication cycle. Altogether, yeast is estimated to contain up to 400 origins, some of which are capable of directing autonomous replication to plasmids (autonomously replicating sequences or ARS). However, only 10% of the yeast origins of replication have been identified (Herrick & Bensimon, 1999).

Essentially, a yeast ARS contains a specific match to the ARS consensus sequence (ACS), which represents a conserved sequence interacting with the origin recognition complex (reviewed in Newlon, 1997). Besides the ACS up to four so-called B-elements may be located within the B domain, the DNA flanking an origin of replication. However, B domains show little apparent sequence similarities among different ARS elements (Lin & Kowalski, 1997).

Grigoriev (1998) and Gierlik *et al.* (2000) did not find significant correlations between base skew and annotated ARS elements of the yeast *S. cerevisiae*, although a correlation was found at the ends of the yeast chromosomes. In these papers, only the positions of few annotated ARS elements were taken into account. We combine both computational prediction of ARS elements and analysis of the CG-skew to study the correlation of these origin prediction methods in yeast.

Methods and algorithms

Four different base skews are calculated by a Perl program in order to predict the location of origins of replication (Lobry, 1996; Freeman *et al.*, 1998): The CG-skew $(C-G)/(C+G)$, the AT-skew $(A-T)/(A+T)$, the purine

excess (the sum of all purines minus the sum of all pyrimidines $(A+G)-(T+C)$ encountered in a walk along the sequence up to the point plotted), and the keto excess $((G+T)-(A+C)$ analogous to the purine excess). CG- and AT-skew are measured for sliding window sizes of 10000 bp.

We detect functional DNA elements with a regular expression based approach. Since B elements in yeast are less conserved than the ACS element, we predict origins of replication in yeast by a Perl program searching for matches to the ACS, 5'-WTTTAYRTTTW-3', taken from the TRANSFAC database (Wingender *et al.*, 1996).

Implementation and results

All programs have been written in Perl under the LINUX environment. The CG-skew is used to predict the origin of replication as well as the termination site of *E. coli* (Figures 1 & 2) as previously shown by Lobry (1996).

For *S. cerevisiae*, the base skews of all chromosomes are compared to computationally predicted ARS sites (see for instance CG-skew, Figure 3). We relate the locations of all origins of replication predicted by base skews to positions of ACS elements.

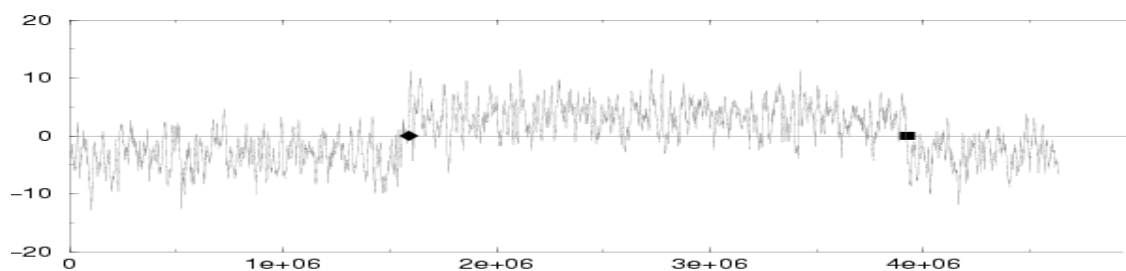


Figure 1. The CG-skew for the prokaryote *E. coli* changes sign near the location of the origin of replication (box) and the termination site (diamond; positions taken from Freeman *et al.*, 1998).

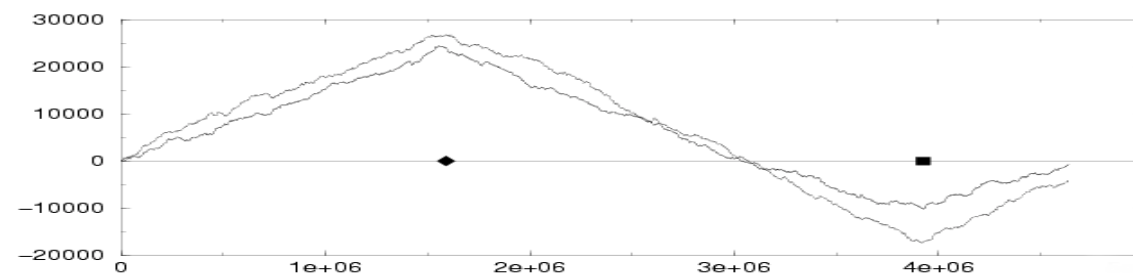


Figure 2. Purine (black graph) and keto excess (gray graph) have their minimum near the location of the origin of replication (box) and the maximum near the termination site (diamond) in *E. coli* (Freeman *et al.*, 1998).

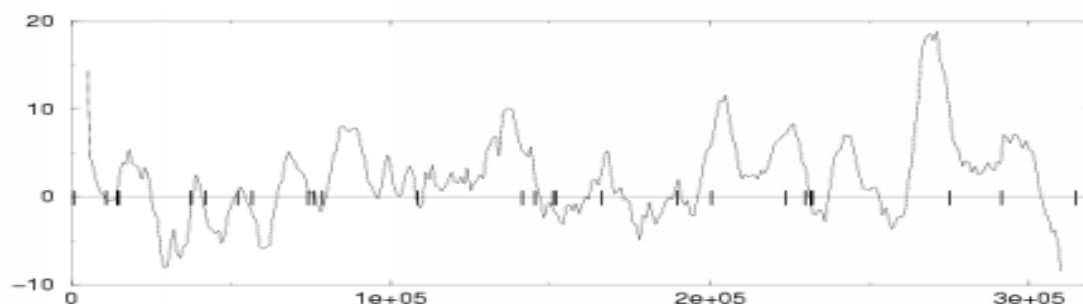


Figure 3. The CG-skew of chromosome 3 of the eukaryote *S. cerevisiae* changes sign frequently. Vertical ticks indicate predicted ARS elements.

Discussion

We study the locations of all origins of replication in yeast. In particular, we analyze whether or not there is an excess of predicted ARS sites nearby an origin predicted by base skew analyses. Such a coincidence would indicate a possible origin of replication. The candidates are compared with annotated ARS.

References

1. Beletskii,A., and Bhagwat,A.S. (1996). Transcription-induced mutations: increase in C to T mutations in the non-transcribed strand during transcription in *Escherichia coli*. *Proc. Nat. Acad. Sci. USA*, **93**, 13919-13924.
2. Francino,M.P., Chao,L., Riley,M.A., and Ochman,H. (1996). Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, **272**, 107-109.
3. Freeman,J.M., Plasterer,T.N., Smith,T.F., and Mohr,S.C. (1998). Patterns of genome organization in bacteria. *Science*, **279**, 1827.
4. Gierlik,A., Kowalczyk,M., Mackiewicz,P., Dudek,M.R., and Cebrat,S. (2000). Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? *J. theor. Biol.*, **202**, 305-314.
5. Grigoriev,A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, **26**, 2286-2290.
6. Herrick,J., and Bensimon,A. (1999). Single molecule analysis of DNA replication. *Biochimie*, **81**, 859-871.
7. Kunkel,T.A. (1992). Biological asymmetries and the fidelity of eukaryotic DNA replication. *Bioessays*, **14**, 303-308.
8. Lin,S., and Kowalski,D. (1997). Functional equivalency and diversity of *cis*-acting elements among yeast replication origins. *Mol. Cell. Biol.*, **17**, 5473-5484.
9. Lobry,J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660-665.
10. Marians,K.J. (1992). Prokaryotic DNA replication. *Annu. Rev. Biochem.*, **61**, 673-719.
11. Mrázek,J., and Karlin,S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA*, **95**, 3720-3725.
12. Newlon,C.S. (1997). Putting it all together: building a prereplicative complex. *Cell*, **91**, 717-720.
13. Wingender,E., Dietze,P., Karas,H., and Knuppel,R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238-241.
14. Zhang,C.T., and Zhang,R. (1991). Analysis of distribution of base in codon in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, **19**, 6313-6317.

BASIO: A SOFTWARE SYSTEM FOR SEGMENTATION OF BIOLOGICAL SEQUENCES INTO DOMAINS WITH HOMOGENOUS COMPOSITION

¹*Ramensky V.E., ¹Makeev V.Ju., ²Roytberg M.A., ¹Tumanyan V.G.

¹Engelhardt Institute of Molecular Biology, Moscow, Russia

²Institute of Mathematical Problems of Biology, Puschino, Russia

e-mail: ramensky@imb.ac.ru

*Corresponding author

Keywords: DNA, sequence, composition, domains

Resume

Motivation:

Assessing compositional organisation is an important step in DNA sequence analysis. Functionally important sequence regions are often biased in their local nucleotide composition from the average composition of the whole sequence, or separate regions with a more uniform composition. Besides that, many search tools use local sequence composition as a reference point in statistical tests; thus preliminary segmentation to the regions with uniform composition can improve performance of such tools.

Results:

We have developed the BASIO (Bayesian Approach to Sequence segmentatIOn) system, which allows one to segment the sequence into regions with homogenous nucleotide composition at different length scales. We consider a sequence as a series of independent random Bernoulli segments. A Bayesian estimator is used to calculate likelihood of a partition and of a boundary between segments. A parameter is set to control the length scale of resulting segmentation.

Availability:

The BASIO package is available free of charge as a set of executables for Windows 9x,NT, Linux and SGI Irix from <http://www.imb.ac.ru/combio/basio>. The source code can be obtained from authors on request.

Introduction

The heterogeneous composition of DNA sequences is a long-discussed issue. The organization of genomes is arranged over a wide range of length scales, and different functional regions are believed to be associated with the domains of a specific composition. Among other examples one can call middle range clusters in eukaryotic sequences, particularly GpC islands, open reading frames in budding yeast chromosomes, long G+C-richer regions in human genomes, which contain the major part of coding sequences. Heterogeneity of DNA sequences at least partially responsible for correlation found at many length scales. Assessing such correlation depends significantly on the statistical model of DNA sequence. Different statistical models of DNA were compared by for human and *E. coli* and the results in the two cases were remarkably similar. In this study it was demonstrated that a given local base composition tends to persist over a scale of at least kilobases or tens of kilobases. Thus a multidomain model, in which different parts of the genome are modelled by different stochastic processes, provides a reasonable first approximation.

Algorithms and implementation

A symbolic sequence over an alphabet Ω of L letters is considered as a series of segments. Each segment is characterised with counts $\mathbf{n} = (n_1, \dots, n_L)$ and is modelled as a random series of the Bernoulli type. For each configuration tested for the optimum, the Bernoulli probabilities are estimated from the counts. The measure of the statistical homogeneity of the a block is its marginal likelihood, which reflects the overall probability of obtaining the given sequence in the two stage random process. First, the composition σ is picked up according to the prior distribution, and then the sequence is generated in the Bernoulli random process with the letter probabilities σ . If $p(\sigma)$ is the uniform distribution on the surface of the simplex S , then

$$P(\mathbf{n}) = \frac{(L-1)!}{(N+L-1)!} n_1! \dots n_L! \quad (1)$$

Since we consider the segments as independent, the complete likelihood of the sequence segmentation into k segments with known boundary location writes

$$P = \prod_k P_k(\mathbf{n}_k). \quad (2)$$

This quantity is optimised over the set of all possible boundary configurations yielding the optimal segmentation. Since the total value of the optimization functional is the product of the values for the blocks, the dynamic programming can be used for the efficient optimization.

The optimal segmentation usually yields too short blocks. If one tries to study segmentation of a longer scale, one need to remove boundaries that separates segments with close composition. This can be done in a two ways. First, the problem of segmenting the sequence into the fixed number of segments can be studied via introducing an additional multiplier β^k in (2), which corresponds to assigning penalty to each boundary added. The other way is to consider all possible partitions that retain the particular boundary and to calculate the probability of this boundary using the partition function approach. The results on boundary filtration obtained via these two techniques agree for the majority of samples.

Acknowledgements

This study has been partially supported by Russian Human Genome Program 18/48 (V.E.R., V.Ju.M., V.G.T), and 99-0153-F(018) (M.A.R.), INTAS grant 99-1476, and MGRI project 244.

CAN GENETIC ALGORITHMS ASSIST IN GENOMIC RESEARCH?

Weston P.S.

UK HGMP Resource Centre, Cambridge, UK

e-mail: pweston@hgmp.mrc.ac.uk

Keywords: genetic algorithms, large search spaces, genomic research

Resume

Motivation:

Genetic algorithms (Goldberg, 1989) appear to be useful in areas where the solution search space is so large that more conventional alternatives are not feasible. The system described here shows them in operation.

Results:

The software program GATool has been developed. The program demonstrates constraint satisfaction, where the weighting applied to a problem sub-area affects the importance of that requirement being met by a candidate solution, and the generation of mathematical statements which evaluate to a user-supplied value. The Travelling Salesman Problem - defining the shortest possible route between a number of towns, all of which must be visited - is being implemented. Users can alter parameters to investigate the impact of changes on performance.

Availability:

the software is accessible over the Internet for free to registered users of the UK HGMP Resource Centre. Its graphical user interface requires an X-Window capable display. It can be run by typing "gatool" at the HGMP Resource Centre's telnet menu prompt at <telnet://menu.hgmp.mrc.ac.uk/>.

Introduction

Genetic algorithms are problem solving methods based on behaviours observed over time in populations. They use three fundamental mechanisms:

1. Selection for reproduction by fitness
2. Crossover of characteristics
3. Mutation of attributes

These processes are applied to successive generations of a population of candidate solutions, until a satisfactory solution is evolved.

Methods and algorithms

A population of candidate solutions is randomly generated. Fitness is calculated for each population member according to its closeness to the solution of the problem being represented. The two fittest members of the population are selected for reproduction. Their characteristics are passed on to two new population members, who replace two of the least fit candidate solutions. The characteristics of the parents are split between the two offspring, such that offspring 1 may have 30% of parent 1's characteristics and 70% of parent 2's, while offspring 2 is 70% parent 1 and 30% of parent 2; the proportion depends on the location of the randomly generated crossover points. In order to maintain a certain level of diversity, as the population of candidate solutions passes through successive generations some characteristics are randomly mutated.

Implementation and results

The software is written in Perl, and its graphical user interface uses the Tk module. Mutation rate is set at 0.02, two crossover points are used, and the default population size is 20. Performance is dependent on the parameters specified by the user, and because of the randomness inherent in the method of generating a solution runtime will vary on each invocation. However, sample runs on default settings indicate that the current version of the program can solve a 16-weight constraint specification in an average of 32.1 generations, with a solution taking an average of 4.6 seconds to evolve.

Discussion

This tool is being created to facilitate the exchange of ideas between computer scientists interested in genetic algorithms and molecular biologists with intimate knowledge of genomic discovery challenges. As development continues, more aspects of the program will become user-configurable; the intention is that this program will become a framework inside which both algorithms and data representations can be evaluated.

Acknowledgements

I would like to thank my colleagues at the HGMP Resource Centre for their advice, assistance, and encouragement, particularly Phil Gardner, Martin Bishop, Duncan Campbell, Gary Williams, Peter Tribble, Lee Cave-Berry, and Marc Botcherby, and my former colleague, Michael Rhodes.

Reference

1. Goldberg, D.E. (1989) Genetic Algorithms in Search, Optimisation and Machine Learning. Addison Wesley, Reading, Mass.

A DATABASE OF GENETIC TEXTS WITH LATENT PERIODICITY (LPD)

**Chaley M.B., Korotkov E.V.*

Centre "Bioengineering" RAS, Moscow, Russia

e-mail: mariam@biengi.ac.ru

*Corresponding author

Keywords: latent periodicity regions, nucleotide sequences, GenBank analysis, LPD database

Resume

Motivation:

A periodicity in genes or amino acid sequences frequently causes a particular secondary structure of proteins, α -helix, β -sheet or coiled coil structure and etc. Recently existence of a latent periodicity has been also shown in DNA and protein sequences. A biological meaning of the latent periodicity is not clear yet, though it is supposed the latent periodicity to be a consequence of multiple ancient duplications of some DNA fragment and the latter bases divergency. Perhaps the latent periodicity of gene or amino acid sequence influences a protein spatial structure. The latent periodicity of nucleotide sequences has not been strongly studied. So, a database on the latent periodicity in DNA/RNA may at least answer the questions how frequently, and where the latent periodicity is present in known base sequences. An analysis of data revealed may help in turn to understand why the latent periodicity in DNAs is needed.

Results:

We have combined the search results for the latent periodicity regions in nucleotide sequences of the GenBank database into a LPD database (Latent Periodicity Database). LPD database is the first base on the latent periodicity in DNAs/RNAs. This database provides the next possibilities:

1. to acquire information about the latent periodicity regions in a sequence with proper GenBank Accession code
2. to pick out all LPD records those with particular length of the latent periodic unit
3. to receive data about base statistics over all sites of the latent period
4. to analyze all possible lengths of the latent period for the region of latent periodicity

Availability:

Publication of the LPD database in Internet is planed to the end of 2000 year. Demo version may be available on request via e-mail: mariam@biengi.ru.

Introduction

A reason why the latent periodicity has arose may be the multiple duplications of some DNA fragment accompanied by strong divergency or a process of internal convergence in DNA sequences induced by the need to support a functional spatial structure of DNA strand or their encoded protein. A biological meaning of the latent periodicity and a relationship of gene latent structure with protein have been discussed in details in (Chaley et al., 1999). It is quite probable for gene latent periodicity to condition some kind of a protein sectoring needed for assembling of structural or regulatory complexes with other proteins or with DNA/RNA.

Method

In general, a periodicity of nucleotide sequences may be classified as homologous (perfect), eroded (imperfect) and latent. A definition of DNA/RNA latent periodicity and a method for its searching have been described earlier (Korotkov, Korotkova, 1996). Describing the latent periodicity, one can say only about statistical sufficiency for particular nucleotides to be present in each site of the periodic unit. For instance, the latent periodicity $\{(G/C/T)(C/T)N(G/A)(T/A)\}_n$ implies that bases G, C and T are prevailing at the first position of the latent periodic units, C and T are the most probable at the second position, at the third – any base may occur and so on.

Implementation and results

The first release of a database on genetic texts with latent periodicity – LPD database has been constructed to give an information about sequence latent periodicity of DNAs/RNAs from the GenBank. Delphi-4 programming tool have been used to create the LPD interface.

The LPD database contains the sequences having latent periodicity of 2, 3, ..., and so on but less than 1001 bases. The database consists of discrete records, which are identified by different LPD Accession codes. Figure 1 shows a structure of LPD record.

LPD ACCESSION A00001

GenBank ACCESSION AB000381

Z= 7.80

LATENT PERIOD = 6

COORDINATES: 2235 - 2361

SQ CCTCTGGACA TACATTGAGC CCTGATCCTG GATATATATT GAGCCCCGCT

SQ CCTGGTCACA CACTGAGCCC TGATCTGACA CTGAGCCTGG ATTCTGATCA

SQ CATACTGAGA CCCATCCCCA CACTTGG

Figure 1. An example of the LPD database record.

Besides LPD Accession code, each record of the LPD database contains GenBank Accession code of a sequence where latent periodicity region has been found, co-ordinates of this region in the GenBank, a length of periodic unit, and Z value showing a statistical sufficiency of the latent periodicity found. The details of Z counting have been discussed elsewhere, see, for example, Chaley et al. (1999). The whole sequence of the latent periodicity region is also shown.

Because the latent periodicity region has been analysed on a presence of all possible latent period lengths ranging from 2 to L/2 bases where L is the length of this region, a spectrum of Z values versus all analysed latent periods is also shown to user (Fig.2)

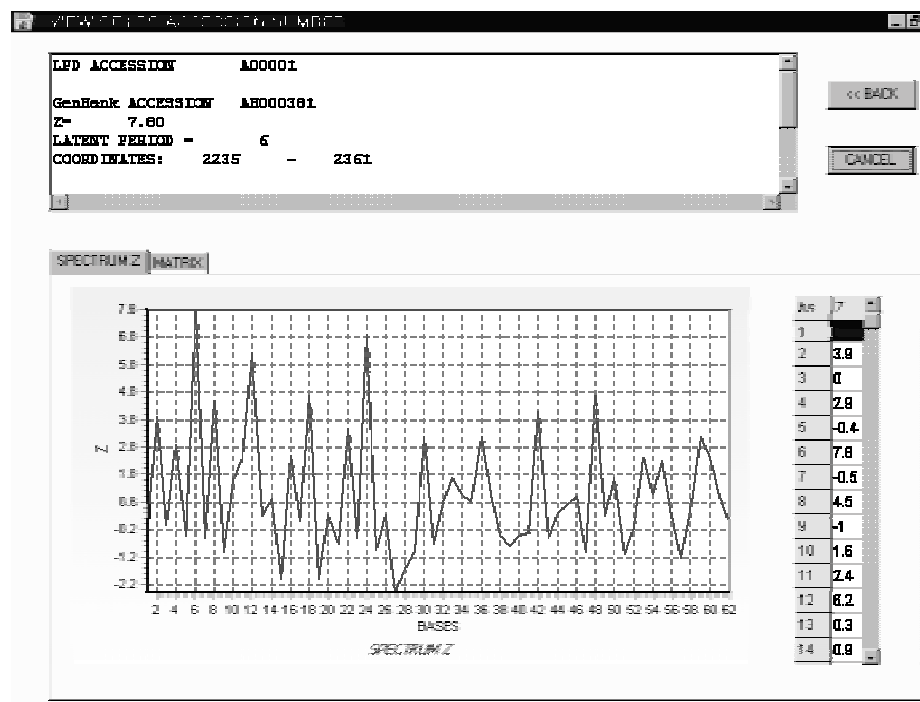


Figure 2. A spectrum of Z value according to analysed period length for the latent periodicity region.

Besides that, a matrix of A, T, C, G quantities at each site of the latent period and a histogram showing the deviation of A, T, C, G base composition at each site from casual one according to χ^2 are also shown for each LPD record (see Fig.3).

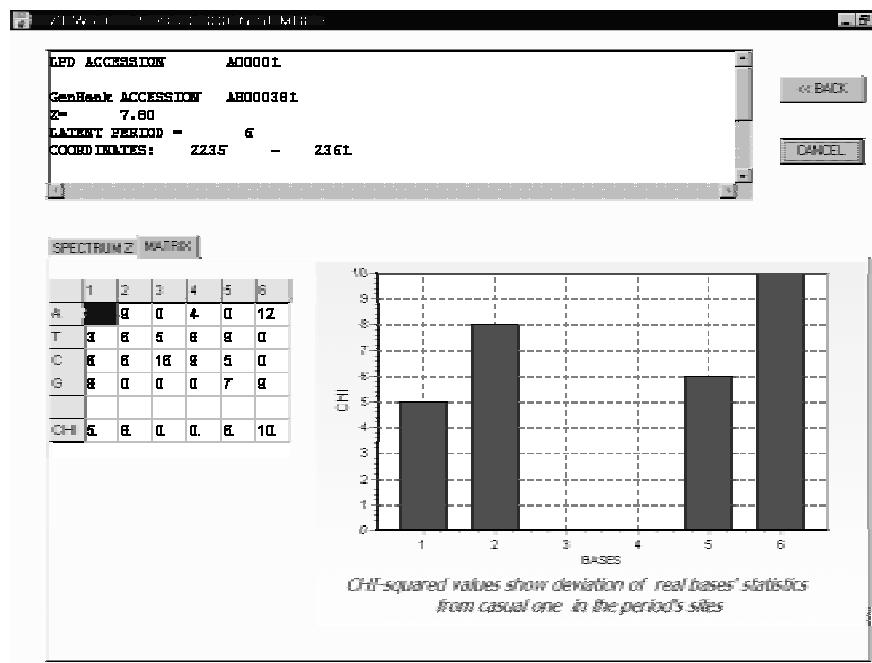


Figure 3. Base statistics over all sites of the latent period.

The LPD database lets a user to search the latent periodicity regions according to GenBank Accession code or a length of latent period (2, 3, ..., 1000 bases), or it lets to search directly with LPD Accession code.

Currently the LPD database continues to grow. Recent volume of the database counts about 18 Mb, more than 45000 records corresponding mainly to primate sequences. With the results accumulation on the search of latent periodicity, it is expected that the LPD volume will increase up to 500 Mb. The next step in database elaboration will be its publication in Internet.

Discussion

We have created the first version of LPD database on the latent periodicity of nucleotide sequences. Information given by this database is additional to description of structural peculiarities of the sequences in the GenBank database. It should be expected the LPD database to be of interest for the users searching similarity to query sequences in GenBank. It will also provide to extend our understanding of biological meaning of the latent periodicity.

References

1. Chaley, M.B., Korotkov, E.V., Skryabin, K.G. (1999) Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples. DNA Res., 6, 153-163.
2. Korotkov, E.V. and Korotkova, M.A. (1996) Enlarged similarity of nucleic acid sequences. DNA Res., 3, 157-164.

DNA SEQUENCE ASSEMBLY ALGORITHMS BASED ON CLUSTERING APPROACHES

Elloumi M.

Computer Science Department, Faculty of Sciences of Tunis, and Faculty of Economic Sciences and Management of Tunis, El Manar 2092 Tunis, Tunisia

e-mail : Mourad.Elloumi@fsegt.rnu.tn

Keywords: DNA sequence assembly, clustering approaches, exact algorithms, approximation algorithms, contigs overlaps, complexities

Resume

We present an exact *DNA Sequence Assembly* (DSA) algorithm and two approximation ones. The approximation DSA algorithms are based, respectively, on Zahn's clustering approach and Lu and Fu's one. Our DSA algorithms proceed within two steps : (i) During the first step, we construct the *Best Set of maximum weight Contigs* (BSC). (ii) Then, during the second step, we *order* the *Maximum Weight Contigs* (MWC) of the BSC, according to their overlaps order. Both of the two subproblems dealt with, respectively, during the first and during the second step are NP-complete ones. Our exact algorithm solves the first subproblem in a time of the order of $O(n^n)$, where n is the number of the strings, then, solves the second subproblem in a time of the order of $O(m^m)$, where $m=5n/2^\circ$ is the number of the MWCs. Our approximation algorithms solve the first subproblem in a time of the order of $O(n^2 \cdot l^2)$, then, solve the second subproblem in a time of the order of $O(m^2 \cdot l^2)$, where l is the length of a string.

THE EVOLUTION OF REGULATORY FAMILIES IN ARCHEA AND EUBACTERIA: A COMMON ORIGIN OF TRANSCRIPTIONAL REPRESSORS

*Ernesto Perez-Rueda, *J. Collado-Vides*

Centro de Investigacion sobre Fijacion de Nitrogeno, Universidad Nacional Autonoma de Mexico, Mexico

e-mail: collado@cifn.unam.mx

*Corresponding author

Keywords: transcription factor, families, evolution, bacteria, helix-turn-helix motif, transcription regulation

Resume

An exhaustive collection of 314 transcriptional DNA-binding proteins have been identified in *Escherichia coli*, based on literature search and computational predictions. They were identified and analyzed in 17 eubacteria and 6 archaea genomes. Around 900 new transcription regulators with the helix-turn-helix (HTH) as the main DNA-binding motif were identified in addition to the 314 of *E.coli*. Based on the distribution in different organisms of these regulators, an evolutionary reconstruction of the 20 different regulatory families that group all these regulators, is presented. A super-group of proteins with a common origin is proposed. These proteins have their HTH motif at the N-terminus. Evidence is shown supporting the notion of a super-group of proteins with their HTH motif at the N-terminus that share a common origin. The supergroup is formed by repressor proteins of eight families sharing functional and structural features. Similar evidences suggest that the LysR proteins have different structural restrictions and different origin. This other group, although with the HTH also in the N-terminus, has constitute mostly dual proteins able to activate several genes and repress their own gene. These results shed light on the origin of transcription regulation in bacteria and archaea.

USING LOCUS-SPECIFIC DATABASES OF HUMAN MUTATIONS FOR ESTIMATING PERNUCLEOTIDE RATE OF SPONTANEOUS MUTATION

Kondrashov A.S.

National Center for Biotechnology Information, Bethesda, USA
e-mail: kondrashov@ncbi.nlm.nih.gov

Keywords: mutations, diseases, estimate

Resume

Data on molecular nature of disease-causing human mutations provide a new, important opportunity to estimate m , per nucleotide rate of spontaneous mutation. Currently, sufficiently large locus-specific data bases are available for over 20 human loci. Estimating m at a locus requires knowledge of a) per locus mutation rates, often available with acceptable precision for dominant and X-linked recessive diseases, b) the sequence of the locus and, thus, of the number of nucleotides whose mutations can create in-frame stop codons (target size), and c) the fraction s from all disease-causing mutations at the locus due to various nonsense nucleotide substitutions. Direct estimates of $m \sim 2 \times 10^{-8}$ obtained in this way are in good agreement with indirect estimates obtained by comparison of Homo and Pan orthologous pseudogenes. Thus, the total number of spontaneous de novo mutations per zygote in our species is over 100.

COMPUTER MODELING OF EVOLUTION OF THE GENETIC DIVERSITY OF INTERACTING POLYGENIC SYSTEMS AND PATTERNS OF MOBILE GENETIC ELEMENTS IN THE COURSE OF SELECTION FOR THE QUANTITATIVE CHARACTER

^{1,2,*}Ratner V.A. ¹Yudanin A.Ya., ²Egorova A.V.

¹Institute of Cytology & Genetics, SB RAS, Novosibirsk, Russia

e-mail: ratner@bionet.nsc.ru

²Novosibirsk State University, Russia

*Corresponding author

Keywords: computer model, MGE pattern, polygenic system, selection, inbreeding, modifier, tree of similarity

Resume

The computer system [1]

The system of computer programs for modeling of population dynamics of interacting patterns of polygenes and mobile genetic elements (MGEs) in the course of selection for the quantitative character was developed. The multi-loci pattern always is very interesting stochastic subject in population, requesting the special means of description. Actually this is stochastic (Monte-Carlo) system of modeling, accounting of the main sources of random and directed changes of patterns: recombinations, MGE transpositions and excisions, genetic drift, different deterministic «trends» of selection, and the types of interaction between polygenes and MGE copies in the finite population. The model permits to watch «from inside» the dynamics of all population characteristics, inaccessible for direct measurements in experiment: the frequencies of the polygenes and MGE copies, their average heterozygosity, proportions of adaptive and random fixations, coefficient of inbreeding, heritability, etc. Moreover, this system of modeling could be used for checking of different hypotheses of interaction between polygenes and MGE copies. We simulated by the system the real experiments of L.A.Vasilyeva et al. on the truncation selection of the quantitative character «*radius incompletus*» in populations of *Drosophila melanogaster* [4].

The Computer Model [2]

Basing on this computer system, the computer model of population dynamics of the polygenic system of additive character and MGE pattern under directed truncation selection for this character was developed. The results of computer modeling are in good accordance with the experimental data. It was shown, that MGE-modifiers were quickly and adaptively fixed (or lost) together with the modified polygenes, and MGE-markers and independent copies were fixed (or lost) so fast, but random. The anomalous fast fixation of MGE-pattern, actually observed, may be explained not only by acting of MGE-copies on the polygene expression, but also by fast increase of all-genomic inbreeding. The method of specific labeling of all initial haploid genomes, tracing the origin of every genomic segment of every individual (in model) and direct calculation of mean portion of identical meetings per locus per generation, was used to show the main result. Under strong selection for quantitative character, controlled by polygenic system, and under high value of multiple progeny production in the finite population, the coefficient of nonsystematic inbreeding increased quickly up to the values of 0.7-0.9 for 15-20 generations. This corresponds to the high level of homozygosity (random, as a rule) of all the loci with MGE among them. It was shown, as a result, that active alleles of polygenes, being modified by MGE-copies, were selected and fixed among the first segments and with maximal probability. The adaptive homozygotization of polygenes and MGE-modifiers, and random homozygotization of MGE-markers and independent copies, and of all the other parts of genome, occurred. These results grounded the hypothesis of the «pattern- champion» of polygenes, formulated earlier for explanation of results of selection-genetic experiments.

The building of the trees of pattern similarity [3]

The building of the trees (dendrogrammes) of pattern similarity is convenient, being approbated in the theory of molecular evolution as method of representation of their diversity, useful for quantitative estimates and comparisons. The method UPGMA was used. The trees of similarity of MGE patterns, and also of identity labels for consequent generations, were built. It was shown that under selection the diversity of these patterns dramatically decreased in the first generations; then up to 10 generation population is represented by one big «family»; and in the final (50-th) generation there is only one group of tightly related genotypes (see Fig. 1).

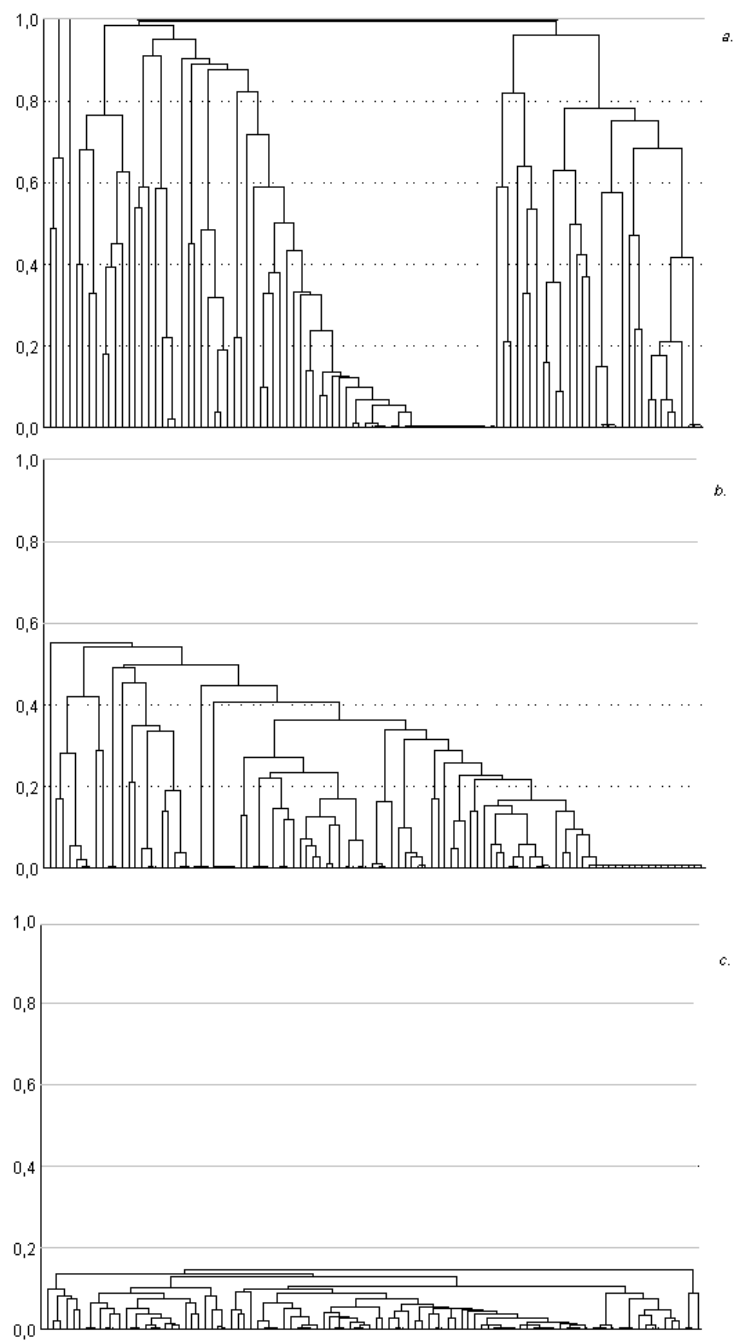


Figure 1. Trees of similarity of the identity label patterns for consequent generations: 4-th (a), 10-th (b), 50-th (c). The endpoints correspond to different patterns of the sample. The ordinate is the distance between patterns measured by the portion of the differences.

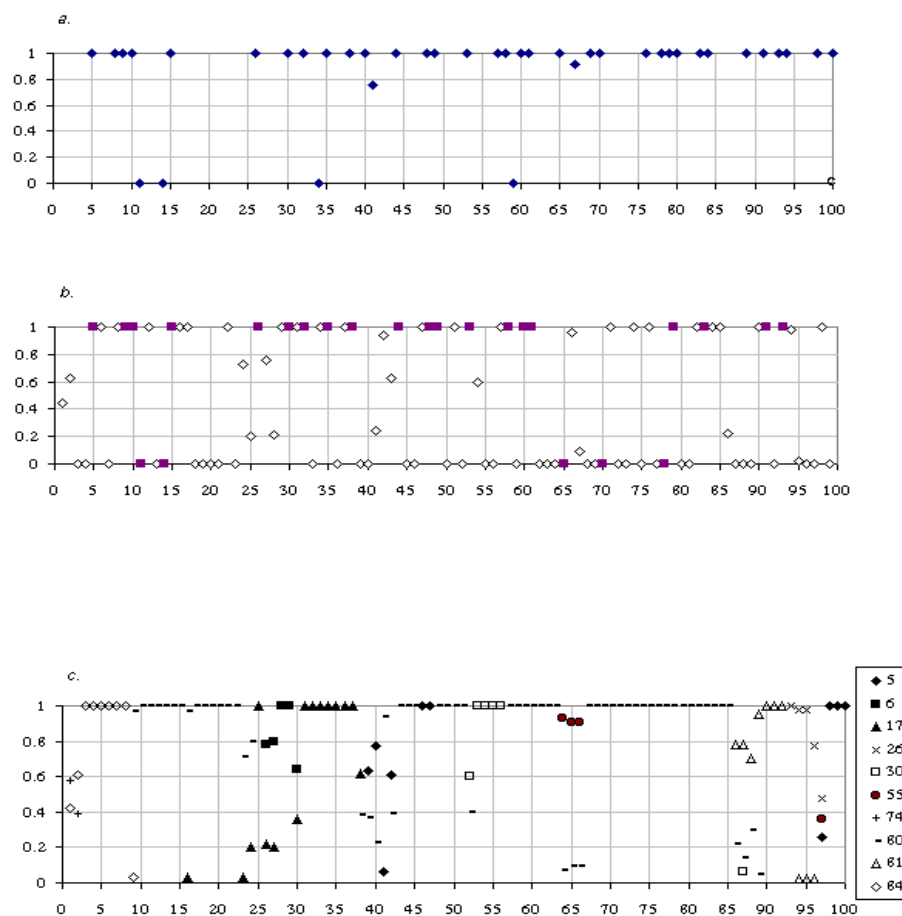


Figure 2. The final pattern-consensuses (after 50 generations of positive selection: a) - polygenes, the frequencies of active alleles; b) - copies of mobile elements; c) - identity labels; different symbols correspond to segments of different origin; the numbers of ancestor segments are indicated. The abscissa contains the order numbers of segments; the ordinate is the population frequency of corresponding subjects.

The final united pattern-consensus contains of different blocks including the segments of common origin (common labels of identity), each of them containing the active allele of polygene modified by the MGE copy (Fig.2).

This work was partially supported by grants of RFBR (N 97-04-49232 и 00-04-49499) and SSTP of RF Ministry of Education «Russian Universities – Basic Studies» (N 1760).

References

1. Ratner V.A., Yudanin A.Ya. (2000a) Computer system for modelling of the population dynamics of patterns of polygenes and mobile genetic elements under truncation selection for quantitative character. *Genetika (Russ)* 36, 3, 399-406.
2. Ratner V.A., Yudanin A.Ya. (2000b) The truncation family selection and non-systematic inbreeding result in fast fixation of the patterns of mobile genetic elements in computer, *Genetika (Russ.)* 36, 3, 407-412.
3. Ratner V.A., Yudanin A.Ya., Egorova A.V. (2000), *Genetika (Russ.)*. 36
4. Ratner V.A. Self-reproduction of assemblies of macromolecules: comparative analysis of a problem. *Computational Technologies*, 2000, V.5., p.120-127.
5. Ratner V.A., Vasilyeva L.A., Bubenshchikova E.V., Antonenko O.V. (2000), this volume.

PATTERNS OF MOBILE GENETIC ELEMENTS (MGEs) GENOMIC LOCALIZATION: INDUCTION OF TRANSPOSITIONS BY STRESS FACTORS, RESPONSE TO SELECTION AND POSSIBLE EVOLUTIONARY CONSEQUENCES

^{1,2,*}Ratner V.A., ^{1,2}Vasilyeva L.A., ¹Bubenshchikova E.V., ¹Antonenko O.V.

¹Institute of Cytology & Genetics, SB RAS, Novosibirsk, Russia

e-mail: ratner@bionet.nsc.ru

²Novosibirsk State University, Russia

*Corresponding author

Keywords: mobile genetic elements, pattern, polygenes, induction of transpositions, stress, selection, inbreeding, evolution of patterns

Resume

The mobile genetic elements (MGEs) occupy up to 10-30% of genomic DNA of animals and above 50% - of plants. The MGE genomic system participates in expression of genes and polygenes, their structural and regulatory variability, as a result – in response to selection of Mendelian and quantitative characters, and as a sum total – in evolutionary process. The original laboratory and isogenic lines of *Drosophila melanogaster* with Mendelian mutation *radius incompletus* were chosen as adequate experimental model of investigation of phenomena. The expression of this mutation is controlled by specific polygenic system. The patterns of some MGEs (412, etc.) were revealed by method of *in situ* hybridization of the probes, containing the copies of these MGEs, with DNA of polytene chromosomes of drosophila larvae salivary gland cells. The stress factors were: external treatments (heat shock (HS), cold shock (CS), heavy heat shock (HHS), γ -irradiation, ethanol, etc.) and genetic crosses (inbreeding, outbreeding, isogenization, etc.). The response of MGE patterns to stress treatments of males was investigated, the phenomenon of induction of retrotransposon (412, B104, etc.) transpositions being proved. The phenomenon of induction of novel genetic variability of polygenes by MGE transpositions was shown. The response of MGE patterns to strong truncation selection of *radius incompletus* character (the sum length of two fragments of the broken wing radial vein) was investigated in selection-genetic experiments. This phenomenon was demonstrated as such itself, the contributions of random and selective factors to final MGE patterns-consensuses after selection completion were analyzed. The copies of MGE-modifiers acting on the polygene expression were revealed.

Induction of MGE transpositions

The males from isogenic lines were treated by temperature factors (HS, CS, HHS) and γ -irradiation; then they were crossed with non-treated females of the same lines; transpositions were revealed at F1 larvae by *in situ* hybridization. As a whole, the induced transpositions occurred with the rates $\sim 10^{-2} - 10^{-1}$ events per segment per spermium per generation. These values ought to be estimated as enormous! The similar estimations of the rates were found after induction in the course of isogenization of lines. The «hot» segments of transpositions were found in all induction experiments. As an example, in isogenic line # 51 after treatment by HHS, the 70% of all MGE transpositions were localized in two segments (43B and 97DE) of Bridges's cytological map.

Induction of the polygene variability

In the isogenic line # 51 without HHS treatment, the response to selection for quantitative character was inefficient, i.e. the genetic variability of polygenes was absent. After HHS-treatment of the males the selection became high efficient during the 50 generations. The only consequences of HHS treatment were induced transpositions of MGEs. It means that they induced the variability of polygenic system.

Response of MGE pattern to selection for quantitative character

The selection after HHS was accomplished in (+)- and (-)-directions, in 3 replicates, in isogenic line # 51 with effective size $N_e = 160$. The 35 novel heterozygous MGE positions had appeared after HHS-induction in the sample of 85 individuals. In the course of selection 26 of them expressed themselves as independent copies and markers, but 9 – as selective MGE 412 copies. It was surprising that the second group contained all the «hot» segments of induction (43B, 97DE, etc.). It is suggested that MGE-modifiers of polygenes were among them. As a result, the final MGE patterns-consensuses of (+)- and (-)-selection at 50-th generation contained the random and selective components. The most important feature of selection dynamics is the fast elimination of polymorphism of all MGE copies: after 10-20 generations the MGE patterns became actually homozygous.

This effect could be explained by very powerful selective inbreeding under strong truncation selection. The similar features of the response to selection were conformed in all different independent selection experiments.

The mechanisms of transposition induction

It is suggested that genetic system of response to HS and CS are the molecular mechanisms of the temperature (HS, CS, HHS) and inbreeding induction. The motifs of functional sites of HS-system were found into DNA-sequences of retrotransposons. However, the MGE copies are not active components of HS-system. They are rather passive genetic elements accepting the shifting-on signal only because they contain the corresponding sensitive regulatory sites. Furthermore, the HS-system respond not only to HS, but also to tens different treatments enforced the cell concentration of the proteins with defective spatial conformation. The natural factors (temperature, inbreeding, virus infection, etc.) and antropogenic ones (heavy metals, ethanol, different chemicals and poisons, etc.) are among them. γ -irradiation induced probably the two-stranded DNA breaks, those could be healed by copies of retrotransposons. The influence of outbreeding could be explained probably by heterozygous state of MGE copies. The isogenization is combination of outbreeding with very fast inbreeding.

Possible evolutionary consequences

The revealed facts and phenomena grounded the new concept of variability and evolution of genomes with participation of MGE system. The MGE transpositions in the cells of the male germ line are first of all related with the «unpacked» loci of these cells, those being functional in germ line. It is probable that these loci played the role of polygenes of the expression systems of different traits, - the vein formation of the wing. In this case transpositions produce the «soft» modification of the adjacent polygenes, that rather being not «rough» their defects, but regulatory variability. The action of enhancers, insulators and different functional sites of transcription regulation could be molecular mechanisms of modification. Under induction such changes became mass in population and multiple in genomes. It means that being in stress conditions of existence, populations respond on them by the bursts of induced MGE transpositions, and through them – by the bursts of regulatory variability of genes and polygenes. The factors of induction may be natural and antropogenic. The HS and CS are probably the frequent factors in conditions of cold winter and hot summer of Northern EurAsia. One such case was found by us in the course of selection experiment. The chemical and radiation contamination and common usage of ethanol could be different factors of induction in many species including humans. The fast non-systematic inbreeding and outbreeding could be important in small populations. The strong truncation selection in the small populations leads automatically to the strong all-genomic inbreeding, that devastates the reserve of hereditary variability and could lead to inbred degeneration. It is important that always such events occur independently from selected character. Meanwhile, induction of transpositions rapidly restores this reserve. Therefore, the populations could quickly respond by regulatory variability to stress changes of conditions of existence, and quickly reorganize the species norm of characters. From the other side, the finite population could overcome not any values of induction. There is evident upper limitation of the role of transpositions, those could be compatible with viability. Therefore, the induced bursts of transpositions are permissible only for short time intervals. The inducible populations have probably the increased rates of speciation and molecular evolution. The transposed retrotransposons must also obtain the accelerated processes of mutations and molecular macro-evolution. As a result, the stress treatment becomes though short, but important factor of evolutionary process. In the frames of Molecular genetic regulatory systems of the cells and organisms, MGE could play the role of «movable cassettes of functional sites», participating in «soft» reorganization of genetic regulation. The tracks of such regulatory reorganization were found into the genomes of many higher plants and animals, including humans.

The work was supported by grants of ISF (N RAS300), RFBR (NN 97-04-49232 и 00-04-49499), SP «Integration» (N 618) и SSTP of the Ministry of Education of RF «Russian Universities – Basic Studies» (N 1760).

References

1. Vasilyeva L.A., Bubenshchikova E.V., Ratner V.A. (1999) Heavy heat shock induced retrotransposon transposition in *Drosophila*. *Genet. Res., Camb.*, 74 (2): 111-119.
2. Vasilyeva L.A., Ratner V.A. (2000) Heavy heat shock (HHS) induced new genetic variability in the polygenic system of a quantitative trait in *Drosophila melanogaster*. *Genetika. (Russ.)*, 36, N 4, 493-499.
3. Vasilyeva L.A., Bubenshchikova E.V., Antonenko O.V., Ratner V.A. (2000) Response of the TE 412 localization pattern to truncation selection of a quantitative trait in an isogenic line of *Drosophila melanogaster* after heavy heat shock. *Geneika (Russ.)*, 36, N 6, 774-781.
4. Vasilyeva L.A., Ratner V.A. (2000) The polygenic system of the quantitative character *radius incompletus* in *Drosophila*: genetic features, interaction with other genes, evolutionary properties. In: «The Modern Concepts of Evolutionary Genetics» (Eds. V.K.Shumny and A.L.Markel). Novosibirsk: IGC SB RAS, 132-144. (Russ.).
5. Ratner V.A., Vasilyeva L.A. (2000) Mobile genetic elements (MGEs): «selfish DNA» vs. functional elements of genome? там же, 145-170. (Russ.).

EVOLUTION OF THE CODE AND THE EARLIEST PROTEINS. RECONSTRUCTION FROM PRESENT-DAY SEQUENCES

Trifonov E.N.

The Weizmann Institute of Science, Rehovot, Israel
e-mail: edward.trifonov@weizmann.ac.il

Keywords: proteins, evolution

Resume

Hidden periodical (GCU)_n pattern in extant mRNA, and predominant (GCT)_n repeat in the triplet expansion diseases suggest that the GCU triplet and its 9 point change derivatives could have been the first codons. The earliest six amino acids (A, D, G, P, S, and T) are suggested by their chemical simplicity, by experiments of Stanley Miller, and by association with more ancient aminoacyl tRNA synthetases of class II. Strikingly, all these amino acids, indeed, are encoded by the above mentioned codons (1).

After, thus, establishing the earliest group of the amino acids, one may try to reconstitute a whole chronology of the amino acids, by exploiting all other known criteria of evolutionary age of the amino acids. 40 such different criteria and hypotheses about chronological order of appearance of amino acids in the early evolution are summarized in consensus ranking, earliest first (2, 3): A, G, D, V, S, E, P, L, T, N, K, R, I, Q, C, H, F, M, Y, W

Due to consensus nature of the chronology it has several important properties not visible in individual rankings by any of the initial criteria. Nine amino acids of the Miller's imitation of primordial environment are all ranked as topmost (A, G, D, V, S, E, P, L, T). This result does not change even after several criteria related to Miller's data are excluded from calculations.

One may expect that in the composition of the ancient proteins the earliest amino acids would dominate. This is confirmed by matching prokaryotic and eukaryotic protein sequences. The glycine content of the matching residues is especially high. The glycine may, then, serve as a measure of the time since the separation of compared species – glycine clock. This approach is applied to over 400 pairwise sequence alignments of proteins of 6 major kingdoms. The evolutionary tree is derived, where the kingdoms separate consecutively from the central stem in the order Eubacteria, Archaea, Protocista. Fungi, Plants and Animals come last, by an apparent trifurcation (3). The glycine clock tree is consistent with the consensus tree topology of the kingdoms derived by traditional techniques.

The above amino-acid chronology allows to reconstruct the chronology of codons as well. Three striking features are revealed: (i) the codons providing the most stable codon-anticodon interactions appear first, (ii) the new codons appear as complementary pairs and (iii) the new codons are derived processively from the chronologically earlier codons by wobble mutations and complementary copying.

Two essentially independent amino-acid alphabets are suggested by the above evolutionary scheme, for two complementary coding strands of the earliest small genes. The Glycine family includes amino acids encoded by triplets with purines in central position - G, D, S, E, Q, N, R, K, C, H, Y and W. The Alanine family consists of amino acids A, V, P, S, L, T, I, M and F, with pyrimidines in the central positions of their codons. One may speculate that after the earliest genes were, presumably, fused to form longer molecules, the encoded protein sequences contained a mosaic of short patches of residues from two different alphabets (4). This sequence organization may, perhaps, still be recognized in extant proteins.

Analysis of the sequence-wise independent crystallized protein structures reveals that all globular proteins are built of standard size closed loops, 25-30 residues, consecutively arranged along the protein sequences (5). This structural organization of proteins is independent on presence and positioning of the secondary structure elements which are, thus, indeed, secondary. The loop of 25-30 amino acids is a polymer-statistical optimum for ring closure and is likely to represent a major stage in the early evolution of proteins.

References

1. Trifonov, E. N., Bettecken, T., Sequence fossils, triplet expansion, and reconstruction of earliest codons. *Gene* 205, 1-6 (1997)
2. Trifonov, E. N., Elucidating sequence codes: three codes for evolution. *Annals NY Acad. Sci.* 870, 330-338 (1999)
3. Trifonov, E. N., Glycine clock: Eubacteria first, Archaea next, Protocista, Fungi, Planta and Animalia at last. *Gene Therapy and Molecular Biology* 4, 313-322 (1999)
4. Trifonov, E. N., Leap Into Life's Beginnings. Tracking the chronology of amino acids. *Science Spectra* 20, 62-71 (2000)
5. Berezovsky, I. N., Grosberg, A. Y., Trifonov, E. N., Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters* 466, 283-286 (2000)

EVOLUTION OF PLANT REGULATORY SEQUENCES

*Goebel U., Wiehe T., Mitchell-Olds, T.

Max-Planck-Institute of Chemical Ecology, Jena, Germany

e-mail: goebel@stargate.ice.mpg.de

*Corresponding author

Keywords: regulatory regions, phylogenetic footprint, sequence-property-mapping, evolution of noncoding regions

Resume

Motivation:

The pattern of sequence conservation in a comparison of two homologous genomic DNA fragments is routinely used to identify functionally important regions [2,3]. The rationale behind this approach is to detect the presence of a function via the traces of purifying selection. For sake of simplicity it is assumed that DNA sequences present themselves to selection as strings of characters, with fitness differentials approximately proportional to the edit distance of two strings. In the case of regulatory regions, which are evaluated by selection on the level of physicochemical properties of the DNA molecule, the success of conservation based approaches depends on how far the sequence-property mapping is from being one to one (existence or not of *neutral mutations* [5]). We propose to account for this mapping by comparing symbolic sequences consisting of structural states instead of the raw sequences, and discuss an example of very different degrees of conservation on the level of sequence and structure.

Results:

We have preliminarily analyzed the 5' flanking regions of two homologous endochitinase loci from *Brassica napus* and *Solanum tuberosum*. The results suggest that structural conservation is a valuable aid in the assessment of weak sequence similarity in regulatory elements. The ratio of the abundance of fragments which show a similar pattern of the parameter Major-Groove-Width [6] in the flanking region to the abundance in the first exon is about two times the ratio of sequence conserved fragments, and thus better discriminating between the two regions of the gene.

Introduction

Ponomarenko et al. [1] have shown that binding sites of transcription factors can be characterized in terms of physicochemical properties of the DNA molecule, without explicit reference to the nucleotide sequence. One way to utilize this fact is to scan an unknown sequence for the occurrence of the typical profile of a known factor. Here we take a different view and scan the alignment of two homologous sequences for *phylogenetic footprints* of conservation of a physicochemical property, which allows the identification of unknown functional elements.

Methods and Algorithms

As a first approach to assessing sequence similarity on the level of derived properties of a sequence, we have transcribed DNA sequences into symbolic sequences: Consecutive characters in the symbolic sequence correspond to consecutive sliding windows in the input. The alphabet of the symbolic sequence consists of a fixed number of arbitrary but different characters, each of which denotes a bin in the range of the (real valued) property which is to be coded. A given sliding window is assigned the character which belongs to the mean value of the property in the window. Alignment of the real and symbolic sequences is done with the alignment tool sim96 [4], which reports all matching fragments above a certain score, even if they overlap.

Results

Two homologous *endochitinase loci* from *Brassica napus* (GenBank acc. no. M95835) and *Solanum tuberosum* (X15494) were compared with respect to the derived property Major Groove Width [6]. The range was divided into ten bins of width 0.334, corresponding to the symbolic characters {A,B,C,D,E,F,G,H,I,K}. The window size was 6. The overall (raw) sequence identity of the pair of loci is 47% in the flanking region and 62% in the first exon. There are 19 (possibly nondisjunct) gap free aligned fragments of percent identity $\geq 60\%$ in the flanking region, and two long disjunct alignments in the exon. This disparity of patterns becomes more pronounced if we proceed to the symbolic sequences. Here, there are 106 (possibly nondisjunct) gap free alignments of percent identity $\geq 60\%$ in the flanking region and 6 alignments (which do not cover all of the sequences) in the exon (for a comparison of sequence and structural similarity of the full length sequences, see Fig. 1). In addition the flanking region is characterized by longer runs of identical states.

One of the aligned fragments in the flanking region demonstrates the different evolutionary dynamics at the sequence and structure level:

The symbolic sequence alignment is gap free at 80% identity:

```
B. napus          GFEDCCCCCBAAABBBAAAAABBB
S. tuberosum     GFEDCBCCCCBABBBBBBAABAAABB .
```

A gap free alignment at the sequence level has only 28% identity:

```
B. napus   GCGCTCATATCATAATTACTTTTAA rev=TTAAAAGTAATTatgatatgAGCGC
tuberosum ATCACCCATTATTTGTGATTCATCG rev=CGatgAatCACAAATAatgGGTGAT,
```

but the sharing of some displaced words between the sequences (marked in the reverse strand by lowercase letters) may point to a more complex evolutionary history. A query of PlantCARE [7] with the homologous fragments revealed possible similarities of the reverse strand of Brassica with the site AT~ocs-element from *Arabidopsis thaliana*

```
B. napus      TTTAAAAGTAATTatgA---TatgAGCGC---
AT~ocs-element  -----ATCTTatgTCATTGatgACGACCTCC
```

and of the reverse strand of *Solanum* with PS~cab-CMA2_1 of *Pisum sativum*

```
S. tuberosum  ---CGatg-AATCACAAATAatgGGTGAT[A]
PS~cab-CMA2_1  CACACatgGAA---N(18)---atgATATGA .
```

(note the two ATG elements, which also occur in AT~ocs-element). Both matches are weak by themselves, but the structural conservation corroborates both the hypothesis that this is indeed a regulatory site and that the fragments are truly homologous in the sense of sharing a common ancestor. The discrepancy between the alignment on the level of sequence and structure may indicate that mutations have been compensated for at remote sites.

Discussion

The results presented show that meaningful information can be extracted from phylogenetic footprints of properties of a DNA sequence such as physicochemical states. We plan to extend the approach by incorporating properties other than major groove width. Overall we want to better understand the relation between evolution on the level of sequence and on that of structure, with the goal of supporting the assessment of weak sequence similarities in noncoding regions.

References

1. J.V. Ponomarenko et al.: Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics* **15** 654-668 (1999).
2. Laurent Duret and Philipp Bucher: Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology* **7** 399-406 (1997).
3. Christopher B. Burge and Samuel Karlin: Finding the genes in genomic DNA. *Current Opinion in Structural Biology* **8** 346-354 (1998).
4. <http://globin.cse.psu.edu/globin/html/software.html#sim96>
5. Peter Schuster and Peter F. Stadler: Sequence redundancy in biopolymers. A study on RNA and protein structures. In: Gerald Myers, ed. *Viral Regulatory Structures. SFI Studies in the Sciences of Complexity. Advances in HIV and HPV virus research*, Vol. XXVIII: 163-186, Addison-Wesley, Reading MA, (1998).
6. Karas H, Knuppel R, Schulz W, Sklenar H, and Wingender E: Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput. Appl. Biosci.* **12** 441-446 (1996).
7. S. Rombauts, P. Dehais, M. Van Montagu and P. Rouze: PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res.* **27(1)** 295-6 (1999).

A NEW VERSION OF SYNAP COMPUTER PROGRAM FOR LOGICAL MODELING OF PHYLOGENY

*¹Baikov K.S., ²Zverev A.A.

¹Central Siberian Botanical Garden, SB RAS, Novosibirsk, Russia
e-mail: kbaikov@mail.ru

²Department of Botany, Tomsk State University, Tomsk, Russia
e-mail: zverev@ecos.tsu.ru

*Corresponding author

Keywords: new version, computer program, method SYNAP, logical modeling, phylogeny, object, vector

Resume

Since 1994 we have developed a new method for logical modeling phylogeny. Its name SYNAP is produced from the first letters of the term 'synapomorphy' of W. Hennig (1966). The method is developed as a combination of the best logical achievements in phylogenetic systematics. Algorithms of Wagner (1961), Camin & Sokal (1965), Li (1990) as well as computer programs PAUP, Hennig86, PHYLIP were re-analyzed.

Introduction

Contemporary methods and computer programs give means to infer phylogeny and to estimate evolutionary relationships on the base of molecular and morphological data. Study of biodiversity using logical modeling of phylogeny allows to systematize our knowledge of biodiversity and consider its development in dynamics, as a process.

Methods and algorithms

The SYNAP computer program was compiled by Andrew Zverev (Tomsk State University, Russia). The program is based on SYNAP method (Baikov, 1996, 1999). A history of creating and development of the program were described in details earlier (Baikov, 1999). The program SYNAP is developed in Clipper 5.2 and needs in about 0,5 Mb of disk space, steadily works on PC 286 with 1 Mb RAM. In case of processing the small data matrix all process may be performed without use of a hard disk. The compression of results reduces their sizes in 10-100 times.

The program has several versions which can conditionally be divided into two groups. The first group has a textual format of data (ASCII-codes) and is operated with the help of a command file (versions from SYNAP066 till SYNAP113). The second group has special format of initial data and completely is operated with the help of the screen menu (version SYNAP150 and above). In all versions of the program the manipulator "mouse" is not supported.

Implementation and results

Main menu of SYNAP computer program version 3.0 is shown on Fig. 1. With command "Data / Open" you can open a file you have edited and saved previously. Then choose it from the list. With command "Data / New" you can create a new data matrix (Fig. 2). You can give it a file name, title of data matrix, author's name, numbers of objects and vectors, some comments.

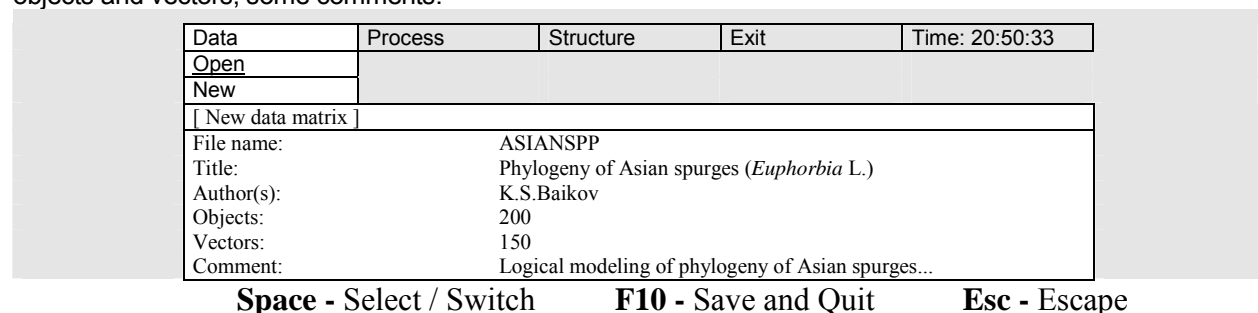


Figure 1.

A new data format was elaborated. Data matrix consists of numbered rows (objects) and numbered columns. Textual label (up to 30 symbols) and description (up to 255 symbols) may be attached to an object. Description (up to 255 symbols) and weight (is equal one, two or three) may be given to a vector. Description of the active vector is placed in special window under the data (fig. 2).

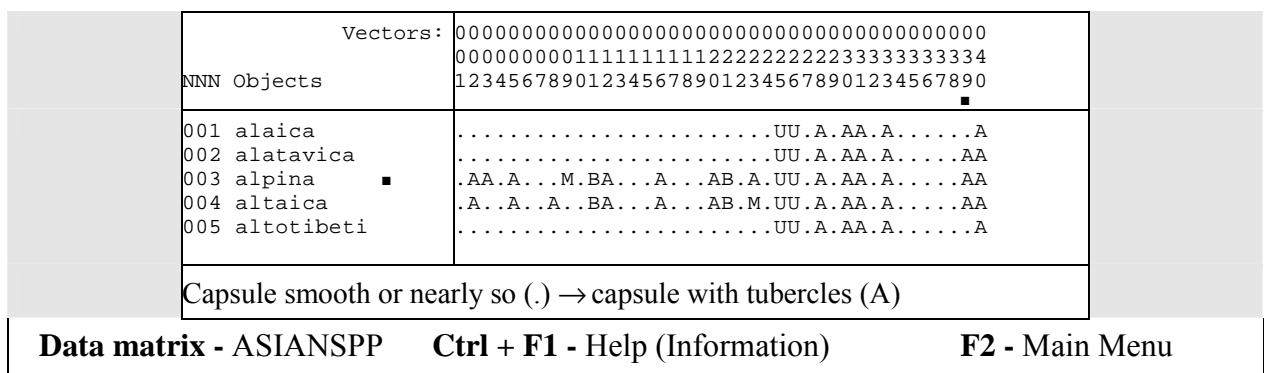


Figure 2.

In comparison with other programs of the phylogenetic analysis for personal computers, such as PAUP (Swofford, 1985), Hennig86 (Farris, 1988) and PHYLIP (Felsenstein, 1993) SYNAP has some important peculiarities. So, the complemented palette of codes (logic variants of the description of attributes) allows to take into account and to process practically all suitable cases. Logical rules of processing are offered for each code. That increases the information quality in resulted evolutionary scheme and opportunity of its explanation. Concepts of "a new attribute" and "an elementary evolutionary vector" are developed to specify synapomorphy of Hennig (1966) as similarity in a new attribute arisen after one and the same phylogenetic event (Baikov, 1997).

Some important options for creating and editing data matrix are shown on Fig. 3. Only in the version you may add a new object or a new vector, move any one to the other place, weight a vector, exclude lists of active objects and active vectors out of file with the resulted phylogenetic scheme.

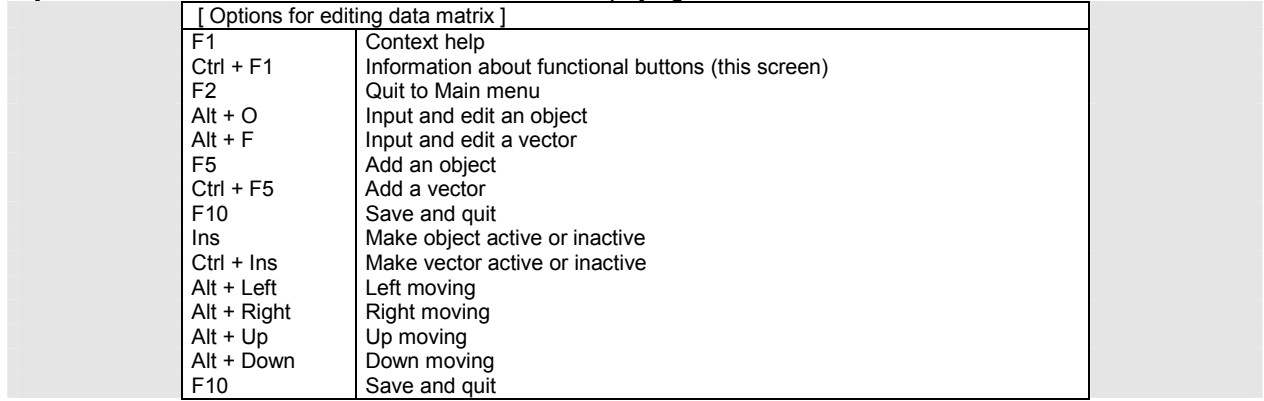


Figure 2.

Objects connect together according to the maximum of index of potential relationship which is equal to the sum of identical vectors. They are marked in step-by-step protocol by N-code and S-code (Fig. 4). In case of weighed vectors the weighed index will be calculated. Special rules for taxa connection in case of equal index are improved. Using the protocol, you may read it and change your mind about vector direction, its weight and so on.

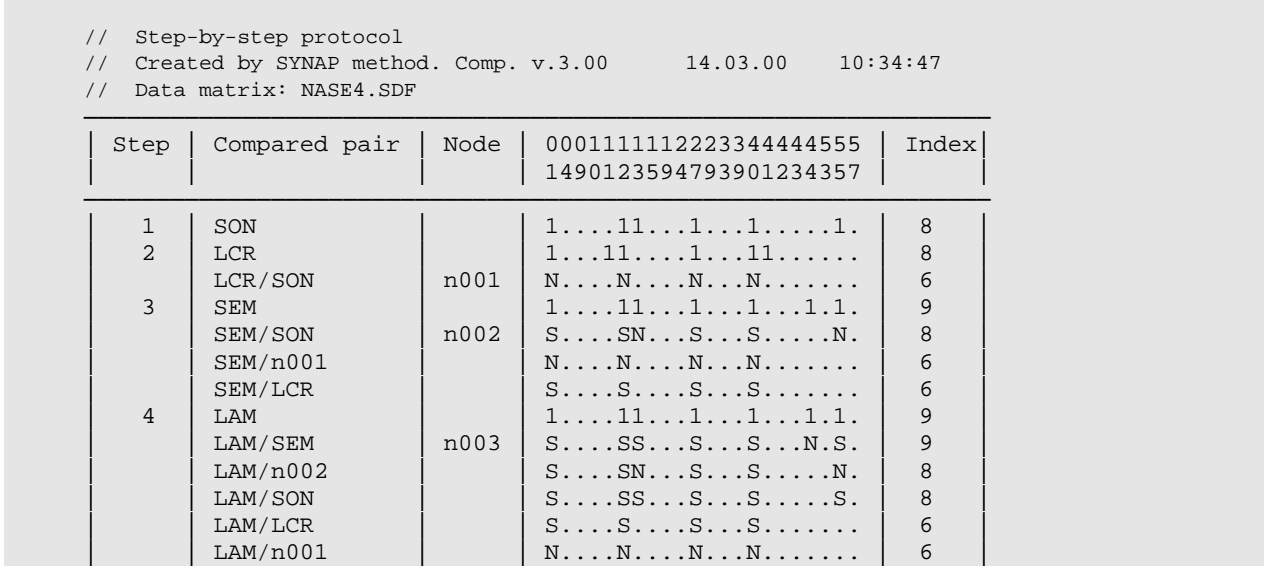


Figure 4.

All procedure is performed on comfortable graphics and interactive mode. You may mark tree fragments by textual labels of taxa and numbers of vectors (Fig. 5). Special information on parallelisms and reversions as well as some statistics and weight of characters are given too (without letter = unique, p = parallel, r = reversion). Lists of active taxa and active vectors may be placed after the scheme. The target file for tree demonstration in graphics, using program Component 2.0 for Windows is available also.

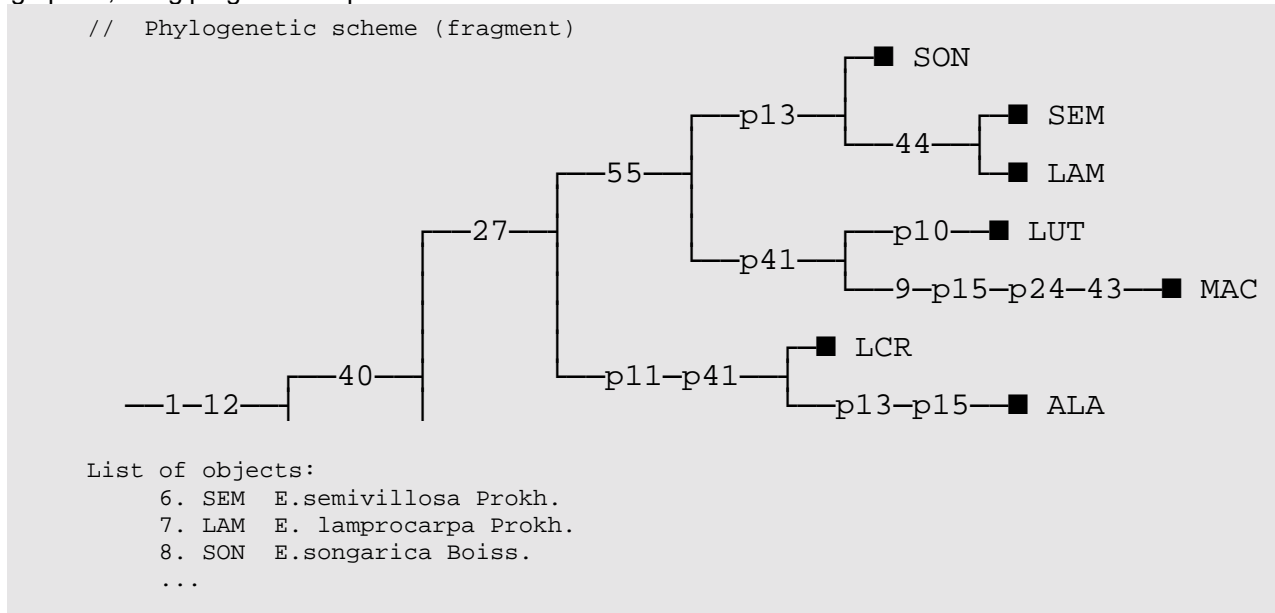


Figure 5.

Acknowledgements

The research was supported by grants of the International Science Foundation (RA7000, RA7300), Russian Foundation for Basic Researches (98-04-49459, 99-07-90222), Presidium of Siberian Branch of Russian Academy of Sciences (integrated grant IG-17 and grant for the young scientists).

References

1. Baikov K.S. (1996): SYNAP - A New Algorithm for Phylogenetic Reconstruction. *Journal of General Biology*, **57**, 2: 165-176. (in Russian).
2. Baikov K.S. (1997): Reconstruction of Phylogenesis as a Method of Studying and Conserving the Biodiversity. *Siberian Journal of Ecology*, **1**: 29-33.
3. Baikov K.S. (1999a): Logical modeling of phylogeny: problems and prospects. *Trudy ZIN RAS*, **278**: 45-47 (Russian), 48-52 (English).
4. Baikov K.S. (1999b): Basic modeling of phylogeny using method SYNAP. Novosibirsk. (in Russian)
5. Camin J.H., Sokal R.R. (1965): A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 3: 311-326.
6. Farris J.S. (1988): Hennig86 reference. Version 1.5. New York.
7. Hennig W. (1966): Phylogenetic systematics. Urbana: Univ. Illinois Press.
8. Li C. (1990): A new method for cladistic analysis - Median Elimination Series (MES). *Acta Phytotax. Sinica*, **28**, 1: 34-53.
9. Swofford D.L. (1985): PAUP. Phylogenetic Analysis Using Parsimony, version 2.4.0. Illinois.
10. Wagner W.H.Jr. (1961): Problems in classification of ferns. *In: Recent advances in botany*. Toronto: Univ. Toronto Press: 841-844.



**SECTION 4.
BIOINFORMATICS OF DNA, RNA,
AND PROTEIN STRUCTURE.
STRUCTURAL GENOMICS**

STRUCTURE AND FORMAT OF THE EnPDB DATABASE ACCUMULATING SPATIAL STRUCTURES OF DNA, RNA AND PROTEINS

*Grigorovich D.A., *Ivanisenko V.A., Kolchanov N.A.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: salix@bionet.nsc.ru

*Corresponding author

Keywords: PDB, proteins, SRS, databases

Resume

Information about spatial structure of DNA, RNA, and proteins is stored in the Protein Data Bank (Bernstein). This databank is the only worldwide known official source accumulating scientific information on the known spatial structures of macromolecules. In PDB, the information is contained on amino acid content of a protein, which is called the primary structure (represented by the consequence of symbols), information on the secondary protein structure (local packaging of polypeptide chain in the space in the form of alpha-helices and beta-threads), along with information about coordinates of atoms forming the protein. The typical protein contains up to several thousands atoms, each of them being supplied by coordinates X,Y,Z. An access to this information for non-commercial organizations engaged in fundamental research is free and may be provided via the Internet. However, it should be noted that at the moment of PDB development, the number of deciphered structures was not too large. With this respect, the task of an automated search and access to information on the spatial structures was not so topical. The developers of PDB made an accent to completeness of representation of structural data, methods, and conditions of experiments, in which these data were obtained. Additionally, the functional peculiarities of macromolecules are described. The structure of the base itself was chosen in an optimal way for the primary data accumulation and took into account the possibility of adding the novel information. For example, the information for different macromolecules in PDB is divided into separate files. Besides, there is a seria of reserve fields, which were supposed to provide "evolution" of the database, following the evolution of experimental and theoretical technologies for extraction of novel knowledge.

During the several recent years, a qualitative sharp progress has occurred in technique of synthesis, isolation, and crystallization of biological macromolecules. As a result, the number of deciphered structures has increased significantly. Currently, the number of such structures attained more than 10000 and constantly grows in number, increasing more than twice each year. To provide an effective search in such huge bulk of data is possible only by using modern computer technologies. Among these technologies for search and access to molecular-biological databases is the SRS. Effective application of SRS is directly depends upon the extent of formalization of the initial data. With this respect, at the first stage, the task was formulated to make a transformation of the PDB database into the EnPDB database, which has the more formalized form of data representation. This task includes the following special cases:

- classification of information stored in PDB;
- extraction of valuable significant information;
- development of an algorithm for transformation;
- realization of the algorithm in the form of the program-converter.

Besides, the additional tasks were set to create the links between the EnPDB database with the other databases and to enable addition of the novel information, which will expand the possibilities of a search.

An example of the EnPDB database entry performed by the SRS technique is shown in Fig. 1. Fig. 2 exemplifies the variant of the query, which could be addressed to the EnPDB database. In the Fig.3, an information is given on the fields (indices) of the EnPDB database. The EnPDB database is available via the Internet by the address: <http://srs5.bionet.nsc.ru/srs5bin/cgi-bin/wgetz?-fun+Pagelibinfo+-info+ENPDB>

Taking into account contemporary volumes of information stored in the PDB, the calculation of a single characteristic out of the list given above for all the proteins from PDB is an extremely difficult time-consuming task requiring many hours, days, or even weeks of calculations by quick-operating computers. Therefore, one more important task, which we consider within the frames of the given project, is to develop a seria of daughterly databases and their integration with the EnPDB. Each of daughter databases contains the information on particular peculiarities of proteins stored in the EnPDB database.

In future, we plan to develop such programs for analysis of spatial protein structures as structural alignment of protein surfaces, active sites; and to perform their integration with the EnPDB database.

SRS Query Form Page - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

JmolTitle

JmolRef

JmolVolume

JmolYear

Resolution

ChainAmount

ChainSizes

HelixAmount

SheetAmount

DnaRnaAmount

ProteinAmount

HetAmount

Heterogen

Separate multiple values by & (and), | (or), ! (and not)

Готово Местная intrасеть

Figure 1. An example of a query for the search of information in the SRS system.

QueryResult - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

ID 1A1F (RasMol, 3D Atlas)
 HEADER COMPLEX (ZINC FINGER/DNA)
 DATE 10-DEC-1997
 TITLE DSNR (ZIF268 VARIANT) ZINC FINGER-DNA COMPLEX (GACC SITE)
 COMPOUND MOL_ID: 1; MOLECULE: DSNR ZINC FINGER PEPTIDE; CHAIN: A; FRAGMENT: ZINC FINGER; BIOLOGICAL_UNIT: MONOMER; MOL_ID: 2; MOLECULE: DUPLEX OLIGONUCLEOTIDE BINDING SITE; CHAIN: B, C; ENGINEERED: YES
 SOURCE MOL_ID: 1; ORGANISM_SCIENTIFIC: MUS MUSCULUS; ORGANISM_COMMON: MOUSE; MOL_ID: 2; SYNTHETIC: YES;
 SYNTHESIS MOL_ID: 1; EXPRESSION_SYSTEM: ESCHERICHIA COLI;
 EXPRESSION_SYSTEM_STRAIN: BL21 (DE3); EXPRESSION_SYSTEM_PLASMID: PDSNR; MOL_ID: 2;
 KEYWORD COMPLEX (ZINC FINGER/DNA), ZINC FINGER, DNA-BINDING PROTEIN
 TECHNIQUE X-RAY DIFFRACTION
 AUTHOR H.ELROD-ERICKSON,T.E.BENSON,C.O.PABO
 RESOLUTION 2.1
 LINK EMBL [M19643](#) [M20157](#) [M22326](#) [M28844](#) [M28845](#)
 LINK_PIR [A29883](#) [A32065](#) [J50304](#)
 LINK_SWISS-PROT [EGR1_MOUSE](#)
 LINK_TRANSFAC [T00244](#) [T00455](#)
 LINK_TRRD4 [EGR1](#)

Готово Местная intrасеть

Figure 2. An example of an entry 1A1F of the EnPDB database in the SRS system.

Data-fields in SRS

Name	Short Name	Type	No of Keys	No of Entry References	Indexing Date	Status
<u>ID</u>	id	ID	4280	4280	7/30/99	ok
<u>Header</u>	hdr	string	723	8494	7/30/99	ok
<u>Date</u>	dte	int	1009	4280	7/30/99	ok
<u>Title</u>	ttl	string	5103	35052	7/30/99	ok
<u>Compound</u>	cpd	show				not indexed
<u>Molecule</u>	mol	string	2650	10672	7/30/99	ok
<u>Synonym</u>	smm	string	1442	3778	7/30/99	ok
<u>EC</u>	ec	string	388	1835	7/30/99	ok
<u>BioUnit</u>	bun	string	126	941	7/30/99	ok
<u>Gene</u>	gen	show				not indexed
<u>MolSource</u>	nls	string	3	4280	7/30/99	ok
<u>Source</u>	src	string	3281	21770	7/30/99	ok
<u>Synthesis</u>	snt	string	1785	13189	7/30/99	ok
<u>Keyword</u>	kw	string	2891	29035	7/30/99	ok
<u>Technique</u>	tch	string	22	9090	7/30/99	ok
<u>Author</u>	aut	string	4668	16316	7/30/99	ok
<u>Jrnl</u>	jrn	show				not indexed
<u>JrnlAuthor</u>	jau	string	8117	34509	7/30/99	ok
<u>JrnlTitle</u>	jti	string	6962	90638	7/30/99	ok
<u>JrnlRef</u>	jre	string	362	8885	7/30/99	ok
<u>JrnlVolume</u>	jvo	string	391	8168	7/30/99	ok
<u>JrnlYear</u>	jye	int	32	7362	7/30/99	ok
<u>Remark_1</u>	jrn	show				not indexed
<u>Resolution</u>	res	real	174	3461	7/30/99	ok
<u>ChainAmount</u>	cha	int	22	4277	7/30/99	ok
<u>ChainSizes</u>	chs	int	587	9061	7/30/99	ok
<u>HelixAmount</u>	hla	int	130	4280	7/30/99	ok
<u>SheetAmount</u>	sha	int	117	4280	7/30/99	ok
<u>DnaRnaAmount</u>	dra	int	8	4277	7/30/99	ok
<u>ProteinAmount</u>	pra	int	23	4277	7/30/99	ok
<u>HetAmount</u>	hta	int	46	4280	7/30/99	ok
<u>Heterogen</u>	htg	string	2317	15613	7/30/99	ok

Figure 3. The list of fields (indices) in the EnPDB database. For indices, the number of various values and the total number of values are indicated. Some other internal service information is given.

The work was supported by the Russian Foundation for Basic Research (grants Nos 98-07-91078, 00-04-49252).

MODEL OF PCR KINETICS

Titov I.I.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: titov@bionet.nsc.ru

Keywords: polymerase chain reaction, kinetics, renaturation, computer simulation

Resume

Motivation:

Despite the fact that PCR is a powerful tool in modern biology and medicine, its quantitative model describing product accumulation at later stages of amplification is not developed yet.

Results:

Kinetics of standard amplification is treated as recurrent scheme with a variable amplification factor $H(n)$ introduced. At each cycle number n , the $H(n)$ is obtained from solving a system of equations of chemical kinetics, which takes into account the reassociation and destruction of PCR product. The amplification limit is calculated as a fixed stable point $H=1$ of the transformation. A good agreement has been found between the results obtained under the proposed model and the known experimental data. The model can be straightforwardly generalized for use in more realistic PCR computer simulation.

Introduction

PCR is a widely used technique for specific amplification of the target DNA to reasonable quantities. For successful PCR performance an optimal pair of primers is required. The procedure of selection of an appropriate set of primers is implemented in a wide range of computer packages (e.g., [1, 2]) for molecular biology. However, even good-fitted primers require optimal reaction conditions. When approaching high product concentrations, the PCR efficacy is limited by cooperative action of a number of factors (renaturation of the product, polymerase and primers deficiency, etc. [3]), providing a permanent reduction of the amplification factor $H(n)$. Its departure from the maximal value of 2 can be significant. When the value of the amplification factor is close to 1, further amplification does not provide any substantial increase in the product yield ("plateau" effect).

The Theory

Key factors of the plateau effect. Product accumulation is limited by many factors. Among them, renaturation and destruction are most significant ones, as none of the modern PCR techniques can get rid of them. The renaturation of a product inhibits primer annealing and strand synthesis. The destruction of a product is responsible for the plateau effect *per se*, otherwise the product content would have been growing infinitely, though slowly. While the product concentration, C , is high, a strand can be annealed in the process of amplification, and then this annealed strand can be cleft by Taq polymerase possessing 5'-3' exonuclease activity. The higher is the concentration of the product, the higher the rate of the process. The destruction process equilibrates the synthesis when the amplification is at its limit.

A formal description of amplification kinetics. PCR efficacy is usually characterized by some constant amplification factor [4], which in fact is the geometrical mean of the partial amplification factors (each of them describes the product accumulation at a single cycle). Instead, we introduce a variable amplification factor, H , at a cycle n in the recurrent form:

$$C_{n+1} = H(C_n) C_n \quad (1)$$

Obviously, the value of H depends on the PCR parameters and can be calculated within the frame of the scheme of chemical reactions proceeding at each cycle. Under the model being described, the H is obtained in the analytical form. A universal amplification factor allows me to describe the PCR kinetics within the entire range of product concentrations. The transformation (1) has been found to have two fixed points, the unstable ($C = 0$) and the stable ($H = 1$), the latter corresponding to the amplification limit.

The kinetic scheme of a cycle. As limiting factors of PCR, here I consider only renaturation of complementary product strands (with a second order rate constant, k_a) and destruction of renatured product strands (with a first order rate constant, k_d). The renaturation is assumed to block successful synthesis on either of reassociated strands. For a pair of complementary strands, I calculate the probabilities P_i of the following processes:

- each strand reproduces a complementary strand;
- the pair of strands is annealed and neither of them is cleft;

c) the strands are annealed, then either of them is cleft;

d) the strands are annealed, then both are cleft.

The amplification factor. It can be readily obtained by summing over a-d) contributions:

$$H = \sum_i H_i P_i, \quad (2)$$

where $H_1 = 2$, $H_2 = 1$, $H_3 = 0.5$, $H_4 = 0$. It is convenient to go to dimensionless variables $u = k_a C \tau_0$, $T = \frac{\tau_0}{\tau}$, $D = k_d \tau$. (C is the initial concentration of complementary pairs at the beginning of each cycle, τ is the duration of the phase of synthesis, τ_0 is the period of annealing-free synthesis. One can simply estimate τ_0 through the product length L and the rate of nucleotide incorporation k_i : $\tau_0 = \frac{L}{k_i}$).

Renumbering cycles by dimensionless concentration u (1), one can obtain a general expression:

$$H = \frac{2}{u+1} + \int_0^1 dx u \frac{e^{D(Tx-1)}}{(1+ux)^2}. \quad (3)$$

As can be seen from Eq. 3, in case the product concentration is small (the "dilute limit"), $H = 2$ and it grows exponentially. This increase entails the monotonous decrease of the amplification factor that finally approaches the value $H=1$.

It is instructive to consider the low T limit ("rapid" synthesis), where the amplification factor is expressed algebraically:

$$H = \frac{2}{u+1} + \frac{ue^{-D}}{u+1}. \quad (4)$$

Comparing with experimental data.

As it is difficult to obtain a reliable value of the amplification factor in the dilute limit (small product concentrations can hardly be detected), I set it equal to 2 throughout. It could be estimated from the dependence of the cycle number required for a given product yield on the initial concentration. However, such a "delay" experiment would produce a very rough estimate of the amplification factor. In particular, a non-specific amplification increases the delay. In [4], the decrease of the initial copy number by 3 orders of magnitude has resulted in a delay of 10 cycles, which corresponds to $H=1.995$. In this process, the non-specific product was not detected, what was the case while further decreasing of the initial product content. Meanwhile, the regression of dependence [4] within the whole range of initial product concentration corresponds to the apparent amplification factor 2.15.

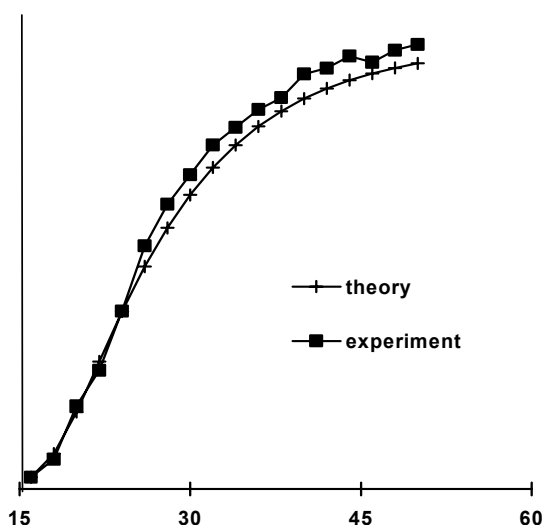


Figure 1. The product accumulation versus cycle number: the theory and the experiment of Higuchi et al [4] with polymerase excess. The experimental fluorescence curves showing no dependence on polymerase concentration (7.5U and 10U) were averaged and then the background fluorescence was subtracted from. The calculations were carried out with Eqs. (1, 4). Note that the only adjustable parameter was the specific fluorescence of a product.

To test the model, I have used the data on amplification kinetics published in [4, 5]. First, since synthesis temperature has a dramatic impact on the rate of nucleotide incorporation k_i [6] and hence on T , I have

estimated the T values by interpolating the known temperature dependence of k_i [6]. The values have been found to be low ($T \approx 0.08$ for kinetics given in [4] and $T \approx 0.03$ - in [5]). Secondly, using the approximation (Eq. (4)), I have reconstructed the amplification kinetics. A good agreement has been found (Fig.1) between calculations performed by my model and the experimental amplification kinetics [4] with excess of polymerase.

Using Eqs. (1, 4) and processing the data [4] and [5], I have estimated the destruction rate as $k_d \approx 10^{-2}$ and $k_d \approx 4 \cdot 10^{-3}$ (sec molecule) $^{-1}$, correspondingly. The difference of the estimates is likely due to the non-excess polymerase concentration (2.5U) of [5].

Discussion

The main limitation of my model is the assumption on the ultimate specificity of the primer binding. In other words, it is assumed that some computer tools [1, 2] already optimize the primers. In addition, this simple theory considers neither primers nor polymerase deficiency nor its aging. (The latter decreases the amplification limit and results in apparent reduction of the product yield after plateau.) However, these factors, together with a given distribution of non-specific primer binding sites and other PCR parameters (K and Mg ions concentrations, etc.), can be incorporated into the model by direct numerical solving of phenomenological equations of chemical kinetics at each cycle. While losing speed of analytical solutions (3, 4), this brutal approach would make it possible to simulate more realistically not only standard PCR dynamics but, e.g., asymmetric amplification. In doing so, one can optimize the PCR by varying the initial parameter set. Such computer system for PCR dynamics simulation may be a useful addition to any computer tool searching for optimal primers.

Acknowledgments

The stimulating discussions with M.I.Voevoda proved to be very useful.

References

1. W. Rychlik and R.E. Rhoads (1989) *Nucl. Acids Res.* **17**: 8543-8551.
2. L.S. Davidow (1992) Selection of PCR designed mismatch primers. *CABIOS* **8**: 193.
3. A.D. Sardelli (1993) Plateau effect - understanding PCR limitations. *Amplifications: A Forum for PCR Users* **9**: 1-5.
4. R. Higuchi, C. Fockler, G. Dollinger and R. Watson (1993) Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *BioTechnology* **11**: 1026-1030.
5. W. Rychlik, W.J. Spencer and R.E. Rhoads (1990) Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucl. Acids Res.* **18**: 6409-6412.
6. A. Landgraf and H. Wolfes (1993) Taq Polymerase. In: *Methods in Molecular Biology* v.16: Enzymes in Molecular Biology, ed. M.M. Burnell. Humana Press Inc. Totowa, NY, pp.31-58.

INFORMATION SYSTEM 'HIV VACCINE DEVELOPMENT'

**Belova O.E., Bazhan S.I.*

State Research Center of Virology and Biotechnology «Vector», Koltsovo, Russia

e-mail: belova@vector.nsc.ru

*Corresponding author

Keywords: information system, hypertext, Internet, vaccine, HIV-1

Resume

Motivation:

To date there have been many works concerning design of vaccine; but there are still many problems in vaccine development. We have attempted to provide a guide to immunobiology-driven HIV vaccine elaboration - the information system 'HIV Vaccine Development'.

Results:

The pilot version of information system is presented. The system integrates information about the current approaches to the development of HIV vaccines. It gives also the reviews of different aspects of HIV molecular biology, virology and immunology. The system can be used also as an educational one.

Availability:

The system is available via Internet at <http://www.vector.nsc.ru/hivvac>.

Introduction

Considerable effort is now being focused on evaluating the vaccine approaches for preventing HIV infection. A variety of vaccine constructs and strategies have been explored. Every known strategy for making such a vaccine has been evaluated, but even the most promising candidate formulations to emerge are at best years from human use. Despite over a decade of intensive efforts, an HIV vaccine remains elusive.

Although the creation of new vaccines is never easy, the development of a vaccine to prevent HIV vaccine, or even for that matter, to delay or temper the devastating impact of AIDS has proven especially difficult. A major hurdle of vaccine development remains the astonishing variability of the virus, especially for the envelope sequence. Further, there is still uncertainty about how HIV is transmitted and what kinds of immunity (cellular or humoral responses) are most effective against the virus. And finally, there is the facet that more than any single factor hinders progress toward a vaccine: an absence of suitable animal model for HIV disease.

So the task is to take advantages of what has been learned about design of vaccines to try to find actually safe and effective forms of vaccines, prospective systems of vaccine delivery and optimal schedules of immunization. The developing of information system, integrating knowledge about HIV vaccine design, construction and trials, and data on HIV biology, virology and immunopathology can help in solving this actual task.

Methods

Modern databases of biological information have become a valuable tool for study. A list of databases and an amount of included information rapidly increases [1-3]. The development of databases is of special productive in those domains of knowledge where the data may be easily formalized. Bibliographic databases PubMed (National Library of Medicine, NIH, USA), DNA Vaccine Web-site (Academy of Sciences, USA) or available on Internet genetic databases of protein and amino acid sequences, such as EMBL and SWISSPROT, are examples.

It can be said with assurance that most of the data published in literature can not be formalized and because of this are not accumulated in databases. This fact results from the dynamical growth of knowledge and as a consequence it is difficult to describe them in any previously designed rigorous system suitable for computer processing.

One of the perspective lines of databases development is connected with creation of specialized (i.e. subject-oriented) information systems with different modes of data presentation. Last time such databases came to be developed in different areas of biology. They offer to integrate not only formalized but also nonformalized information. However, their advancement is retarded because of the necessity of previous analysis, extension and systematization of data in some specialized area. Database TRRD containing data on transcription regulation of genes providing a control of a row of concrete physiological systems (interferon-inducible genes,

genes of lipid metabolism, erythroid-specific genes [4-6]) is an example of such subject-oriented database. Another example is ImMunoGeneTics database [7], specializing on immunoglobulins.

Notice that known HIV Molecular Immunology Database (Los Alamos National Laboratory, USA) and AIDS Vaccine Web-site (USA, National Institute of Allergy and Infection Diseases) do not provide the desired analytical information devoted to analysis of problems and ways of HIV vaccine elaboration.

To elaborate the information system 'HIV Vaccine Development', we rejected formalization of the traditional technique and selected the mode that can be characterized by the two main concepts of hierarchical thesaurus and computer encyclopedia [8].

The system is made on the basis of hypertext technology. The text processing automation is connected in last time with hypertext technology which is actively developed now [9]. Such technology allows to structure the big volume of text information, to link the received structural elements to constructions which reflect the text semantics and to realize procedures of access to need user information on the base of existing links.

Hypertext principle of database development in combination with exceptional potentialities inherent in computer sorting of text comes to light new possibilities of creation of new technology of data storage, exploit and integration.

Some important peculiarities of hypertext organization of databases essentially reduce the work load on elaboration of databases and information systems [10]. In particular, this technology does not require a severe data structuring, so it is possible to include in database conceptual and experimental information, which can not be formalized in present time. Thus hypertext is used for integration and structuring of huge number of different information present in literature and computer databases and nets. In the same time the rigorously structured information suitable for searching procedures can be also organized as a hypertext.

Implementation and results

The work on the system 'HIV Vaccine Development' was began in 1998-1999 years as one of the projects of Russian program 'Vaccines of new generation and medical diagnostic systems of the future'. First version of informational-analytical system 'HIV Vaccine Development' is available via Internet at <http://www.vector.nsc.ru/hivvac/>.

The purpose of information system elaboration – integration of contemporary knowledge in HIV immunology and vaccinology. System may be used for solving tasks connected with HIV vaccines design, constructing and trial, in particular, with design of new synthetic immunogenes, candidates HIV vaccine, stimulating T- and B-cellular responses to HIV-1. Furthermore, the systems gives a overview of different aspects of HIV molecular biology, virology and immunology and may be used as educational system. The expected users of system – investigators in area of vaccine construction, students.

The system includes data on different aspects of HIV molecular biology, virology, immunology and immunopathology and besides, has access to the similar international databases via context links – Internet's addresses.

The hypertext system consists of abstracts describing the key notions combined by means of context notes that supplement one another. These could be called 'structured analytical reviews'. System structure can include tree-like hierarchical structures, tables and figures. The records of all sections are embedded to the hypertext links and thus it makes navigation easy and quickly. System is also designed to store bibliographic references.

To facilitate the users' work the system is designed by analogy with structure of printed issues. The detailed content is the main entrance into the system. Entering in necessary section user exerts navigation on nets of interconnected records (hypertext net) selecting necessary information.

The system is augmented with hypertext reviews, prepared in system of Web-sites generation (Front Page Express). HTML format was used as a basic, because of Internet is the main mean of access to information system. The figures are present as GIF-images (in GIF formats).

The information-analytical system contains the divisions on HIV molecular biology, virology, immunology and immunopathology, on known strategies of vaccine application, design of artificial T-and B-cellular immunogens, candidate HIV vaccine. Due to the lack of space, we have to skip some material and give further only a brief description of sections. However, the complete details are available at <http://www.vector.nsc.ru/hivvac/>.

Molecular biology. This section contains the description of structure-functional organization of HIV-1 genome, characteristics of structural and regulatory genome elements, mRNA and protein processing, models of regulation of HIV-1 transcription and replication, cellular and viral factors which take part in these processes, etc.

Immunobiology. The section contains the description of T- and B-cellular immunity and stimulation of humoral and cellular immune responses under HIV-infection. HIV strategies permitting to avoid humoral response are considered.

Models of HIV immunobiology and immunopathology. The data on immunopatogenesis of HIV infection and the proposed mechanisms of HIV inducible immune disorder are described.

Development of HIV vaccines. This part reviews the biology and specifications of vaccines, the various technologies used to make antibody and cellular response vaccines, and reasons why the making of an effective and safe HIV vaccine has proven so difficult.

Design of synthetic HIV vaccines. The topic of this section is approaches to enhance or to potentiate the immune response to vaccines. The section includes description and analysis of strategies for construction of an effective new generation multivalent vaccines. Special attention was paid to design of polyepitope CTL vaccines, to criteria of cellular epitope selection and to technologies of epitopes assembly in one polypeptide. The questions of vaccine delivery are considered.

Testing of HIV vaccine. In this section we examine the demonstrated or potential performance of candidate HIV vaccines and strategies developed today in the context of the contemporary specifications for an AIDS vaccine. Methods of elucidating of immunobiological properties of vaccines, side effects, the efficiency of immunization are discussed.

The text is written in English. The data were taken from scientific publications and reviews and from the experience on vaccine study and construction which have been accumulated in SRC VB 'Vector'.

Discussion

We have attempted to present and analyze the processes, as well as the products, of contemporary HIV vaccine research.

Further we plan to elaborate the information system in two directions. The first is connected with update and supplement of analytical hypertext information. The second direction- the addition of complex of programs for searching, analysis and visualization of HIV-1 protein and amino acids sequences. The program complex will allow to select and analyze the optimal CTL-epitopes from basic HIV-1 antigens, perspective for construction of synthetic immunogens, candidates HIV-1 vaccine.

References

1. Wallace, J.C. and Henikoff, S. PATMAT: a searching and extraction program for sequence pattern and block queries and databases. *Comput. Appl. Biosci.* 8, 249-254 (1992).
2. Bairoch, A. The Enzyme data bank. *Nucleic Acids Res.* 22, 17, 3626-3627 (1994).
3. Holm, L. and Sander, C. Parser for protein folding units. *Proteins.* 19, 256-268 (1994).
4. Anan'ko, E.A., Bazhan, S.I., Belova, O.E. and Kel, A.E. Mechanisms of transcription of the interferon-induced genes: a description in the IIG-TRRD information system. *Mol. Biol.*, 31, 592-605 (1997).
5. Ignat'eva, E.V., Merkulova, T.I., Vishnevskii, O.E. and Kel, A.E. Regulation of transcription of genes of lipid metabolism: a description in the TRRD database. *Mol. Biol.* 31, 4, 575-591 (1997).
6. Podkolodnaya, O.A. and Stepanenko, I.L. Mechanisms of transcription regulation of the erythroid-specific genes. *Mol. Biol.* 31, 4, 562-574 (1997).
7. Lefranc, M.-P. IMGT Locus on Focus. *Exp. Clin. Immunogenet.* 15, 1-7 (1998).
8. Alexandrov, A.A. Molecular-genetic database and software development. In *Ratner V.A. and Kolchanov N.A. (eds.), Reports of International Conference 'Modeling and Computer Methods in Molecular Biology and Genetics'*. Nova, New York. (1992).
9. Nielsen, J. *Hypertext and Hypermedia*. WA: Academic Press (1990).

RNA-POLYMERASE – PROMOTER RECOGNITION. SPECIFIC FEATURES OF ELECTROSTATIC POTENTIAL OF “EARLY” T4 PHAGE DNA PROMOTERS

^{1*}*Dzhelyadin T.R.*, ¹*Sorokin A.A.*, ¹*Ivanova N.N.*, ¹*Sivozhelezov V.S.*,
¹*Kamzolova S.G.*, ²*Polozov R.V.*

¹Institute of Cell Biophysics, Pushchino, Russia

²Institute of Theoretical and Experimental Biophysics, Pushchino, Russia

e-mail: dzhelyadin@pbc.iteb.serpukhov.su

*Corresponding author

Keywords: electrostatic potential, RNA-polymerase, T4 phage DNA promoter, recognition

Introduction

RNA-polymerase (RNAP) performing transcription of genetic information is able to promptly and precisely locate promoter sites of DNA. This protein consists of α , β , β' , and σ subunits, and recognizes a specific, about 100 nucleotide-long region responsible for DNA-RNAP binding and complex formation. A question arises on what determines RNAP affinity to a particular fragment of DNA molecule. Analysis of the nucleotide sequences of promoter sites allowed two canonical consensus hexanucleotides to be identified and some non-canonical elements to be proposed for specific promoter groups. However, these sequence peculiarities of promoter regions did not allow to explain the high specificity of RNAP interaction with its numerous promoters (Liebing H.D. and Ruger W, 1989). Some other determinants should be involved in the process of RNAP–promoter recognition. The possible answer may lie in physical properties of promoter DNA regions determined by their sequences.

For both nonspecific protein binding to DNA molecule and the process of recognition of specific binding sites by their proteins, electrostatic interactions play an important role caused by polyelectrolyte nature of DNA. Earlier we have developed calculation techniques allowing to investigate the character of electrostatic potential distribution around long DNA fragments (Polozov R.V. et al., 1999) and proteins (Fedoseyev A.I. et al., 1992).

This work aims to identify electrostatic determinants in promoters.

Electrostatic potentials were determined for double-stranded regions of DNA with nucleotide sequences of “early” phage T4 DNA promoters. The latter were selected because their biochemical properties are well studied. Particularly, quantitative data were published on the influence of ADP-ribosylation of RNAP α -subunit on the strength of these promoters (Kosh T. et al. 1995). Electrostatic potentials of the C-terminal domain of α -subunit were also calculated, both for the native and the ADP-ribose-modified protein.

Methods

A full-atom model of DNA molecule was used with atom coordinates of base pairs taken from Landolt-Bornstein 1989, and atomic charges from Zhurkin V.B. et al. 1980. Dependence of the geometry parameters of DNA helix on its nucleotide sequence (Ponomarenko M.P. et al. 1997) was taken into account in the model. Atom coordinates of the C-terminal domain structure of α -subunit were taken from the PDB databank (Jeon Y.H. et al. 1995). That structure, as well as the same structure with ADP-ribose covalently attached, was optimized using the package AMBER.

Electrostatic potential around α -subunits, both native and ribosylated, was calculated by solving the Poisson-Boltzmann equation using the method described in Fedoseyev A.I. et al. 1992, and depicted on the points belonging to the surface 5.5Å away from the van der Waals surface of the native subunit. The points corresponding to positive potential values were colored light gray, negative black, and within 0.5kT/q of zero, white.

Electrostatic potential around DNA helix was calculated as described in Polozov R.V. et al. 1999. For further analysis, the two-dimensional potential distribution was averaged using a dynamic window 31Å wide along the helix axis and fully covering the cylinder circle.

Results and Discussion

ADP-ribose covalently binds to Arg-265 of RNAP (Fig.1) (Kosh T. et al. 1995). Since ADP-ribose contains two phosphate groups, it markedly influences the electrostatic potential distribution around the α -subunit as seen from Fig.2. The potential distribution around the native α -subunit shows that the positive-potential region is

located in the left-hand part of Fig.2a and corresponds to location of the Arg-265 (Fig.1a), which is the target of ADP-ribosylation (Fig.1b). For the ribosylated α -subunit, this region is mostly occupied by negative or neutral

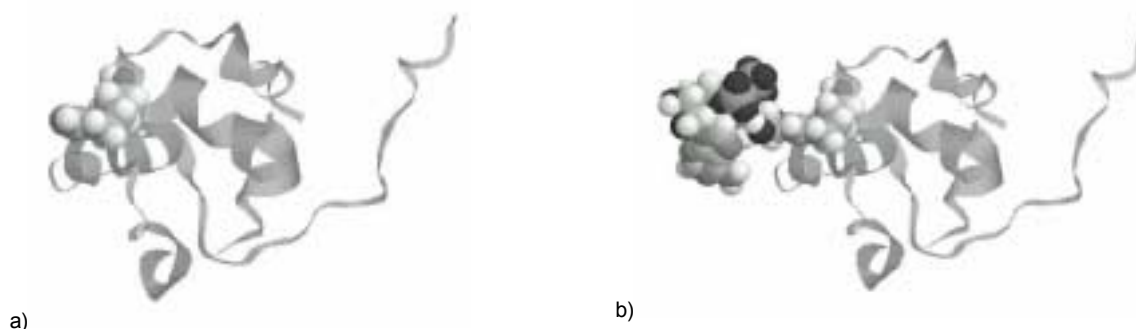


Figure 1. Three dimensional CTD structure of the α -subunit of RNA polymerase of *E.coli* : a) native, b) ADP-ribosylated. The Arg-265 residue, native or modified, is shown as van der Waals spheres, the rest of protein as ribbons.

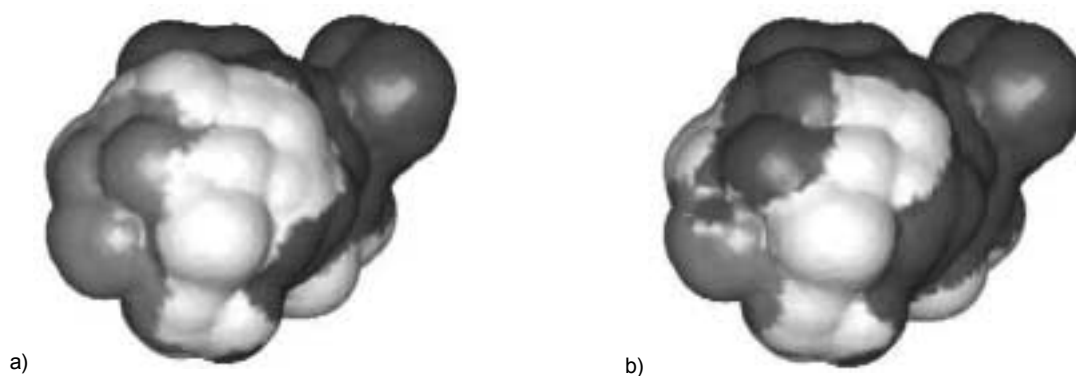


Figure 2. Distribution of electrostatic potential around the α -subunit of RNA polymerase of *E.coli*: a) native, b) ADP-ribosylated.

potential (Fig.2b).

One can thus suggest that the specific features in electrostatic potential distribution will manifest for those promoters whose strength is essentially affected by ADP-ribosylation of RNAP α -subunit.

It is known that "early" phage T4 promoters may be divided into three groups by their functional response to ADP-ribosylation of the α -subunit: the first group is not affected by the modification, the second interacts better with the modified protein, and the third is more active with the native protein (Kosh T. et al. 1995).

The promoter P164.5 which is the most active with the native protein has two wide negative-potential areas circularly surrounding the DNA in the left-hand part of the promoter region. In Fig.3d, this corresponds to two gorges in the 1100Å and 1250Å regions along the Z axis.

The promoters P161.1 and P54.4 (Figs.3a and 3b), whose activity increases by RNAP ADP-ribosylation, have positive peaks at 1100Å, also circularly surrounding the DNA, and a circle of negative potential in the 1250Å region. The presence of the positive peak in electrostatic profiles of these promoters can explain stronger affinity of those regions to the modified protein carrying the additional negative charge.

It is known that the affinity of the promoter P50 remains unchanged upon ADP-ribosylation of the α -subunit. Thus, as could be expected, this promoter does not have noticeable specific features in either 1100Å or 1250Å regions along the Z axis (Fig.3c).

The data obtained indicate that there is good correlation between the patterns of electrostatic potential distribution in promoter regions of T4 DNA and promoter strength variation caused by ADP-ribosylation of RNAP α -subunit.

Thus, investigation of the electrostatic potential around DNA may identify characteristic features of the electrostatic potential possibly determining functional properties of the genome.

This work was supported by the Russian Foundation for Fundamental Research (grant 99-04-48177).

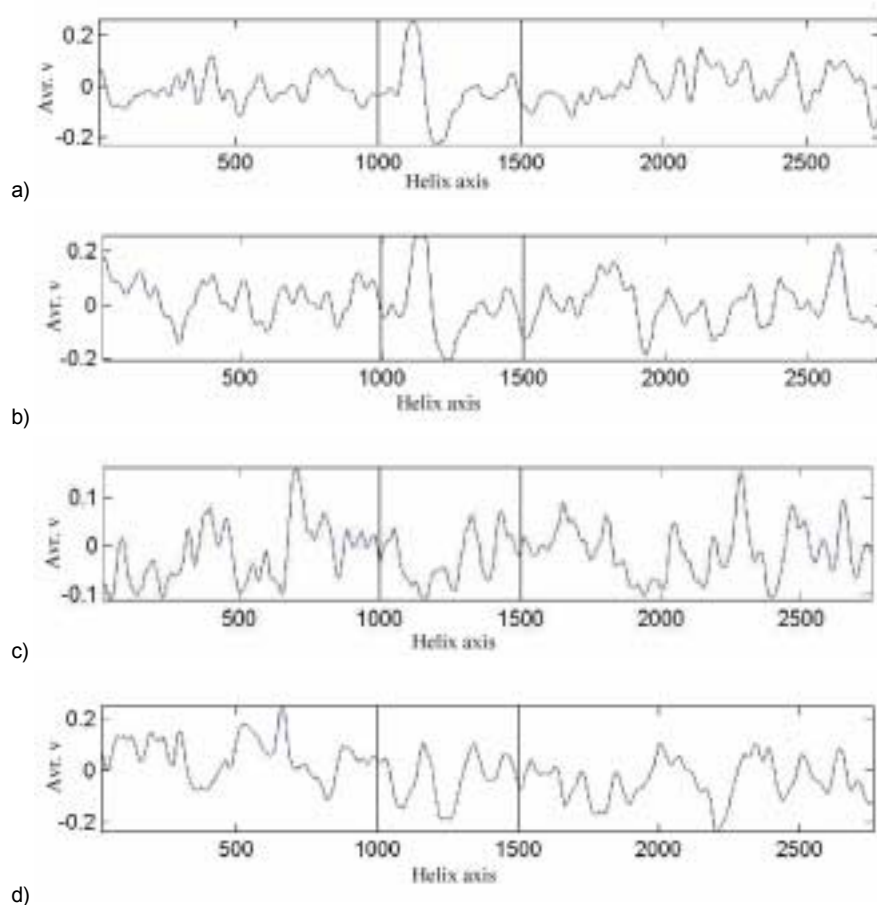


Figure 3. Distribution of electrostatic potential along the double-helix axis of DNA containing promoters: a) P161.1 promoter; b) P54.4 promoter; c) P50.0 promoter; d) P164.5 promoter.

References

1. Fedoseyev,A.I., Lazarev,P.I., Sivozhelezov,V.S., Chernyaev,E.V., Petrenko,I.I., Purtov,S.V., (1992) Mathematical modelling of 3D protein molecule potential in nonlinear media. In: *Proceedings of the "Physique en Herbe" Congress, Marseille, July*.
2. Jeon,Y.H., Negishi,T., Shirakawa,M., Yamazaki,T., Fujita,N., Ishihama,A., Kyogoku,Y., (1995) Solution structure of the activator contact domain of the RNA-polymerase alpha subunit, *Science*, **270** (5241):1495-1497.
3. Kosh,T., Raudonikiene,A., Wilkens,K., Ruger,W., (1995) Overexpression, purification, and characterization of the ADP-ribosyltransferase (gpAlt) of bacteriophage T4: ADP-rybosylation of E.coli RNA-polymerase modulates T4 "early" transcription. *Gene Expression*, **4**, 253-264.
4. Landolt-Bornstein (1989) Numerical Data and Functional Relationships in Science and Technology, Ed. W. Saenger, *New Series VII/1b, Berlin, Springer-Verlag*.
5. Liebing,H.D. and Ruger,W., (1989) Transcription from early promoters of T4 phage DNA, *J.MolBiol*, **208**,517-536.
6. Polozov,R.V., Dzhelyadin,T.R., Sorokin,A.A., Ivanova,I.I., Sivozhelezov,V.S., Kamzolova,S.G., (1999) Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences. *J. Biomolec. Struct.Dyn.*, **16**(6), 1135-1143.
7. Ponomarenko,M.P., Ponomarenko,Iu.V., Kel',A.E., Kolchanov,N.A., Karas,H., Wingender,E., Sklenar,H., (1997) Computer analysis of conformational features of the eukaryotic TATA-box DNA promoters., *Mol. Biol. (Mosk)*, **31** (4),733-740.
8. Zhurkin,V.B., Poltev,V.I., Florent'ev,V.L., (1980) Atom-atomic potential functions for conformational calculations of nucleic acids. *Mol Biol (Mosk)* **14**(5), 1116-1130.

STRUCTURE-BASED TARGET PREDICTION OF TRANSCRIPTION FACTORS

*¹Sarai A., ¹Selvaraj S., ¹Prabakaran P., ²Kono H.

¹RIKEN Tsukuba Institute, Tsukuba, Japan

²Department of Chemistry, University of Pennsylvania, Philadelphia, USA

e-mail: sarai@rtc.riken.go.jp

*Corresponding author

Keywords: target prediction, transcription factors, protein-DNA complex

Resume

Motivation:

Genome projects have produced complete genome sequences of more than 30 organisms. These data have opened up the possibility of systematic analysis of gene regulation at the genome level. Also, structural data of protein-DNA complexes have been increasing rapidly. Here, we attempt to utilize the structural information for predicting target sites of transcription factors.

Methods:

We made a statistical analysis of structural database of protein-DNA complex, and derived empirical potential functions for the specific interactions between bases and amino acids. Then, we used a sequence-structure threading to find target sites of transcription factors in real genome sequences and to examine the relationship between structure and specificity in protein-DNA recognition.

Results:

The statistical potentials showed different characteristics for different combinations of bases and amino acids, indicating that these potentials can be used to quantify specificity in the base-amino acid interactions. We calculated Z-score for a given protein-DNA complex against random sequences by the sequence-structure threading. The application to several transcription factors showed that the specificity represented by the Z-score is quite high for some transcription factors, and that the energy potentials are sufficiently sensitive to detect the target sites of some transcription factors. We also demonstrated the ability of the present method in quantifying the effects of cooperative binding and conformational changes of protein and/or DNA and distinction between cognate and non-cognate binding. We have also derived the contact potential by using computer simulation of base-amino acid interactions, and found that the calculated potentials agree well with the statistical potentials.

Introduction

Regulation of gene expression in higher organisms is achieved by a complex network of transcription factors and their target genes. Explosive amount of sequence information of genes and transcription factors from genome analyses are presenting a great challenge to bioinformatics. Finding target genes for transcription factors at the genome level will lay a basis for the analysis of gene regulatory network. Sequences contain rich amount of biological information, and comparison of known binding sequences is among the most commonly used methods for the target prediction. However, the accuracy of this method is rather limited, dependent on the quality of sequence information used. Furthermore, because transcription factors usually bind to multiple target sequences and regulate multiple genes in a cooperative manner, the target prediction is a rather complicated problem. In order to tackle this problem, we need to utilize as much information as possible. Structural data on protein-DNA complexes contain valuable functional information as well. Due to the progress of X-ray crystallography and NMR spectroscopy techniques, structural data on the protein-DNA complexes have been rapidly increasing and more than 300 complexes have been registered in the Protein Data Bank (PDB). Here, we show that these structural data can be used for the target prediction. As the structural information is independent of sequence information, it can complement the sequence-based approach.

Methods

We extracted interacting pairs of bases and amino acids from a refined set of non-redundant protein-DNA complexes (Kono and Sarai, 1999). The distant-dependent statistical potentials for the specific base-amino acid interactions were derived from the spatial distributions of C α atoms of amino acids around a base. The potential function for each pairs of base and amino acid in a particular protein-DNA complex was summed to derive a total potential energy. By threading a set of random DNA sequences onto the template structure, we calculated

the Z-score of the specific sequences against the random sequences, which represent the specificity of the complex. The threading procedure was also applied to the real genome sequences in order to find potential target sites.

For the derivation of the potential by computer simulation, we generated C α position and conformations of amino acid and calculated interaction energies (Pichierri *et al.*, 1999). Conformational ensemble was Boltzmann averaged to derive interaction free energy, and free-energy contour maps were generated for the distribution of C α position around base. For the conformational sampling, we used exhaustive sampling, canonical and multicanonical Monte Carlo sampling methods (Pichierri *et al.*, 1999; Sayano *et al.*, 2000).

Results and Discussion

We evaluated Z-score by calculating energy against 50,000 random DNA sequences, and the average Z-score calculated for 9 different protein-DNA frameworks was -3.1, which means that there are potentially one or two DNA sequences that are better fit to the framework among 1,000 random sequences (Kono and Sarai, 1999). The threading procedure was used to find target sites of transcription factors in real genome sequences. As an example of such applications, we could identify the experimentally-verified binding sites of the transcription factor MATa1/ α 2 in the promoter of *HO* gene successfully.

We also applied this method to examine the relationship between structure and specificity in protein-DNA recognition quantitatively (Kono and Sarai, 1999; Sarai and Kono, 2000). We tested the method by calculating the scores for cognate and non-cognate complexes registered in PDB. The method could distinguish the two structures as a difference in the Z-scores. Thus, the subtle differences in specificity of these structures could be detected by the analysis of energy potentials. The effect of cooperative binding was examined by comparing the monomer and heterodimer complexes of MATa1/ α 2 transcription factors. We found that the heterodimer binding enhance the specificity in a non-additive manner. This result indicates that the conformational changes introduced by the heterodimer binding play an important role in enhancing the specificity. We have also examined the effects of DNA deformation, and showed that these structural effects have significant consequences to the specificity.

The accuracy of the structure-based method in the target prediction is still limited because of the limited number of available structural data. In order to complement the method, we have used computer simulations to derive contact potential between bases and amino acids. The interaction free-energy maps derived from the calculations for different pairs of base and amino acid have shown different specificity, and agreed with the statistical potentials derived from the structural data. We are trying to estimate these data for all the combination of bases and amino acids and, together with the statistical potentials, apply to the prediction of target sequences.

The increase in the structural data, together with the information from computer simulation, will make the structure-based method promising for the accurate target prediction of transcription factors. This method can also be applied to proteins of unknown structure having substantial sequence similarity to known proteins, on the basis of which structures can be modeled and binding sites can be predicted. Because the structure-based method is independent of the other methods relying on sequence information, it can complement those methods. Thus, the combination of these methods would enable one to predict target sites and genes of transcription factors more accurately, and contribute significantly to the functional genomics in the era of post-genome science.

References

1. Kono,H. and Sarai,A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* 35, 114-131.
2. Sarai,A. and Kono,H. (2000) DNA-Protein Interactions: Target Predictions. In *Handbook of Computational Biology* eds. Crabbe, Drew and A. Konopka, Marcel Dekker Inc. New York.
3. Pichierri,F., Aida,M., Gromiha,M. and Sarai,A. (1999) Free energy maps of base-amino acid interaction for protein-DNA recognition. *J. Am. Chem. Soc.* 121, 6152-6157.
4. Sayano,K., Kono,H., Gromiha,M. and Sarai,A. (2000) Multicanonical Monte Carlo Calculation of Free-Energy Map for Base-Amino Acid Interaction. *J. Compt. Chem.*, in press.

REGIONS OF POTENTIAL INTERACTIONS IN RNA MOLECULES AS FOUND BY COMPUTER SEARCH

Shabalina S.A.

NCBI, NLM, National Institute of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, USA
e-mail: shabalin@virgo.nlm.nih.gov

Keywords: RNA-RNA interactions, computer analysis, pattern of complementarity, E.coli 16S rRNA

Resume

Motivation:

Intermolecular interactions between RNA molecules in the course of translations may be more important than is generally accepted. Some experimental evidences support this assertion. Intermolecular hybridization experiments (Sarge and Maxwell, 1991; Matveeva and Shabalina, 1993; Hu et al., 1999) showed that mammalian rRNAs hybridize with mRNAs. Smardo and Calvet (1987) have found stable intermolecular hybrids between of human rRNAs and tRNAs.

Results:

Computer analysis of sites of the possible mRNA-rRNA interactions in eukaryotes showed an extensive potential for the formation of stable molecular hybrids between both large ribosomal RNAs, on the one hand, and many mRNAs, on the other hand (Matveeva and Shabalina, 1993). However, no long complementary regions were found. Instead, our analysis revealed several sites on rRNA which have unspecific mRNA-binding ability. Therefore, we suggested that mRNA-rRNA interactions follow "the multiple contact model" and occur due to the formation of duplexes between short complementary sites scattered over the sequences (Nechipurenko et al., 1995). We introduced a term "clinger-fragments" for such potential sites of intermolecular interaction between rRNAs and mRNAs. The applicability of the multiple contact model to RNA-RNA interactions in prokaryotes, in particular to *E. coli*, is discussed in this presentation.

Availability:

The program is available on request (shabalin@virgo.nlm.nih.gov).

Introduction

Complementary interactions between nucleic acid molecules play the critical part in replication, transcription, recombination, splicing, and DNA repair. In addition to the classical codon-anticodon interactions between mRNAs and tRNAs, these also include interactions between ribosomal RNAs and the short regions of mRNAs, known as Shine-Dalgarno sequences and located close to their 5'-ends. This communication presents the results of computer-assisted search for the potential double-helical hybrid structures formed between regions of either the same or different molecules of 16S rRNA, mRNA and tRNA of *E. coli*.

Methods and algorithms

Free energy of a hybrid structure was calculated as the sum of terms which correspond to all its consecutive pairs of dinucleotides. For each pair of dinucleotides, its free energy under a given temperature was calculated using the standard formula $\Delta G(T) = \Delta H - T\Delta S$ from their enthalpy and entropy, which were obtained experimentally by Frier *et al.* (1986). In some cases I used Boltzmann statistical weight of a structure, $p(T)$, which can be calculated from its free energy $P(T) = \exp[\Delta G(T)/RT]$.

Complementary sites with the energy of interaction or statistical weight lower than some threshold were selected. Different threshold values allowed us to find the sites of different lengths. The results were presented as dot-matrices. On the basis of these data, the distribution of the number of sites, complementary for a pair of RNA molecules, along one of them can be found. For every nucleotide of 16S rRNA the number of complementary sites on the mRNA and tRNA molecules was determined and 16S rRNA sites which have the maximum number of complementary sites in the mRNA and tRNA sequences were found.

Implementation and results

Comparison of profiles of 16S rRNA complementarity to different mRNAs showed that in all cases maxima of complementarity are located in more or less the same regions on 16S rRNA. This pattern is apparently a general one. The overall distribution of the degree of complementarity to 100 mRNAs that encode very different proteins along 16S rRNA in *E. coli* is strikingly similar to the earlier reported distribution of complementarity to murine mRNAs along the murine 18S rRNA (Matveeva and Shabalina, 1993).

It is of obvious interest to see whether the patterns of complementarity to 16S rRNA are similar for mRNAs encoding protein that belong to different functional groups. The analysis of profiles of 16S rRNA complementarity to various groups of mRNAs (encoding heat-shock genes, enzymes, etc.) showed that overall

patterns were similar in all these cases; in particular the maxima of complementarity almost coincided. Thus, the major clinger-fragments on 16S rRNA are universal for different genes. The sites of 16S rRNA that are potentially involved in intermolecular interactions with mRNA are presented in the Table 1. The total number of clear-cut clinger-fragments complementary to mRNA sequences is 17. These fragments are G,C-rich and frequently contain GGC, CGG, CCG, and GCC trinucleotides. The regions of mRNAs which are complementary to the 16S rRNA clinger-fragments usually contain inexact repeats. Monte-Carlo experiments were performed to assess the statistical significance of the clinger-fragments.

An analogous analysis of complementarity of 16S rRNA to tRNAs of *E. coli* produced similar results. The number of clinger-fragments for tRNA on 16S rRNA molecule was approximately the same and, strikingly, most of them coincided with the clinger-fragments for mRNAs (Table 1). This coincidence is hardly accidental and may be relevant to the translation process. Intermolecular contacts may be formed either successively or simultaneously through the formation of the triplex structure tRNA-rRNA-mRNA.

Discussion

There are some experimental data on *E. coli* rRNAs that are relevant to my computer analysis (McCarthy and Brimacombe, 1994). Most of the data were obtained either by site-directed cross-linking between 16S rRNA and mRNA analogs or by *in vitro* intermolecular hybridization between 16S rRNA and mRNAs. Current experimental data confirm that at least 3 of 16 theoretically predicted clinger-fragments are, in fact, involved in interactions with mRNAs. It will be very interesting to collect more data on the other potential clinger-fragments. In particular, those peaks of complementarity that are located in single-stranded regions of 16S rRNA may be, indeed, involved in intermolecular interactions, perhaps in binding of RNA matrices by the ribosome, which may facilitate the process of translation.

References

- Frier S. M., Kierzek R., Jaeger J.A., Sugimoto N., Caruthers M.N., Neilson T., Turner D.H. (1986). Improved free-energy parameters for prediction of RNA duplex stability. *Proc.Natl.Acad.Sci. USA*, **83**, 9373-9377.(1)
- Hu,M. C-Y., Tranque,P.,EdelmanG.M.,Mauro V.P. (1999). rRNA-complementarity in the 5'untranslated regionof mRNA specifying the Gtx homeodomain protein: Evidence that base-pairing to 18S rRNA affects translational efficiency. *Proc.Natl.Acad.Sci.USA*, **96**,1339-1344.
- Matveeva,O.V., and Shabalina,S.A. (1993). Intermolecular mRNA-rRNA hybridization and the distribution of potential interaction regions in murine 18S rRNA. *Nucleic Acids Research*, **21**, 1007-1011.
- McCarthy, J.E.G., and Brimacombe R. (1994). Prokaryotic translation: the interactive pathway leading to initiation. *TIG*, **10**,402-407.
- Nechipurenko,Yu. D., Popov ,N. V., Isaev, M. A., Shabalina ,S. A., Matveeva ,O. V. (1995). Multiple contact model describing interaction of mRNA with rRNA sites during translation processes. *Biofizika*, **40** (6), 1208-1213.
- Sarge K. D., and Maxwell E. S. (1991). Evidence for a competitive - displacement model for the initiation of protein synthesis involving the intermolecular hybridization of 5S rRNA, 18S rRNA and mRNA. *FEBS Letters*, **294**, 234-238.
- Smardo F. L. Jr. and Calvet J. P. (1987). Human glutamate tRNA forms stable hybrids *in vitro* with 28S ribosomal RNA. *Nucleic Acids Res.*, **15**, 661-681.

Table 1. Clinger-fragments for mRNAs and tRNAs in *E. coli* 16S rRNA.

Position of a clinger-fragment	Sequence of a clinger-fragment	Relevance of a clinger-fragment to	
		mRNA	TRNA
35-46	GCUGGCGGCAGG	+	-
102-111	UGGCGGACGG(G)	+	+
314-324	ACUGGAACUGA	-	+
338-353	UCCUACGGGAGGCAG	+	+
440-449	AGCGGGGAGG	+	+
515-541	UGCCAGCAGCCGCGGUAAUACGGAGG	+	+
580-587	GCAGGCGG	+	-
611-616	CCCCGG	+	+
719-726	CGGUGGCG	+	+
732-740	GGCCCCUG	+	+
877-888	ACCGCCUGGGGA	+	+
924-932	GGGGGCCCG	+	+
978-993	CCUUACCUGGUCUUGA	-	+
1126-1143	CCAGCGGUCCGGCCGGG	+	+
1159-1176	GCCAGUGAUAAACUGGAG	+	+
1277-1283	GGACCUC	-	+
1303-1313	GGAUUGGA	-	+
1381-1387	UCCCGGG	+	+
1399-1404	CGCCCG	+	-
1516-1524	GAACCUGCC	+	+
1533-1539	ACCUCCU	+	+

STRUCTURAL FEATURES OF mRNA 5'UTRs OF EUKARYOTIC GENES EXPRESSED AT HIGH AND LOW LEVELS

*Vorobiev D.G., Titov I.I., Kochetov A.V., Kolchanov N.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: denis@bionet.nsc.ru

*Corresponding author

Keywords: eukaryotic mRNAs, translation efficiency, secondary structure, statistical analysis

Resume

Motivation:

It was shown previously (Kochetov et al., 1998) that 5'-untranslated regions (5'UTRs) of mRNAs of high and low expression eukaryotic genes (HE and LE, respectively) differ by context characteristics: leader sequences of HE mRNAs are shorter, more asymmetric by complementary nucleotides (G/C and A/U) content and they contain less false start codons. These data enable us to suppose that 5'UTRs of HE genes are characterized by more stable secondary structure (SS).

Results:

In the paper presented, the hypothesis was verified: HE mRNA 5'UTRs are likely to form more stable secondary structure in comparison with the low expression ones. It was also found that LE 5'UTRs have more stable SS than randomly generated sequences of the same length and nucleotide content. These facts allow us to assume that higher stability of LE 5'UTR secondary structure may play an important role in the control of regulatory gene expression.

Introduction

A large bulk of experimental data evidences that translation efficiencies of eukaryotic mRNAs may vary within wide range (Ray et al., 1983; Kochetov and Shumny, 1998). The translation initiation, which is known to include the process of scanning along the 5'UTR by 40S ribosomal subunit, is strongly influenced by 5'UTR features. For example, it was shown experimentally that the hairpin-like structures might exert negative effect to the rate of 40S ribosomal subunit moving along mRNA. The extent of a hairpin negative effect on eukaryotic mRNA translation *in vivo* depends upon its stability and localization within the molecule (Kozak, 1994; Vega Laso et al., 1993).

Methods and algorithms

5'UTR sequences were extracted from the EMBL databank. We have used only the genes with experimentally detected transcription start, this fact being verified by annotating literature sources.

In order to predict secondary structure, we have applied the software program "Fitness" developed by us previously for prediction of RNA's SS. This program is based on genetic algorithm described elsewhere (Titov et al., 2000). For calculation of SS energy, this program uses thermodynamical parameters from Turner compilation (Turner et al., 1988). For each sequence from the sample, we have calculated the energy E of the most stable SS and the energy E_{hairpin} of the most stable hairpin. To characterize the favor of nucleotide content during SS formation, we have calculated the following parameters: (1) G+C-content, (2) the measure of misbalance between complementary nucleotides G and C, calculated as

$D_{GC} = |P_G - P_C| / (P_G + P_C)$, and (3) the weighted content of complementary pairs, or energy capacity given

$$\text{as } E_{\text{capacity}} = 9P_G P_C + 3P_A P_U + 2P_G P_U,$$

where P_A, P_U, P_G, P_C are the shares of corresponding nucleotides in a sequence. We have also calculated z-score of the values E and E_{hairpin} :

$$Z(E_{\text{nat}}) = (E_{\text{nat}} - \langle E_{\text{rand}} \rangle) / \sqrt{\text{Disp}(E_{\text{rand}})},$$

where E_{nat} is the energy of SS in a natural sequence, E_{rand} – the energy of SS in a random sequence of the same length and nucleotide content.

The values of characteristics E_{capacity} , E , E_{hairpin} , and G+C-content in the samples examined were found to be normally distributed (according to Kolmogorov-Smirnov test, data not shown). Statistical analysis was performed

by the software package Statistica 4.5 (Statsoft™). In order to compare the mean values of two Gaussian distributions, we have applied the Student's criterion (t-test).

Results

We have compiled the samples of 5'UTR for two groups of genes, i.e., HE and LE genes. The HE group contains 92 genes, the expression products of which are produced in cells in a large amount (e.g., actins, tubulins, HSPs, etc.). We based on the assumption that mRNAs corresponding to these genes should be effectively translated. The LE gene group includes 50 genes referring to regulatory protein synthesis (e.g., transcription factors, growth factors, etc.). Their expression is exerted under strict control (Chen and Shyu, 1995; Pahl and Baeuerle, 1996), since excessive production causes strong disorders. The samples were subdivided according to taxon ranging: monocot plant genes, dicot plant genes, and mammalian genes.

The SS was predicted for the natural and random mRNAs leader sequences. If the length of a leader exceeds 250 nucleotides, then only first 250 nucleotides were taken into analysis.

The testing of the algorithm' stability was performed in the set of randomly generated sequences with the equal content of A, T, G, C nucleotides in the interval ranging in length from 20 to 300 nucleotides. As was expected, the values of energies in optimal SS and their dispersions were linearly dependent upon the length of the sequence. Testing the dependency of energy from the G+C-content under the fixed length of the sequence has also proved the stability of the algorithm. The free energy of predicted SSs reduced with the growth of G+C-content.

Dependence of the secondary structure parameters from the length of 5'UTR. As is known, the mean values of the SS' energy linearly depend upon the length of a sequence (Fontana et. al., 1993). It occurred that LE gene leader (with the length of 236.6 ± 265.7 nucleotides) is in average about three folds longer than HE gene leader (with the length of 87.3 ± 50.0 nucleotides).

Dependence of the secondary structure of leader mRNA on the 5'UTR contextual features. At the next step, we have tried to determine the impact of nucleotide content into stability of SS. We have compared the values of G+C-content, misbalance in G and C contents, and energy capacity E_{capacity} in HE and LE gene groups. The results of this comparison are shown in Table 1. The data obtained give the evidence that in two taxonomic groups, in monocots and mammals, the values of energy capacity of leaders in LE genes are significantly ($p < 0.05$) higher than in HE genes. Therefore, nucleotide context of LE genes is more favorable for secondary structure formation.

Table 1. Characteristics of nucleotide content and SS of the 5'UTRs of mRNA from the samples of HE and LE genes. The significance of the difference between the groups of HE and LE genes.

	dicots		monocots		mammals		total	
	HE	LE	HE	LE	HE	LE	HE	LE
Volume of a sample	44	22	15	11	33	17	92	50
G+C-content, %	37.1 ± 7.7	35.9 ± 6.7	56.2 ± 8.8	56.4 ± 11.0	53.7 ± 12.5	62.3 ± 11.7	46.2 ± 13.1	49.4 ± 15.4
	the difference is insignificant		the difference is insignificant		$p < 0.05$		the difference is insignificant	
Misbalance by G and C	0.30 ± 0.22	0.32 ± 0.23	0.35 ± 0.21	0.17 ± 0.1	0.21 ± 0.16	0.17 ± 0.11	0.28 ± 0.2	0.24 ± 0.19
Energy capacity	1.31 ± 0.23	1.28 ± 0.3	1.65 ± 0.21	1.92 ± 0.36	1.86 ± 0.43	2.15 ± 0.53	1.57 ± 0.4	1.72 ± 0.56
	the difference is insignificant		$p < 0.05$		$p < 0.05$		the difference is insignificant	
Energy z-score of complete SS	- 0.19 ± 1.22	- 0.9 ± 1.43	0.15 ± 0.5	0.31 ± 1.02	0.26 ± 1.1	- 0.26 ± 1.57	0.03 ± 1.11	- 0.42 ± 1.46
	$p < 0.05$		the difference is insignificant		the difference is insignificant		$p < 0.05$	
Z-score of the energy of the most stable hairpin	- 0.12 ± 1.09	- 1.24 ± 1.78	- 0.46 ± 1.49	0.14 ± 0.71	0.28 ± 1.04	- 0.02 ± 1.08	0.06 ± 1.03	- 0.65 ± 1.58
	$p < 0.01$		the difference is insignificant		the difference is insignificant		$p < 0.01$	
Energy per nucleotide	- 0.03 ± 0.06	- 0.05 ± 0.05	- 0.05 ± 0.09	- 0.11 ± 0.05	- 0.07 ± 0.07	- 0.17 ± 0.08	- 0.05 ± 0.06	- 0.11 ± 0.08
	the difference is insignificant		$p < 0.01$		$p < 0.01$		$p < 0.01$	

Comparative analysis of the secondary structure parameters in 5'UTRs and random sequences. Since LE mRNA 5'UTR features potentiate formation of the stable secondary structure (they are longer and contain closer concentrations of complementary nucleotides), these factors could make the major impact in the

difference in secondary structure stability between 5'UTRs of HE and LE mRNAs. However, apart from these parameters, SS stability can depend on the content of repeats. For accounting this factor, we have tried to eliminate the effects of nucleotide content and the length of a sequence. With this aim, for each natural sequence, we have calculated z-score of SS' characteristics. The mean z-score values calculated for the samples of HE and LE genes were compared to each other. The results of comparison are given in Table 1. It was found that secondary structure is more stable in the group of LE genes in comparison with the random sequences (mean z-score value is < 0). The difference of mean z-score values between HE and LE gene groups is significant ($p < 0.05$) for the samples of monocot genes and for the united sample of all taxa.

Thus, the factor determining the difference in SS' energy between HE and LE genes is the length of a leader sequence. However, the differences still occur if we consider the energy normalized per 5'UTR length. The main impact into these differences is produced by G+C-content in mammals; by misbalance in C and G nucleotides – in monocots. In dicots, the difference appears due to the properties of a sequence, which are related not to the total nucleotide content, but to the consequence of nucleotides.

An increased SS stability in the leader sequences of LE genes in comparison to random sequences was unexpected phenomenon. We tried to determine the putative cause of this fact. In each sequence, we made the search for the most stable hairpin without defects and supplied with a loop. It occurred that the stem of such a hairpin is present in SS approximately in 60 % of cases (data not shown). As a consequence, z-score value of the SS' energy significantly ($R^2=0.45$, $\alpha < 0.01$) correlates to the z-score value of the most stable hairpin. We have found that in the sample of LE genes of monocots and in the united sample of LE genes, z-score values of the most stable hairpin significantly ($p < 0.05$) less than in the corresponding samples of HE genes. Besides, by modulo, z-score value of the most stable hairpin (equalling to -1.24 in monocots and to -0.65 in the united sample) is higher than the z-score values of the total sample of SS sequences (-0.9 and -0.42, respectively). This means that the basic factor determining the general stabilization of SS in the group of LE genes compared to random sequences is the presence in the 5'UTR of inverted repeat, which forms the hairpin with increased stability.

Discussion

Functionally active mRNA probably should be translated with a certain efficiency, hence, the presence of «strong» negative signals is unlikely. Therefore, the parameters of 5'UTRs of mRNAs of LE genes should be adopted to support translation. From the other hand, the sample of HE genes is insufficiently characterized. The situation is possible when we use the sequence of a gene referring to the multi-gene family, which makes the small impact into the total protein synthesis. The 5'UTR parameters for such a gene may differ from the values typical for HE genes. Hence, the difference between the samples of HE and LE genes could be little and revealed only by statistical methods. This very situation we observe in our case.

Unexpectedly, the 5'UTRs of LE mRNAs form more stable SS than random sequences. The results of computational analysis showed that LE mRNA 5'UTRs form more stable secondary structure than HE ones. With this respect, our results may be interpreted as supporting the evidence that SS of leaders in LE genes has the functional significance. This functional significance may be realized through supporting translational activity of mRNAs at a low level and, therefore, deleterious excessive production is prevented.

Acknowledgements

The work was partially supported by the Russian Foundation of Basic Research (grants Nos 99-07-90203, 00-04-49229, 00-07-90337). The authors are grateful to Galina Orlova for the help in translation the manuscript. Alex Kochetov was supported by SD RAS grant for young scientists.

References

1. Chen, C.-Y.A. and Shyu, A.-B. (1995) Trends Biochem. Sci., **20**, 465-470.
2. Fontana W., Konings D., Stadler P., Schuster P. (1993) Biopolymers, **33**, 1389-1404.
3. Kochetov, A.V., Ischenko, I.V., Vorobiev, D.G., Kel, A.E., Babenko, V.N., Kisselev, L.L., Kolchanov, N.A. (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. FEBS Lett., **440**, 351-355.
4. Kochetov, A.V. and Shumny, V.K. (1998) Influence of the mRNA structure on the translation initiation process in plant cells. Advances In Current Biology (Russ), **118**, 754-770.
5. Kozak, M. (1994). Determinants of translational fidelity and efficiency in vertebrate mRNAs. Biochimie, **76**, 815-821.
6. Pahl, H.L. and Baeuerle, P.A. (1996) Control of gene expression by proteolysis. Curr. Opin. Cell Biol., **8**, 340-347.
7. Ray, B.K., Brendler, T.G., Adya, S., Daniels-McQueen, S., Miller, J.K., Hershey, J.W.B., Grifo, J.A., Merrick, W.C., Thach, R.E. (1983) Role of mRNA competition in regulating translation: further characterization of mRNA discriminatory initiation factors. Proc. Natl. Acad. Sci. USA, **80**, 663-667.
8. Titov I. I., Ivanisenko V. A., Kolchanov N. A. (2000) Fitness - a WWW-resource for RNA folding simulation based on genetic algorithm with local minimization. Comp. Techn., *in press*.
9. Turner, D.H., Sugimoto, N., Freier, S.M. (1988) RNA structure prediction. Ann. Rev. Biophys. Biophys. Chem., **17**, 167-192.
10. Vega Laso, M. R., Zhu, D., Sagliocco, F., Brown, A. J., Tuite, M. F., McCarthy J.E. (1993) Inhibition of translational initiation in the yeast *Sacharomyces cerevisiae* as a function of the position and stability of hairpin. J. Biol. Chem., **268**, 6453-6462.

MASS ANALYSIS OF RNA SECONDARY STRUCTURES USING A GENETIC ALGORITHM

**Titov I.I., Vorobiev D.G., Kolchanov N.A.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: titov@bionet.nsc.ru

*Corresponding author

Keywords: translation, translation initiation, RNA secondary structure, statistical analysis

Resume

Motivation:

Effects of the secondary structure of mRNA leader sequences on translation rate are exemplified experimentally. However, no statistical evidence on the role of secondary structure (unlike the average composition effect) at the pre-elongation stage is yet available, since the leader sequences are extremely heterogeneous in both their lengths and nucleotide composition.

Results:

We have demonstrated that distribution of relative deviations Z (z-scores) of optimal secondary structure energies of random sequences can be approximated by the normal distribution in most cases, justifying application of standard statistical tests. Two patterns were found while analyzing z-scores of 5'UTRs of plant and mammalian genes with high (H) and low (L) expression levels. In the first (dicot plants) pattern, these regions display relatively uniform composition but deviate considerable from random sequences in stability of their secondary structures. As expected, these deviations of L genes are more pronounced. In the second pattern, typical of mammalian genes, the leader regions are heterogeneous in their average GC composition within each subset and differ from one another even more drastically. However, characteristics of secondary structure of these regions differ insignificantly from the random sequences. The effects observed suggest two possible scenarios of expression regulation. The first scenario involves local functional elements based on secondary structures in the leader regions. The second scenario implies a generalized control of the genes with similar expression levels. Its global order requires a new level of gene organization (for example, their localization in isochores), what supports a selection hypothesis of compositional transition in higher organisms.

Introduction

Recent success in sequencing provided bioinformatics with a possibility to analyze unexampled data massifs. The tremendous volume of sequences obtained allows weak dependencies, insignificant in case of small-volume samples, to be revealed. As a rule, the sequences analyzed are heterogeneous (according to definition [1], as they are composed of statistically distinct subgroups due to evolutionary dependence, spotty sequencing, dependencies of control indicators on processing algorithms and inessential parameters of sequences, etc.), impeding obtaining reliable conclusions.

This work describes the computer analysis of secondary structures of mRNA leader regions using a program based on genetic algorithm (GA). The algorithm speed [2] allows mass calculations of secondary structures to be performed. The energy of secondary structure in parameterization of Turner *et al.* [3] is employed as a control indicator. The difficulties encountered while analyzing are of general character, in particular, they are independent of the function optimized. Let us consider two difficulties in detail:

1. GA does not guarantee the optimal solution. In addition, the optimization itself depends not only on the algorithm settings, but also on parameters of a sequence, in particular, its length and nucleotide composition.
2. More important is the fact that the energy of the secondary structure itself depends on the same parameters.

Therefore, it is actually difficult to form a uniform set of data (where the distinctions between the sequences are limited by the effects of factors inessential for the analysis and may, thus, be considered as random).

Typical of the structural properties of RNAs and proteins is a strong dependence even on the average alphabetical content, whereas the diversity of structures and functions is determined by the order of symbols. Thus, it is expedient in mass analysis to compare naturally occurring sequences with the sequences displaying similar average compositions and random order of symbols (the "null hypothesis"). Since the distributions of control indicators are unknown *a priori*, the technique of relative deviations (z-scores) is frequently used. Lattice proteins, conventional proteins and their structures generated by threading, and RNAs divided into small

subsequences were the objects of such analysis. Typical of sequences in these examples are equal lengths, which is unusual for natural data. In addition, the conclusions made are qualitative and not statistically substantiated.

Calculation of secondary structure

The program Fitness [2] based on the GA was used in this work. The release used has the following additions compared with the previous release described in detail in [2]: (a) the secondary structures considered now contain no thermodynamically unfavorable helices and (b) inhibition of stems composed by partially overlapping stems (the so-called running loops [4]) is canceled.

As a result of (a), the algorithm finds the optimal or close to optimal structure of sequences up to 300 nt long (Fig. 1). In the previous release, the "high frequency" noise on the surface of free energy trapped the sequences longer than 150–200 nt in the local optima.

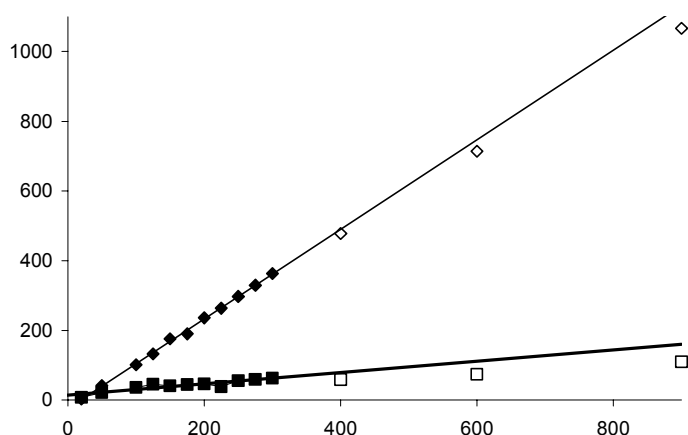


Figure 1. Dependence of free energy (diamonds) and the square root of its dispersion (squares) of the optimal secondary structure of random sequences of the same composition on their lengths (averaged over five sequences). Deviations for long sequences (transparent symbols) from the linear approximation for short sequences (less than 300 nt; dark symbols) may serve as a criterion of optimization quality.

Although addition (b) fails to provide an essential improvement of the optimum reached, it is very important for the accuracy of the structure prediction. Our calculations using random sequences have demonstrated that the fraction of running loops increases with sequence length (Table 1), as one would expect. Due to (b), the search space for the algorithm increases considerably, but the calculation time does not. Depending on the alphabet, the optimization speed is ordered as $GC < AU < AUGC$, complying with a smoother landscape of the free energy for alphabets with a higher size and lower energy of complementary bonds [5].

Table 1. Calculation time (P-II 300 MHz) and characteristics of the secondary structure of random sequences of 2 lengths and 3 compositions (averaged over 5 sequences). Sets of 50 random sequences for each length and alphabet were used to test the normality of z-score distributions through 3 standard statistical procedures (significance level p is indicated for each test). In this case, GA was run once for each sequence. The only case of non-normality (by one test of 3) is shown bold.

Alphabet	Length (nt)	Number of iterations	Calculation time (min)	Running loop frequency (%)	Free energy of the secondary structure (kcal/mole)	Kolmogorov-Smirnov	Lilliefors	Shapiro-Wilk	Normality rejected
AU	100	324 ±319	0.73 ±0.87	0.25±0.19	-5.4 ±2.0	$p > 0.2$	$p > 0.2$	$p < 0.43$	No
AU	300	228 ±136	7.0 ±4.2	0.29 ±0.12	-20.5±3.7	$p > 0.2$	$p > 0.2$	$p < 0.24$	No
AUGC	100	259 ±250	0.35 ±0.45	0.15 ±0.14	-11.9±3.8	$p > 0.2$	$p < 0.2$	$p < 0.40$	No
AUGC	300	221 ±107	3.1 ±1.6	0.18 ±0.08	-41.9±5.6	$p > 0.2$	$p > 0.2$	$p < 0.63$	No
GC	100	203 ±116	0.42 ±0.26	0.31 ±0.15	-63.4±5.0	$p > 0.2$	$p > 0.2$	$p < 0.38$	No
GC	300	463 ±119	8.3 ±20	0.39 ±0.10	-200.0±7.5	$p > 0.2$	$p < 0.1$	$p < 0.0014$	Yes

Z-score technique

Relative deviation (or z-score) of a random variable x is determined as

$$Z = \frac{\text{Observed}(x) - \text{Expected}(x)}{\sqrt{\text{disp}(x)}}$$

Our calculations (Fig. 1) suggest that distributions of the free energy z-scores of random sequences not exceeding 300 nucleotides are statistically uniform with respect to the length and compositions of these sequences. It is evident that the z-score distribution itself belongs to the class of limit distributions and important that it can be approximated by the normal distribution in most cases (Table 1). Thus, we may use standard statistical procedures for testing the hypotheses on higher or lower stability of the secondary structures of the sequences under study compared to the random sequences of the same lengths and compositions. Essentially, heterogeneity of the sample studied with respect to these parameters is allowed.

Relative stability of the secondary structures of mRNA 5'-untranslated regions of high- and low-expressed genes

Statistical analysis described above addresses the question to what extent the secondary structure of natural RNAs is important. Genes displaying high and low expression levels (selection of sequences is detailed in the work of Vorobiev *et al.*, this issue) are most convenient objects for such a study, as they allow the analysis to be reduced to comparison.

The samples of H and L genes appeared to be considerably heterogeneous in both their primary and secondary structures (Table 2). However, two patterns may be separated. In the first (mammals), H and L genes differ considerably in their average compositions: L genes display an essentially higher fraction of GC nucleotides (compare the structure energies of 50% and 100% GC sequences in Table 1). In addition, the deviations of secondary structure stabilities of H and L genes are opposite (insignificantly, from the random sequences; significantly, from one another).

Table 2. Characteristics of primary and secondary structures of mRNA leader regions of H and L genes. Data of small and non-significant set of monocot plants are not shown separately (but contribute to the overall picture). Significant deviations are shown bold ($p < 0.05$).

	Dicots		mammals		All (including monocots)	
	H	L	H	L	H	L
Sample volume	44	22	33	17	92	50
G+C content, %	37.1±7.7	35.9±6.6	53.7±12.5	62.3±11.7	46.2±13.1	49.4±15.4
Z(E)	-0.19±1.22	-0.9±1.43	0.26±1.1	-0.26±1.57	0.03±1.11	-0.42±1.46

In the second pattern (dicot plants), the sequences are essentially more uniform in their composition both between and within the H and L sets. Compared with the mammalian genes, they are GC-poor and display smaller GC variations despite the greater sample volume. A weak effect of z-scores of secondary structure stability is, in contrary, significant. The pooled sample of H and L genes follows the same pattern.

Discussion

The hypothesis on translation efficiency is most popular (compared with transcription maximization and mutation bias hypotheses) while interpreting the usage of synonymous codons in different genomes. Higher rates of both translation and transcription may result in a higher level of gene expression. Optimization of the rates of successive (stationary) processes (for instance, translation initiation and elongation of peptide chain) through selection would lead to their coordination with respect to efficiency. The easiest way to reach this is to control the factors common for all the processes. On the one hand, their rates should be coordinated, as long as the yield of peptide chain is a limiting factor. On the other, these processes are not akin, including the mRNA regions involved. This leads to differences in the corresponding regulatory mechanisms. Rare codons within the coding region slow down the ribosome movement, as does secondary structure within leader region [6].

The energy of RNA secondary structure depends mainly on the average nucleotide composition (Table 1), as confirmed through processing a great number of random sequences [5]. The difference between the structural characteristics of coding regions of high- and low-expression genes in one of the first observations [7] and further application of only context analysis of leader regions [8] is connected with the domination of this effect. (Individual regions of a pre-mRNA differ insignificantly in their nucleotide content [9].) The technique of z-scores allows the effect of the secondary structure to be separated from the trivial consequence of average composition, e.g. environmental temperature.

Our observations suggest two possible scenarios of expression regulation — local and global. The local scenario is observed in plants and connected with differential modulation of leader regions by secondary

structure elements, playing the role of conventional functional elements. The global scenario is typical of higher organisms and based on generalized control of the genes with similar expression levels through close localization in isochores. A compositional transition to warm-blooded isochore structure [10] may have made an emergence of such a "blockwise" regulatory mechanism easier.

Acknowledgements

We are grateful to G. Chirikova for translating this manuscript into English. The work was supported by the INTAS-RFBR grant 95-0653 and Integrational Project of SB RAS No 66.

References

1. Feller W. An introduction to probability theory and its applications. John Wiley & Sons, 1970.
2. Titov I.I., Ivanisenko V.A., Kolchanov N.A. (2000) *Comp. Tech.* **5** (in press).
3. Turner D.H., Sugimoto N., Freier S.M. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17**, 167.
4. Studnicka G.M., Rahu J.M., Cummings I.M., Salser W.A. (1978) *NAR* **5**, 3365.
5. Fontana W., Konnings D.A.M., Stadler P.F., Schuster P. (1993) *Biopolymers* **33**, 1389.
6. Kozak M. (1994) *Biochimie* **76**, 815.
7. Ischenko I.V., Kel A.E., Omelyanchuk L.V., Kolchanov N.A. (1993) In: *Computer analysis of genetic macromolecules. Structure, function and evolution.* (World Scientific Publishing; Eds: H. Lim and N.A. Kolchanov), pp.156-167.
8. Kochetov A.V., Ischenko I.V., Vorobiev D.G., Kel A.E., Babenko V.N., Kisselev L.L., Kolchanov N.A. (1998) *FEBS Lett.* **440**, 351.
9. Clay O., Cacciò S., Zoubak S., Mouchiroud D., Bernardi G. (1996) *Mol. Phylogenet. Evol.* **5**, 2.
10. Bernardi G., Bernardi G. (1986) *J. Mol. Evol.* **24**, 1.

TRANSTERM - A DATABASE OF RNA COMPONENTS AND MOTIFS

**Jacobs G.H., Stockwell P.A., Brown C.M.*

Department of Biochemistry, University of Otago, Dunedin, New Zealand

e-mail: grant.jacobs@stonebow.otago.ac.nz

*Corresponding author

Keywords: translational regulation, mRNA, untranslated regions, computer database, WWW computer tool

Resume

Motivation:

While there are well-known DNA, protein and tertiary structure databases, there are few established databases of mRNAs and their flanking untranslated regions (UTRs). As evidenced by the poor correlation between mRNA levels and protein expression, regulation of translation is a biologically important event. Motifs in mRNAs, particularly within the UTRs, are important for control of translation and transport of mRNAs. We present a database of mRNAs, their associated UTRs and motifs on the World Wide Web (WWW).

Results:

We have recently enhanced our database of transcripts, Transterm, to include:

Taxonomy data from the NCBI Taxonomy database to assign entries to species (taxas)

Tools to retrieve the archived database files. These contain pre-calculated values such as codon usage tables, consensus of the initiation and termination regions and the information content of these regions, parameters of the coding regions (eg. Nc and GC3 values), the initiation and termination codons used.

A flexible tool search the database for patterns (motifs). Searches can be restricted to specific species and specific regions of the mRNAs (5'-UTR, CDS, initiation region, etc.) The motifs can either be those stored within Transterm, or entered by users.

A tool to allow users to search their own sequences for motifs.

A tool to locate a species of interest by searching through the Transterm taxonomy information

All the tools described are available via a new interactive WWW interface. Links to related databases and tools are provided. Transterm contains mRNA sequence data from all annotated divisions of Genbank and all genomes available from the NCBI genome site.

Availability:

Transterm is available via the WWW at <http://uther.otago.ac.nz/Transterm.html>.

Introduction

Considerable recent attention has focussed on regulation of transcription, particularly with the available of many genome sequences. Relatively little attention has been paid to regulatory sequences with mRNAs which have roles in mRNA regulation (levels of translation and degradation) and transport (1-7).

Transterm somewhat misleading name comes from its original origin as a database of translation termination signals. Over time it has evolved to incorporate the complete mRNA, with CDS and both UTRs.

Methods and algorithms

With the exception of a locally-modified version of RossOverbeek's pattern searching algorithm (*scan_for_matches*, written in C and Perl), all software has been written in-house. Extraction of sequence data from the Genbank database relies on *indexdb* and *fishterm*, software written in Pascal by Dr. PA Stockwell. Download of the databases, construction of the local version of the NCBI taxonomy database, scripts to generate the Transterm database are written in C, Perl and Unix shell scripts. Most of these have been written by Dr. G.H. Jacobs with a few portions written by Mark Schrieber and Mark Dalphin (a former member of the group). The WWW interface uses HTML and Javascript and is almost entirely CGI-driven (in Perl). The WWW interface was written by Dr. Jacobs. Interaction with the *postgres* relational database is via SQL, mostly within Perl code.

Several computers housed within the Department of Biochemistry are utilised by this project. The Genbank databases are sourced from the local bioinformatics server (a Sun UltraSparc 70). The *indexdb* and *fishterm* software is run from a DEC Alpha workstation. The WWW interface and *postgres* relational database is run from a Linux-based PC-workstation.

Implementation and results

Transterm is constructed from releases of Genbank (available locally), the NCBI Taxonomy database (<ftp://ncbi.nlm.nih.gov/pub/taxonomy>), and Genbank-format genomes (available from NCBI: <ftp://ncbi.nlm.nih.gov/genbank/genomes/>).

The species-level taxids and names are extracted the NCBI taxonomy database into local files, for use by Perl scripts.

Genbank files are divided into the species found, retaining only those entries with valid coding sequences. Additional check on database formatting errors are made, such as the exclusion of entries without taxids and the like. Genomes are retained as separate "units", so that each genome is a single selectable unit for examination via the WWW interface.

Each species or genome has its data parsed to generate FASTA-format files of each of the regions of the mRNAs:

- The coding sequence (CDS), including the initiation and termination codons,
- The 5'-UTR,
- The 3'-UTR,
- The initiation region, and
- The termination region.

These regions are summarised in Figure 1.

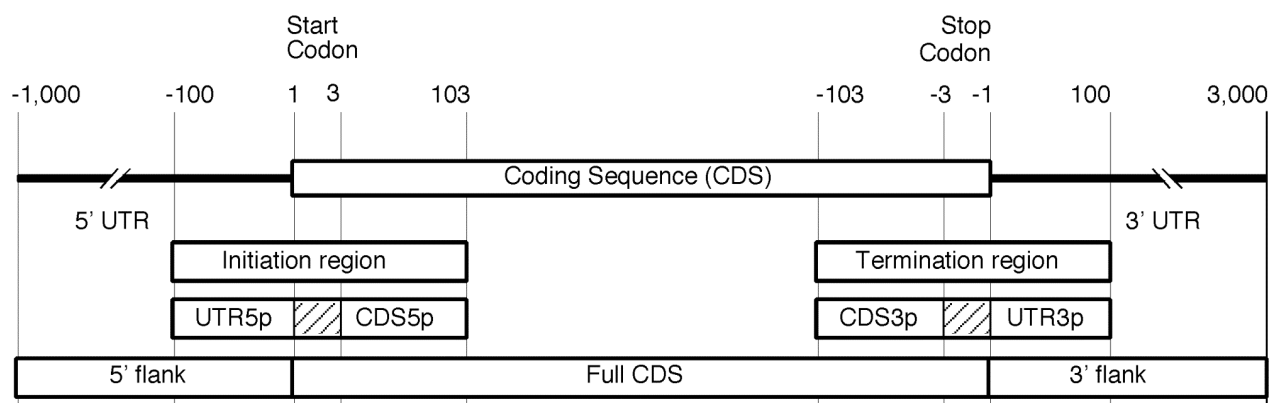


Figure 1. mRNA regions used within Transterm.

At present the UTRs are strictly-speaking flanking regions in many cases. (To locate unknown motifs, it is useful to have a large dataset of UTRs, even if the termini of these UTRs has not been established.) Current work is underway to distinguish the experimental known UTRs within our flanking regions.

5' flanking regions are considered to be the up to 1000 bases prior to the first base of the initiation codon. 3' flanking regions are considered to be the up to 3000 bases following the last base of the termination codon. Each putative UTR is "limited" by the presence of adjacent coding sequences in Genbank. If another coding sequence is encountered prior to the maximum length of our stored flanking regions, the flanking region is terminated at that point.

The initiation and termination regions are considered to be the up too 100 bases on either side of the initiation or termination codon (limited by adjacent coding sequences as described for UTRs). All coding sequences must be longer than 100 bases.

Three datasets of sequences are made: (i) all sequences with no duplicates removed ("redundant"), (ii) with duplicates removed by the Cleanup algorithm(5) and (iii) with duplicates removed by previous Transterm criteria, using the immediate start and stop contexts(1). For genomes, only redundant data is used, as genomes are treated as a complete whole.

The files are retained for users to search. Also from these files a variety of calculations are performed and archived (available from the WWW interface). Each entry has information on the start codon, termination codon, GC3, Nc and CAI values, length of the CDS, and database reference information. Each species has an alignment of the initiation and termination regions, a consensus of initiation and termination regions, a table of

the information content of initiation and termination regions, a codon usage table, the Genbank entry names for that species and the scientific name of the species.

Transterm is periodically updated; at the time of writing an upgrade based on Genbank(8) release116 was underway. The previous release contained data from over 10,000 species, including 20 genomes (3 eukaryotic). The updated release is anticipated to be considerably larger.

Discussion

We have presented an interactive database of information about mRNA sequences, their components and regulatory motifs we believe will be a valuable resource centre for biological scientists. Development of this database is continuing, in particular the incorporation of annotation relating to functional activity of the mRNAs. The database is being used in house as a data-mining tools alongside experimental molecular biology.

Acknowledgements

The authors wish to thank the support of the Marsden Fund (Royal Society of New Zealand) and the University of Otago.

References

1. Jacobs,G.H. Stockwell, P.A., Schrieber,M.J. Tate,W.P. and Brown,C.M. (2000) *Nucleic Acids Res*, 28, 293-295.
2. Dalphin,M.E., Stockwell,P,A., Tate,W.P. and Brown,C.M. (1999) *Nucleic Acids Res.*, 27, 294-294
3. Brown, C.M., Dalphin,M.E., Stockwell,P.A. and Tate,W.P. (1993) *Nucleic Acids Res.*, 21, 3119-2123
4. Brown, C.M., Stockwell, P.A., Trotman, C.N. and Tate, W.P. (1990) *Nucleic Acids Res.*, 18, 6339-6345
5. Pesole G, Liuni S, Grillo G, Ippedico M, Larizza A, Makalowski W, Saccone C. (1999) *Nucleic Acids Res.*, 27, 188-191
6. Grillo, G., Attimonelli, M., Liuni, S. and Pesole, G. (1996) *CABIOS*, 12, 1-8
7. Kochetov,A.V., Ponomarenko,M.P., Frolov,A.S., Kisselev,L.L., Kolchanov,N.A. (1999) *Bioinformatics*, 15, 704-12
8. Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A. and Wheeler, D.L. (1999) *Nucleic Acids Res.*, 27, 12-7

CRASP: SOFTWARE PACKAGE FOR ANALYSIS OF PHYSICO-CHEMICAL PARAMETERS OF ALIGNED SEQUENCES OF PROTEIN FAMILIES

Afonnikov D.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: ada@bionet.nsc.ru

Keywords: amino acid sequences, co-adaptive substitutions, multiple sequence alignment, amino acid characteristics

Resume

At present, the data on aligned primary sequences of protein families are being accumulated very rapidly. One of promising approaches to analysis of these data is to study correlations between amino acid substitutions at positions of protein sequences [1-3]. The information obtained could be valuable for revealing the peculiarities of the structure and function of the proteins under study. In relation to this topic, the main goal is to develop publicly available methods and software for correlation analysis of protein sequences.

During analysis of co-adaptive substitutions, two important tasks appear. The first is to study correlations of amino acid substitutions at positions of a protein in order to reveal the pairs of residues that are related due to functional interactions (e.g., steric contact). The resulted information could be used for prediction of possible contacts between the residues [4,5].

The second task is to reveal and analyse the conserved integral physico-chemical characteristics of a protein [6,7]. As the examples of these characteristics may serve the total charge of a protein molecule, the volume of its hydrophobic core, hydrophobic moments of alpha helices, etc. The constancy of these characteristics in the course of evolution implies that they are responsible for the key features significant for protein structure and function. Co-adaptive substitutions may be one of possible mechanisms for supporting the constancy of these characteristics, together with invariance of residues at positions of a protein and conserved substitutions of residues to those similar by physico-chemical properties [6].

In the present communication, we introduce the software package CRASP designed for analysis of physico-chemical pairwise correlations between amino acid substitutions at positions of aligned sequences of a protein family. In this package, we have realized an approach that is based on revealing of conserved physico-chemical characteristics of a protein obtained on the basis of information about correlations of substitutions of amino acids at the pairwise positions of a protein. The software package is available via the Internet at <http://wwwmgs.bionet.nsc.ru/programs/CRASP>.

By using the software package developed, it is possible to reveal the pairs of protein positions, at which amino acid substitutions occur in a co-adaptive manner; the approach is based on estimation of correlation coefficient between the values of a certain physico-chemical parameter at a pair of positions of multiply aligned proteins. Besides, the package enables the following procedures:

- to analyze the obtained information on pairwise correlations between substitutions;
- to reveal both conserved and variable integral physico-chemical characteristics of a protein;
- to evaluate how conservation (or variability) of revealed characteristics is produced by co-adaptive substitutions of residues.

It should be noted that the software package consists of two modules: the program for analysis of correlations of substitutions at the pairs of protein positions and the program for analysis of integral protein characteristics. In principle, the user can address to each of these two modules independently (in particular, in case the physico-chemical characteristics is *a priori* known by a user). The block-scheme of the software package CRASP is shown in Fig. 1.

General methods and algorithms

The basic methods for evaluation of dispersion, covariation between the values of physico-chemical characteristic of a protein, correlation coefficients between them, and estimation of significance level of dependencies revealed are described in our previous papers [8,9]. Below, we shall briefly describe the realization of these methods in the software package CRASP.

Analysis of pairwise correlations of amino acid substitutions. To reveal correlations of amino acid substitutions at positions of a protein, an analysis is performed of the following protein physico-chemical characteristics: volume of a side group, charge, hydrophobicity, etc. These parameters are chosen for analysis by the user from the current version of the database on physico-chemical amino acid properties. At present, this database contains

the set of 36 characteristics. It is supposed that these characteristics reflect the interactions between the

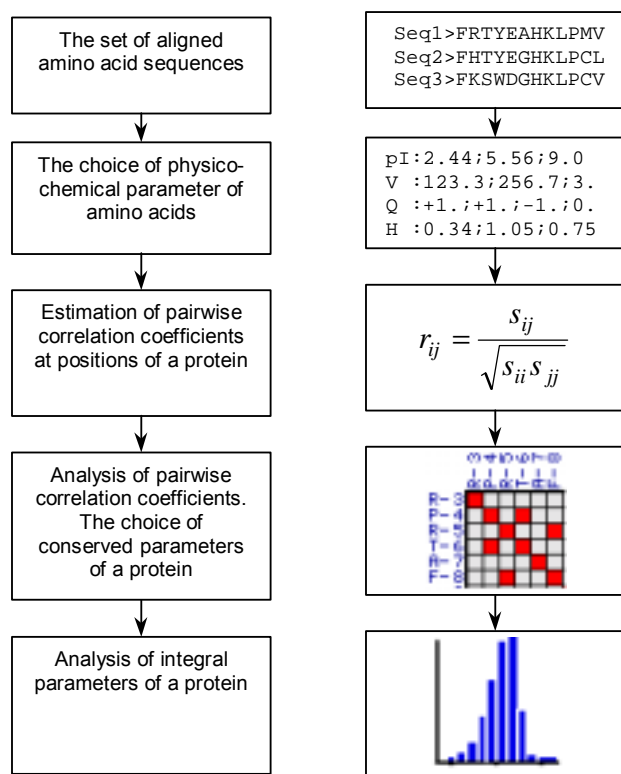


Figure 1. Block-scheme of the CRASP package.

residues in a protein globule. Hence, the revealed regularities between the values of physico-chemical characteristics of amino acids may indicate to existence of specific functionally important interactions between the residues.

For analysis, the set of aligned amino acid sequences of a protein family is used. Next, each type of amino acid within alignment matrix is substituted to corresponding value of the chosen physico-chemical characteristics of amino acids. As the measure of relationships between amino acid substitutions at positions of a protein (columns of the resulted numerical matrix), we use the values of both linear correlation coefficient between the values of physico-chemical characteristics and partial correlation coefficients [8,9]. The partial correlation coefficients enable to evaluate the extent of direct relationships between the pair of protein positions, providing that the residues in the rest positions are fixed.

Data weighting. To take into account the evolutionary relationships of the sequences analyzed, it is possible to use various methods of data weighting in the software package developed. In the CRASP package, the possibility is realized to use the following data weighting schemes: weight calculation following the method by Vingron and Argos [10]; data weighting by accounting phylogenetic relationships in a protein family [9], or

application of weight coefficients input by a user. A user produces the choice of the weighting method.

Analysis of conserved integral characteristics of a protein. Integral characteristics of a protein F is determined as a linear combination of values of a certain physico-chemical property at positions of a protein [9]. For example,

for the sequence with the index k , we get

$$F_k = \sum_{i=1}^L c_i f_{ki},$$

where c_i 's are some real numbers reflecting the impact of the residue at the i -th position into the value of integral characteristic F_k , L is the length of a sequence, f_{ki} is the value of physico-chemical property f for the residue at position i of the sequence k . As the measure of the constancy of the value F , for the set of sequences of a protein family, we use its sample dispersion $D(F)$ [9]. In order to reveal conserved integral characteristics, we suggest to use information about pairwise correlations between the values of physico-chemical property at positions of alignment. Currently, the stage of revealing characteristics is not automated, so, a user makes the choice of integral characteristics personally. As the possible variants of such characteristics, we suggest to use the values of physico-chemical characteristic at the groups of such positions that have the maximal in absolute value correlation coefficients. To reveal such groups, an approach based on cluster analysis is realized in the package CRASP. For all positions analyzed, the procedure of clusterization is made on the base of the following

measure of relatedness of the pair of positions i, j

$$d_{ij} = 1 - |r_{ij}|,$$

that is, this distance is more the less, the most correlated are two positions of a protein. Clusterization is performed according to the nearest neighbor method. The results are given as a binary tree. Each node of this tree unifies two sets of positions (leaf nodes of a tree that are the daughter nodes for a given node); the location of a node in the dendrogram corresponds to the maximal correlation coefficient between all possible pairs compiled of positions from two different sets. In such a way, it is possible to reveal both pairs of correlating positions and the groups of positions, where substitutions are not accidental.

To evaluate the extent of the constancy of the characteristics F , the value of its dispersion is compared to the value expected under assumption of independent substitutions at positions of a protein D_{exp} . This value may be estimated by the formula:

$$D_{\text{exp}}(F) = \sum_{i=1}^L \sum_{j=1}^L c_i c_j r_{ij} D(f_i) D(f_j) = \sum_{i=1}^L c_i^2 D(f_i), (1)$$

where r_{ij} are the linear correlation coefficients of a physico-chemical property at a pair of positions ij (for independent amino acid substitutions, we set $r_{ij}=0$), and $D(f)$ is the value of a sampling dispersion of the value of physico-chemical property f at position i .

In the software package CRASP, for evaluation of the F value constancy, the Monte-Carlo procedure is also used. Based on the model of evolution of physico-chemical protein characteristics, suggested earlier in [9], a large number of samples (up to 10000), with the size equaling to the size N of the set under analysis, is generated in the package CRASP. These samples are obtained by means of the Gaussian distribution of L independent variables, each is being characterized by the mean value and dispersion equal to their estimates for physico-chemical property of amino acids at positions of a protein. For each of these random samples, we estimate dispersions of a characteristic F , $D_{\text{rand}}(F)$ and $D_{\text{exp}}(F)$, and calculate the value $\lambda = D_{\text{rand}}(F)/D_{\text{exp}}(F)$. For the samples obtained, we evaluate the distribution of values $D_{\text{exp}}(F)$, detected by the formula (1). While testing the hypothesis $D(F) \ll D_{\text{exp}}(F)$, the fraction of the samples p such that $D_{\text{rand}}(F) > D_{\text{exp}}(F)$, is an estimate of the significance threshold of false positive estimates. Thus, $1-p$ is an estimation of the significance level of the constancy of physico-chemical character F . Let, for example, during the modeling of 1000 samples, there were

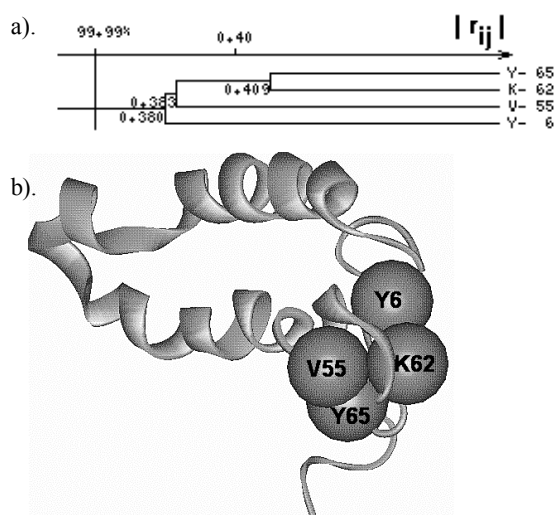


Figure 2. A group of DnaJ domain' positions, residue substitutions at which occur in a co-adaptive manner.

a). Representation of correlated dependency between positions 6, 55, 62 and 65 as a binary tree. The correlation coefficient absolute values are given.

b). Positioning of residues at positions 6,55,62,65 in the spatial structure of DnaJ domain. Coordinates are taken from the work [13] (PDB code - 1bq0). Dark spheres show the side groups of residues at positions indicated above.

found 10 out of them such that $D_{\text{rand}}(F) > D_{\text{exp}}(F)$. In this case, we consider that the constancy of the characteristic F is produced by co-adaptive substitutions of residues at the 99% significance level. Analogous estimations of significance level may be done for verifying the hypothesis $D(F) \gg D_{\text{exp}}(F)$.

Analysis of J-domain 's sequences

By means of the CRASP software package, we have analyzed a seria of protein families (CREB - AP-1 [9], homeodomains [14], experimental data on phage display [15,16]). In the work given, we have performed an analysis of multiple alignment of the J-domain sequences (extracted from the Pfam database, record PF00226 [11]). These domains are located in the N- terminal part of the DnaJ proteins, which are referred to chaperone system [12]. From the sample, the identical sequences and the sequences with a large number of deletions were eliminated, the threshold of variability at positions was equal to 7 different types of amino acids. The results of analysis have revealed that the group of positions corresponding to the numbers 6, 55, 62 and 65 demonstrate significant correlation coefficients between each other according to the value of isoelectric point of amino acids pI (Fig. 2a). Notably, position 65 has positive correlation coefficients with the other positions from this group. On the contrary, positions 6, 62, and 55 have mutually negative correlation coefficients. Interestingly, within the spatial structure of DnaJ domain, the residues at these positions are closely located (Fig. 2b). We have supposed that due to compensatory effect (negative correlations), the total isoelectric point value of amino acids at positions 6, 62, and 55 ($pI_{[6,62,55]}$) is conserved. On the other hand, the positive relationships of pI values at these positions with position 65 should support the constancy of the difference between the values $pI_{[6,62,55]}$ and $pI_{[65]}$. Thus, we suggest that in general, for this group of positions, the co-adaptive substitutions of residues will provide for the constancy of the characteristic $F = pI_6 + pI_{55} + pI_{62} - pI_{65}$. An analysis proved that dispersion of this value $D(F)$ in the sample of DnaJ domains equals to 2.87, which is ~ 2 fold less than the value $D_{\text{exp}}(F) = 5.73$. Numerical modeling gave the evidence that out of 10000 random sets, neither had the $D_{\text{rand}}(F)$ less than $D(F)$. So, we may conclude that the constancy of the characteristics, which we have found, appears due to co-

adaptive amino acid substitutions (at the 99.99% significance level). It may be supposed that the constancy of this characteristic supports stability of mutual packaging of N- and C-terminal regions of the DnaJ domain.

Acknowledgements

The work is supported by the Russian Foundation for Basic Research (grants Nos 98-07-91078, 99-04-49879 and INTAS № INTAS-96-1787). The author is grateful to Galina Orlova for translation of the manuscript into English.

References

1. Shindyalov, I.N., Kolchanov, N.A. and Sander, C. (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Prot. Eng.*, 7, 349-358.
2. Göbel, U., Sander, C., Schneider, R. and Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Prot. Struct. Funct. Genet.*, 18, 309-317.
3. Neher, E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci USA*, 91, 98-102.
4. Thomas, D.J., Casari, G. and Sander, C. (1996) The prediction of protein contacts from multiple sequence alignments. *Protein Eng.*, 9, 941-948.
5. Ortiz, A.R., Kolinski, A. and Skolnick, J. (1998) Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignment. *J. Mol. Biol.*, 277, 419-448.
6. Lim, V.I. and Ptitsyn, O.B. (1970) On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins. *Mol. Biol. (USSR)*, 4, 372-382.
7. Gerstein, M., Sonnhammer, E.L., and Chothia, C. (1994) Volume changes in protein evolution. *J. Mol. Biol.*, 236, 1076-1078.
8. Afonnikov D.A., Kondrakhin Yu.V., and Titov I.I. Detection of correlating DNA-binding positions in CREB and AP-1 family transcription factors. *Mol. Biol. (Mosk.)*, vol. 31, pp. 741-748.
9. Afonnikov, D.A., Oshchepkov D.Yu., Kolchanov N.A. Estimation of variances and covariances of protein physico-chemical characteristics in families of homologous sequences. *Computational Technologies*, in press (2000).
10. Vingron, M. and Argos, P. (1989) A fast and sensitive multiple sequence alignment algorithm. *CABIOS*, 5, 115-121.
11. Sonnhammer E.L., Eddy S.R., Birney E., Bateman A. and Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, 26, 320-322 (1998).
12. Bukau, B., Horwich, A.L. The Hsp70 and Hsp60 chaperone machines. *Cell*, 92, 351-366 (1998).
13. Huang, K., Flanagan, J.M., Prestegard, J.H. The influence of c-terminal extension on the structure of the "J-domain" in E. Coli DnaJ. *Protein Sci.*, 8, 203 (1999).
14. Afonnikov D., Novel functional features of DNA-binding domain of the "homeodomain" class revealed by analysis of correlations of amino acid substitutions in its positions. This issue.
15. Afonnikov D., Wingender E. Statistical relation between positions of the alpha-helix in the Zinc-finger DNA-binding domain: results from the phage display data analysis *Proc. of the international conference on bioinformatics of genome regulation and structure. Novosibirsk, August 24-31, 1998, V.2, p.380-383.*
16. Valuev V.P., Afonnikov D.A., Kashinskaya Ju.O., Orlov Yu.L. The ASPD Database on synthetic peptides. *Computational Technologies*, in press (2000).

NOVEL FUNCTIONAL FEATURES OF DNA-BINDING DOMAIN OF THE “HOMEODOMAIN” CLASS REVEALED BY ANALYSIS OF CORRELATIONS OF AMINO ACID SUBSTITUTIONS IN ITS POSITIONS

Afonnikov D.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: ada@bionet.nsc.ru

Keywords: amino acid sequences, co-adaptive substitutions, DNA-binding, homeodomain

Introduction

The study of the sets of homologous sequences of iso-functional proteins is one of the most important methods of analysis in molecular biology. Under application of this method it is supposed that the function and spatial structure of a protein is conserved in the course of evolution [1]. This means that physico-chemical parameters of a protein, which are responsible for specific packing of a polypeptide chain along with the functional features of a protein in the course of evolution, should also be sustained at a constant level [2]. Revealing of such conservative characteristics in analysis of homologous sequences could essentially increase our knowledge on function, structure, and evolution of the proteins analyzed.

One of the possible mechanisms providing for the constancy of conservative characteristics is an existence of co-adaptive amino acid substitutions in a protein sequence [3]. These are substitutions in pairs or groups of positions of a protein, which are evolved in a related manner (co-adaptively) because of the interactions between residues. In particular, for the residues, closely located in a protein globule, these substitutions may be of compensatory nature [3-6].

In the present paper, we have made an analysis of co-adaptive substitutions at pairs of positions of DNA-binding domain of the "homeodomain" class [7]. We have used an approach based on analysis of correlations between the values of physico-chemical characteristics of amino acid residues at the protein positions in the protein family [9,9]. We have made a comparison of the results obtained with the data reported by Clarke [10], who has earlier used for analysis of co-adaptive substitutions in homeodomains an approach based on the information theory. The results of our analysis, have enabled also novel functional feature of homeodomains, namely, the constancy of the total charge of residues in the contact region of two alpha-helices.

Materials and methods

The sample of homeodomain sequences is based on the data extracted from the Pfam database [11], record PF00046. After removing the sequences with the large amount of deletions or insertions, along with identical sequences, the sample contained $N=372$ proteins (the sample is available via the e-mail by request). Positions of multiple sequence alignment with the number of different amino acids less than 5 were not taken into analysis. Thus, the number of positions analyzed is equal to $L=47$.

In order to evaluate dispersions and co-variations of physico-chemical characteristics of a protein, we have applied an approach suggested earlier [9]. A phylogenetic tree of homeodomains, which is necessary to evaluate these parameters, was constructed by the program CLUSTALW [12]. As the measure of relatedness of residue substitutions at the pair of a protein positions, we have used partial correlation coefficients [9],

$r_{ij} = -a_{ij} / \sqrt{a_{ii} a_{jj}}$, where a_{ij} are the elements of a matrix reverse to covariation matrix. Its value allows to estimate a direct relationship between positions i and j of the protein. As a physico-chemical characteristic of amino acid residues, we used isoelectric point value pI of amino acids [13], this value being a characteristic of amino acid's charge.

In order to evaluate robustness (stability) of correlation coefficients obtained, we have generated 300 samples, which contained the sequences chosen at random from initial sample. The size of the reduced sample was determined randomly too and was set within the limits from $0.8N$ to N . For each reduced sample, the pairwise partial correlation coefficients of physico-chemical characteristics of amino acids at positions of a protein were estimated. Then the following parameters were calculated: 1) rs_{mean} , the mean value of correlation coefficient r_{ij} for particular pair of positions and 2) rs_{disp} , the dispersion of the ratio of correlation coefficient r_{ij} to the expected root-mean square deviation of the correlation coefficient value at independent positions of a protein $rs = r_{ij} / \sqrt{1/(N-1)}$. The definition of a parameter rs is based on the fact that dispersion of distribution of

correlation coefficient values for independent variables is determined as $D_0=1/(N-1)$ [14]. The more is dispersion $r_{S_{disp}}$, the less reliable is correlation coefficient evaluation.

Results and discussion

The obtained partial correlation coefficient estimates we have compared, first of all, to the data by Clarke [10] (see Table 1). Out of 16 pairs of the most correlating residues detected in the work [10], 6 have significant ($P>95\%$) robust estimates of correlation coefficients of isoelectric point values of amino acids, according to our data. Among this number, following [10], two pairs of residues form salt bridges (19-30;17-52), three pairs act in a network of DNA interactions (44-46;46-47;46-54), whereas the pair of residues 8-25 demonstrates no direct interactions; and functional role of correlations is unclear in this case. For three pairs of positions, due to our estimates, correlation coefficients are significant, but, however, are not sufficiently robust (26-46, 26-44 positions are involved in the network of DNA interactions; 31-42 - act in competing interactions with DNA phosphate group). The rest 7 pairs indicated by Clarke [10], have no significant correlations according to our data. Among them are four pairs (6-47, 12-50, 8-43, 28-54), for which there was found no functional interpretation in the work [10].

Sufficient agreement between two different approaches for detection of co-adaptive substitutions proves the evidence that that such substitutions in homeodomains have really occurred in the course of evolution. Interestingly, the best coincidence was found for the residues forming salt bridges.

By our analysis, we have also found pairs of positions that have significant correlation coefficients but that were not registered in the work [10] as correlating ones. The most significant and robust of them are shown in Table 2.

Two pairs of residues (52-42 and 50-41) are located at the alpha-helix H3 of a homeodomain, which fits into the major DNA groove (see Fig. 1). More likely, correlations of these residues are determined by their interactions with DNA. The pair of residues 29-4 has no direct contact with DNA. However, these residues are located near by phosphate groups of the opposite DNA threads. Interestingly, the location of these residues is similar to that of the pair 8-25. Functional role of these correlations stays unclear. Nevertheless, one may suppose that correlation of residues in these positions is governed by their interaction with the phosphate skeleton of DNA and support the constancy of a homeodomain orientation relatively DNA.

The more interesting are the correlations of isoelectric point values of amino acids at positions 37-15 and 18-15 (Table 2). Analogously to the pair of residues 19-30, they are located in the region of the interface of two helices, H1 and H2, of a homeodomain (Fig. 2). Between these positions are prevailing the significant negative correlation coefficients. The functional role of these three pairs of residues is related, probably, to stabilization of the helices packaging due to electrostatic interactions. All these residues are located closely enough to form the salt bridges (the distance between the charged atoms equals from 2.9 to 4.0 Å and satisfies to the salt bridge criterion [16]).

It was suggested that such interactions might provide for the constancy of the total pl value due to co-adaptive substitutions of residues at these positions of homeodomain. An analysis has revealed that in the sample under study, dispersion D of the total values pl at these positions equals to 46.56. This value is essentially less than could be expected by random for independent substitutions of residues at positions considered. The expected value D_{exp} could be estimated as the sum of dispersions pl at the indicated positions of a protein, that is, equal to 70.22. The ratio D_{exp}/D equals to 1.53.

To evaluate the significance of detected difference between D_{exp} and D , we have performed the Monte-Karlo test within the framework of the model of evolution implemented to estimate the variances and co-variances of

Table 1. Results of calculation of correlation coefficients for the most related pairs of positions of the homeodomain described in [10].

Pair of positions	r	$r_{S_{disp}}$
19-30	-0.333	0.237
28-54	-0.085	0.080
44-46	-0.169	0.131
8-43	+0.079	0.126
46-47	+0.112	0.103
26-46	-0.255	0.789
26-47	+0.020	0.420
44-47	+0.094	0.263
46-50	+0.027	0.290
26-44	+0.267	0.871
12-50	+0.030	0.080
17-52	-0.273	0.289
31-42	-0.295	0.833
6-47	-0.021	0.069
8-25	+0.327	0.260
46-54	-0.110	0.285

For each pair, the partial correlation coefficient of isoelectric point value r of an amino acid and an estimation of its robustness $r_{S_{disp}}$ are given. The pairs of positions with the most robust correlation coefficients exceeding the 95% significance level ($|r_j|>0.11$) are marked by bold. The estimation of correlation coefficient was considered to be robust if the value $r_{S_{disp}}$ falls among the 95% estimates less in values ($r_{S_{disp}}<0.405$). The numeration of positions is given according to [10].

Table 2. The pairs of residues of a homeodomain, which have significant ($P>99.99\%$) correlation coefficients r .

Pair of positions	r	$r_{S_{disp}}$
37-15	-0.274	0.142
18-15	-0.228	0.197
29-4	-0.274	0.101
52-42	-0.226	0.263
50-41	+0.218	0.260

For these pairs, the estimates of $r_{S_{disp}}$ are also shown.

protein physico-chemical characteristics [9]. We have generated 1000 samples of independent Gaussian variables, such that their means and dispersions equal to those at positions of the sample of homeodomains. Within these samples, we have evaluated the parameters D and D_{exp} . As was found, in neither random sample, the value D was less than in real proteins. The ratio D_{exp}/D did not exceed the value calculated for the real sample of homeodomains too.

Since pI value characterizes an amino

acid charge, a conclusion follows from the data presented that the value of the total charge in positions from the inter-helical contact region of a homeodomain is retains at the constant level in the course of evolution. This fact may reflect the functional importance of this characteristic. It may be supposed that the constancy of a charge provides stability of H1 and H2 homeodomain helices packing, and, therefore, stability of DNA-protein complex in a whole.

Acknowledgements

The work is supported by the Russian Foundation for Basic Research (grants Nos 97-04-49740, 98-07-91078, 99-04-49879). The author is grateful to Galina Orlova for translation of the manuscript into English.

References

1. Chothia C. and Lesk A.M. The relation between divergence of sequence and structure in proteins. *EMBO J.*, 5, 823-826 (1986).
2. Gerstein, M., Sonnhammer, E.L. and Chothia, C. Volume changes in protein evolution. *J. Mol. Biol.*, 236, 1076-1078 (1994).
3. V.I. Lim and O.B. Ptitsyn, On the constancy of the hydrophobic nucleus volume in molecules of the myoglobins and hemoglobins, *Mol. Biol. (USSR)*, 4, 372 (1970).
4. Göbel U., Sander C., Schneider R. and Valencia A. Correlated mutations and residue contacts in proteins. *Prot. Struct. Funct. Genet.*, 18, 309-317 (1994).
5. Shindyalov I.N., Kolchanov N.A. and Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Prot. Eng.*, 7, 349-358, (1994).
6. Altshuh D., Lesk A.M., Bloomer A.C. and Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.*, 193, 693-707 (1987).
7. Burglin, T.R. A comprehensive classification of homeobox genes. in *Guidebook to the homeobox genes*, Duboule D. Ed. A Sambrook & Toose Publication at Oxford University Press, 1994, pp. 27-71.
8. Afonnikov, D.A., Oshchepkov D.Yu., Kolchanov N.A. Estimation of variances and covariances of protein physico-chemical characteristics in families of homologous sequences. *Numerical Technologies*, in press (2000).
9. Afonnikov, D.A., Kondrakhin, Yu.V. and Titov, I.I. Revealing of correlated positions of the DNA-binding region of the CREB and AP-1 transcription factor families. *Mol. Biol. (Russian)*, 31, 741-748 (1997).
10. Clarke N.D. Covariation of residues in homeodomain sequence family. *Prot. Sci.*, 4, 2269-2278 (1995).
11. Sonnhammer E.L., Eddy S.R., Birney E., Bateman A. and Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, 26, 320-322 (1998).
12. Thompson, J.D., Higgins, D.G. and Gibson, T.J. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap-penalties and weight matrix choice. *Nucl. Acids Res.*, 22, 4673-4680 (1994).
13. White, A., Handler, P., Smith, E.L., Hill, R.L. and Lehman, I.R. *Principles of Biochemistry*, vol. 1., McGraw-Hill, Inc (1978).
14. Anderson, T.W., An introduction to multivariate statistical analysis, John Wiley & Sons Inc., NY (1958).
15. Kissinger C.R., Liu B., Martin-Blanco E., Kornberg T.B. and Pabo C.O. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. // *Cell*. 1990. V. 63. P. 579-590.
16. Barlow D.J. and Thornton J.M. Ion pairs in proteins. // *J. Mol. Biol.* 1983. V. 168. P. 867-885.

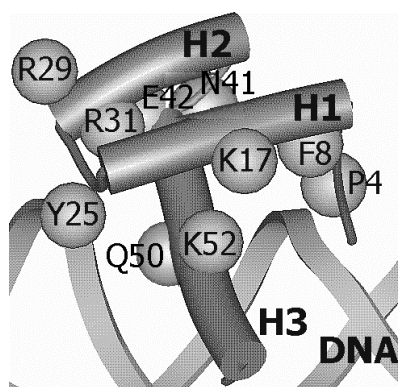


Figure 1. The spatial structure of the homeodomain-DNA complex. The coordinates of a complex are taken from the work [15]. The designations of alpha-helices are given. The spheres denote the side groups of residues in positions with significant correlations of amino acid substitutions (Tables 1,2).

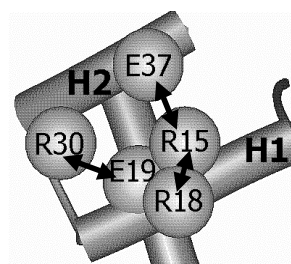


Figure 2. Schematic representation of residues in the region, where H1 and H2 helices of homeodomain are in contact. The pairs of residues that are substituted "co-adaptively" are linked by arrows.

ANALYSIS OF STRUCTURAL MOTIFS IN PROTEINS

¹Jianghong A.N., ²Wako H., ¹*Sarai A.

¹RIKEN Tsukuba Institute, Tsukuba, Japan

²School of Social Sciences, Waseda University, Tokyo, Japan

e-mail: sarai@rtc.riken.go.jp

*Corresponding author

Keywords: protein, motif, structural motif, three-dimensional structure, Delaunay tessellation

Resume

Motivation:

The function of proteins is usually achieved by particular conformations of their substructures, which are called motifs, but the motifs are often represented as patterns of sequence, instead of three-dimensional structures (3D structures). In order to investigate the relationship between function and motif more quantitatively, a novel representation of structural motifs of proteins was proposed.

Methods and Results:

A program uniquely coding a 3D structure by a set of digital string has been developed. Patterns from PROSITE were first mapped to the sequences of proteins in PDB. Then, the 3D structures corresponding to the sequence motifs were coded by the program. The analysis showed that the conformations of the structural motifs were well characterize by these codes.

Availability:

The program is available on request from the authors.

Introduction

The function of proteins is usually associated with some specific sites, such as active sites in enzymes and ligand binding sites. Thus, the 3D conformations of these sites (motifs) are one of the most important representations of the function. However, the motif information has been mostly collected in a form of sequence patterns or profiles, as in popular motif databases such as PROSITE (Bairoch et al., 1997), Blocks (Henikoff & Henikoff, 1994) and PRINT (Attwood et al., 1998). On the other hand, a large number of structures of protein are known: the number of entries in the Protein Data Bank (PDB) (Bernstein et al., 1977) is over 10,000 now, and the number will increase further by the progress of structural genomics. Detection of motifs from structural aspect will become an important and interesting subject.

Methods and algorithms

Collecting structural motifs from PDB

By mapping sequence patterns of PROSITE to the 3D structures of PDB, the corresponding structures of the patterns can be collected. We call these corresponding local structures structural motifs. Because one motif can be found in many proteins, a set of local structures is assigned to it. We also collect the structural motifs from the ligand binding sites. The common motifs are collected from the PDB structures with the same or similar ligands.

Coding structural motifs

The coding program of structural motifs is based on a novel method called Delaunay tessellation (Wako et al., 1998), which uniquely divides the interior space of a structural motif into non-overlapping volume elements, named Delaunay tetrahedron whose vertices are the C α atom positions. Then one unique code (string of digits) is assigned to each tetrahedron according to the vertex residues and four surrounding tetrahedrons. Therefore, each structural motif can be represented as a set of codes.

Analysis of structural motifs

For each pattern, its structural motifs collected from different PDB entries were investigated. In order to analyze the conformations of structural motifs, distance between any two code sets was measured using the following formula:

$Distance(A,B) = 1 - |A \cap B| / |A * B|$, where A, B are the code sets and the distance is between 0 and 1.

Then, all structural motifs for each pattern were clustered according to the distance. We observed that some patterns have almost stable conformations but other patterns change their conformations in different proteins.

Figure 1 shows an example of clustering of the legume_lectins alpha-chain signature, which is a 10 residues pattern (PROSITE code PS00308). In this example, the motif has very similar conformations in many proteins, whereas in a few rear cases it has different structures. The result is almost consistent with Kasuya's investigation, in which the conformations of PROSITE patterns were checked and the rmsd values were calculated (Kasuya & Thornton, 1999).

Discussion

Structural flexibility of proteins introduces some noises in the coding. In order to represent the structural motif properly the coding program should be refined further. We are now applying the method to ligand binding sites of proteins, for which we expect better results because the conformations of specific ligand binding sites may be more stable. The codes can be used to detect new structural motifs from protein if the whole structure of the protein was coded. Furthermore, because the codes represent the 3D structures directly, motifs extending over more than one chain can be detected as well.

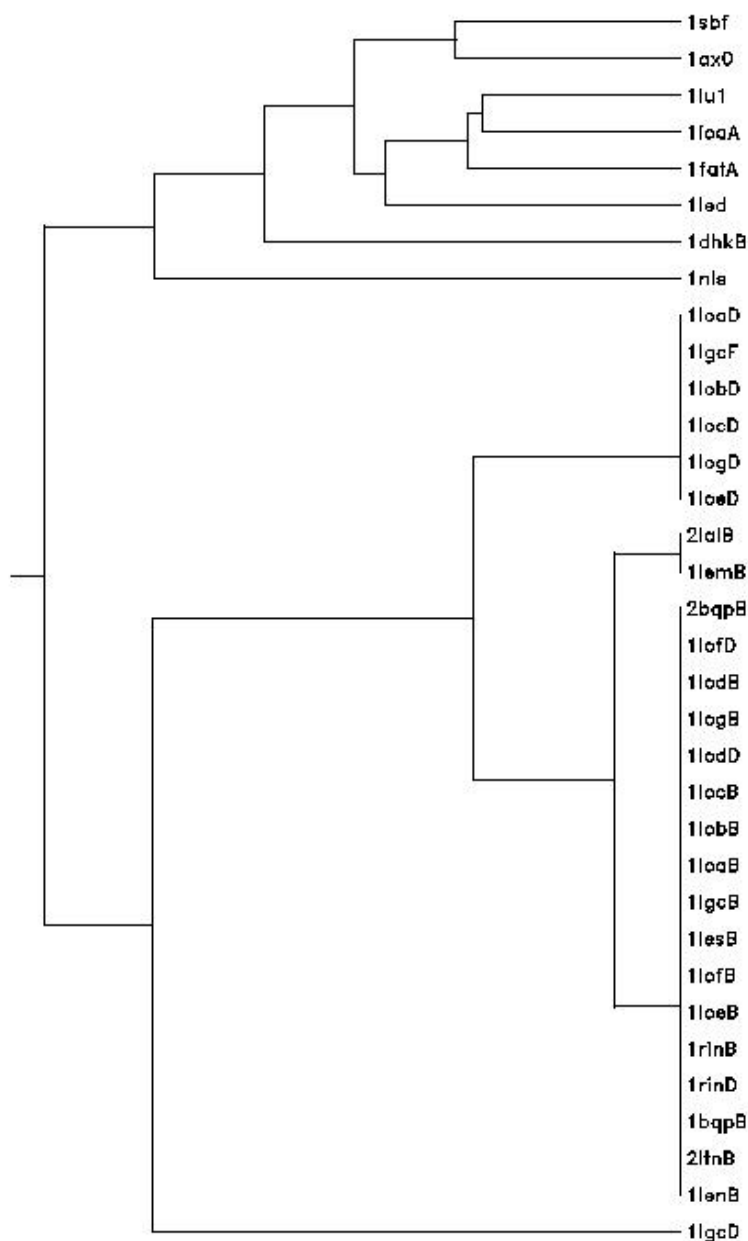


Figure 1. Clustering of the legume_lectins alpha-chain signature, which is a 10 residues pattern (PROSITE code PS00308). The labels stand for the PDB entries in which the motif is found. The length of branches is in proportion to the distance. In three groups, the distance among the members is equal to zero.

References

1. Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, 25,217-221.
2. Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, 19, 97-107.
3. Attwood,T.K., Beck, M.E., Flower, D.R., Scordis, P. and Selley, J. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* 26, 304-308.
4. Bernstein,F.C., Koetzle, T.F., Williams, G.J.B., Meyer,E.F., Brice, M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112, 535-542.
5. Wako,H. and Yamato,T. (1998). Novel method to detect a motif of local structures in different protein conformations. *Protein Eng.*, 11, 981-990.
6. Kasuya,A. and Thornton,J.M. (1999). Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, 286, 1673-1691.

'IN SILICO' ANALYSIS OF POINT MUTATION EFFECTS ON THE CODING REGION OF HUMAN BETA GLOBIN GENE

**¹Arrigo P., ²Ivaldi G., ³Cardo P.P.*

¹C.N.R. Istituto Circuiti Elettronici, Genova, Italy

²E.O Ospedali Galliera, Laboratorio di Genetica Umana, Genova, Italy

³Universita' degli Studi di Genova Dipartimento Scienze della, Salute, Italy

e-mail: arrigo@ice.ge.cnr.it

*Corresponding author

Keywords: biosequences, molecular screening, gene expression, mutations, disease, energy landscape

Resume

The completion of the human genome sequencing tasks leave the problem to investigate the functionality of the biosequences open; the discovery of the properties, that are closely related to the primary structure of informational macromolecules, is a very important task for the fall out of the genomic research in the molecular screening of the diseases. In the new phase of the international genome initiatives greater attention is addressed to study the differences in the gene expression.

The mutational event are one of the more extensively studied source of variability: a single base variation can dramatically modify, by a structural change, the level expression both influencing the efficiency of the informational flow and by modification of the molecular phenotype.

These very small nucleotide changes are acquiring relevance in the molecular screening of many diseases. Many variations are not lethal for an organism, But they can modify a specific pathway or can favour subsequent mutations. In the present paper we have applied a computer assisted approach to build the potential landscape of a gene specific cDNA coding sequence. The 'landscape' approach is a physical approach to characterise the molecular structure under a conformational dynamics, for instance the protein folding. An energy landscape represent the energy peaks and valley that are correlated to the change in the metastable conformation. The complexity of a landscape depend by the number atoms in the molecule. The distribution of peak and valley seem to be not uniform. In the present paper we acquired the concept of landscape in order to analyse the effects of point mutations.

The proposed method allows to have a global view of the sequence and consequently it is quite easy to detect the effects of small mutational events. The comparison, between the mutated profile and the reference one, allows to recognise the changes induced by a single base variation inside the coding region; these modification can help to evaluate the efficiency of translation of the messenger.

In the present paper we have analysed a sample of 57 pathological human beta haemoglobin cDNA chains, these mutations are commonly known like functionally 'unstable', because they show abnormal physico-chemical properties. These variants has been selected according the following constrains: they are characterised by a single base substitution (not deletion or insertion) and they are supported by a biochemical screening in the italian population.

The results of our analysis point out that a single base modification can induce long range modification in the pseudo-potential landscape of a sequence: the 'in silico' results agree with the experimental evidences about the relevance of the two coding region ends in the traslational stability. The processed mutation can be classified into two distinct groups: the first shows a downstream shift of the minimal potential domain at the 5' end, the second group present an insertion of an unusual domain strongly associated with the initiator codon.

These findings suggest the possibility to integrate the traditional classification of haemoglobin variants with information about the structural modification induced by the event. This 'in silico' approach allows also to select mutation specific signature and it can help to plan the experimental activity for for mRNA stability essay.

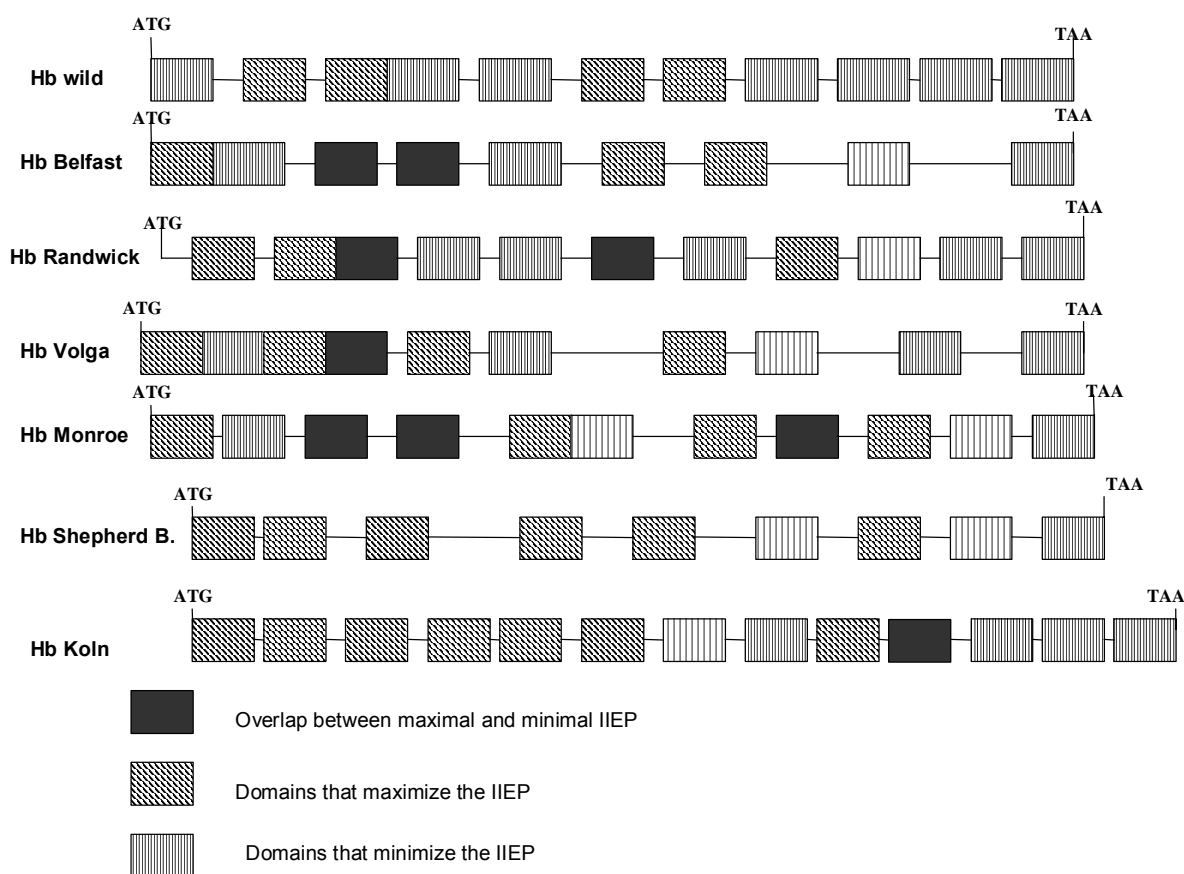


Figure 1. Sample of IIEP landscape for Hb variants.

References

1. P.Hieter, M.Boguski: 'Functional genomics: it's all how you read', *Science* vol.278 pp.601-602 1997.
2. Gu Z.,Ladeana H.,Kwok P.: 'Single nucleotide polymorphism hunting in cyberspace', *Human Mutation* vol. 12 221-228 1998.
3. Hardison R.C. et al.: 'A Syllabus of Human Hemoglobin variants (1996) via the World Wide Web'. *Hemoglobin* vol.22(2) pp. 113-127 1998.
4. Toronen P. et al.: 'Analysis of gene expression data using self-organizing maps', *FEBS Letters* vol.451 142-146 1999.
5. Arrigo P. et al.: 'Can functional regions of proteins be predicted from their coding sequences? The case study of G-protein coupled receptors'. *Gene*. 1998 Oct 9;221(1):GC 65-110.
6. Liebhaber S.A et al. : ' Translation inhibition by an m-RNA coding region secondary structure is determined by its proximity to the AUG initiation codon' *J. Mol. Biol.* Vol.226

DESIGN AND IMPLEMENTATION OF THERMODYNAMIC DATABASE FOR PROTEIN-NUCLEIC ACID INTERACTIONS

Prabakaran P., An J., Gromiha M.M., Selvaraj S., Uedaira H., ¹Kono H. and *Sarai A.

RIKEN Tsukuba Institute, Japan

¹Department of Chemistry, University of Pennsylvania, USA

e-mail: sarai@rtc.riken.go.jp

*Corresponding author

Keywords: thermodynamic data, database, protein-nucleic acid interaction

Resume

Motivation:

There has been a tremendous explosion of structural and thermodynamic data for biomolecules and their interactions attempting to understand their structure, function and properties. Protein-nucleic acid interaction plays an important role in the regulation of gene expression. Although there are biomolecular databases documenting details of sequence and structural aspects, there exists no database for protein-nucleic acid binding. Thus, we have decided to develop the Thermodynamic Database for Protein-Nucleic Acid Interactions to help researchers understand the mechanism of protein-nucleic acid recognition.

Results:

The database contains thermodynamic data, experimental conditions, structural information of proteins, nucleic acids and of the complex, and literature information. All other archives needed for acquiring a thorough knowledge about the protein-nucleic acid recognition such as PDB, NDB, PMD, EC, PIR, ProTherm, 3DinSight and NCBI PUBMED are hypertext-linked. The World Wide Web (WWW) interface allows users to search data based on various conditions, to extract information about protein, nucleic acid, and of the complex and their conformational changes upon binding, and to visualize molecular structures and their interactions.

Availability:

The database is freely accessible through the Internet (<http://www.rtc.riken.go.jp>).

Introduction

For the past few decades, due to the developments in molecular biology and computer technology, attempts have been made to catalog and systematize the accumulation of biological data through various databases and archives. For example, PDB (Bernstein *et al.*, 1977) and NDB (Berman *et al.*) for three-dimensional structures, SWISS-PROT (Bairoch and Apweiler, 1997), PIR (George *et al.*, 1997), EMBL (Stoesser *et al.*, 1997) and GenBank (Benson *et al.*, 1997) for sequences, and other functional databases like PROSITE (Bairoch *et al.*, 1997) etc., along with different visualization and analysis tools are available. However, there is a paucity for databases which describe the stability, specificity and energetic components of biomolecular interactions. Recently, Pfeil (1998) compiled a set of thermodynamic data on protein stability and folding. We have already taken the effort to develop and maintain the on-line Thermodynamic Database for Proteins and Mutants (ProTherm) (Gromiha *et al.*, 1999, 2000). Herein, we have constructed the Thermodynamic Database for Protein-Nucleic Acid Interactions, which contains several thermodynamic data of binding, experimental conditions, structural information, and literature information efficiently interfaced with major on-line resources related to the data available in the database. A WWW interface helps user to search the data with required conditions and get results with desired sorting formats.

Design and Implementation

Each entry of the database is referred by a serial number and has six distinct sections with the following information.

1. *Protein information:* name, source, fragment and sequence of the protein, enzyme code (EC), Protein Information Resource (PIR) code, Protein Data Bank (PDB) code for wild and mutant structures, information about monomeric, oligomeric states, ProTherm number, details of mutation with mutant residue name, number and secondary structure at the mutant sites, accessibility and Protein Mutant Database (PMD) number.
2. *Nucleic acid information:* name, source and sequence of the nucleic acid, information on mutation and sequence of mutant nucleic acid, GenBank number and Nucleic Acid Database (NDB) number.

3. *Complex Information*: codes for PDB and NDB, Protein-Nucleic Acid Complex Database number, details of ligand molecules, accessibility of relevant mutant residue in the complex and conformational changes of protein and nucleic acid upon binding.
4. *Experimental condition*: temperature (T), the pH value, details about buffer, ions additives and experimental method.
5. *Binding data*: dissociation constant (K_d), association constant (K_a), Gibbs free energy (ΔG), enthalpy change (ΔH) and heat capacity change (ΔC_p) for wild and mutant entities, stoichiometry, activity (k_m and k_{cat}).
6. *Literature*: reference, authors, keywords and remarks.

Thus, for each binding data at a given experimental condition, all other relevant information for the data are given in a single table as mentioned above and are properly cross-linked with the major biological databases through the corresponding codes or entry numbers. In addition, the mutation sites and surrounding residues, base-amino acid interactions etc., can be visualized and analyzed through 3DinSight, an integrated relational database and search tool for the structure, function and properties of biomolecules (An et al., 1998).

Search, Display and Sorting options

A WWW interface can be used to search data for various conditions with different sorting options for output according to user's purpose and convenience. The binding database can be searched in a variety of ways as below.

- i. retrieving data for a particular protein by its name, source, sequence or PDB code along with the specification of mutation type, secondary structure and the accessibility (ASA in % or \AA^2) range
- ii. retrieving data for a nucleic acid by mentioning the name, source, sequence or NDB code
- iii. retrieving the complex information by providing ligand name, conformational changes in protein or nucleic acid on binding, or PDB code
- iv. extracting data based on various experimental conditions, namely, method, T , pH , buffer, ion or additives and a particular range of values can be specified
- v. selecting data based on various binding parameters such as K_d , K_a , ΔG , ΔH and ΔC_p with a preferred range of values and
- vi. searching by author name, keywords and year of publication

Finally, the output format can be specified by selecting various display options and by sorting with ASA, T , ion concentration, K_d and year of publication.

Discussion

The current state of the database with more than 1000 entries by manual input, followed by the complete checking provides the opportunity to search and analyze the thermodynamic data of protein-nucleic acid binding. The secondary structure and solvent accessible surface area at the mutation sites, base-amino acid interactions, comparison of conformation in the complex and that of free protein and nucleic acid and visualization of the molecules etc., supplemented with the thermodynamic data are the characteristic features of the database. Hence, the Thermodynamic Database for Protein-Nucleic Acid Interactions will be a very useful database to help researchers understand the mechanism of protein-nucleic acid recognition.

References

1. An, J., Nakama, T., Kubota, Y. and Sarai, A. (1998) 3DinSight: an integrated database and search tool for the structure, function and properties of biomolecules. *Bioinformatics*, 14, 188-195.
2. Bairoch, A. and Apweiler, R. (1997) The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.*, 25, 31-36.
3. Bairoch, A., Bucher, P. and Hofmann, K. (1997) The PROSITE database, its status in 1997. *NAR.*, 25, 217-221.
4. Benson, D.A., Boguski, M.S., Lipman, D. and Ostell, J. (1997) GenBank. *Nucleic Acids Res.*, 25, 1-6.
5. Berman, H.M., Zardecki, C., Westbrook, J. (1998) The Nucleic Acid Database: A resource for nucleic acid science. *Acta Crystallogr D* 54, 1095-1104.
6. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112, 535-542.
7. George, D.G. et al. (1997) The protein information resource (PIR) and the PIR-international sequence database. *Nucleic Acids Res.*, 25, 24-27.
8. Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H. and Sarai, A. (1999) ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.*, 27, 286-288.
9. Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., Prabakaran, P. and Sarai, A. (2000) ProTherm, Version 2.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, 28, 283-285.
10. Pefeil, W. (1998) Protein Stability and Folding: A Collection of Thermodynamic Data. Springer, NY.
11. Stoesser, G., Strek, P., Tuli, M.A., Stoehr, P.J. and Cameron, G.N. (1997) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, 25, 7-14.

TOWARDS A STRUCTURAL BASIS OF HUMAN NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS

^{1,2,3}*Sunyaev S.*, ³*Ramensky V.*, ^{1,2}*Bork P.*

¹European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany

²Max Delbrück Center for Molecular Medicine (MDC) Robert-Roessle-Strasse 10, D-13122 Berlin, Germany

³Engelhardt Institute of Molecular Biology, Vavilova 32, 117984 Moscow, Russia

*Corresponding author

Keywords: SNP, disease, mutations, proteins, genetic and phenotypic variations

Resume

About 90% of human genetic variety has been ascribed to single nucleotide polymorphism (SNP) allelic variants with a frequency higher than one percent. Due to the application of high throughput SNP detection techniques, the number of identified SNPs is growing rapidly and this allows detailed statistical studies. This also applies to SNPs that affect the amino acid sequence of a gene product (non-synonymous SNPs); they complement the large body of literature on mutations, causing Mendelian diseases, that represent the usually rare non-synonymous mutations with an allele frequency far below one percent.

To understand the relation between genetic and phenotypic variations, it is essential to assess the structural consequences of the respective non-synonymous mutations in proteins. To quantify how often a disease phenotype can be explained by a destructive effect on protein structures or functions, we have mapped known disease mutations onto known three-dimensional structures of proteins. The results were compared with a control set of substitutions observed between these proteins and their closely related homologues from other species which are unlikely to cause severe effects on the phenotype. With the knowledge about the structural properties of these two sets, we have also mapped a large number of non-synonymous SNPs (which are usually thought to be neutral or only cause minor phenotypic effects) onto protein structures. This enables us to estimate a lower limit for the quantity of non-synonymous SNPs with likely phenotypic effects which is an important baseline for the current efforts to identify SNPs associated with multifactorial human disorders.

The three data sets needed for the comparative analysis: (1) disease-causing mutations, (2) substitutions between close homologues in human and other species and (3) human non-synonymous SNPs, as well as structural characteristics of corresponding proteins were extracted from public databases.

As a result of the comparison of disease-causing mutations with between-species substitutions in the same set of proteins we found that disease-causing mutations are much more likely to occur at sites with low solvent accessibility. In fact, 35% of 551 disease-causing mutations from our dataset affect buried sites while only 9% of 225 substitutions between species do. This indicates that disease-causing mutations often affect intrinsic structural features of proteins. To increase the discrimination between the two sets we also took into account possible interaction sites. Overall, about 70% of the disease-causing mutations are located in sites likely to be structurally and functionally important, namely sites with less than 5% solvent accessibility or in β -strands, active sites, sites involved in disulphide bonds or evolutionary conservative sites (defined as sites with HSSP variability parameter VAR <10). In contrast, in the same set of proteins only 17% of substitutions observed between human sequences and closely related homologues from mammalian species are located at these sites.

Unexpectedly, the fraction of polymorphic sites located in structurally and functionally important regions (defined as described above) was 45%, which is significantly higher than the 24% in the case of the interspecies variation when considering proteins from the dataset of polymorphic sites (P -value of the χ^2 -test equals to 0.00013). In this set we observe the abundance of immune system-related proteins with high β -strand content; this fact explains the 17% vs. 24% difference for two protein sets. The result above suggests that a significant fraction of human protein allelic variants is represented by amino acid substitutions with a strong impact on protein structure, function, stability or folding. These variants are normally eliminated during long evolutionary times as can be seen from the comparison with the interspecies variation. One would expect, that variants under pressure of purifying selection tend to have a lower allele frequency. Indeed, for non-synonymous SNPs we observe a correlation between allele frequency and fraction of occurrence in structurally and functionally important regions. The observation that many non-synonymous SNPs are likely to have a phenotypic effect may be considered as indirect evidence that common amino acid variants may contribute to genetic risk of common human disorders (so called the common disease-common variant hypothesis).

In summary, there is a surprisingly high fraction of non-synonymous SNPs that affect structure and, probably, function of proteins. This implies that a considerable fraction of the non-synonymous SNPs have indeed some (probably negative) effect on the phenotype. The allele frequency distribution makes it evident that variants in structurally important sites are not selectively neutral. Taking these observations and given the progress in structural genomics as well as in large scale SNP discovery, the comparative analysis of structural properties of protein allelic variants such as described above should have an important role in the pre-selection of candidates for disease-association studies and will help in the explanation of phenotypic effects.

THE DATABASE ASPD ON EXPERIMENTS WITH APPLICATION OF PHAGE DISPLAY TECHNIQUE

**Valuev V. P., Afonnikov D.A., Petrenko O., Beylina A.G., Lokhova I.V., Grigorovich D.A., Fokin O.N., Ivanisenko V.A.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: valuev@bionet.nsc.ru

*Corresponding author

Keywords: phage display, database, synthetic peptides, antibodies

Resume

Motivation:

Phage display is one of the most intensively evolving and promising techniques of today's molecular biology. It allows for exhaustive search for synthetic peptides performing certain function, for example, antibodies, DNA-binding domains etc. At present time it is necessary to put in order accumulated data and make it available both for theoretical analysis and for experiment planning.

Results:

A database is developed that comprises description of more than 100 experiments. This database is searchable and navigable under SRS. The analysis of data is possible in order to retrieve pairwise correlations, build sequence profiles, predict the activity etc.

Availability:

<http://www.mgs/mgs/systems/fastprot/>

Introduction

Phage display is one of the most intensively evolving and promising techniques of today's molecular biology. First this method was proposed in 1985 (Smith, 1985). Since then the number of works with application of phage display is increasing steadily and by now equals to about 1000 papers.

The essence of the method consists of use of filamentous bacteriophage as a means for peptide library display. These libraries consist of short (of length of about 10 or in extreme case, few tens of amino acids) peptides or proteins, where all or some positions are fully or partly randomized. Obtained this way libraries contain $\sim 10^7 - 10^9$ sequences. They are exposed to the phage capsid surface by means of inserting corresponding DNA fragment in the gene coding for the capsid protein. Then the phage selection on affinity for binding with specific targets (receptors, antibodies, DNA regulatory regions etc.) and amplification of the bound phage is carried out. To recover the phage particles bound the target molecules are bound beforehand with biotin and after incubation with phage library are exposed to the streptavidin-coated dish. After several selection/amplification cycles the best binding phages can be isolated and corresponding peptides can be sequenced.

The filamentous bacteriophage contains a cyclic one-stranded DNA molecule. Its genome counts about 6500 nucleotides and encodes for 10 proteins, 5 of which are structural elements of the capsid (pIII, pVI, pVII, pVIII, and pIX), 3 are necessary for DNA synthesis and 2 perform poorly understood assemblage functions.

A phage particle looks like an elongated flexible tube made of several thousands (2700 for fd phage and 5000 for M13) of copies of pVIII protein. About 5 copies of each of minor capsid proteins pIII and pIV are found on one end of the tube and pVII and pIX on the other. Inserts are made into pIII and pVIII. The first consists of 406 amino acids and includes two domains. The second consists of 50 amino acids and has 4 structural units. In both cases difficulties at insertion of large enough polypeptides are overcome by means of helpers or phage gene duplications (Hill and Stockley, 1996).

The phage display method is widely used for different purposes, for example to reveal ligands which activity in vitro can exceed several times the activity of native proteins; to map discontinuous epitopes, recognized by antibodies; to replace non-protein ligands with peptides. This method was applied as well in vivo to recover proteins binding in given tissues. (Pasqualini and Ruoslahti, 1996). It can be a basis for design of novel drugs based on synthetic peptides.

As the phage display technique clears the way to the recovery of great amount of the experimental data in the practically important areas of molecular biology, there comes up a question of storage and systematization of this information with its integration into the system of searchable and navigable information on molecular

biology. The database ASPD (Artificial Selected Peptides Database) being developed by us is meant to solve this problem.

ASPD database format

The database is implemented in standard form meant for SRS (Sequence Retrieval system). SRS is a uniform tool for handling databases on molecular biology, which allows for complex query implementation (including queries to distributed databases), fast search etc. Previously we used SRS when developing the GeneExpress system. The database is split into four parts called tables: the description of the experiment (table 1), the description of the aminoacid sequences of selected peptides (table 2), the description of the phage libraries (table 3), literary references (table 4). The tables are tied by means of the interlinks. Each entry in each table begins with a unique identifier. Then follow links to the tied entries from other tables. The principal content of the table 1 is author's alignment of the peptides obtained in the experiment. As a rule, they are aligned with one or several native proteins, sometimes the consensus sequence is given (the sequences not retrieved in the experiment are ordered with latin letters). In this card is also given general information on the experiments (target – field name Target, native protein, performing the same function – field name Template).

In the table 2 is given the information on the sequences bound to one specific target. The target description is given (field names Target and Target specification) and the descriptions of each sequence from the table 1, which include the number of sequenced clones with the given sequence, and, if the affinity of binding of

```

Identifier PH1VV001
Lib_reference PH3VV001
Lit_reference PH4VV001
Target monoclonal antibody GV1A8
Template NMWKNDMVEQMHEDIISLWDQSLKPCVKLT
Number_of_sequences 024
Alignment
01;.....asIVSLWDsl.....
02;..asTDVRQQLGWREGVVGLWDRHsl.....
03;....asNRDYDSRSDIWSIWSLGRERsl...
04;.....asYPRETAIQLWsl.....
05;.....asDOTHERAWHLWQGTVSLTRPsl.
06;.....asTGKDSVWWLWGVNsl.....
07;.....asVQERVEVYGLWGGIGNLSADsl.
08;.asHDADAKEGPKRRSASELWGPsl.....
09;.....asRSSILERIWGGsl.....
10;..asVSSPFDFSAHSPIEGLWAGEsl.....
11;.....asELEENMFCSWKLVGYSCRGPsl...
12;asRGDVPAASARGSVSIRDLWRsl.....
13;.....asYKHGTVRMLWPGGGVRVADGsl
14;..asVSMKLEDKPRTALFDLWQATsl.....
15;.....asESMDRRRGVWELWGTPSKSTRsl..
AA;..NMWKNDMVEQMHEDIISLWDQSLKPCVKLT.
//

```

Figure 1. Example of an entry from the table 1.

```

Identifier PH2VV001
Align_reference PH1VV001
Lib_reference PH3VV001
Lit_reference PH4VV001
Target monoclonal antibody GV1A8
Target_specification
Number_of_cycles 5
Affinity_units Mx10e-8
Seq_identifier 01;66
Sequence IVSLWD
Number_of_Clones 1
Number_of_Cycles 6
Number_of_Clones 2
Affinity 71
//

```

Figure 2. Example of an entry from the table 2.

sequence to target was determined, the affinity value . Table 3 comprises the description of the phage libraries. These description contain the name of the phage used, the name of the capsid protein (3 or 7), the DNA insert sequence, the general form of the aminoacid sequence of the exposed peptide, the volume of the library.

Database content

At present the database ASPD contains information on over than 100 experiments with application of the phage display technique from 70 papers. The substantial part of the data is on antibodies (including the ones to the HIV), there is also information on various hormones, DNA-binding domains (in particular, zinc fingers), discontinuous epitopes etc.

```

Identifier PH2VV001
Lit_reference PH4VV001
Phage_name M13
Phage_protein 3
DNA_sequence gcc agt GCA TCA XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK XXK
XXK XXK TCG CTA aca ggt ggg tct
Template A-S-X-X-X-X-X-X—X-X-X-X-X-X-X-X-X-X-X-X-X-S-L
Library_volume 1.0E+8
//

```

Figure 3. An example of an entry from the table 3.

Table 1. Entry examples from the ASPD.

Entry name	Class, to which belong peptides under investigation
RNA-binding zinc fingers from transcription factor IIIa	DNA and RNA binding domains
Erythropoietin (EPO)	Hormones
Autoantibodies to the thyroid peroxidase	Antibodies
Human growth hormone (hGH)	Hormones
Variable fragment of a murine monoclonal antibody Se155-4, specific for salmonella serogroup B O-polysaccharide	Antibodies
Human CD-4-binding-site anti-gp120 antibody	Antibodies
Monoclonal antibody 3E7 specific for α_0 -subunit of the heterotrimeric G-protein	Antibodies
Domains binding Src, Fyn, Lyn, PI3K, Abl SH3 –domains	Domains involved in the signal transduction
Ecotin, an inhibitor for serine protease	Inhibitors

Analysis tools

The general layout of the ASPD database and supplementary resources is given in figure 4. Data obtained with phage display is of major importance for theoretical investigation of the structure and organization of active sites in proteins. The sequences stored in the ASPD database can be examined in detail with the system for domain recognition (Valuev, Kuropatov, 2000). On the basis of various methods for pattern recognition (linear Fisher discriminant, perceptron, profile and consensus building, markov chains, neural networks) and physico-chemical scales this system allows for the recovery of true functional sites (fig.5) and parameters (such as aminoacid properties in certain positions) that are essential for their functioning. The comparison of the data obtained at the study of functional sites of native proteins with the data of the phage display can turn highly interesting both from the point of view of studying the evolution of these functional sites and from the point of view of elucidating the context of their functioning in vivo. For this purpose for each structure–functional domain described in the database are built consensus sequence, profiles, and other recognition procedures based on different mathematical models.

With the help of the developed by us CRASP software (Afonnikov et al., 1997) was carried out the analysis of pairwise correlations of values of various physico-chemical amino acid properties for all pairs of positions for each described in the database structure-functional domain. From this analysis results a correlation matrix that is available both in text and graphic forms (fig.6). The recovery of significant correlations allows to find

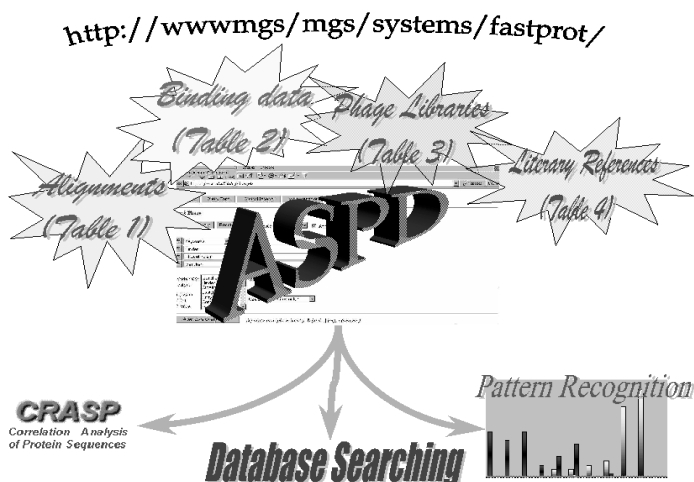


Figure 4. The overall layout of the ASPD database.

structurally and functionally important positions in the sequence, and comparing these data with the data obtained at the analysis of the native proteins can yield significant knowledge on the nature of the protein family. The database ASPD will be also fitted with the module for fast scanning of the main molecular biology databases (Swiss-Prot, PIR, PDB, TrEMBL) with the purpose of recovering patterns in amino acid sequences, similar to the ones obtained through the phage display. Such patterns potentially possess the function for which the phages were selected. Sequences obtained with such scanning can be subject for further investigation with the modules described above. Unfortunately, to date there is no tool for searching for discontinuous stretches of polypeptide chain homologous to the peptides from the ASPD.

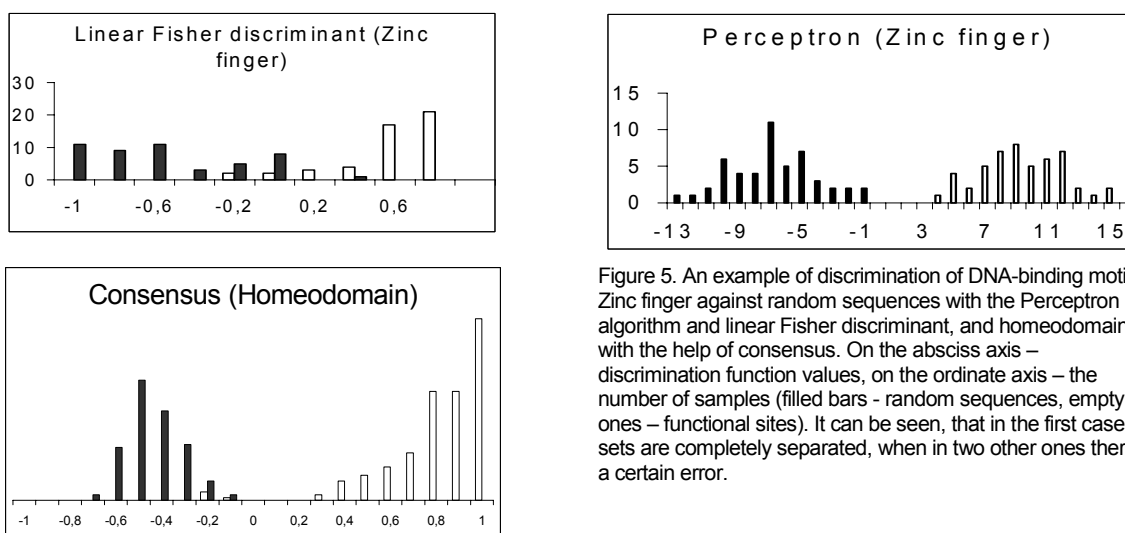


Figure 5. An example of discrimination of DNA-binding motif Zinc finger against random sequences with the Perceptron algorithm and linear Fisher discriminant, and homeodomain with the help of consensus. On the absciss axis – discrimination function values, on the ordinate axis – the number of samples (filled bars - random sequences, empty ones – functional sites). It can be seen, that in the first case the sets are completely separated, when in two other ones there is a certain error.

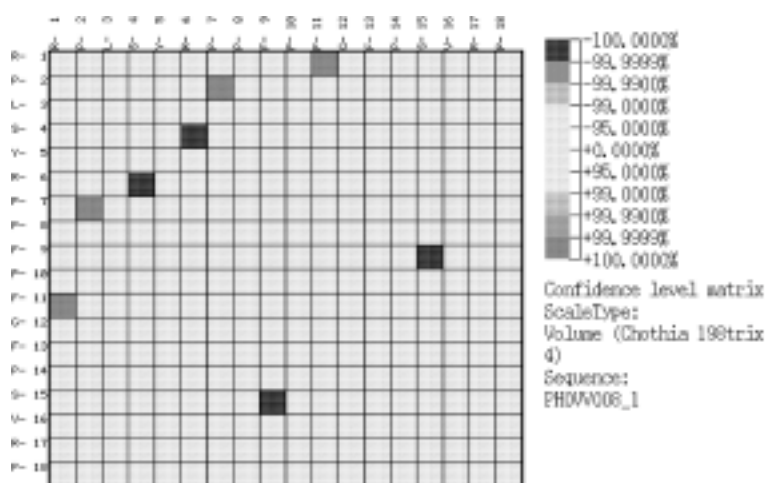


Figure 6. An example of pairwise correlation matrix (correlations of the volume of amino acids in the hapten-binding site linker). There can be seen 4 (the matrix is symmetrical) pairs of significantly correlated positions (1 and 11, 2 and 7, 4 and 6, 9 and 15). Obtained on the phage display data correlation patterns are more reliable and less noisy compared to the ones obtained at the analysis of native protein families, because in the case of the phage display we are dealing with independent samples.

References

1. Smith, G.P. (1985) Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, 228, 1315-1317.
2. Hill H.R. and Stockley P.G. (1996) Phage presentation. *Molecular Microbiology*, 20, 685-692.
3. Pasqualini R. and Ruoslahti E. (1996) Organ targeting in vivo using phage display peptide libraries. *Nature*, 380, 364-366.
4. Afonnikov D.A., Kondrakhin, Yu.V., Titov, I.I. and Kolchanov, N.A. (1997) Detecting direct correlation between positions in multiple alignment of amino-acid sequences. In Mewes, H.W. and Frishman, D. (eds) *Computer science and biology. Genome informatics: Function, structure, phylogeny*, Proc. of the German Conference on Bioinformatics, 1997, pp. 87-98.
5. Valuev V.P., Kuropatov D.A., 2000 Automatic generation of recognition programs for amino acid sequences. *Computational technologies*, 5, Special issue, 67-74.

3-DIMENSIONAL PROTEIN STRUCTURAL CLASS RECOGNITION

**Valuev V.P.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: valuev@bionet.nsc.ru

*Corresponding author

Keywords: protein structure classes, α -, β -, $\alpha+\beta$ -, α/β -protein classes, neural networks

Resume

Motivation:

There is a commonly accepted notion, that there exist several classes of 3-dimensional structure of globular proteins that are determined by predominance of one of the elements of secondary structure. Recognition of these classes starting from aminoacid composition, besides its practical significance, permits to make some conclusions about the nature of this classification.

Results:

Class (in SCOP (Murzin et al., 1995) definition) recognition methods are built by means of neural networks. The accuracy of recognition reached 75% on test set.

Availability:

<http://wwwmgs.bionet.nsc.ru/fastprot>.

Introduction

At present, there are three main structure-functional protein classifications – SCOP (Murzin et al., 1995), CATH (Orengo et al., 1999) and FSSP (Holm and Sander, 1996). They all are represented in the Internet. (SCOP: <http://scop.mrc-lmb.cam.ac.uk/scop/>, CATH: <http://www.biochem.ucl.ac.uk/bsm/cath/>, FSSP: <http://www2.ebi.ac.uk/dali/fssp/fssp.html>). The first one (SCOP) relies on structure-functional basis, that is proteins are united in hierarchical groups starting from their assumed kindred, which manifests itself through similarity of function and structure, or, in case of remote kinship, only structure. The main stages of classification are carried out manually by experts. In doing this some complex eukaryotic proteins can be split into several continuous domains, each of which has homology with a separate protein. The CATH database relies on more automatic approach. Each protein is split into domains that must have a hydrophobic core, and then these domains are classified according to their structural similarity. In FSSP classification all-against-all structural comparison is carried out, and all structures are divided into families by means of clusterization algorithms in the space, where the distance measure is mean square deviation between C $^{\alpha}$ -atoms at best alignment of these structures. Given all the differences among these classifications, they all agree in that at the highest level of hierarchy the overwhelming majority of proteins can be split into few distinct groups, that have neither exactly determined evolutionary kinship nor functional or structural similarity, but only their content at the secondary structure level. This conclusion supports the notion, which appeared long before strict classifications, that all proteins due to physical restrictions on possible ways of folding could be divided into 4 groups: α -helical, β -structural, $\alpha+\beta$ (with alternating α -helices and β -strands) and α/β (proteins, that have different domains with α -helical and β -structural folding) (Chothia, 1984). Nevertheless, though the limits of classes have extended substantially, there exist a considerable number of exceptions (proteins with irregular folding). So, in the CATH classification α/β and $\alpha+\beta$ classes are united into one and the fourth numerous class is distinguished – proteins (more exactly, domains), where dominates irregular folding.

Recognition of protein belonging to one of the 4 classes starting from its aminoacid composition is an important problem, because first it can add to the knowledge on protein structure obtained by other methods, and second, can contribute to the understanding how physico-chemical nature of amino acids determine the secondary structure of protein. This problem was addressed several times before. In the works (Chou and Zhang, 1995; Bahar et al., 1997) for recognition was employed linear Fisher discriminant. Structural classes met following criteria: class of α -helical proteins – content of α -helices more than 40%, β -strands – less than 5%; β -structural class – content of β -strands – more than 40%, α -helices – less than 5%; class $\alpha+\beta$ - α -helices content – more than 15%, content of β -strands – more than 15%, more than 60% antiparallel β -sheets; class α/β - content of α -helices – more than 15%, β -strands – more than 15%, more than 60% parallel β -strands; class ζ (irregular folding) – content of α -helices – less than 10%, content of β -strands – less than 10%. The accuracy of the method exceeded 90% on training set, which included 30 proteins from each of 4 major

classes and 9 proteins from class ζ , and testing on an independent set was not carried out. In the work (Bahar et al., 1997) was applied the method, equivalent to the one used in (Chou and Zhang, 1995) on the same set. The accuracy of recognition of test samples decreased a little (which is due to the fact that (Chou and Zhang, 1995) used slightly modified PDB files (Bahar et al., 1997)), and on independent test set of 62 proteins the accuracy of recognition ranged from 66% to 90% for different classes. In the work (Dubchak et al., 1993) classification was made by means of neural network and the length of sequence was invoked as an additional parameter. The accuracy of recognition on independent set was from 60% to 80%.

We also have used aminoacid composition of a protein for criterium, as the most simple and the most effective from all the simple ones, and neural network for classification. Our work differs from the previous ones not in its methods, but in an effort to handle the problem in a more systematic way. We have applied our method to analyze the SCOP classification of proteins, which comprises all proteins with known 3-dimensional structure, instead of relying on arbitrary criteria in forming classes.

Algorithm and methods

We worked with the SCOP classification, and only with the selection of proteins that have less than 40% homology (file pdb40d_1.37). The number of representatives of each class in this set is given in Table 1. The sets were randomly divided into training and test ones, approximately equal in volume. As features were taken frequencies of occurrence of each aminoacid in the sequence (so totally 20 numbers in the range from 0 to 1 summing up to 1). For discrimination we used neural networks with one hidden layer. There were 20 input neurons, 1 output neuron, and from 3 to 20 neurons in the hidden layer (fig.1). In such a way, each network was trained to recognise only one class (against all the others). For this purpose, we constructed two sets: a set of positive samples (from sequences of proteins belonging to the class to be recognised), and a set of negative samples (from sequences of proteins belonging to other classes). When taking the decision on a protein belonging to this or that class the rule is used, that when the output value is above the threshold (0.5), the sample belongs to the positive class, and when it is below the threshold, the sample is from negative class. At training a slightly modified algorithm of back-propagation of errors was used (for detailed description of the back-propagation of errors algorithm see Rumelhart et al., 1986). One cycle of learning corresponded to submitting all the samples from positive and negative sets. For each sample i the error E_i was calculated according to the formula $E_i = \frac{1}{2}(y_i - d_i)^2$, where y_i – output value of the output neuron, and d_i – its expected value (1 in case of positive sample and 0 in case of negative one). The total error value E was calculated according to the formula

$$E = \sum_{E_i > t} \frac{\eta}{V_j} E_i, \text{ where } t - \text{certain threshold value (we set it at 0.2 level in empirical way), } \eta - \text{learning rate, } V_j -$$

volume of the set j , ($j=1$ for positive set, $j=2$ for negative set). That is when at classification of a sample the error was less than t , this sample was neglected at the network parameters updating, and at summing errors were weighted inversely the volume of the set. After submitting all samples the network parameters were changed

according to the simplest variant of the gradient descent $\Delta w = - \frac{\partial E}{\partial w}$, where w – network parameters.

Learning was stopped when the average error reached certain threshold.

When the number of neurons in hidden layer was changed from 6 to 18 the results virtually did not change. Outside these limits performance decreased (data not shown): with the lesser number of neurons the network lacked capacity for retaining information, with greater number of neurons the number of parameters became comparable with the number of learning samples and overtraining occurred.

The results for each class for the network with 6 neurons in hidden layer are shown in Table 2. We have also applied linear Fisher discriminant to distinguish the same sets. The results are given in Table 3. In table 4 is shown the number of learning cycles for networks with 6 neurons in hidden layer when average error on test set was 0.1.

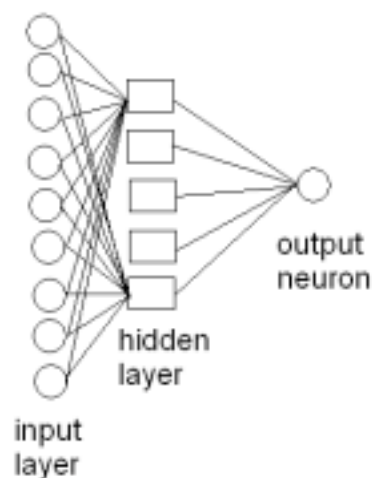


Figure 1. Neural Network.

Results and discussion

So we have made an effort towards building a method for recognition of protein structural classes starting from their aminoacid composition, and we applied our method to all proteins with known 3-dimensional structure, whose homology does not exceed 40%. For this purpose we used the SCOP classification of proteins (Murzin et al., 1995). We didn't succeed in improving the performance of previous works, but that was not our ultimate goal (though our results for the first three classes are not worse than those of others). Unlike other authors, we didn't confine ourselves to selections of few tens of proteins, but applied slightly modified standard methods (neural network and linear Fisher discriminant) to analyze global structural classification of proteins. Probably, some increase in the accuracy can be achieved by introducing some additional features (for example, the length of the chain, as in Dubchak et al., 1993).

The results obtained allow to make some conclusions. First, it was shown, that linear discriminating function, such as linear Fisher discriminant, in general case couldn't distinguish any of the four structural classes against the rest on the basis of aminoacid composition. Nevertheless, application of a non-linear method permits a rather effective classification, which witnesses the fact that predominance of this or that type of folding is really connected with physical reasons stemming from individual aminoacid properties. In the same time, the character of recognition (the number of learning cycles and accuracy on test and training sets) of the three classes (α , β and α/β) being virtually the same, the fourth class ($\alpha+\beta$) resists recognition at the same network parameters, and uniting it with α/β class yields no good result either. This could be probably due to the heterogeneity of this class. This class is also absent in one of the three main structure-functional classifications – CATH (Orengo et al., 1999).

Acknowledgement

The author is grateful to V.A. Ivanisenko for discussion and valuable advice. The work was supported by RFBR grants № 98-07-91078, 99-07-90203 and Integration project of SB RAS No 66.

Table 1. Volume of the classes.

α	261
β	315
α/β	337
$\alpha+\beta$	268
The rest	242

Table 2. The network with 6 neurons in hidden layer.

	α	β	α/β	$\alpha+\beta$	α/β и $\alpha+\beta$
False negatives on test set	30%	25%	24%	37%	33%
False positives on test set	25%	22%	33%	54%	37%
False negatives on training set	22%	23%	12%	5%	23%
False positives on training set	23%	24%	26%	46%	21%

Table 3. Results of classification with linear Fisher discriminant.

	α	β	α/β	$\alpha+\beta$
False negatives on test set	11%	11%	38%	37%
False positives on test set	54%	49%	59%	58%
False negatives on training set	8%	5%	14%	30%
False positives on training set	55%	53%	53%	50%

Table 4. The number of learning cycles for networks with 6 neurons in hidden layer.

α	β	α/β	$\alpha+\beta$	α/β and $\alpha+\beta$
258368	199586	183274	860302	842684

References

1. Bahar I., Atilgan A.R., Jernigan R.L., and Erman B. (1997) Understanding the recognition of protein structural classes by amino acid composition. *Proteins*, 29, 172-185.
2. Chothia, C. (1984) Principles that determine the structure of proteins. *Ann. Rev. Biochem.*, 53, 537-572.
3. Chou K.C., and Zhang C.T. (1995) Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.*, 275-349.
4. Dubchak I., Holbrook, S.R., and Kim, S.H. (1993) Prediction of protein folding class from amino acid composition. *Proteins*, 16, 79-91.
5. Holm L. and Sander C. (1996) Mapping the protein universe. *Science*, 273, 595-602.
6. Murzin, AG., Brenner, SE, Hubbard T, and Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247, 536-540.
7. Orengo C.A, Pearl F.M., Bray J.E., Todd A.E., Martin A.C., Lo Conte L., and Thornton J.M. (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Research*, 27, 275-279.
8. Rumelhart, DE, Hinton, GE, and Williams, RJ (1986) Learning representations by back-propagating errors. *Nature*, 323, 533-536.

BLOCKWISE EVOLUTION OF HEMOSTASIS AND COMPLEMENT FUNCTIONAL SYSTEMS

**Ananko G.G.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: ananko@online.sinor.ru

*Corresponding author

Keywords: adaptive evolution, functional systems, complement, blood coagulation

Resume

Motivation:

Genotypes of organisms are changed in the course of evolution. Both the environment and internal structure of the organism determine the variation mechanisms and the effect of selection. Assuming that natural selection estimates functions of an organism, we considered the genes constituting the functional systems (FS) responsible for these functions, aiming to detect the traces of the effect of organism internal structure on FS evolution.

Results:

It has been found that the frequencies of certain domain types in proteins of hemostasis and complement FS are considerably higher than the expected frequencies, when the exons coding for the domains are randomly distributed in the genome. Twenty types of domains, repeated from 2 to 115 times, are coded for by about 50% of hemostasis and complement FS mRNA. The effect of the initial structure of functional systems on fixation of new exons and genes is demonstrated.

Introduction

The eukaryotic organism is a complex hierarchical system comprising numerous functional systems of different levels. Functional systems are separated according to the results of their function [Anokhin, 1980]. In this work, the functional system (FS) is considered as a group of interacting genes and their products united by the same function. Hemostasis and complement FS were selected as the object. This approach seems most natural, as the selection estimates any structure according to the results of the function of the overall system.

The goal of the work was to detect possible correlations between the existing relations within FS and the mechanisms of adaptive evolution.

A blockwise evolution of the genes of the indicated FS through duplications of DNA loci coding for whole proteins and individual domains and exon shuffling [Gilbert, 1978; Ratner, 1992] is studied in the work.

The FS analyzed belong to protective homeostatic subsystems of the organism. The system of hemostasis (blood coagulation) decreases the blood loss in case of blood vessel impairments and prevents thrombus formation in normal state. The balance of coagulation and anticoagulation factors determines the system function. The complement system protects the organism from alien microorganisms through their lysis.

The FS studied belong to direct-action system, which is reflected in the structure of their control: their functions are controlled mainly by direct interactions of their protein factors with one another, matrix proteins (collagen, fibronectin, etc.), and cell membrane proteins.

Methods and algorithms

Samples of genes of blood coagulation and complement FS were formed using the SWISS-PROT database and published data. A gene was included into the hemostasis FS if there were experimental data on involvement of its protein product in either blood coagulation or blood clot lysis. A gene was included into the complement FS system basing on experimental data on involvement of its product in the complement cascade.

According to Fitch definition [Fitch, 1970], the genes formed through duplication and belonging to one genome are called paralogous. The homologous genes having separated in the course of speciation and, as a rule, performing the same function, are called orthologous [Fitch, 1970]. Exons, coding for protein domains, are capable of duplicating similar to genes; therefore, we consider it permissible to speak about paralogous and orthologous domains too.

The term *domain* designates a stable compact structural unit with the protein that can be separated from its other parts. We call a domain **informative** if it occurs in more than one protein within an FS. The program Tblastx [Altschul et al., 1997] was used to search protein databases for informative domains. Actual frequencies

of each informative domain in proteins of the FS under study (m_i) and the overall sample of available human proteins (n_i) were calculated.

The hypothesis on the coincidence of the frequencies of paralogous domains in the FS genes and the overall sample, as the FS gene sample was random with respect to their structure (with a precision of gene relatedness), was considered as the *null hypothesis* (H_0).

While analyzing, we took into account only the fact of presence/absence of a domain in a gene, leaving aside the number of certain type domains in each gene. According to the hypergeometric distribution, the probability to find m_i genes containing i^{th} domain in a random sample of K genes equals

$$P_i = \binom{K}{m_i} \frac{n_i(n_i-1)(n_i-2)\dots(n_i-m_i+1) (N-n_i)(N-n_i-1)(N-n_i-2)\dots(N-(K-m_i)+1)}{N(N-1)(N-2)\dots(N-K)}$$

where K is the number of proteins in a functional system; n_i , the number of proteins containing at least one domain of a particular type; and N , the number of all the known human proteins (approximately 5000, SWISS-PROT release of November 1999). The following approximate equation was used for the calculations:

$$P_i \approx \binom{K}{m_i} \left(\frac{n_i}{N}\right)^{m_i} \left(\frac{N-n_i}{N}\right)^{K-m_i}$$

The parameter m_i was selected as a control variable; its critical range is $m_i \geq 2$ for all the informative domains. Significance level is 0.01. The null hypothesis was tested for each informative domain. If the condition $P_i (m_i \geq 2 / H_0 \text{ is true}) \leq 0.01$ was met, the null hypothesis was denied.

Results

We included 51 genes into the hemostasis FS and 37 genes into the complement FS basing on the published data and SWISS-PROT-contained information. Totally, 20 types of informative domains were found in both FS: 15 with the hemostasis FS; 9, with the complement FS (Tab.1).

Table 1. Distributions of informative domains in hemostasis (Hem) and complement (Compl) FS in the protein sample studied

Domain type	Number of this type domains in control gene sample	Number of this type domains in FS genes		Number of FS genes containing this type domain, m_i		Number of genes in control sample containing this type domain, n	Probability to find m_i genes containing i^{th} domain in subsample	
		Hem	Compl	Hem	Compl		Hem (K = 51)	Compl (K = 37)
Protease	61	11	7	11	7	61	2.6×10^{-11}	2.9×10^{-7}
EGF	485	61	9	16	9	82	1.1×10^{-16}	6.8×10^{-9}
T1SP	29	6	16	2	6	9	3.6×10^{-3}	7.6×10^{-11}
SCR	151	19	115	3	18	28	2.8×10^{-3}	4.7×10^{-31}
Kringle	58	11		5		9	4.1×10^{-8}	
GLA	12	7		7		12	4.8×10^{-11}	
F1N	15	2		2		4	7.5×10^{-4}	
APPLE	8	8		2		2	1.9×10^{-4}	
VWFC	24	5		3		15	4.9×10^{-4}	
BPT	11	6		2		7	2.2×10^{-3}	
T3SP	35	28		4		5	2.4×10^{-7}	
Fibrin	5	2		2		5	1.2×10^{-3}	
F5/8 C	10	4		2		6	1.7×10^{-3}	
Plastocy	18	12		2		3	4.3×10^{-4}	
Serpin	26	12		12		26	5.1×10^{-17}	
C1Q	9		3		3	9		4.2×10^{-5}
LDLRa	81		7		6	14		1.0×10^{-9}
CUB	16		8		4	8		4.1×10^{-7}
Anaph	9		3		3	5		7.0×10^{-6}
CCFIM	5		5		3	3		1.0×10^{-6}

Four types of domains are common for both systems: protease, EGF, T1SP, and SCR. The fraction of informative domains in the total mRNA of these FS amounts to approximately 50%. Thus, we are basing on an essential part of genes of these FS in our conclusions.

EGF, SCR, LDLRa, protease, and Kringle domains are most numerous, varying from 58 (kringle) to 485 (EGF). The domains of SRC type reside mainly (76%) in the 18 complement system proteins, 10 of them are composed exclusively of SRC domains differing only in their numbers.

The data listed in table demonstrate that actual frequencies of informative domains in the FS (m_i) differ significantly from the expected values (P_i), the truth of the null hypothesis provided. The m_i values observed fall into the critical region (≥ 2); that is, the probability of such event is considerably lower than the critical probability (0.01) for all the informative domains. Thus, *the null hypothesis is denied in all the 20 attempts*. This means that the paralogous domains are mainly localized to the proteins of the FS under study; that is, the exons coding for paralogous domains are distributed unevenly (asymmetrically) relative to the "boundaries" of the functional systems. The "boundaries" were specified through forming the lists of genes included into each system.

Distributions of the domains within groups of related genes were also studied. The number of genes carrying a certain domain type may increase in two ways:

(1) Duplications of the whole genes. To assess the contribution of duplications, all FS genes were divided into groups originating from a common ancestor (according to the published data and SWISS-PROT). The hemostasis system contains 8 groups of related genes (from 12 to 2 genes in each group) and 13 single genes; the complement FS, 5 groups and 7 single genes. The mean copy number of related genes in the hemostasis FS is $51/21 = 2.4$; in the complement FS, $37/12 = 3.1$; and within both FS, $51 + 37/21 + 12 = 2.7$. It is evident that gene duplications played an essential role while forming hemostasis and complement FS. Note also that the copy number of related genes are similar in two different systems.

(2) Intragenic duplications of exons coding for domains with further shuffling. It was found that domains of seven types (EGF, SCR, GLA, LDLRa, T1SP, VWFC, and CCFIM) occur in at least two groups on unrelated genes. This means that the exons coding for one third of informative domains were with a high probability involved into DNA shuffling between unrelated genes.

Increased frequencies of paralogous domains and high copy number of related genes are likely to result from the functional structure of these systems, as the gene samples were formed with respect to their functions.

Discussion

Thus, what are the sources of the revealed asymmetry in domain distribution ?

(1) *Mutational process*. If gene and exon duplications and shuffling occurred in the genes of these FS more frequently than on the average in the entire genome, then modifications of the existing FS genes mainly were affected by the selection, resulting in fixation of novel copies of the existing genes and domains. Increased frequency of blockwise rearrangements may stem from physical proximity of the genes on the chromosome; however, their modern locations may differ considerably from those during the FS formation, hindering estimation of this contribution. Another important factor is the homology of DNA loci. First, the shuffling frequency may increase due to already existing complete or partial homology of a pair of genes [Patthy, 1985]. Second, exon deletions and duplications may arise from an unequal crossover and recombination between repeats of the same family located in introns [Sudhof et al., 1985].

(2) *Selection*. The requirements set forth at the stage of selection to the structure of novel protein may be divided into two constituents:

(a) *External*, imposed by the environment and

(b) *Internal*, imposed by the structure of the functional system itself.

External requirements are always mediated by the organism structure existing at a certain moment. The FS structure whereto a protein is included has the major effect, especially if its function is under the selection pressure. The requirements of the system result from the inner links existing in it at this time.

We consider a *pair of complementary affinity regions of two molecules* as a **link**. This may involve any combinations of DNA, RNA, protein, and metabolite molecules. As we are analyzing the protein structure at the level of domains, let us consider that pairs of domains with complementary regions represent the links. Thus, a *pair of complementary domains is considered as a molecular equivalent of the functional link*. Functions of many domains are known. In addition, evolutionary conservancy of the domain structure of genes confirms their functionality. Thus, fixing a novel gene or domain copy equals in most cases emerging new link (links) between elements of the same or different FS.

The asymmetry found means that fixing the copies of elements already existing in the system of links dominates while forming new FS. This manifests itself as a "propagation" of certain types of domains and their complementary counterparts.

Note that key proteins, their domains, and proteins and domains housing complementary sites are multiply repeated in the corresponding systems. This means that *the trend to use "the own" links repeatedly took place across the entire course of formation of these FS*.

Any structural novelty should either be included into the scheme of existing systemic links or, at least, not impair the links formed. This is more probable provided that a novel FS protein contains at least one functional site complementary to any site of the molecules already existing with the system, and even more probable if the entire "link"—a pair of domains—is doubled. Naturally, the probability of such mutation to be fixed increases too. Thus, one of the reasons of the asymmetry found is **inherent of the systemic character of the living**.

The mechanism underlying the preference of "the own" domains on the background of creating new structures by the selection is, however, vague. Reconstructing a copy of a gene already existing in a system requires less number of changes (mutations) than modifying a copy of an "alien" gene. Thus, what are the parameters used by the selection to "estimate" the efficiency of alternative patterns for creating new functional structures? **The time** necessary for forming new functional structure may represent such parameter. *The selection fixes automatically the result of the quickest formation of a new functional structure*; that is, the pattern realized through least number of adaptive cycles is fixed out of the multiple alternative patterns.

Acknowledgements

The project was supported by the Integration Project of the Siberian Branch of the Russian Academy of Sciences *Simulation of basic genetic processes and systems*. Authors are grateful to Dr. A.Yu. Rzhetsky (Columbia University, USA) for assistance in statistical processing of the results obtained, to Yu.G. Matushkin for helpful criticism, and to G. Chirikova for the help in translation into English.

References

1. Anokhin P.K. (1980) Knot Problems of the Theory of Functional System. Moscow: Nauka.
2. Gilbert W. (1978) Why genes in pieces? *Nature*, **271**, 501.
3. Ratner V.A. (1992) Block-Modular Principle of Organization and Evolution of Molecular Genetic Control Systems (MGCS). *Genetika*, **28**, 5-24.
4. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W, and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.
5. Fitch W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99-113.
6. Patthy L. (1985) Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. *Cell*. **41**, No. 7, 657-663.
7. Sudhof T.C., Goldstein J.L., Brown M.S., and Russel D.W. (1985) The LDL receptor gene: fa mosaic of exons shared with different proteins. *Science*, **228**, 815-822.

PDBSite: A DATABASE ON BIOLOGICALLY ACTIVE SITES AND THEIR SPATIAL SURROUNDINGS IN PROTEINS WITH KNOWN TERTIARY STRUCTURE

**Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: salix@bionet.nsc.ru

*Corresponding author

Keywords: biologically active sites of proteins, tertiary protein structure, databases

Resume

Motivation:

The Protein Data Bank (PDB) database contains an information on the spatial protein structures. Besides, for many proteins there are other data on biologically active sites (i.e., ligand binding regions, enzyme catalytic centers, regions subjected to biochemical modifications, etc.). Development of a database accumulating the data on site features calculated according their tertiary structure and the structure of their surroundings may serve as a base for development of methods aimed at site recognition and at studying of their structure-functional organization.

Results:

A database PDBSite storing the data on biologically active sites contained in the PDB database is developed. PDBSite accumulates amino acid content, structure features calculated by spatial protein structures, and physicochemical properties of sites and their spatial surroundings. PDBSite is a part of FASTProt system, being developed in the Institute of Cytology and designed for analysis of protein structure and function. An analysis of relationships is made, between the parameters of amino acids within sites and their surroundings.

Availability:

PDBSite database is available via the Internet by the addresses

<http://srs5.bionet.nsc.ru:8080/srs5bin/cgi-bin/wgetz?-fun+Pagelibinfo+-info+PDBSITE>

and <http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/>.

Introduction

The data on biologically active protein sites are of extreme importance for solving many problems in molecular biology, biotechnology, and medicine. High specificity of biological activity in proteins is produced by unique structure of active sites that are often organized by a very complicate pattern. In particular, biologically active sites in proteins are often compiled out of remote by primary structure amino acid residues, which form compact clusters in the spatial structure with strictly ordered conformation. Specific structure and conformational parameters of these sites are determined by the structure of their spatial amino acid surroundings. For example, spatial amino acid surroundings of enzyme catalytic centers determine the relief of hollows in catalytic centers of enzymes in a substrate binding regions [1], whereas the residues of antigen determinants of proteins determine their structure by organizing prominent parts at the protein surface [2]. For many natural and mutant proteins, the relationships were found between protein activity and physico-chemical properties of amino acid residues composing the local surroundings of a functional site [3]. The spatial surroundings of biologically active sites may be detected only if the data on tertiary protein structures are available. The database Protein Data Bank (PDB) [4] stores an information on spatial protein structures, many of them are supplied by marking-out of amino acid residues composing biologically active sites (regions binding to ligands, catalytic centers of enzymes, regions exposed to biochemical modification, etc.). The goal of the present paper is to develop a daughter database, PDBSite, on characteristics of biologically active sites stored in PDB database and their spatial surroundings.

To achieve these goals, we have developed original methods, algorithms, and software programs for calculation characteristics of spatial surroundings of sites indicated in PDB, along with calculations of their structural and physico-chemical parameters. The analysis of data stored in PDBSite database has revealed some regularities in distribution of amino acid content within biologically active sites and their spatial surroundings. We have found correlations between physico-chemical parameters of amino acids of the sites and those of their spatial surroundings.

Methods and materials

The PDBSITE database contains the data obtained by the treatment of the following fields of PDB database: HEADER, TITLE, KEYWDS, REMARK 800, SITE, ATOM. For the treatment of information contained in PDB, we have developed the program software for parsing. If a single PDB entry contains the data on several sites, then an individual entry for each site was created. An Internet access to the PDBSITE database is provided by the Sequence Retrieval System (SRS).

The spatial surroundings of sites were calculated in the following way: by using atom coordinates, a parallelepiped including all the atoms of amino acid residues of a site. In what follows, atoms of amino acid residues are excluded from parallelepiped. Spatial surroundings of a site were determined by analysis of all the rest protein residues. As the spatial surroundings of a site we determine as all the rest amino acid residues in case one of constituting them atoms was included inside the parallelepiped.

For characteristics of a site and its surroundings, we take (1) an exposure of each of its residues; (2) average value, (3) sum, and (4) spatial moment of physico-chemical parameters of amino acids, (5) center mass coordinates for each residue, and (6) pairwise distance between mass centers of residues. Additionally, among the site characteristics, we have calculated (8) an indicator of the site discontinuity according to its primary

structure. The indicator of discontinuity of a site was set as $\frac{1}{N} \sum_{i=1}^N (P_{i+1} - P_i - 1)$, where N is the number of

residues, P_i is numerical number the i -th residue of the protein sequence of a site. For calculation of Exposure of amino acid residues, we have applied an approach of immersing the molecule into the cubic lattice (displacement of volume in cubic lattice).

Results and Discussion

Developed PDBSITE database accumulates the data on amino acid content, structural, and physico-chemical characteristics of sites and their surroundings. The list of fields in PDBSITE database and their brief description are given in the Table.

Table. List of fields in PDBSITE database and their brief description.

Field	Description	Field	Description
ID	Identificator	RESNAME	Names of residues in a site and surroundings
PDBID	PDB Identificator	EXPOSE	Exposure of each of the residues of the site and its surroundings
HEADER	PDB HEADER	ORDER	Ordering of physico-chemical parameters of a site and surroundings in the table of values
TITLE	PDB TITLE	AVERAGE	Average value in the table of values of physico-chemical parameters
KEYWORD	PDB KEYWDS	SUM	Sum in the table of values
SITE_DESCR	Description of the site	SPATIAL_MOMENT	Spatial moment in the table of values
LENGTH	Number of residues in a site and in its surroundings	PAIRWISE	Pairwise distance between the residues
EXPOSURE	Average exposure of residues of a site and its surroundings	COORDINATES	C-alpha atoms coordinates of site residues
CHAIN_ID	Chain identifier for a site and its surroundings	CA_ATOMS	Centre mass coordinates of site residues
POS	Positions of residues in a site and its surroundings	COORDINATES	Centre mass coordinates of site residues
		CENTRE_MASS	Centre mass coordinates of site residues
		DISCONTINUITY	Discontinuity of the site

By analyzing relationships between site structure and site surroundings structure, we have obtained the following results. By unifying all the sites into a single group, we have not found the correlations between physico-chemical properties of sites and their surroundings, although slightly pronounced discrepancies in amino acid content were observed. Partition of the sites into the group of active sites and the group of binding sites has changed the result: discrepancies by amino acid content became more expressed. The partition was made as follows: the group of active sites contained all the sites, such that have the word 'active' in the filed SITE_DESCR, whereas the group of binding sites contained the sites extracted according to the presence of the word 'binding'.

We have detected slight correlations for some physico-chemical parameters. In Fig. 1, one can see the dependency between Hydrophilicity of a site and Hydrophilicity of site surroundings for the group of active sites. As can be seen from the Figure, with the growth in Hydrophilicity of a site, the Hydrophilicity of site surroundings

falls. Besides, we have found some difference in amino acid content between the active sites and binding sites groups

The result has drastically changed under stricter partitioning of sites in accordance with their specialization. The alterations in amino acid content became more pronounced both between the above groups of sites and between the groups of sites and their surroundings. In Fig. 3, there is a histogram illustrating the frequencies of amino acids occurrence in the zinc-binding sites and within surroundings of these sites. As can be seen from the figure, only five types of amino acids are most frequent in the zinc-binding sites, although amino acid content distribution is more homogeneous in surroundings of these sites. Notably, the peaks in distributions are not coincident.

We have also found good correlations between properties of amino acids of the sites and those of the site surroundings. Interestingly, the type of physico-chemical property correlating with high frequencies often depends upon the type of site specialization. For example, in Fig.2., one can see the correlation between amino acid Hydrophilicity of zinc-binding sites and amino acid Hydrophilicity of surroundings of these sites.

The results obtained are in a good accordance with the common opinion that function of a site is closely related to the structures of a site and its surroundings. Based on the results obtained, we can suppose that functional role of a site surroundings is of topical value even at the stage of initial recognition of a target, to which the site binds, and the value of the proper site orientation relatively the target. This supposition may be verified by correlations between physico-chemical parameters of the sites and their surroundings. In particular, negative correlations make evidence on possible compensate action of surroundings in response to alterations of a site

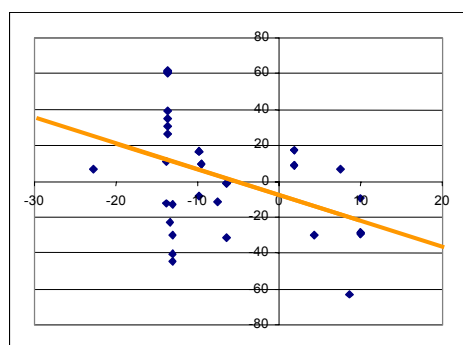


Figure 1. Correlation in Hydrophilicity between amino acids of sites and their surroundings calculated for the group of active sites. $R=-0.4$, confidence level exceeds 95%.

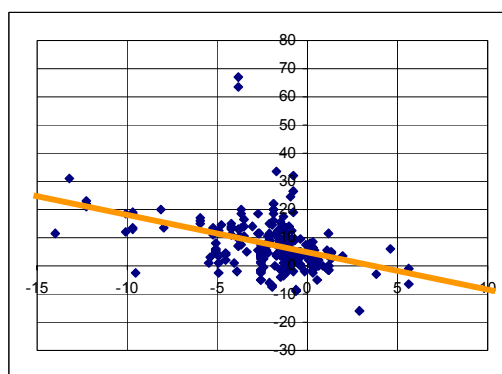


Figure 2. Correlation between Hydrophilicity of amino acids of sites and their surroundings calculated for the group of zinc-binding sites. $R=-0.65$, confidence level exceeds 95%.

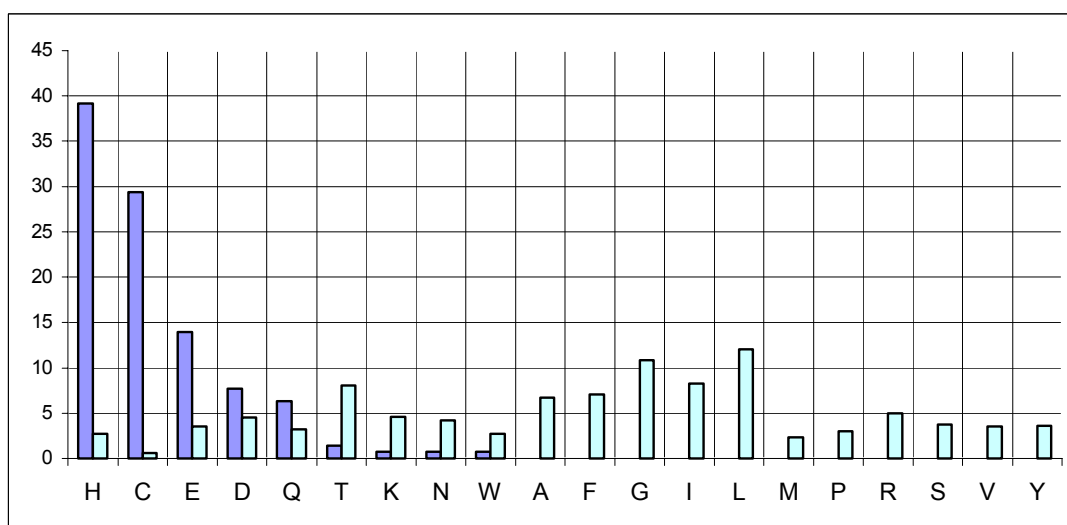


Figure 3. Distribution of amino acid occurrence frequency in sites binding with zinc and in surroundings of these sites. Dark columns correspond to site frequencies, light – to that of surroundings.

physico-chemical parameters. PDBSITE database is a valuable tool for studying structure-functional organization of biologically active sites.

Conclusion

In the Institute of Cytology and Genetics SB RAS, an information system FASTProt is being developed, it accumulates data on the protein structure and function (<http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/>). One of the important goals in developing FASTProt is to provide effective and prompt solving of the tasks aimed at protein structure-functional analysis on the basis of application of integrated resources: software programs for analysis, databases, and daughter databases containing preliminary calculated protein characteristics. Currently, FASTProt includes the database EnPDB on spatial structures of DNA, RNA, and proteins, database on experiments dealing with phage display, the program for analysis of evolutionary correlations between amino acid substitutions in proteins, programs for prediction of protein domains [5]. The task of developing daughter databases has some peculiarities. The data stored in daughter databases should provide an advanced search through parental databases and/or they should be the basis for development of novel algorithms for protein analysis. PDBSITE database integrated into FASTProt is able to aid in solving of the tasks of both types given above. In future, we plan to develop the database of protein spatial patterns of biologically active sites on the basis of analysis of data stored in PDBSITE. Recognition methods resting on patterns or profiles are widely used for site recognition by the primary protein structure [6]. Representation of the sites in a spatial form should essentially improve the recognition ability.

Acknowledgements

The work is supported by Russian Foundation for Basic Research (grants Nos 98-07-91078, 00-04-49252). The authors are grateful to G. Orlova for translation of the paper into English.

References

1. Chothia, C. The nature of the accessible and buried surfaces in proteins. // *J. Mol. Biol.* 1976. V.105. P.1-14
2. Davies, D. R., Cohen G. H. Interactions of protein antigens with antibodies. // *Proc. Natl. Acad. Sci. USA* 1996. V.93. P.7-12.
3. Ivanisenko V.A., Eroshkin A.M. The search of sites containing functionally important substitutions in the sets of related or mutant proteins. // *Molekulyarnaya biologiya (Moscow)*. 1997. T.31. P.880-887. (in Russian)
4. Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T., Tasumi M. The Protein Data Bank: a computer based archival file for macromolecular structures. // *J.Mol.Biol.* 1977. V. 112. P. 535–542.
5. Ivanisenko V.A., Grigorovich D.A., Afonnikov D.A., Kuropatov D.A., Valuev V.P., Kolchanov N.A. An informational system FrameProt on three-dimensional structures of DNA, RNA and proteins integrated with GeneExpress. // *First Russian National Conference on DIGITAL LIBRARIES: ADVANCED METHODS AND TECHNOLOGIES, DIGITAL COLLECTIONS*. October 19-21, 1999, Saint-Petersburg, Russia. P. 175-186.
6. Bairoch A. The PROSITE dictionary of sites and patterns in proteins, its current status. // *Nucl. Acids Res.* 1993. V.21. P.3097-3103

RECEPTOR DATABASE (RDB) AS AN ANALYTICAL TOOL FOR THE DRUG DESIGN

**Nakata K., Takai T., Nakano T. and Kaminuma T.*

Division of Chem-Bio Informatics, National Institute of Health Sciences, Japan

e-mail: nakata@nihs.go.jp

*Corresponding author

Keywords: receptor, internet, world wide web, ligand binding, protein sequence, sequence similarity

Resume

Motivation:

The sequence similarity information on receptor proteins was newly included in RDB. The regulatory genomic signals and regions information, and the binding affinity information for endocrine disruptors were added in RDB, linking to Transfac, TRRD and BADB.

Results:

Including new functions and informations, RDB was extended as an analytical tool for the drug design.

Availability:

RDB is available via Internet at <http://impact.nihs.go.jp/RDB.html>

Introduction

We had developed the receptor database (RDB; Nakata et al., 1999) based on the Internet/World Wide Web (WWW) technology. RDB was constructed so that the system collects data items such as attributes of proteins from distributed data sources of the Internet, and so that it provides various viewing tools effectively, depending on different types of receptor data. Such sources included standard international biological databases; PIR, Swiss Prot, PDB, etc... The Internet/WWW technology enabled us to have powerful viewers for representing retrieved data and knowledge graphically, and also enabled us to link dynamically to ligands and the cell signaling networks database (CSNDB; Takai, et al., 1998).

In this paper, we extended our receptor database (RDB) including new functions and new linking sites. One of new linking sites is Binding Affinity Database for endocrine disruptor (BADB; Kaminuma, et al., in press), which was developed in our laboratory. This database stores experimental data for interaction of exogenous chemicals and biomolecules. The scheme of RDB system configuration is shown in Figure 1.

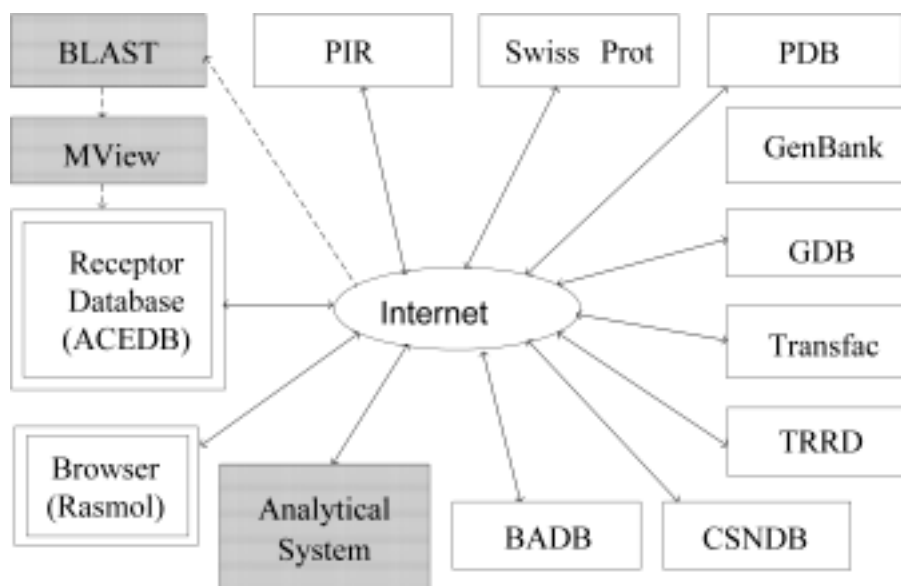


Figure 1. The scheme of system configuration.

Methods and algorithms

Flexibility for data updating which sometimes requires even structural change to data, we used an object-oriented database management system ACEDB (A *Caenorhabditis elegans* Database), instead of relational database. For the on-line modification of data, PERL programs were integrated in the system. For the sequence similarity information, the BLAST search and MView program were executed previously, and the results were included in RDB.

RDB includes the following information:

(a) Functional and structural information of receptor proteins.

(a-1) Amino acid sequence (PIR, Swiss Prot)

(a-2) DNA binding site, ligand binding site and transmembrane region
(with highlighted functional region)

(a-3) Secondary structure prediction

(a-4) Three-dimensional image (PDB)

(a-5) Sequence similarity information (BLAST search and MView)

(b) DNA and gene information

(b-1) DNA sequence (GenBank)

(b-2) gene data (GDB)

(c) Cell signaling information

(c-1) Cell signaling networks (CSNDB)

(d) Cellular molecular interaction

(d-1) Transcription factor information (Transfac) (Wingender et al., 2000)

(d-2) Transcription regulation information (TRRD) (Kolchanov et al., 2000)

(e) Interaction of exogenous chemicals and biomolecule

(e-1) Binding affinity database for endocrine disruptor (BADB)

For example, the detail information on the estrogen receptor is shown in Figure 2.

Discussion

The sequence similarity information is important for the investigation of the structure and function of receptor proteins. Linking to both databases of Transfac and TRRD, the regulatory genomic signals and regions information is available on RDB. Although it is now only limited receptors on BADB, the binding affinity information of endocrine disruptors is effective for a basic research for the drug design.

We are intending for this system to make more useful analytical tool for drug design.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389-3402
2. Brown NP, Leroy C, Sander C. (1998) MView: a web-compatible database search or multiple alignment viewer, *Bioinformatics* 14: 380-381.
3. Kaminuma, T., Takai-Igarashi, T., Nakano, T. and Nakata, K., (in press) Modeling of Signaling Pathways for Endocrine Disruptors, *BioSystems*.
4. Kolchanov NA, Podkolodnaya OA, Ananko EA, Ignatieva EV, Stepanenko IL, Kel-Margoulis OV, Kel AE, Merkulova TI, Goryachkovskaya TN, Busygina TV, Kolpakov FA, Podkolodny NL, Naumochkin AN, Korostishevskaya IM, Romashchenko AG, Overton GC. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000, *Nucleic Acids Res*, 28(1): 298-301.
5. Nakata, K., Takai, T. and Kaminuma, T. (1999) Development of the receptor database (RDB): application to the endocrine disruptor problem, *Bioinformatics*, 15, 544-552.
6. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prus M, Reuter I, Schachere F. (2000) TRANSFAC: an integrated system for gene expression regulation, *Nucleic Acids Res*, 28(1): 316-319.

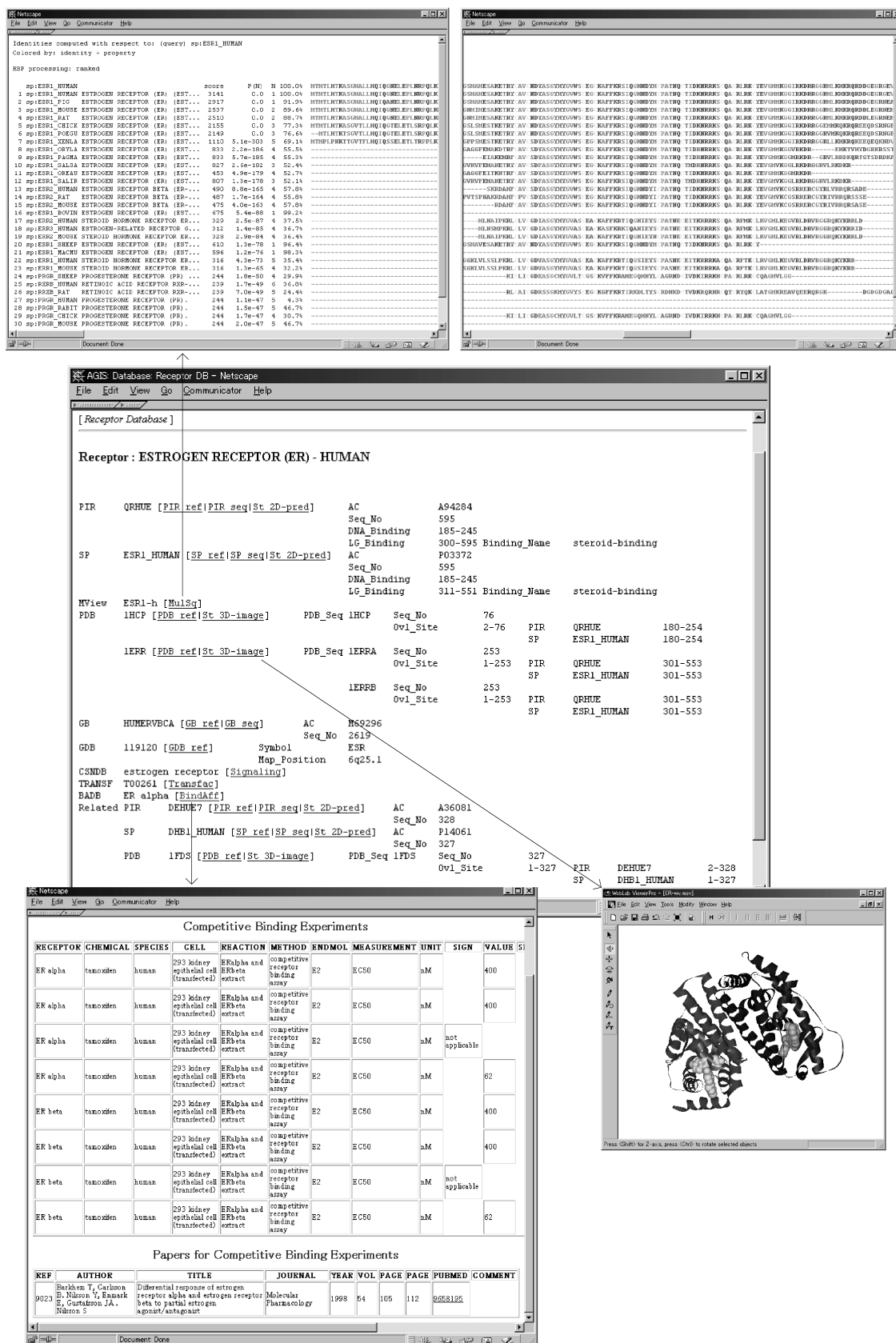


Figure 2. The detail information on the estrogen receptor.

PROTEIN PRIMARY SEQUENCES AS MARKOV CHAINS

¹Mitra Chanchal K, ²Sen Arusharka

¹Department of Biochemistry, School of Life Sciences University of Hyderabad, India
e-mail: ckmsl@uohyd.ernet.in

²Department of Mathematics and Statistics, School of Mathematics and Computer Information Sciences, University of Hyderabad, India
e-mail: asensm@uohyd.ernet.in

*Corresponding author

Keywords: protein, primary sequence, Markov chain, sequence analysis, regularity, protein structure

Resume

Motivation:

The protein primary sequence is believed to contain all necessary information for the overall three dimensional folded structure and the functional properties of the protein. Although several empirical algorithms exist that predict the overall folded structure of the protein based on its primary sequence, none has a theoretical basis. We have tried to locate and identify some of the order or regularity present in protein primary sequence using a simple Markov model. A first order Markov model does not show any order or regularity and therefore a higher order model has to be considered. The analysis has been performed using the SwissProt protein sequence databank. The primary motivation came from the simple observation that the amino acid composition differ significantly at various positions of the protein sequence, particularly in the initial region of the sequence and tend to stabilise afterwards (Figure 1).

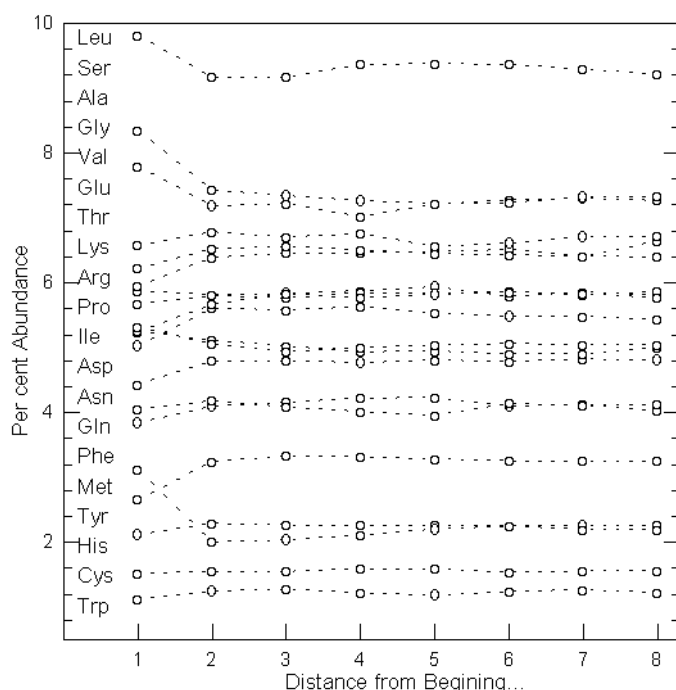


Figure 1. The variability of the amino acid composition along the sequence length. For this experiment, we have chosen all complete sequences longer than 511 residues. The length has been divided into 8 blocks each of length 64 residues. The amino acid composition has been computed in each block independently. The amino acid composition practically stabilizes after a few amino acids from the beginning. In other words, the amino acid composition in the early parts of the sequences are abnormal. On the abscissa, the number represents blocks of 64 residues from the beginning. The amino acid labels on the body of the graph show in correct order (for the first data point), the graphs for the respective distribution. We have also repeated the same graph after skipping the first 1-2 residues (they are often abnormal) but no significant differences can be seen.

Results:

We have already computed the first order Markov dependence for all the protein sequences available in the database. We do not find the expected behaviour and the results suggest that first order dependence is not sufficient to explain the observations. Statistical analysis of the results are shown in Figure 2.

Results for the higher order dependence are yet to be compiled and will be presented.

Methods and algorithms

The computations to check for Markov dependence has been carried out as follows:

All fragments (partial sequences) have been eliminated.

All sequences smaller than 512 residues have been ignored.

The frequencies for all possible pairs are computed (n_{ij})

The frequencies for all the residues are computed (n_i)

A test statistic $Z = \sum_{i,j=1}^{20} \frac{(n_{ij} - n_i n_j / n)^2}{n_i n_j / n}$ is computed as where n is the total number of residues in the given sequence.

The distribution of Z is shown in Figure 2.

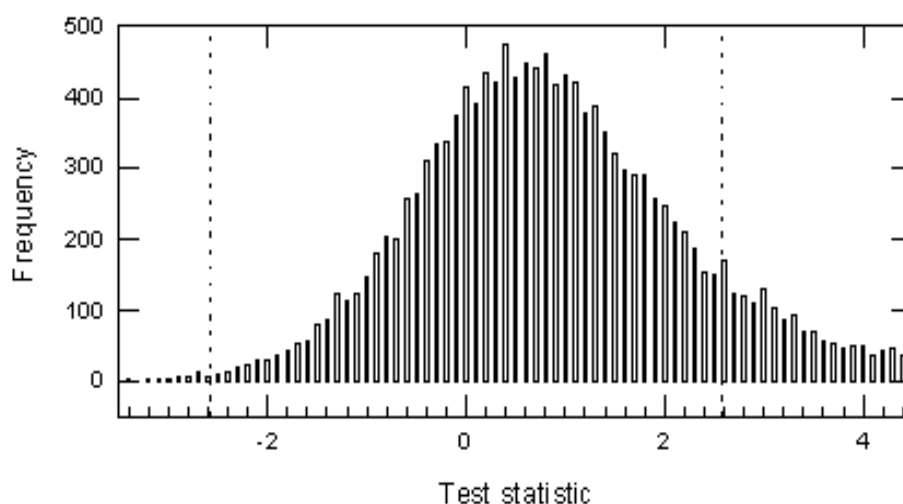


Figure 2. The frequency distribution of the test statistic for the test for independence vs first order Markov dependence in protein sequences. The vertical dotted lines indicate the 99% confidence limit for the test statistic, viz., ± 2.58 . The shape of the distribution appears close to normal but the K-S test for normality resulted in rejection, with p -value < 0.0001 . In addition, the mean of the observed distribution is shifted from zero.

The computations for equilibrium distributions have been done as follows:

All fragments (partial sequences) have been eliminated.

All sequences smaller than 512 residues have been ignored.

All possible pairs (i.e., 20×20) have been counted and stored in a matrix. This matrix is referred to as the pair-frequency matrix.

A row with all zeros is eliminated.

Each element of this pair frequency matrix is divided by the sum of the corresponding row.

The matrix is transposed. The solution of the transposed matrix using the following equation:

$$\begin{cases} (P' - I) \cdot x \\ 1' \cdot x \end{cases} = \begin{cases} 0 \\ 1 \end{cases}$$
 where P' is the transposed matrix gives us the equilibrium distribution of the 20 different amino acid residues.

The equilibrium distributions are all equal to $1/20$, i.e., 0.05 , suggesting that all the amino acid residues are equally likely to appear in a sequence in the steady state (near the tail end of the sequence). This result is contrary to common experience.

Discussion

The first order Markov analysis is not suitable for analysis of protein sequences and higher order preferences must be invoked. Alternatively, we can confidently say that it is the long range preferences that play the major role in the protein structure and function.

Acknowledgement

This work has been made possible by a grant from the University Grants Commission, Government of India. Receptor Database (RDB) as an analytical tool for the Drug Design

FROM GENOMES TO PROTEIN SPACE

**Peitsch Manuel C., Schwede Torsten, Diemand Alexander. and Guex Nicolas*

Glaxo Wellcome Experimental Research S.A. and Swiss Institute of Bioinformatics,
16, ch. Des Aulx, 1228 Plan-les-Ouates, Switzerland

e-mail:mcp13936@glaxowellcome.co.uk

*Corresponding author

Keywords: protein folding & structure, bioinformatics

Resume

Understanding the function and the physiological role of proteins is a basic requirement for the discovery of novel medicines (small molecules) and "biologicals" (protein-based products) with medical, industrial or commodity applications. Although the draft sequence of the complete human genome is about to be ready, humankind is very far from understanding the function and the physiological role of the gene products it encodes. Indeed, being able to read the letters and the words is disconnected from understanding their meaning. Therefore, the attention of many biologists is now shifting to the Functional Analysis of the genome. Functional Analysis, the first major step after genome sequencing and gene identification must rely on a combination of technologies. Consequently, new experimental approaches, and their automation for large-scale applications, will need development. Concurrently, and in order to maximise the value of large data sets, one will witness the development of new data mining methods and mathematical models for biological processes simulation. A protein's function is tightly linked to its three dimensional (3D) structure. As residues located far apart in the primary sequence can be very close in space, and only a few residues are generally responsible for a protein's function, insights into the 3D structure of a protein can represent a key component of the Functional Analysis process. Consequently, an atomic level 3D representation to assign roles to specific residues is a major asset, both for planning experiments and explaining observations. As the experimental elucidation of these 3-D structures by X-ray crystallography or NMR is often hampered by difficulties, protein modelling has been developed to provide a faster route to structural information. Indeed, the known 3-D structures can be used to build structural models for related family members using comparative protein modelling methods. To overcome the difficulty in building protein models by non-experts, we have designed the SWISS-MODEL server and its front-end the Swiss-PdbViewer (DeepView) which together form a protein modelling environment freely available on the Internet. To further explore the value of automated protein modelling methods and to build a database of all possible model structures we have, in collaboration with Silicon Graphics Inc, submitted 211,000 protein sequences (SWISS-PROT/trEMBL) to the SWISS-MODEL server and generated over 65,000 model structures. To this end Silicon Graphics has deployed a 64-processor Silicon Graphics CRAY Origin2000 server with 32 Gb of memory. This collection of models is available over the World Wide Web and is constantly improved (now over 80,000 model structures). This project is now ongoing and will yield ever increasingly reliable 3D structures for the scientific community.

DATABASE OF PATTERNS PROF_PAT, USED TO DETECT LOCAL SIMILARITIES

^{1*}*Bachinsky A.G.*, ²*Grigorovich D.A.*, ¹*Naumochkin A.N.*, ¹*Nizolenko L.Ph.*, ¹*Yarigin A.A.*

¹Research Institute of Molecular Biology, Koltsovo, Russia

e-mail: bachin@vector.nsc.ru

²Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: odip@bionet.nsc.ru

*Corresponding author

Keywords: protein families, patterns, motifs, similarity search, database

Resume

Motivation:

When analysing novel protein sequences, it is now essential to extend search strategies to include a range of 'secondary' databases. Pattern databases have become vital tools for identifying distant relationships in sequences, and hence for predicting protein function and structure. The main drawback of such methods is the relatively small representation of proteins in trial samples at the time of their construction. Therefore a negative result of an amino acid sequence comparison with such a databank forces a researcher to search for similarities in the original protein banks. We developed a database of patterns constructed for groups of related proteins with maximum representation of amino acid sequences of SWISS-PROT in the groups.

Results:

Software tools and a new method have been designed to construct patterns of protein families. By using such method, a databank of protein family patterns, PROF_PAT, is produced. This bank is based on SWISS-PROT (rl.38) and TrEMBL (rl.11), and contains patterns of more than 14,000 groups of related proteins in a format similar to that of the PROSITE. Motifs of patterns, which had the minimum level of probability to be found in random sequences, were selected. Flexible fast search program accompanies the bank. The researcher can specify a similarity matrix (the type PAM (PAM, BLOSUM and other). Variable levels of similarity can be set (permitting search strategies ranging from exact matches to increasing levels of "fuzziness").

Availability:

The Internet address for comparing sequences with the bank is:

http://wwwmgs.bionet.nsc.ru/mgs/programs/prof_pat/. The local version of the bank and search programs (approximately 50 Mb) is available via ftp: ftp://ftp.bionet.nsc.ru/pub/biology/vector/prof_pat/, and ftp://ftp.ebi.ac.uk/pub/databases/prof_pat/. Another appropriate way for its external use is to mail amino acid sequences to bachin@vector.nsc.ru for comparison with PROF_PAT 1.3.

Introduction

Up to now, the main method of suggesting possible functions of the newly deciphered amino acid sequences has been to search them for similarity with sequences available in protein banks such as PIR (Barker et al., 1999), SWISS-PROT (Bairoch and Apweiler, 1999) and others. As these banks grow larger, such comparisons become more promising but at the same time more time-consuming. In addition, in the case of distant proteins the search for global similarity of complete sequences may fail to show a positive result, because the conservative blocks responsible for their special functions may prove to be relatively short and scattered all over the sequence. This may be why a number of works appeared in the last few years, aimed at the selection of sites in groups of related proteins. These sites are representative of a protein family as a whole, and both identify new proteins and refine structural and functional properties of those already known. Such databases as PROSITE (Hofmann, et al., 1999), BLOCKS (Henikoff and Henikoff, 1991, Henikoff, et al., 1999), PRINTS (Attwood et al., 1999) are among the most well known and accessible via Internet. There is also a number of other similar databases i.e. PFAM (Bateman, et al., 1999), SBASE (Murvai, et al., 1999), IDENTIFY (Nevill-Manning, et al. 1998).

We have devoted our efforts to develop a technique and construct patterns for the greatest possible number of proteins belonging to the SWISS-PROT+TrEMBL (Bachinsky et al., 1996, 1997). We are convinced that if a secondary bank is not really representative, it would not be widely used. It is because of negative results in the comparison of a sequence with this bank force the user to consult other banks or make direct comparisons of the sequence with large banks of sequences.

Methods and algorithms

The selection and concurrent alignment of related protein groups

All full-length sequences of prototype banks that had more than 30 amino acids in lengths were combined in one file. In order to select groups of related proteins, a special program was written based on FASTA 2.0 (Pearson, 1994). The sequences similar in the sense of FASTA form a primary set of related proteins. Pairwise similarity of the proteins belonging to a set was assessed by the program CLUSTALV (Higgins et al., 1992). Then, if not all pairs of proteins had 30% similarity, the set was divided into subsets so that all pairwise similarities were at least 30%. Thus, more than 14,000 subsets or groups were obtained, containing more than 100,000 sequences.

Proteins of every subset were aligned together. The files containing aligned sequences were supplemented with two fields: DE - description(s) of proteins forming the group, and KW - key words (mainly the union of values of field KW for proteins falling into the set). Patterns were constructed based on such aligned families.

The construction of patterns of protein families. We will regard the combination of motifs that represent relatively conservative intervals of positions of aligned proteins of the family as a pattern of a family of related proteins.

The motifs of patterns are represented by ambiguous words of the type:

K-[D,E] - F - [I,V] - C - X - [A, S, T] - X - [M, N, D]... Thus, an initial pattern of a protein family is an ordered combination of non-overlapping motifs of the type $r:A_1-A_2-A_3-...-A_n$. Here r is position number of an aligned group of proteins (the trial sample), where the motif begins, A_j is a set of amino acids, located in $r+j-1$ position of the trial sample. For a passive position $A_j = X$: any amino acid is acceptable.

Comparison of amino acid sequences with patterns

The searches for exact matching between amino acid sequences' fragments and pattern motifs

The main algorithm for comparing an amino acid sequence with the pattern database uses the modification of finite automaton of Aho-Corasic (Aho and Corasic, 1975), constructed based on a set of samples, which are to be searched for in the input text. The automaton is presented as an oriented tree-like graph, where nodes are states of the automaton and arcs are admissible transitions from some states to the others, marked with symbols from the alphabet S of the amino acids' designation. The automaton works in cycles. In every cycle one more symbol of a text is read,

which determines the automaton's transition from the current state into a new one. The automaton's behaviour is characterised by three functions: function of transitions $G(s,a)$; rejections' function $F(s)$ and output function $O(s)$. The values of these functions are calculated once when constructing the automaton based on a given set of samples. In Fig. 1. the functions of the automaton constructed on the set of samples $R=\{r_1, r_2, r_3, r_4, r_5\} = \{HE, SHE, HIS, HER, HERS\}$ are illustrated.

When constructing the automaton in every motif, 4 neighbouring positions are chosen (the core of a motif), having minimum value of the product P_i and containing no passive positions. Then this core is

converted into exactly determined words of length 4 that act as samples in constructing automaton. If coincidence of a current fragment of an input sequence and one of the automaton samples is observed, comparison is performed (up to the first non coincidence) of all the other motif positions from the list of the output function, and the corresponding fragments of the sequence (the stage of extending the core). According to the results of this stage, the final decision is made on whether there is similarity or not.

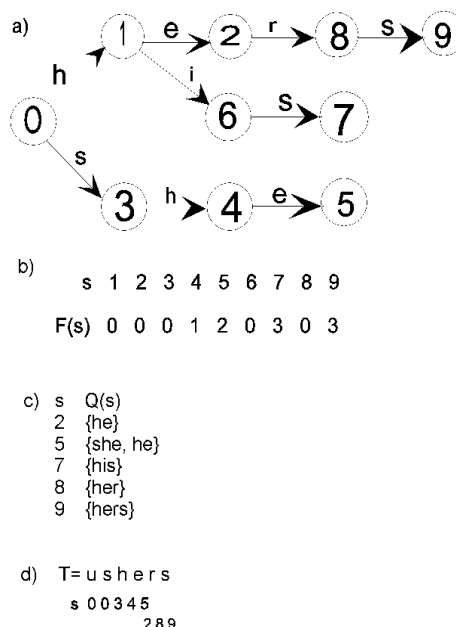


Figure 1. Illustrations of the functions of the Aho-Corasic automaton constructed on the set of samples $R=\{r_1, r_2, r_3, r_4, r_5\} = \{HE, SHE, HIS, HER, HERS\}$: a) graph representation of transition function $G(s,a)$; b) rejections' function $F(s)$; c) output function $O(s)$; d) automaton's transition from state to state if the input text is "ushers".

The search for distant similarity

To reveal a distant similarity, the algorithm of comparison is modified. The user specifies the matrix of similarity of amino acid residues (e.g., using the one from families PAM, BLOSUM, etc.) and D - the level of similarity within the limits of motif. For all states of the automaton, the function of rejection is set to zero. Besides, a sequence as a whole does not input to the automaton, but specially processed words.

The comparison of patterns with the parent banks

To examine the recognising ability of the patterns and exclude certain motifs, which are non-specific for a given family, all patterns were compared with all the proteins of the SWISS-PROT+TrEMBL. In the routine comparison between patterns and the banks, only exact similarities of two or more motifs per a pattern were registered, i.e. the cases when fragments of amino acid sequences belong to the motifs. The similarity is regarded as 'positive' one, if at least one of the two following conditions is met. 1. Query sequence belongs to the trial sample. 2. All words of one of the DE fields of the pattern (the names of the proteins forming the family) are present in the field DE (protein name) of the sequence. The similarity is considered 'conditionally positive' (UNKNOWN), if at least one of the DE or KW words of the pattern coincides with one of the words determined in fields DE and/or KW of the sequence. Thus, proteins are defined as conditionally related if they possess some common function (e.g., hydrolases, dehydrogenases, oxidoreductases, etc.) or some specific features of their structure (for instance, transmembrane segments). All other cases of similarity are regarded as false positive. As a result of comparison with the bank, a pattern bank entry is created, similar in its structure to entries of PROSITE.

Implementation and results

In version 1.3. presented here, the total number of motifs in more than 13,000 patterns is over 200,000 with specificities varying from one expected false positive prediction in 10^8 tests and higher. The total combined length of patterns is about 2,000,000 positions.

To find a distant similarity, a very fast flexible comparison procedure is employed, using the modified algorithm of Aho-Corasic (Aho and Corasic, 1975), various matrices of similarity/distance for amino acid residues, the predetermined grade of similarity between a fragment of an amino acid sequence and a pattern motif.

Patterns identify nearly 130 thousand of amino acid sequences of SWISS-PROT+TrEMBL as having shown 'positive' or 'conditionally-positive' similarity. In the latter case, the similar sequences, not included into the trial samples, are usually identified.

Almost all sequences of the trial samples are recognised by all motifs of the corresponding patterns. Certain violations of this rule are due only to the presence of non-standard symbols in the particular sequences of the trial samples that have fallen into the intervals of positions represented by pattern motifs.

A number of cases of false-positive similarity may be divided into two classes. Sometimes it is a really chance similarity. However, sometimes two or more pattern motifs show similarity to the fragments of a certain sequence; the order of the fragments' locations often correspond to that of the motifs' locations, which increases even more the certainty that the similarity is not random. In most cases, false-positive similarity is revealed with sequences described only as products of some genes, and this information is not included into descriptions of patterns.

All patterns were searched in 1480 new sequences of TrEMBL more recent version described as ORFs. 419 were undoubtedly identified. Some other comparisons of the PROF_PAT and other secondary banks show that PROF_PAT exceeds the most popular banks PROSITE, PRINTS, and BLOCKS under such index as number of patterns and motifs. The more pattern motifs show similarity to the sites of a query sequence, the higher would be the likelihood that the amino acid sequence is related to proteins of the trial sample (Henikoff and Henikoff, 1991), especially if the motifs' order coincides with that of the sample proteins.

Thus, we have constructed a bank of patterns for protein families, representing about 2/3 of the full-length protein sequences of the bank SWISS-PROT, release 38 and TrEMBL, release 11. The fast flexible search program for close and distant similarity provides comparisons of amino acid sequences of interest with the bank of patterns in the interactive mode. The PROF_PAT technology update has been developed and tested, so the new versions of PROF_PAT will be created following each new versions of SWISS-PROT+TrEMBL.

References

1. Aho,A.V. and Corasic,M.J. (1975) Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM*, **18**, 333-340
2. Attwood,T.K., *et al.* (1999) PRINTS prepares for the new millennium.. *Nucleic Acids Res.*, **27**, 220-225.
3. Bachinsky,A.G., *et al.* (1996) A new release of a bank protein family patterns PROF_PAT 1.0.: A technology of construction and programs of fast search. *Molecular Biology (Russian)*, **30**, 1409-1419.

4. Bachinsky,A.G. *et al.* (1997) A bank of protein family patterns for rapid identification of possible functions of amino acid sequences. *Comput. Applic. Biosci.*, **13**, 115-122.
5. Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.
6. Barker,W.C., *et al.* (1999) The PIR-International Protein Sequence Database. *Nucleic Acids Res.*, **27**, 39-43.
7. Bateman,A., *et al.* (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins *Nucl. Acids Res.*, **27**, 260-262.
8. Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucl. Acids Res.*, **19**, 6565-6572.
9. Henikoff,J.G., Henikoff,S. and Pietrokovski,S., (1999) New features of the Blocks Database servers. *Nucl. Acids Res.*, **27**, 226-228.
10. Higgins,D.G., Bleasby,A.G. and Fuch, R. (1992) CLUSTAL V: Improved software for multiple sequence alignment. *Comput. Applic. Biosci.*, **8**, 189-191.
11. Hofmann, K., *et al.* (1999) The PROSITE database, its status in 1999; *Nucleic Acids Res.*, **27**, 215-219.
12. Murvai,J., *et al.* (1999) The SBASE protein domain library, release 6.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **27**, 257-259.
13. Nevill-Manning, C.G., Wu, T.D. and Brutlag, D.L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc.Natl.Acad.Sci.*, **95**, 5865-5871.
14. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. in Griffin A.M., Griffin H.G., (eds) *Methods in Molecular Biology. Computer analysis of sequence data. Part 1.* Humana Press, Totova. **24**, pp.307-331.

ESTIMATION OF THE ENTROPY CHANGE UPON H-BOND FORMATION IN PROTEINS

*Rakhmaninova A.B., *Mironov A.A.*

State Scientific Center for Biotechnology NII Genetika, Moscow, Russia

e-mail: mironov@genetika.ru

*Corresponding author

Keywords: H-bond, protein folding, loop, entropy

Resume

Introduction

We are doing the computer modeling of free loops in proteins. Previously we have shown that helices in many proteins, in particular in DNA-binding proteins such as the phage 434 repressors (2cro,1r69), the TrpR repressor (2wrp), are longer than generally accepted (Grigor'iev et al., 1999). Thus it has been interesting to estimate the free energy change for formation of additional H-bonds and to compare it with the estimates for the energy cost upon the corresponding change of the loop conformation (Grigor'ev et al., 1997). It is especially important since there exist different, often conflicting, opinions about the energy cost of the H-bonds and the role of the latter in the protein folding (Stickle et al, 1992; Sippl, 1996; Yang and Honig, 1995). Here we report estimates of the formation probability for various H-bonds in the protein main-chain and calculate the corresponding entropy loss.

Model

We consider the poly-Gly-chain, $(-\text{NH}-\text{CH}_2-\text{CO}-)_n$, and the poly-Ala-chain approximated as $(-\text{NH}-\text{C}(\text{C})\text{H}-\text{CO}-)_n$, where n is the number of units. Using the Monte-Carlo procedure we select at random the values of the ϕ and ψ angles of all residues in the chain and then estimate the probability of the chain states with hydrogen bonds $H(i \rightarrow i+n)$, where $n=2,3,4,5$ according to the standard H-bond nomenclature (Kabsch & Sander, 1983). The results are expressed as the chain entropy loss: $T\Delta S = -kT \ln W_{S+H}/W_S$, where W_{S+H} is the number of the sterically allowed chain states with H-bond and W_S is the number of all sterically allowed chain states encountered in the experiment. This value is directly comparable with other energy parameters of the polypeptide chain.

The standard geometrical criteria for the H-bond formation are used (Stickle et al, 1992). Three modeling regimes are considered dependent on the choice of the minimum allowed interatomic distances that are defined as:

(R1) the "normal threshold" distances in proteins (Schultz & Schirmer, 1979);

(R2) the "minimal threshold" distances in proteins (Schultz & Schirmer, 1979);

(R3) the "minimal threshold" distances minus 0.1 Å.

In all cases the minimum allowed distance between O and H atoms connected by an H-bond is 1.8 Å.

Results

The Ramachandran plot for the dipeptide $-\text{C}-\text{CO}-\text{X}-\text{NH}-\text{C}$, $\text{X}=\text{Gly}$ or Ala , with the step 1° shows that small variations in the definition of the steric hindrances drastically influence the sterically allowed region of the plot (the plot for $\text{X}=\text{Ala}$ is given in Fig.1). At that, the size of the $\alpha(\text{R})$ region changes at most 1.4-fold and thus has little influence on the estimates of $T\Delta S$ for the right-handed turns. On the contrary, the correct estimate of the probability of left-handed turns is impossible, as it completely depends on the definition of the steric hindrances.

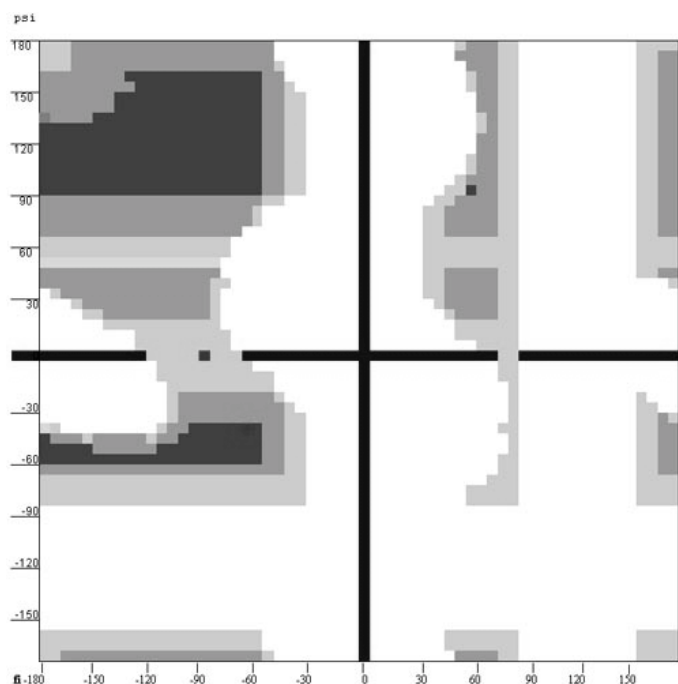


Figure 1. The Ramachandran plot for dipeptide C-CO-NH-C(C)H-CO-NH-C by different criteria for the steric hindrances. Shades: darkest: (R1), intermediate: (R2), lightest: (R3). The centers of the α (R) and of the transitional α/β regions are marked.

As shown by analysis of the influence of various steric hindrances on the formation probability of right-handed turns with different H-bonds in poly-Ala-chains, this probability decreases as the criteria for the steric hindrances toughen. The results for the case $H(i \rightarrow i+4)$ are shown in Fig. 2: compare the plots for (R1), (R2) and (R3). Moreover, the long-range steric interactions that are not taken into account by the Ramachandran plot, also significantly decrease the probability of the H-bond formation, see Fig.2. Therefore it seems that the popular methods based on the formalism of the Zimm-Bragg theory (Yang and Honig, 1995) underestimate the entropy change upon the H-bond formation.

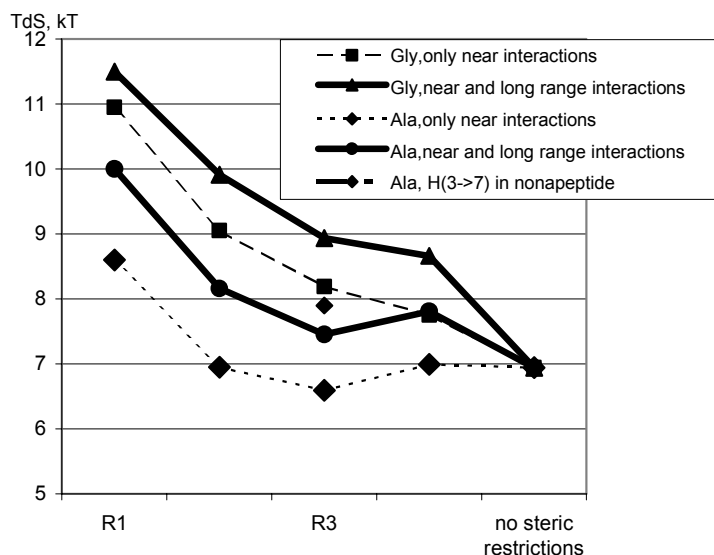


Figure 2. The entropy loss by the formation of the single right-handed $H(i \rightarrow i+4)$ -bond in Gly- and Ala-pentapeptides. R1-R4: see the model description.

Simulations of initiation and formation of the $3/10$ (R)-helix and α (R)-helix produce the following results. The minimal value of $T\Delta S$ in poly-Ala-chain is $7.9kT$ for initiation of the α (R)-helix and $6.0kT$ for initiation of the $3/10$ (R)-helix. Thus, to fix the first turn of the α (R)-helix, ΔH of a single H-bond should exceed $8kT$ ($4,6$ Kcal/mol). This seems to be unrealistic. The energy cost for consequent H-bonds in helices is much lower: for formation of the second H-bond $T\Delta S$ is $4kT$, and for formation of third H-bond, $3kT$ (Fig. 3).

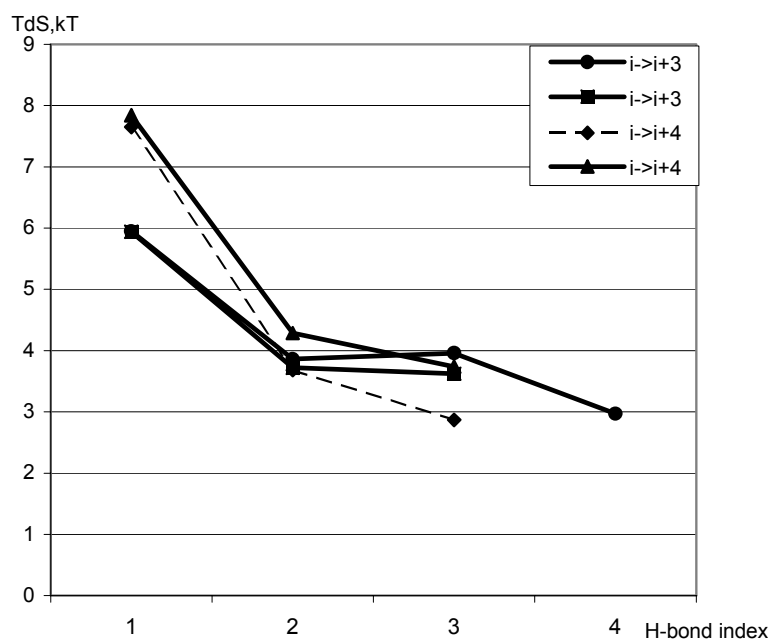


Figure 3. The entropy loss at 3/10(R)- and α (R)-helix initiation and propagation in Ala-octapeptides. Results of two different experiments are given.

It has been suggested that the 3/10(R)-helix can be an intermediate for initiation of the α (R)-helix (Sung, 1994). Indeed, $T\Delta S$ in poly-Ala-chain for formation of a forked H-bond ($i \rightarrow i+3$, $i \rightarrow i+4$) is $8.8kT$ through $9.0kT$. Assuming that such bonds 1.5-fold more than single H-bonds (Stickle et al., 1992), we obtain minimum requirement for ΔH of a single H-bond that can be reduced to $6kT$ (3.5 Kcal/mol).

Thus, we have for the first time estimated the influence of the long-range interactions on the probability of the H-bond formation. The obtained estimates for the entropy change of the polypeptide chain during helix initiation and propagation strongly contradict the assumption that H-bonds are an important factor stabilizing the protein secondary structures. At the same time the addition of one residue to an already stabilized helix is probably energetically favorable, since its cost in our estimates is $2kT$ through $3kT$.

References

1. Grigor'iev, I.V., Mironov, A.A. and Rakhmaninova, A.B. (1999) Refinement of helix boundaries in α -helical globular proteins. *Mol. Biol.*, 33, 206-214.
2. Grigor'iev, I.V., Derevyanko, S.V, Rakhmaninova, A.B. and Mironov, A.A. (1997) Estimation of the free energy of loops in globular proteins. *Mol. Biol.*, 31, 911-916.
3. Stickle, D.F., Presta, L.G., Dill, K.A. and Rose, G.D. (1992) *J. Mol. Biol.*, 226, 1143-1159.
4. Sippl, M.J. (1996) *J. Mol. Biol.*, 260, 644-648.
5. Yang, A.-S. and Honig, B. (1995) *J. Mol. Biol.*, 252, 351-365.
6. Kabsch, W. and Sander, C. (1983) *Biopolymers*, 22, 2577-2637.
7. Schultz, G.E and Schirmer, R.H. (1979) *Principles of Protein Structure*, Manheim, Springer.
8. Sung, S.-S. (1994) *Biophys. J.*, 66, 1796-1803.

THE SEARCH OF REGIONS IN HIV-1 PROTEINS THAT HAVE LOCAL SIMILARITIES WITH HUMAN PROTEINS

**Bazhan S.I., Bachinsky A.G., Maksyutov A.Z.*

SRC VB "Vector", Koltsovo, Novosibirsk region, Russia

e-mail: bazhan@vector.nsc.ru

*Corresponding author

Keywords: HIV-1, local similarity, epitopes

Resume

Motivation:

The potential ability of virus antigens to cause autoimmune responses and subsequent disease is known. It has been also shown that autoantibody production revealed in HIV-infected patients may be one of the key factors of AIDS pathogenesis. Therefore design of chimeric antigens, combining potentially protective epitopes both humoral and cellular immune responses, should be carried out with some restrictions. It is necessary to reveal and exclude from structure of a resulting antigen viral epitopes having similarity to human proteins.

Results:

Fragments in HIV-1 proteins having high similarity with human proteins have been revealed. Profiles of occurrence frequencies of fragments of human proteins in the HIV-1 proteins were constructed for estimation of potentiality of inclusion HIV-1 epitopes in a structure of a mosaic polyepitope immunogen.

Introduction

The lacking of effective vaccines and explosive increase of HIV-infected individuals in Russia and the World require new approaches to development of new effective anti-HIV vaccines. According to current view, one of the most promising approaches to developing of the next generation of effective and safe vaccines is based on the identification of T and B cell epitopes inside of viral proteins and creating on their basis synthetic polyepitope anti-HIV vaccines. Such vaccines should be free of many defects inherent to vaccines based on viable attenuated and whole-inactivated pathogen or natural subunits.

Many regions in HIV proteins have been identified, which have determined pathogenic properties. Therefore it is necessary to exclude such regions when designing of polyepitope antigens. It is known the potential ability of virus antigens to cause autoimmune responses and subsequent disease (Dyrberg and Oldstone, 1986; Solinema and De Camilli, 1995). The antibodies or cytotoxic T lymphocytes (CTLs) directed against a virus can cross-react with human proteins causing damage of cells resulting in development of illnesses. It has been shown that autoantibody production revealed in HIV-infected patients may be one of the key factors of AIDS pathogenesis (Nakamura, Nakamura, 1992). Therefore it is necessary to reveal and exclude viral epitopes having similarity with human proteins from a structure of designing chimeric polyepitope antigen.

The main purpose of the study was to search fragments of HIV-1 and human proteins, which are locally similar. Identification of such regions could lead to understanding of HIV immunopathological effects. The research was directed on searching of that regions, similarity of which is enough for induction of cross-reacting antibodies or CTLs or for competitive replacement of human proteins by similar fragments of virus proteins, causing interfering of macroorganism homeostasis.

Methods and algorithms

For search of local similarity between virus and human proteins all human proteins containing in banks SWISS-PROT (rl.38, JUL-1999)+TrEMBL (rl.12, NOV-1999) were chosen. There were 14,823 of such proteins. The comparison was carried out with use of the SIM program (Huang and Miller, 1991) received on EBI server.

Modified matrix ASM (antigenic similarity matrix) was used as a measure of amino acid residue similarity (Maksyutov et al., 1987). The matrix is modified by subtraction from all elements of number 75.

Parameters of the program SIM were: Gap-Open Penalty - 100, Gap-Extension Penalty - 20.

Some programs of the data preparation and analysis of results were specially written in Perl.

Implementation and results

Profiles of occurrence frequencies of fragments of human proteins in the HIV-1 proteins were constructed for estimation of potentiality of inclusion HIV-1 epitopes in a structure of a mosaic polyepitope immunogen. The

variants with number of matched positions not less than 7 and their part not less than 70 % were accepted (Fig. 1). Because such fragments covered the most part of positions of HIV-1 proteins, it was necessary to set some threshold, excess of which means that given region has statistically significant number of cases of local similarity. Random amino acid sequences were synthesized as a set of random chosen short fragments (of length 1-4) from real human proteins. Assuming Poisson distribution for number of cases of local similarity for concrete region, the estimation $N_0 = 6$ was received so that number of cases of local similarity (N) is nonrandom with probability 0,99 when $N > N_0$.

Table 1. Some cases of local similarities between HIV-1 and human proteins.

Similarity: human/HIV-1	DESCRIPTIONS: human/HIV-1
323 PRRARPGMELEERLLL -- : 844 PRRIRQG LERILL	DE ZONADHESIN (FRAGMENT). Versus ENV:
318 KERKLVDCHRELEK - 237 KWRKLVDF RELNK	DE RAD50. Versus POL:
1403 GLKXXX VIVIPVG - : 266 GLKXXXSVTVLDVG	DE VON WILLEBRAND FACTOR PRECURSOR. Versus POL:
231 DELELELAENRLLTE : : 464 EEAELELAENREILKE	DE NUMA PROTEIN. Versus POL:
470 DVNKQLEEAQQKI - 531 DV KQLTEAVQKI	DE BK125H2.1 (FRAGMENT). Versus POL:
363 KKKMKLK VKKSRE - 26 KKKYKLVHIVWASRE	DE TRANSCRIPTION FACTOR IIIA (FACTOR A). Versus GAG:
966 KVTTEAKLKKLEEEQ - : 95 KDTKEA LDKIEEEQ	DE MYOSIN HEAVY CHAIN, NONMUSCLE TYPE A. Versus GAG:
209 DIITEED KSKKKGQ - : 102 DKIEEEQNKSKKKAQ	DE LENS EPITHELIUM-DERIVED GROWTH FACTOR. And TRANSCRIPTIONAL COACTIVATOR P52. Versus GAG:
889 KGRPARFLDS PEP : - 442 KGRPGNFLQSRPEP	DE KIAA0620 PROTEIN (FRAGMENT). Versus GAG:
154 LQQRPEPT PEE : -- 461 LQSRPEPTAPPEE	DE THYROID HORMONE RECEPTOR ALPHA-2. Versus GAG:
274 DSHIRAALSIIERRKKR STGV : : - 78 DWHLGQGVSIWRKKRYSTQV	DE BONE MORPHOGENETIC PROTEIN 3 Versus VIF
37 AARERRRRARQERLRQ - 37 ARNRNRNRWR ERQRQ	DE CALDESMON (CDM). Versus REV:
524 QPPGLERLWLEGNPWDCG : - 74 QLPPLERLTDCNE DCG	DE INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN COMPLEX ACID LABILE CHAIN Versus REV:
510 SSPKTAWMNC MKKC : - 16 SQPKTACTNCYCKKC	DE TYROSINE KINASE. versus TAT:
250 SRLDPTGTFEKEMIGR -- 77 SRGDPTGP KEMAGR	DE 2,4-DIENOYL-COA REDUCTASE, versus TAT:

A question is open about measure of similarity, which is sufficient for an induction of cross-reactive immune response to reject potentially autoantigenic regions. It is necessary to note that potentially autoantigenic regions identified on the basis of proposed method give the information only of recommendation character because the importance of defined similarities should be evaluated in experiments. It is necessary to note that some regions of HIV-1 proteins with low values of similarity profiles ($N < N_0$) can have high local similarity with some concrete human proteins. Apparently, such regions should be also excluded when artificial immunogens, candidates of anti-HIV vaccines, are designing. In these cases the special attention should be paid to the role, which the human protein plays in organism homeostasis.

Some most interesting peptide sites, and also descriptions of human similar proteins are listed in Table 1.

Search of locally similar regions of HIV-1 and human proteins carried out earlier by other authors (Cantalloube et al., 1994) has revealed only one of fragments presented in the Table 1. The complete spectrum of found similarities contains many additional fragments, which could help to explain pathogenic mechanisms of HIV infection.

In conclusion we would like to emphasize that similarities between sequences of HIV-1 and human proteins give a key to rational design of safe polypeptide anti HIV-1 vaccines because allow to exclude epitopes, which could play a role in the pathogenesis of AIDS. Accordingly, anyone new vaccine constructions when designing, both chimeric and subunit, should be checked on the presence of locally similar regions with human proteins. Thus valid strategy is required for designing of anti-HIV vaccine to exclude undesirable epitopes that potentially can cross-react with normal host cell proteins.

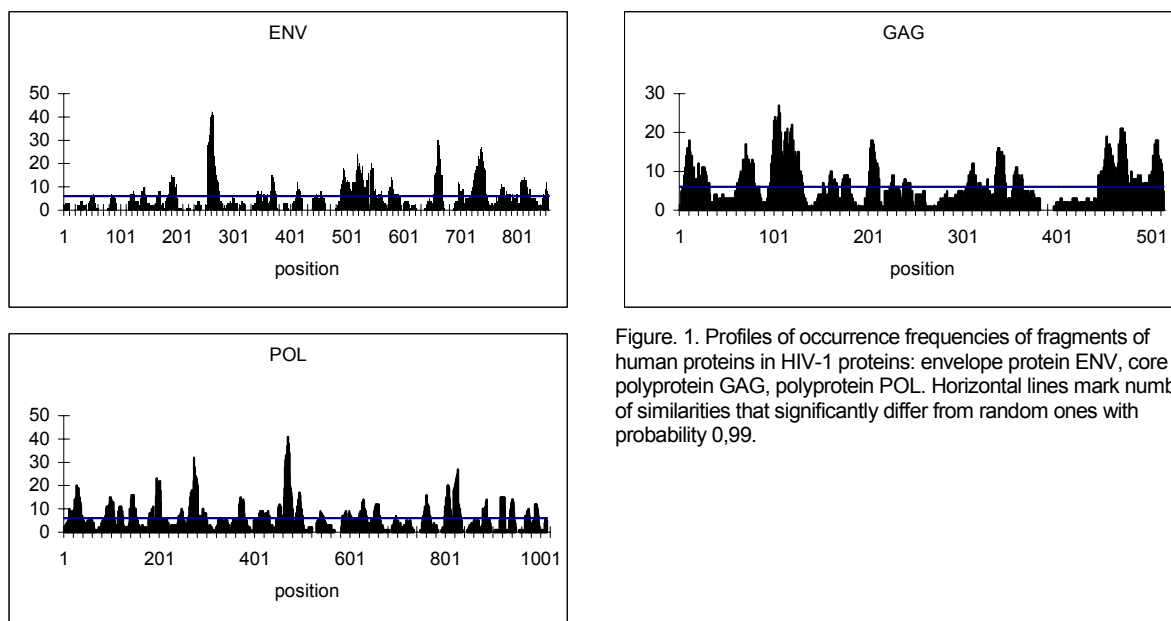


Figure. 1. Profiles of occurrence frequencies of fragments of human proteins in HIV-1 proteins: envelope protein ENV, core polyprotein GAG, polyprotein POL. Horizontal lines mark number of similarities that significantly differ from random ones with probability 0,99.

References

1. Cantalloube H.M.J., Nahum C.E. and Zagury J.F. (1994) Screening of protein sequences databases by Automat for search of host sequences integration and/or autoimmune disorders induction by retroviruses. *Biomed & Pharmacother*, **48**, 17-25.
2. Dyrberg T. and Oldstone M.B.A. (1986) Peptides as Probes to Study Molecular Mimicry and Virus-Induced Autoimmunity. *Curr. Top. Microbiol. Immunol.* **130**, 25-37.
3. Maksyutov A.Z., Eroshkin A.M. and Kulichkov V.A. (1987) Method for calculating immunochemical cross-reaction between homologous proteins. *Molecular Biology*, **21**, 30-37.
4. Nakamura M.C. and Nakamura R.M. (1992) Contemporary concept of autoimmunity and autoimmune diseases. *J. Clin. Lab. Anal.*, **6**, 841.
5. Solinena M. and De Camilli P. (1995) Coxsackieviruses and diabetes. *Nature Med.* **1**, 25-26.
6. Huang X. and Miller W. (1991) A Time-Efficient, Linear-Space Local Similarity Algorithm, *Advances in Applied Mathematics*, **12**, 337-357,

L-ZIP MOTIF AS A PROBABLE DIMERIZATION MOTIF OF LTB4 RECEPTOR

*Lukashev V.A., Lukashova V.V., Rola-Pleszczynski M., *Stankova J.*

Sherbrooke University, Quebec, Canada

e-mail: stankova@courrier.usherb.ca

*Corresponding author

Keywords: LTB4 receptor, GPCR, leucine zipper motif, dimerization

Resume

Motivation:

Computer modeling to predict the probable sites in LTB4 receptor that might be responsible for the dimerization of this protein.

Results:

The regions potentially involved in the formation of spatially neighboring domains within one receptor subunit as well as between two or more receptor subunits were determined.

Introduction

Dimerization or oligomerization of cell surface receptors has been shown to occur after binding of several polypeptide hormones, cytokines and growth factors to their receptors. These include protein-tyrosine kinase receptors, cytokine receptors, antigen receptors, receptors for tumor necrosis factor (TNF) and related factors, and serine/threonine kinase receptors. G protein-coupled receptors are believed to signal as monomers. However, recent evidence suggests that not only GPCRs could form dimer/oligomer structures, but their dimerization may also play an important role in signal transduction (1). Hebert et.al. provided the first direct evidence that beta₂-adrenergic receptors can exist in dimeric form (2). Several structurally distinct GPCRs have also been demonstrated to form homodimers, including glutagon receptor (3), δ -opioid receptor (4), dopamine 2 receptors (5). Using different molecular and biochemical approaches, it was clearly demonstrated that muscarinic m3 (6) as well as CaR receptor (7) can be present as dimers on the cell surface. Co-immunoprecipitation experiments also provided the first physical evidence for GABA_BR1 and GABA_BR2 receptor heterodimerization (8).

Leukotriene B₄ is a powerful mediator, that represents one of the products of 5-lipoxygenation of arachidonic acid, secreted by activated phagocytic cells, mainly by neutrophils. The LTB4 receptor, originally cloned by Yokomizo et. al (9), is a member of the G protein-coupled receptor superfamily.

In the present report, we employed computer modeling programs to predict the probable sites in LTB4 receptor that might be responsible for dimerization.

Materials and Methods

The search for high homology regions in the LTB4 receptor with different members of the GPCR family as well as with other proteins employed the POISK and MOTIV programs described previously (10,11). The programs involve the scanning of the profile of coordinate numbers of the protein of interest and comparative analysis with such profiles of other proteins over the protein primary structure database PIR, release 34.0 (12).

Results and Discussion

Peptide segments of various lengths of the LTB4 receptor were compared with primary amino acid sequences from the PIR database using computer programs MOTIV and POISK. Several high homology regions (\approx 40 %) were found between the LTB4 receptor and different proteins from the G-protein coupled receptor (GPCR) superfamily as well as other cell receptors. The sequence of the LTB4 receptor from amino acid 75 to 151 is thought to be one such continuous high homology region.

Another interesting domain was detected in the third extracellular loop (we designated it as "4-out"). This peptide segment may be considered as a putative "L-Zip-motif". Such a motif has been shown to be important in the dimerization of some transcription factors necessary for binding to DNA. Several works showed the importance of the zipper region for Myc dimerization specificity (13, 14). Moreover, substitution of negatively charged glutamic acid residues with a polar and a nonpolar amino acids in Myc transcription factor was shown to be essential for Myc homodimerization (15). Fig. 1A shows the comparative analysis of "L-Zip-motif" of the LTB4 receptor with those found in the virus glycoprotein gp41 from several HIV-1 strains. A single amino acid

substitution of the most conserved isoleucine in L-zip motif of gp41 was shown to abrogate envelope –mediated cell fusion and virus infectivity but did not interfere with protein transport and oligomerization (16). A similar motif (homology with some mismatches) was determined in other proteins (Fig. 1B). High degree homology (and, probably, structural analogy) in the region containing the “L-Zip-motif” was observed in LTB₄ receptor, IGFBP-2, heat shock protein 101 and others. Two structurally similar “L-Zip-motifs” were also discovered in the β_3 -adrenergic receptor and in the homeobox containing gene transcription factor LBX-1.

We performed the search of peptide segments in the LTB₄ receptor that could arrange themselves in close proximity. Our results pointed out the probable regions involved in the formation of spatially neighboring domains within one receptor subunit as well as between two or more receptor subunits.

We discovered probable spatial contact between two first extracellular (“2-out”) loops. Spatial affinity was determined to exist between the region of residues from 80 to 90 of the LTB₄R and the same peptide segment in the opposite direction. Similar results were obtained for the receptor region containing “4-out” loop – ARs 250-266. We could predict spatial interaction of two “4-out” loops in the opposite direction). It is interesting to notice that this region consists of a probable “L-Zip-motif”. Spatial contact can be attributed to this part of the domain. The peptide fragment “4-out” was predicted to be largely α -helical in the region next to the TM domains. In conclusion, Fig. 2 shows a model of the simultaneous interaction of peptide segments “2-out”, “4-out” and “4-out”. We propose that two subunits of LTB₄R could form a dimer due to spatial contact of peptide segments “2-out” with “2-out”, “4-out” with “4-out”, and “4-out” with “2-out”. In the tertiary structure of LTB₄R, peptide segments “2-out” and “4-out” could arrange themselves in close proximity and thus form a spatial domain. Briefly, two molecules of LTB₄R could interact by their “4-out” loops in opposite direction and by “2-out” segments in both directions.

Since “2-out” and “4-out” loops form a cluster, it would be interesting to observe the relative arrangement of transmembrane (TM) domains II and III with TM domains VI and VII. Our results, presented in Fig. 3, indicated a strong spatial affinity between TM domains III and VI. Moreover, TM domain III could form a cluster with TM VII. In this model, the “2-out” loop could be found in close proximity to the “4-out”, the TM III domain might make contact with TM VII and TM VI domains in forward and reverse directions, respectively. Analysis of affinity of TM domains II and IV (data not shown) confirmed our model of TM domain interactions. TM domain III seemed to be a “core” interaction domain. Hebert et al. showed that peptides derived from the sixth transmembrane domain (TMVI) reduced the amount of β_2 -adrenergic receptor dimerization in a dose-dependent manner. Similar results were obtained for a peptide derived from TM VI or TMVII of the D2 dopamine receptor. The GpA dimerization motif, proposed for the glycophorin A receptor, was found to be partially conserved within the adrenergic receptor family. In the M5 metabotropic glutamate receptor it was demonstrated that receptor dimerization is mediated by disulfide bond formation and that the domain involved is the large extracellular N-terminus (17). Two conserved extracellular Cys residues play key roles in the formation of disulfide-linked muscarinic m3 dimers (6). In conclusion, different domain may determine GPCR dimerization and maintain the specificity of this process. Further research is required to confirm our prediction.

References

1. Hebert T.E., Bouvier M.(1998) *Biochem. Cell. Biol.* 76, 1-11
2. Hebert T.E., Moffet. S., Morello J.-P., etc.(1996)*J. Biol. Chem.*271, 16384-16392
3. Herbeg J.T., Codina J., Rich K.A., etc. (1984) *J. Biol. Chem.* 259, 9285-9294
4. Cvejic S., Devi L.A. (1997) *J. Biol. Chem.* 272, 26959-26964
5. Gordon Y.K.Ng., O'Dowd B.F., Lee S.P.,etc. (1996) *Biochem and Biophys. Res. Com.* 227, 200-204.
6. Zeng F.-Y., Wess J. (1999) *J. Biol. Chem* 274, 19487-19497
7. Bai M., Trivedi S., Brown E.M. (1998) 273, 23605-23610
8. Marshall F.H., Jones K.A., Kaupmann K., Bettler B. (1999) *TIPS*, 20, 396-399
9. Yokomizo T., Izumi T., Chang K., Takuya Y., Shimizu T (1997) *Nature* 387, 620-624
10. Lkashev V.A., Bausk N.V., Mazalov., etc. *Instr. And Methods in physics Res.* (1995) 359, 259-262.
11. Lukashev V.A.,Kulichkov V.A. *Biophysika* (1990) 35, n2, 236-241
12. SWISS-PROT Protein Sequence Data bank release 34.0 October 1996
13. Amati B., Brooks M.W.,Levy N.,Littlewood T.D.,etc.(1993)*Cell*, 72, 233-245
14. Marchetti A., Abril-marti M., Illi B., etc. (1995)*J. Mol. Biol.* 248, 541-550
15. Soucek L., Helmer-Citterich M., Sacco A., etc. (1998) *Oncogene* 17, 2463-2472
16. Dubay J.W., Roberts S.J., Brody B., Hunter E. 1992 *J. of Virology* 66, 4748-4756
17. Romano C., Yang W.-L., O'Malley K.L. (1996) *J. Biol. Chem.* 271, 28612-28616

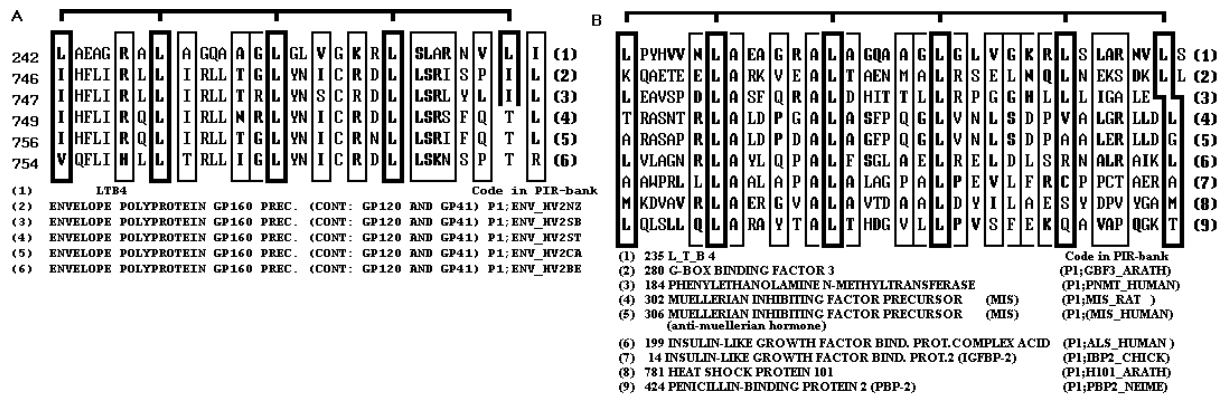


Figure 1. Leucine zipper region of LTB4 receptor. Sequence alignment of several HIV-1 isolates (A) and several proteins (B) shows a conserved nature of L-zipper motifs with that of LTB4 receptor.

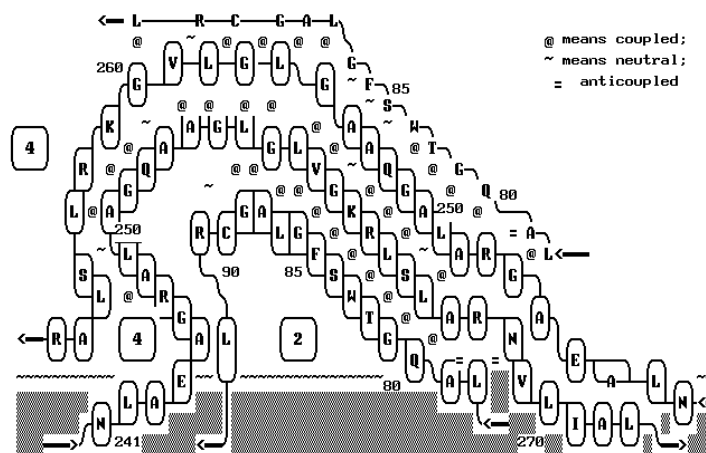


Figure 2. The model of probable spatial interaction of the "4-out" loop with "4-out" and "2-out" loops in LTB4 receptor.

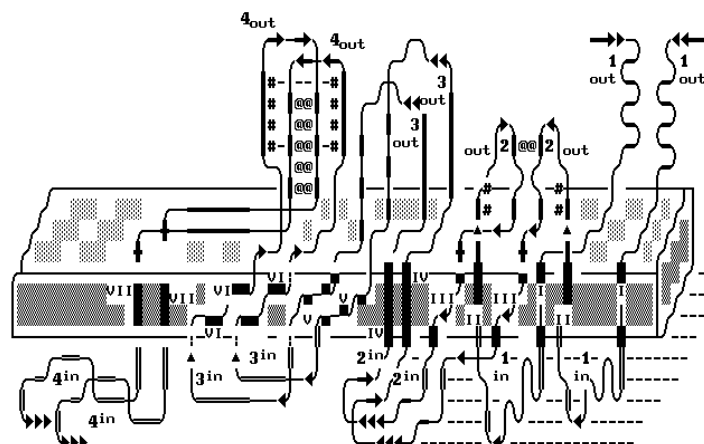


Figure 3. The probable interaction of transmembrane domains TM III, TM VI and TMVII in LTB4 receptor.

INVESTIGATION OF THE AMINO ACID SEQUENCES OF MYCOBACTERIUM TUBERCULOSIS COMPLETE GENOME WITH PROTEIN FAMILY PATTERNS BANK PROF_PAT 1.3

**Nizolenko L.Ph., Kozhina E.M., Yarigin A.A., Bachinsky A.G.*

SRC VB "Vector", Koltsovo, Russia.

e-mail: bazhan@vector.nsc.ru

*Corresponding author

Keywords: protein families, patterns, data banks, amino acid sequences, protein comparison, Mycobacterium tuberculosis, complete genome

Resume

Motivation:

Despite the availability of effective short-course chemotherapy, recent years have seen increased incidence of tuberculosis in developing as well as in industrialized countries. At least 1/3 of the world population is infected by Mycobacterium tuberculosis. Tuberculosis was declared a global emergency. Radical measures are needed to solving this problem. The combination of genomics and bioinformatics has the potential to generate information for elucidation the unusual biology of this airborne disease agent, Mycobacterium.tuberculosis.

Results:

Were investigated 3924 amino acid sequences of Mycobacterium tuberculosis strain "H37Rv". The similarity to proteins with known function or to big groups of hypothetical proteins was confirmed for more then half of this sequences. 30 proteins which unknown function were identified as a members of PROF_PAT families

Introduction

A number of works appeared in the last few years in the bioinformatic, aimed at the selection of sites (patterns, blocks, motifs) in groups of related proteins, which are representative of a protein family as a whole, at making "secondary data banks" of this objects and at their use both to identify new proteins and to refine structural and functional properties of those already known.

When this objects are constructed, the completeness of representation of the prototype bank proteins is very important. Otherwise, the negative result of some amino acid sequence test makes us to compare it directly with all "primary" protein bank.

According to this characteristic bank PROF_PAT surpasses other similar ones. Besides, it is highly sensitive and specific. Moreover, it allows to investigate not only a few e protein sequences, but large groups of them, even as large as all amino acid sequences, translated from complete genomes of microorganisms.

Methods and Algorithms

3924 amino acid sequences of Mycobacterium tuberculosis strain "H37Rv", presented in the Internet (Cole et al. 1998) were investigated.

Protein family patterns and the bank of this patterns PROF_PAT constructed by original technology (Bachinsky et al. 1996). Amino acid sequences become the members of the protein family, when the similarity level for all pair combinations exceeds 30%.

The version of the patterns bank PROF_PAT 1.3 constructed on the basis of the 38th release of the SWISS-PROT bank and 12th release of TREMBL, contains patterns of 16139 groups of related proteins including more then 139000 amino acid sequences.

The main algorithm of comparison of amino acid sequences with bank PROF_PAT exploits the Aho-Corasic finite automation (Aho A.V., Corasic M.J. 1975, Bachinsky A.G. et al. 1996). For M.tuberculosis proteins analysis the similarity matrix PAM250 and the level of similarity 70% were used (Bachinsky et al.,1999). The degree of closeness of tested amino acid sequence and PROF_PAT family was defined by parametr "Score" = $\lg P$, where P - the probability of chance similarity between sequence fragments and motifs of pattern, which identify them.

Implementation

We have investigated amino acid sequences of open reading frames of Mycobacterium tuberculosis strain "H37Rv" complete genome by protein family patterns bank PROF_PAT. The similarity to proteins with known

function or to big groups of hypothetical proteins, described by other investigators (Cole S.T. et al. 1998) was confirmed for 2057 sequences.

For 1310 sequences the “negative results” were confirmed. It means, that comparison neither with PROF_PAT, nor with some other banks (EMBL, TREMBL, PROSITE, SwissProt (Cole S.T. et al. 1998)) no detected similarity with any proteins group.

502 amino acid sequences of *M.tuberculosis*, described by Cole S.T. et al. (1998) similar to some proteins with appointed function, do not show any similarity with patterns of bank PROF_PAT.

The most interesting we consider to be two sets of amino acid sequences of *M.tuberculosis*.

First one presented in the table 1. There are sequences, which similarity to any proteins except hypothetical ones with unknown function was not described up to now. Every member of this set is identified by one (or more) PROF_PAT protein family pattern. And there are proteins in these families with determined or predicted function.

Second set consist of amino acid sequences described by Cole S.T. et al. (1998) similar to known proteins. However, the results of comparisons of these sequences with PROF_PAT bank are in disagreement with their data. Some of them seems to be more close to family of hypothetical proteins.

For sequences, presented in table 2, our results really contradict Cole S.T. et al. (1998). We have compared this sequences with proteins of the last 1999 release of SWISS-PROT and TREMBL directly and found them to be much more similar to proteins from corresponding PROF_PAT families, then to proteins, described as their relatives before.

In this way, on a level with confirmation of known data, we have described 30 proteins of *Mycobacterium tuberculosis*, which function was unknown up to now or, probable, was identified incorrectly, to be a members of PROF_PAT families. And so we can propose a new function for them.

Thus, expediency of bank PROF_PAT use was demonstrated again. It allows to get new information for assumption of structural and functional similarity for distinct proteins as well as for large groups of amino acid sequences.

Table 1. The most reliable cases of *M.tuberculosis* amino acid sequences identification by protein family patterns bank PROF_PAT.

Sequence	m*/n**	Description of the bank PROF_PAT protein family	Sequence description (Cole S.T. et al. 1998)
Rv0040c	12/13 - Score=143.82	PROLINE RICH 28 KD ANTIGEN	Unknown
Rv0496	13/14 - Score=142.04	DNA-BINDING PROTEIN	Similar to hypothetical proteins
Rv0650	12/12 - Score=124.27	BACTERIAL INTERNALIZATION PROTEIN, GLUCOKINASE (EC 2.7.1.2)	Unknown
Rv1173	10/13 - Score=82.195	BIOTIN SYNTHETASE RELATED PROT.	Hypothetical protein
Rv1466	5/5 - Score=53.535	DNAG, RPOD, CPOA, SIGMA 42 PROTEIN (DTDP-4-KETO-L-RHAMNOSE REDUCTASE)	Similar to ORF's downstream of sigma factors ORF3 downstream of RpoD
Rv1488	9/10 - Score=63.462	MEMBRANE PROTEIN STOMATIN-LIKE PROTEIN	Hypothetical protein
Rv1504c	6/14 - Score=40.424	POLYSACCHARIDE BIOSYNTHESIS PROT.	Hypothetical protein
Rv1540	12/12 - Score=133.05	PROBABLE PSEUDOURIDINE SYNTHASE	Member of the yabO/yceC/yfil family of hypothetical proteins
Rv1711	11/11 - Score=90.333	RIBOSOMAL LARGE SUBUNIT PSEUDOURIDINE SYNTHASE B (PSEUDOURIDYLATE SYNTHASE) (URACIL HYDROLYASE)	Similar to a large family of hypothetical proteins
Rv2165c	8/8 - Score=83.888	MRAW PROTEIN (YLLC PROTEIN)	Strong similarity to hypothetical proteins. Belongs to the YABC, YLXA family
Rv2182c	9/9 - Score=104.18	SAM-DEPENDENT METHYTRANSFERASE	Unknown integral membrane protein
Rv2188	9/9 - Score=104.18	PUTATIVE TRANSMEMBRANE PROTON-DEPENDENT TRANSPORT PROTEIN, LIPASE	Unknown integral membrane protein
Rv2188	18/18 - Score=191.15	GLYCOSYL TRANSFERASE	Similar to several hypothetical proteins
Rv2511	10/10 - Score=107.08	OLIGORIBONUCLEASE (EC 3.1.-.-)	Equivalent to <i>M. leprae</i> cosmid L383; similar to <i>E.coli</i> prot. in psd-amib intergenic region.

Sequence	m*/n**	Description of the bank PROF_PAT protein family	Sequence description (Cole S.T. et al. 1998)
Rv2604c	8/8 - Score=76.644	IMIDAZOLEGLYCEROL-PHOSPHATE SYNTHASE, SUBUNIT H, AMIDOTRANSFERASE, HISH PROTEIN	Unknown but highly similar to M. leprae G466810 HISH, also eg to YAAE_BACSU P37528 hypothetical 21.4 kd protein
Rv2794c	10/10 - Score=100.03 9/9 - Score=99.225	LANZ6 IRON-CHELATING COMPLEX SUBUNIT L-PROLINE 3-HYDROXYLASE,	Similar to two hypothetical proteins from Streptomyces sp.
Rv2852c	26/26 - Score=281.8	MALATE:QUINONE OXIDOREDUCTASE (EC 1.1.99.16) (MALATE DEHYDROGENASE (ACCEPTOR)) (MQO)	Unknown, similar to YOJH_ECOLI P33940 hypothetical 54.3 kd protein
Rv2961	7/7 - Score=83.278	TRANSPOSASE	Similar to hypothetical proteins
Rv3024c	17/17 - Score=140.64	TRNA (5-METHYLAMINOMETHYL-2-THIOURIDYLATE)-METHYLTRANSFERASE	Similar to hypothetical proteins
Rv3586	14/14 - Score=135.57	DNA-BINDING PROTEIN	Similar to hypothetical proteins

*m - The number of motifs, which identify amino acid sequence.

**n - The number of motifs in the pattern.

Table 2. A certain cases of disagreement of M.tuberculosis amino acid sequences identification by protein family patterns bank PROF_PAT and data published before.

Sequence	m*/n**	Description of the bank PROF_PAT Protein family	Sequence description (Cole S.T. et al. 1998)
Rv0126	26/26 - Score=301.8	TREHALOSE SYNTHASE, A3(2) GLYCOGEN METABOLISM CLUSTER I	Possible glycosyl hydrolase similar to possible maltase; also to proteins associated with amino-acid transport
Rv0554 (bpoC)	14/14 - Score=147.79	THIOSTERASE	Similar to BPA2_STRAU P29715 non-haem bromoperoxidase bpo-a2
Rv1939	7/7 - Score=71.562	ACTINORHODIN POLYKETIDE DIMERASE, OXIDOREDUCTASE	Similar to nitrilotriacetate monooxygenase
Rv3538 (ufaA2)	9/9 - Score=94.939	17-BETA-HYDROXYSTEROID DEHYDROGENASE TYPE IV	Unknown fatty acid methyltransferase similar to G886104 cyclopropane mycolic acid synthase (CMA1)
Rv3752c	4/4 - Score=42.195	NITROGEN FIXATION PROTEIN	Unknown, probable member of the cytidine and deoxycytidylate deaminases family

*m - The number of motifs, which identify amino acid sequence.

**n - The number of motifs in the pattern.

References

1. Aho A.V., Corasic M.J. (1975) Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM.*, 18, P.333-340.
2. Bachinsky A.G., Yarigin A.A., Gusev V.D., Naumochkin A.N., Nemitnikova L.A., Nizolenko L.Ph., Kulichkov V.A. (1996) A new version of a bank of patterns of protein families PROF_PAT 1.0. *Molecular biology (Rus)*, 30, P. 1409-1419.
3. Bachinsky A.G., Yarigin A.A., Naumochkin A.N., Nizolenko L.Ph., Kulichkov V.A. (1999) Network release of the protein family pattern bank PROF_PAT 1.1. *Molecular biology (Rus)*, 33, P. 873-880.
4. Cole S.T., Brosch R., Parkhill, J., Garnier T., Churcher C., Harris D., Gordon S.V., Eiglmeier K., Gas S., Barry III C.E., Tekaia F., Badcock K., Basham D., Brown D., Chillingworth T., Connor R., Davies R., Devlin K., Feltwell T., Gentles S., Hamlin N., Holroyd S., Hornsby T., Jagels K., Krogh A., McLean J., Moule S., Murphy L., Oliver S., Osborne J., Quail M.A., Rajandream M.A., Rogers J., Rutter S., Seeger K., Skelton S., Squares S., Squares R., Sulston J.E., Taylor K., Whitehead S. and Barrell B.G. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393, P.537-544
5. Cole S.T. et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 396, P.190-198.

EFFECT OF HUMAN NON-SYNONYMOUS SINGLE NUCLEOTIDE POLYMORPHISMS UPON A PROTEIN STRUCTURE

^{1,2,3}*Sunyaev S.*, ^{3*}*Ramensky V.*, ^{1,2}*Bork P.*

¹European Molecular Biology Laboratory, Heidelberg, Germany

²Max Delbrueck Center for Molecular Medicine (MDC) Berlin, Germany

³Engelhardt Institute of Molecular Biology, Moscow, Russia

e-mail: ramensky@imb.ac.ru

*Corresponding author

Keywords: single nucleotide polymorphisms, protein structure

Resume

About 90% of human genetic variety has been ascribed to single nucleotide polymorphism (SNP) allelic variants with a frequency higher than one percent. Due to the application of high throughput SNP detection techniques, the number of identified SNPs is growing rapidly and this allows detailed statistical studies. This also applies to SNPs that affect the amino acid sequence of a gene product (non-synonymous SNPs); they complement the large body of literature on mutations, causing Mendelian diseases, that represent the usually rare non-synonymous mutations with an allele frequency far below one percent.

To understand the relation between genetic and phenotypic variations, it is essential to assess the structural consequences of the respective non-synonymous mutations in proteins. To quantify how often a disease phenotype can be explained by a destructive effect on protein structures or functions, we have mapped known disease mutations onto known three-dimensional structures of proteins. The results were compared with a control set of substitutions observed between these proteins and their closely related homologues from other species which are unlikely to cause severe effects on the phenotype. With the knowledge about the structural properties of these two sets, we have also mapped a large number of non-synonymous SNPs (which are usually thought to be neutral or only cause minor phenotypic effects) onto protein structures. This enables us to estimate a lower limit for the quantity of non-synonymous SNPs with likely phenotypic effects which is an important baseline for the current efforts to identify SNPs associated with multifactorial human disorders.

The three data sets needed for the comparative analysis: (1) disease-causing mutations, (2) substitutions between close homologues in human and other species and (3) human non-synonymous SNPs, as well as structural characteristics of corresponding proteins were extracted from public databases.

As a result of the comparison of disease-causing mutations with between-species substitutions in the same set of proteins we found that disease-causing mutations are much more likely to occur at sites with low solvent accessibility. In fact, 35% of 551 disease-causing mutations from our dataset affect buried sites while only 9% of 225 substitutions between species do. This indicates that disease-causing mutations often affect intrinsic structural features of proteins. To increase the discrimination between the two sets we also took into account possible interaction sites. Overall, about 70% of the disease-causing mutations are located in sites likely to be structurally and functionally important, namely sites with less than 5% solvent accessibility or in β -strands, active sites, sites involved in disulphide bonds or evolutionary conservative sites (defined as sites with HSSP variability parameter VAR <10). In contrast, in the same set of proteins only 17% of substitutions observed between human sequences and closely related homologues from mammalian species are located at these sites.

Unexpectedly, the fraction of polymorphic sites located in structurally and functionally important regions (defined as described above) was 45%, which is significantly higher than the 24% in the case of the interspecies variation when considering proteins from the dataset of polymorphic sites (P-value of the χ^2 -test equals to 0.00013). In this set we observe the abundance of immune system- related proteins with high β -strand content; this fact explains the 17% vs. 24% difference for two protein sets. The result above suggests that a significant fraction of human protein allelic variants is represented by amino acid substitutions with a strong impact on protein structure, function, stability or folding. These variants are normally eliminated during long evolutionary times as can be seen from the comparison with the interspecies variation. One would expect, that variants under pressure of purifying selection tend to have a lower allele frequency. Indeed, for non-synonymous SNPs we observe a correlation between allele frequency and fraction of occurrence in structurally and functionally important regions. The observation that many non-synonymous SNPs are likely to have a phenotypic effect may be considered as indirect evidence that common amino acid variants may contribute to genetic risk of common human disorders (so called the common disease-common variant hypothesis).

In summary, there is a surprisingly high fraction of non-synonymous SNPs that affect structure and, probably, function of proteins. This implies that a considerable fraction of the non-synonymous SNPs have indeed some (probably negative) effect on the phenotype. The allele frequency distribution makes it evident that variants in structurally important sites are not selectively neutral. Taking these observations and given the progress in structural genomics as well as in large scale SNP discovery, the comparative analysis of structural properties of protein allelic variants such as described above should have an important role in the pre-selection of candidates for disease-association studies and will help in the explanation of phenotypic effects.

ANALYSIS OF HEPATITIS C VIRUS PROTEINS USING SEQUENCE AND PUBLISHED DATA

*Sobolev B.N., Poroikov V.V., Matveev I.V., Olenina L.V., Kolesanova E.F., *Archakov A.I.*

Institute of Biomedical Science, Moscow, Russia

e-mails: boris@ibmh.msk.su, vvp@ibmh.msk.su, matveev@ibmh.msk.su, oleninal@ibmh.msk.su, ekol@ibmh.msk.su, archakov@ibmh.msk.su

*Corresponding author

Resume

Motivation:

Hepatitis C is one of the major human pathogen. HCV investigations as well as developing drug and vaccine are difficult besides of extreme variability of its genome. For revealing the functional characteristics of HCV proteins is needed extended sequence studies with accompany of published experimental results. Our work allowed us to construct the functional maps of HCV proteins and represent these results as database.

Results:

- 1) Based on alignments derived from 4000 HCV polyprotein sequences (whole and fragmental) of different isolates, we detected variable and conservative regions of envelope HCV proteins. We also construct the antigenic and functional maps for the envelope, core and nonstructural proteins of HCV.
- 2) The most variable region of E2 protein HVR1 fragments analysed. We assume that this region adopts strongly definite conformation in relatively independent fold unit and have a certain functional role.
- 3) All collected sequences and functional data have been organised as the database on HCV functional and antigenic maps.

Availability:

The database on HCV functional and antigenic maps will be available via Internet at <http://www.ibmh.msk.su/HCVMAP/> on July 2000.

Introduction

Hepatitis C virus (HCV) is one of the major etiologic agents of parenteral hepatitis [1, 2]. Taking into account the importance of developing drugs and vaccine against hepatitis C HCV protein sequences are studied. The HCV genome contains the only ORF coding the precursor polyprotein and all structural and non-structural proteins are formed by the cotranslational processing of this polyprotein. A lot of sequences (whole polyprotein sequences and fragmental ones) and published experimental data on protein functions are now available. Extreme variability of the HCV genome sequences creates many difficulties in analysis of probable and determined functional sites and antigenic determinants. Our aim is to develop the functional mapping for HCV proteins. Such maps would help in planning further studies.

Methods

Amino acid sequences and literature references were retrieved from the non-redundant database using retrieval systems SRS6 at EBI server and Entrez at NCBI.

Software was developed, which processed the sets of aligned sequences, calculating the alignment profile matrix and searching the patterns in aligned sequences. We applied the additional information (e.g. glycosylation site locations, functionally significant regions) to refine the alignments. The obtained alignment were stored in database as particular sets and processed for further studies by calculating frequency occurrence matrices (profiles). The aligned sequence sets were also used to obtain the blocks corresponded to different functional or antigenic regions. We used collected and obtained data sets to construct the database on functional and antigenic mapping of HCV sequences.

Results and Implementation

Analysis of a large set of E1 and E2 protein sequences (Fig. 1) revealed that a half of E2 and one third of E1 of aligned positions have more than 95% level of identity. These values are considered to be enough for maintaining the common fold and function. On the other hand, the very seldom exchanges can be associated with functionally inactive proteins or defective genomes. Our unpublished data revealed that exchanges, which must result in inactivating the vitally important enzyme functions of HCV proteins, are found in rare cases. 8 and 18 conservative Cys residues were found in E1 and E2 sequences (8 and 18 residues respectively). These residues suggested to form intramolecular but not intermolecular disulphide bonds.

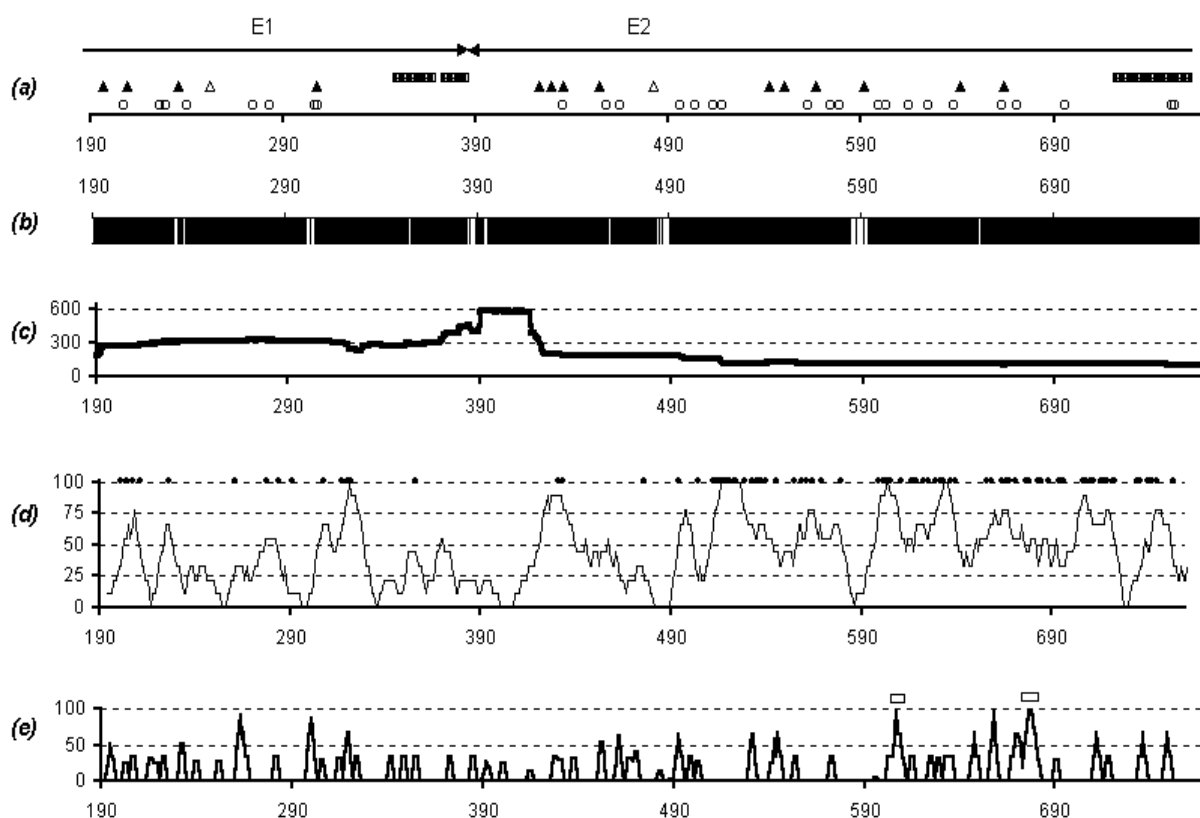


Figure 1. (a) Distribution of the characteristic positions in alignment of E1 and E2 sequences. O – conservative Cys residues. D - conservative glycosylation sites, D - less conservative glycosylation sites, membrane-spanning regions are designated as solid bars. (b) Deletion-containing positions designated by light stretches. (c) The actual height (Y-axis values) of alignment columns. (d) Relative number (in %) of conservative residues (>95% identity) average per 9-residue window. Invariant positions are marked by dots in the upper edge. (e) Relative charge conservativity determined as ratio of the existence of charged residues in a certain alignment position to that of uncharged or oppositely charged residues with averaging per 5-mer window. Two most stressed regions are shown by empty bars.

14 highly conservative glycosylation sites were predicted. However, only one site was presented in all covered sequences. Besides that, two less conservative sites were found. Such difference can result in antigenic specificity in consequence of masking the peptide portion. 9 highly conservative regions were identified in E1 and E2 proteins. Most of them are relatively rich in charged and aromatic residues. They are considered as putative regions, which recognise surface structures of target cells.

The most variable region of envelope area in precursor polyprotein occupies 27 N-terminal residues of E2 protein. Over 500 non-identical HVR1 fragments were analysed. The residue occurrence frequencies in HVR1 positions are presented in fig. 2. We developed the pattern allowing to discriminate HCV sequences containing HVR1 from other sequences at database screening. Based on this result and analysis of highly and moderately conservative positions, we assume that this region adopts strongly definite conformation in a relatively independent fold unit and have a certain functional role; moderately conservative positions of HVR1 could influence on antigenic specificity and participate in receptor recognition.

Developing the profile of the aligned sequence set allowed us to construct antigenic maps of core and envelope proteins accounting the isolate specific character of many antigenic determinants. Applying our map representing system enables to easily evaluate the conservativity of any epitope in roughly quantitative manner. We also mapped some functional data, including the putative fusion peptide and probable receptor-binding sites as well as interferon sensitive determined region (ISDR).

Performing our work we develop the database on functional and antigenic mapping of HCV proteins that includes sets of HCV encoded sequences, alignments, functional maps, and references.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
E ²²	T ⁹⁶	H ³⁵	V ⁴⁴	T ⁶²	G ⁹⁷	G ⁸⁴	S ²⁸	A ⁵⁵	A ⁴⁸	R ⁴²	T ⁴⁰	T ⁵³	S ²³	G ⁵⁵	L ⁵⁰	T ³⁴	S ⁵⁷	L ⁶⁴	F ⁸¹	S ⁴⁸	P ⁴¹	G ⁹⁷	A ⁴⁸	K ²⁹	Q ⁹⁶	N ³⁹
T ²⁰	n	Y ³³	T ⁴³	S ²⁰	A ¹	A ¹⁰	A ¹⁹	Q ¹⁶	G ³⁴	H ¹⁹	A ¹⁷	A ²³	R ¹⁸	S ¹⁵	F ³²	A ³³	G ¹⁸	F ²¹	L ¹⁴	T ²²	S ¹³	e	P ³³	S ²⁶	h	K ³⁶
G ¹⁰	a	R ¹⁴	I ⁵	V ⁶	r	S ²	T ¹⁷	V ¹⁰	S ¹³	K ⁹	N ¹⁰	V ¹¹	Q ¹¹	R ⁸	I ¹⁰	V ²¹	N ¹⁰	I ⁷	S ¹	A ⁷	L ¹²	s	S ¹⁵	Q ¹⁵	r	D ¹⁰
N ⁸	i	T ³	A ¹	I ⁵	c	Q ¹	Q ¹¹	T ⁴	R ¹	Q ⁸	G ⁹	I ⁴	A ¹¹	K ⁵	V ⁴	S ⁶	R ⁴	M ³	n	N ⁶	Q ⁷	a	q	R ¹¹	e	R ⁵
S ⁸	s	I ¹	S ¹	M ¹	e	r	V ¹⁰	M ³	v	Y ⁶	S ⁸	L ⁴	H ⁸	T ⁵	S	l	K ²	p	r	R ³	R ⁶	c	r	A ⁹	a	H ⁴
D ⁶	w	K ¹	l	l	v	v	E ⁴	S ²	L	F ³	D ²	m	Y ⁶	A ³	W	g	T ¹	s	v	Q ²	V ⁵	p	t	N ²	d	q
A ⁵	y	Q ¹	p	a		p	R ²	N ¹	P	N ²	R ²	r	L ⁴	N ²	c	p	A ¹	w	i	V ²	A ³	q	g	H ¹	m	e
R ⁵	h	P ¹	r	c		e	N ¹	P ¹	E	S ²	V ¹	s	T ²	i	h	h	l	v	p	D ¹	H ²	r	n	p	n	p
H ⁴	p	V ¹	c	p		l	h	r	H	M ¹	P ¹	c	F ²	d	q	n	q	g	t	M ¹	T ²	t	h	t	t	t
Q ³	r	f	e	f		i	d			p	H ¹	e	M ²	p	t	r	v	h	y	i	M ¹		l	v	y	
I ¹		n	g	q		k	e		a	l	p	G ²	q	y	c	i	q		p	f		v	m	c		
V ¹		s	m	r		l	k		e	q		N ¹	v		e	c	y		y	i		y		i		
k		g	q	w		m	l		l	i		K ¹	l		k	f			g	y		e		l	s	
m		l	w			c	y		w	c		w	e		q	h			l	e		l		s		
		c	y			g	h		g	e		d			y	p			e	g						v
		m				p	i		c	f		v							f	k						
									d	y		e							h							
									t			p														
									v																	

Figure 2. Occurrence frequency (in %, upper indexes) of amino acid residues in HVR1 positions. Residues with occurrence frequency higher than 5% are shaded. The residues with occurrence lower than 1% are designated by small letters.

Discussion

Though several Internet resources on Hepatitis C problem, including HCV database in Japan [3] are available now, the existing databases do not support the ease way to collect the information on functional mapping of HCV sequences. Starting the work with problem on antigenicity of different HCV isolates we came to necessary of the specialised database on antigenic and functional mapping. We hope that our database and supported software will be useful for further studies directed to developing the drugs and vaccine against HCV.

References

1. Plagemann, P.G.(1991). Hepatitis C virus. Archives of Virology 120, 165-180.
2. Clarke, B. (1997). Molecular biology of hepatitis C virus. Journal of General Virology 78, 2397-2410.
3. <http://s2as02.genes.nig.ac.jp/>

AN APPROACH TO STRUCTURAL ALIGNMENT WITH GENETIC ALGORITHM

*Park S.-J., Yamamura M.

Dept. of Computational Intelligence and Systems Science, Tokyo Institute of Technology,
Yokohama, Japan

e-mail: park@es.dis.titech.ac.jp

*Corresponding author

Keywords: GSA, genetic algorithm, structural alignment

Resume

Motivation:

In the existing methods for the structural alignment, the possible transfer positions and orientations can be ignored, i.e. they are all local searches. In consequence, their results could be failed in the most important sites for proteins. We propose a novel alignment method that is on the basis of the Real-coded Genetic Algorithm: *Genetic Structural Alignment* (GSA). GSA can align much more importance on conserved and active sites with an effective fitness function.

Results:

We confirmed that GSA could realize a global search and the fitness function conserved important sites into the individuals. Furthermore, GSA aligned the simplest globular conformation proteins and complex conformation proteins for the benchmark test. Their results were evaluated by the *accuracy symbols*.

Introduction

The systematic comparison of proteins as a three-dimensional structure is called the structural alignment. The structural alignment is an effective method for searching evolutionary relationships between biological molecules. Many effective algorithms have been developed and utilized which are based on dynamic programming approach (DP-matching) and root square mean deviation (RMSd). It is most likely that they have two problems.

Firstly, they are all local searches, i.e. the possible transfer positions and orientations are ignored. Secondly, the obtained alignments could be failed in biological conserved and active sites, because they emphasize the average of square distances between equivalent C-alpha atoms.

We propose a new approach; let us call GSA. GSA can resolve above-mentioned problems. We compare our method with two existing methods.

Methods and algorithms

The Real-coded GA is adopted as the core of GSA algorithm. GA has several merits in the structural alignment. First, GA can compare with structures in possible spatial positions. Second, the higher precision can be obtained with GA if the complexity of structure was increased or not. Third, GA can design the flexible functions and it has adaptability to any function.

1.Encoding

We can represent individuals with a vector of six parameters real numbers, i.e. three-dimensional translation vector (x, y, z) for the first atom and three rotation angles (alpha, beta, gamma). A chromosome in an initial population could be encoded by:

$$\begin{aligned} -180.0 \leq (\alpha, \beta, \gamma) \leq 180.0 \\ \min(x, y, z) \leq (x, y, z) \leq \max(x, y, z) \end{aligned}$$

where $\min(x, y, z)$ and $\max(x, y, z)$ are the generative space for an initial population.

2.Crossover and Generation Alternation Model

We employed UNDX (unimodal distribution crossover) [Ono and Kobayashi, 1997] as a crossover operator, it is a powerful crossover method in characteristics preservation. Following normal distribution UNDX generates six-dimensional real numbers for two children in a determined area with three parents. In regard to rotation angles, we use complementary angles because the angle of 360.0 degrees presents an equivalent position in geometric spaces. The MGG (minimal generation gap) [Sato, Yamamura and Kobayashi, 1996] is adopted as a generation alternation model.

3. Mutation

Two individuals, who have survived in MGG, will be mutated with a mutation probability. When we assume that $INDI_k$ is mutated, the mutation procedure is described as follows:

1. randomly select an atom in the second structure, i.e. fixed structure
2. the first atom of $INDI_k$ is superimposed on a selected atom in step 1
3. $INDI_k$ is rotated with $\alpha=180.0$, $\beta=0$, $\gamma=0$.
4. Fitness Function

For the estimate of equivalent atoms, we compute distances between each atom in the $INDI_k$ and every atom in the second structure (exception of determined atoms). When we found the nearest atom pair (i, j) , we check their distance $dist(ij)$. If $dist(ij) \leq \delta$ then, that added to a set of equivalent pairs. If $dist(ij) > \delta$ then, it becomes a gap, where δ is a constant for the gap. If $c_i = (a, b)$ is a member of $C = (c_1, c_2, \dots, c_z)$, $dist(c_i)$ indicates $|a - b|$, where C is a set for equivalent atoms. And $g = m - z$ for a gap penalty, where m is the length of $INDI_k$. Hence we define the fitness function f as follows:

$$s = \sum_{i=1}^z \exp(\epsilon \times dist_{c_i}^2)$$
$$f = \frac{s + 1.0}{g + 1.0}$$

where epsilon (< 0) is a constant to adjust the effect of similarity. The delta and epsilon in fitness function f are set to 5.0 and -0.8 , respectively. These values are determined with experiments. The fitness function f is designed that a distance between a pair of atoms is reflected their similarity by exponential function.

The conserved and active sites in similar proteins are should locate in a close geometric position. We would realize a special emphasis on such the important sites. Of course, it is not unique way to do that. We can find the other functions for this approach. GA has adaptability, one of features, to any function. This flexibility in GA is an advantage over the other methods.

Implementation and results

We determined the accuracy symbols as shown Table 1. These symbols can show the important sites, particularly in '*', '|', and '#'. We compare GSA with YSAS (Yale structure alignment server: <http://bioinfo.mbb.yale.edu/Align/>) [Gerstein and Levitt, 1996] and Stralign (structure alignment:

<http://www.hgc.ims.u-tokyo.ac.jp/service/tooldoc/stralign/intro.html>) [Akutsu, 1995]. We will illustrate results of comparative experiments with easy and harder conformations of proteins (Fig. 1).

All experiments ran on SUN SPARC-296Mhz workstation and implemented by C language. The response time for a comparison of 150 average of lengths is approximately 2200 seconds.

Table 1. The Accuracy Symbols for Description of Results.

No.	symbol	$dist(ij)^2$
1	empty	$> \delta$ (or gap)
2	.	> 1.5 and different character
3	:	> 1.5 and same character
4	*	≤ 1.5
5		≤ 1.0
6	#	≤ 0.5

Discussion

As shown in Fig. 1, there are no wide differences between three different methods in the case of easy structures. Nevertheless, it is possible to find a good method when we align the harder structures. In other words, using the RMSd and DP-matching methods are adequate to compare between simplest conformation proteins. However, their accuracies deteriorate into 'not make any importance to functions of proteins', i.e. the number of '.' or ':', when they are compared with complex structures. The symbol '.' is senseless to the biological important sites, because it implies distant positions of equivalent atoms. Thus, it is appreciated that the existing methods fail in catching the nearest pair of atoms. In contrast, GSA makes a success of detecting high structural similarity pairs with independence on the conformations.

Using RMSd and RMS-fitting to rotate a structure and applying DP-matching for estimate equivalent atoms are extremely depending on the protein conformations. Such as the iterative improvement approaches cannot find any novel alignment between complex conformation proteins, because they emphasize the average of square

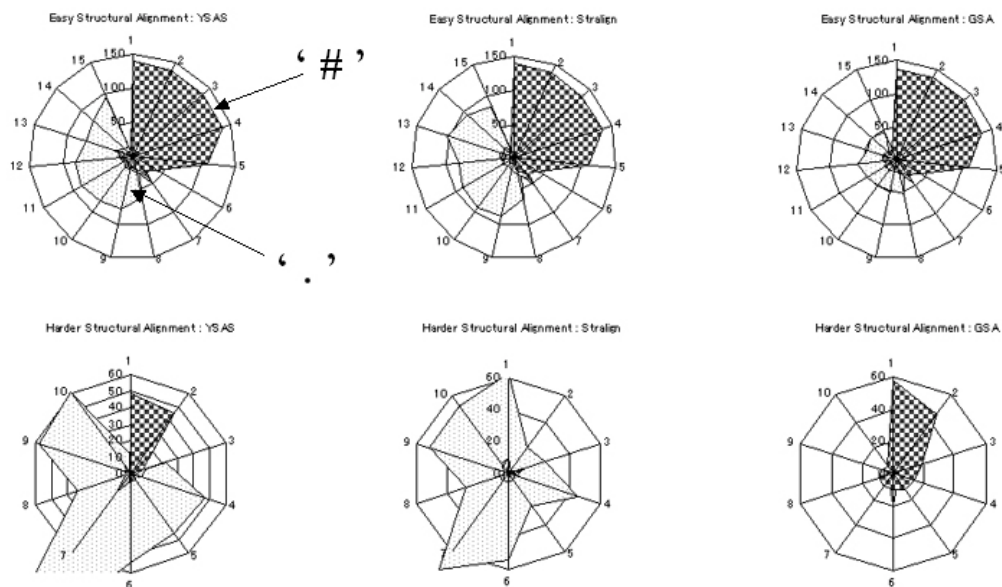


Figure 1. The comparison of accuracy: the globular proteins are easy on their conformations, we aligned 15 pairs of globin proteins. The results of harder structural alignment are obtained from 10 pairs of complex structure proteins. We illustrated the radial graphs with the number of ' ' and '#', each pair is denoted as a radial line.

distances. We confirmed that GSA could realize a global search and the fitness function f conserved important sites into the individuals.

We consider that GSA suitable for MSA (Multiple Structural Alignment). The realization of MGSA (*Multiple Genetic Structural Alignment*) is one of the further issues.

References

1. Ono, I. and Kobayashi, S. (1997) A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, *Proc. 7th Int. Conf. Genetic Algorithms*, 246-253.
2. Satoh, H., Yamamura, M. and Kobayashi, S. (1996) Minimal Generation Gap Model for GAs Considering Both Exploration and Exploitation, *Proc. IIZUKA96*, 494-497.
3. Gerstein, M. and Levitt, M. (1996) Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures, *ISMB96*, 59-67.
4. Akutsu, T. (1995) Protein structure alignment using a graph matching technique, *GIW95*, 1-8.

THEORETICAL MODEL OF INTERACTION: PLATELETE ACTIVATING FACTOR RECEPTOR (PAFR) AND TYROSINE KINASE TYK2

¹*Lukashova V.V., Lukashov V.A., ²Lukashov V.V., ¹Rola-Pleszczynski M., ¹Stankova J.

¹Immunology Division, Department of Pediatrics, Faculty of Medicine, Sherbrooke University, Sherbrooke, Canada

²College of Informatics, Novosibirsk University, Novosibirsk, Russia

e-mail: immunolo@courrier.usherb.ca

*Corresponding author

Keywords: Platelet Activating Factor receptor, tyrosine kinase Tyk2, computer modeling

Resume

Motivation:

Computer modeling to predict the probable sites of interaction: Platelete Activating Factor Receptor and tyrosine kinase Tyk2.

Introduction

The JAK/STAT pathway represents a common signal transduction pathway activated in response to a wide variety of polypeptide ligands. Members of the Jak family tyrosine kinases (Jak1, Jak2, Jak2 and Tyk2) are known to associate with cytokine receptors and become tyrosine phosphorylated following ligand binding. In general, a membrane-proximal region of cytokine receptors appears to be critical for their interaction with Jaks. A proline-rich Box 1 motif is required for interaction of Jak2 with the receptor for erythropoietin, growth hormone and prolactin [1]. In the IFN α R1 component of the IFN alpha receptor, a Tyk2-binding site has been mapped to a membrane-proximal 33 amino acid sequence. The proline-containing Box 1-like domain in IFN α R1 subunit appeared to play only a minor role in Tyk2 binding [2]. Box 2 is a second conserved domain found in the membrane-proximal region of many cytokine receptors such as IL2 receptor bc and G-CSF-receptor b chain [1].

The human PAF receptor was found to directly associate with tyrosine kinase Tyk2 in transiently transfected COS-7 cells [3]. In the present report we employed computer modeling methods to define the probable interaction domains between PAFR and Tyk2.

Methods

Homology between PAFR and Interferon alpha Receptor 1 (IFN α R1) was determined employing computer programs "MOTIV" and "POISK" described previously [4, 6]. Probable interaction domains of PAFR on Tyk2 were evaluated independently, and based on experimental data concerning IFN α R1 interaction with tyrosine kinase Tyk2 [2, 5]. Spatial affinity of different peptide segments in PAFR as well as IFN α R1 and those of Tyk2 was calculated. The search for peptide segments in Tyk2 that could be involved in the formation of an interaction domain with PAFR was also performed.

Results and Discussion

It has been reported that the Tyk2 binding domain is restricted to a 46-amino-acid membrane-proximal region of the IFN α R1 subunit of the IFN alpha receptor. While the proximal half of IFN α R1 is required for maximal binding, amino acids from 486 to 511 seemed to be the most critical for binding. Moreover, by site-directed mutagenesis Yan et al. [2] showed that mainly isoleucine, leucine and the acidic amino acids, aspartic and glutamic acid, were important for binding. This region of the IFN α R1 receptor has a certain homology with the first intracellular ("1-in") loop of PAFR (Fig. 1).

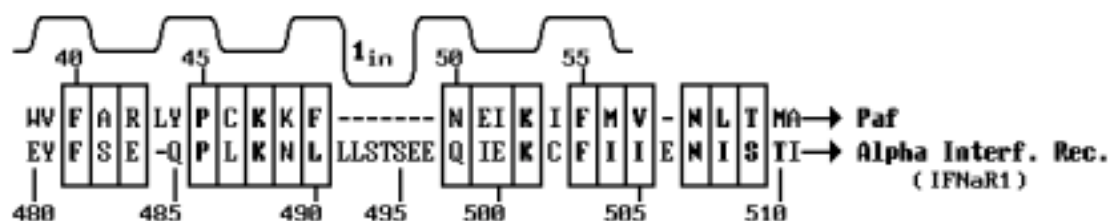


Figure 1. Comparison of the homology regions in the PAFR and the IFN α R1 peptide fragments.

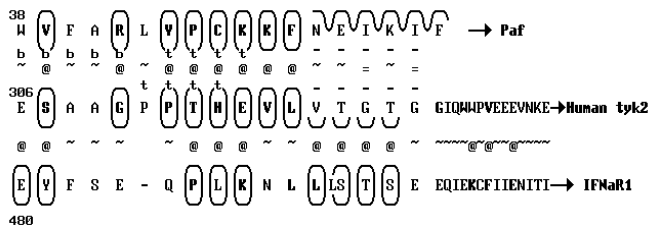


Figure 2. The model of interaction between peptide fragments "1-in" of the PAFR and human Tyk2.

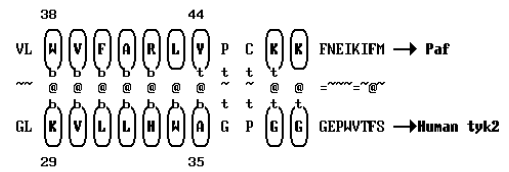


Figure 3. The probable spatial contact of the peptide fragment "1-in" of the PAFR with Tyk2.

We proposed that this peptide segment of PAFR (amino acids 38-60) might interact with Tyk2. Based on calculated spatial affinity, we found two regions in Tyk2 that could interact with the PAFR "1-in" loop. One corresponded to amino acids 306-317 in the JH5 domain of Tyk2. Fig 2 shows the probable interaction of this region with PAFR and IFNar1. Our theoretical prognosis is generally consistent with data of Yan et al [5], who showed that binding of IFNar1 to Tyk2 required a 601-amino acid region of Tyk2 containing the JH3-7 domain. On the other hand GST-Tyk2 fusion protein containing only the JH5 domain bound the IFNar1 cytoplasmic domain poorly and the JH3 and JH6 domains appeared to be the most important in binding to IFNar1 [5]. We discovered another spatial contact between "1-in" loop of PAFR and N-terminal region of Tyk2, amino acid residues from 29 to 35 (Fig.3). Yen demonstrated that the first 53 N-terminal amino acids of Tyk2 were not necessary for IFNar1 binding [5]. However, this may represent a novel binding region with G-protein coupled receptors such as PAFR.

We determined several other domains that may form a high affinity contact with PAFR. Fragments of Tyk2 containing amino acid residues from 196-213 and from 253-270 were found to have a spatial affinity with "1-in" of PAFR. Figure 4 shows the probable interaction of these peptide segments with corresponding regions in PAFR and IFNar1. Our prognosis is in agreement with the fact that GST-Tyk2 fusion protein bearing the JH6 domain (amino acid 128-309) binds IFNar1 effectively [5]. Interestingly, that peptide segment of Tyk2 "196-213" could also interact with the PAFR second intracellular loop ("2-in"). Fig 5 shows the high affinity contact between PAFR (amino acids 124-136) and the region of residues from 200 to 212 in Tyk2. In addition, PAFR domain from 118 to 134 was evaluated to form a 17 amino acid long contact with Tyk2 peptide segment from residue 197 to 213 (Fig. 6). Our estimation of spatial affinity between Tyk2 segment "191-221" and IFNar1 was determined to be consistent with experimental data of Yan [5] and illustrated in Figure 6.

The third intracellular loop ("3-in") also shows a possible 14 amino acid long interaction with the C-terminus of Tyk2 (Fig. 7). This result doesn't correlate with known data concerning the binding of Jak kinases and cytokine receptors. Since this interaction domain is localized in the JH1 kinase domain we can hypothesize that it may be important for correct folding of the protein.

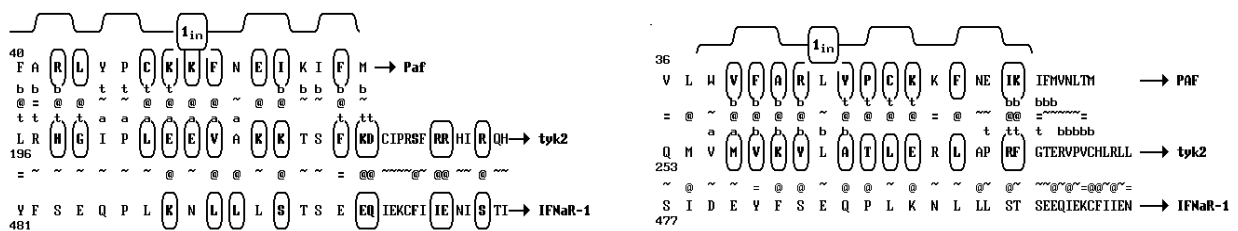


Figure 4. The probable interaction between the first intracellular loop "1-in" of the PAFR and human Tyk2.

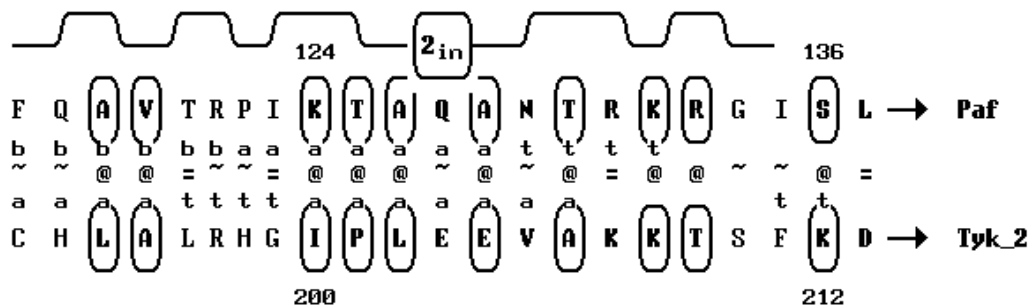


Figure 5. The high affinity contact between the second intracellular loop "2-in" of the PAFR and Tyk2.

We proposed a model of PAFR interaction with tyrosine kinase Tyk2 (Fig.8). We hypothesize that the most powerful contact is localized in the first intracellular "1-in" loop of PAFR. This contact would probably be stabilized by the interaction of the second intracellular loop of PAFR and Tyk2 region (residues 196-212). Mutagenesis of PAFR "1-in" as well as "2-in" loops and/or creation of a chimeric PAFR receptor, where important regions will be substituted, is needed to confirm our theoretical prognosis. The Tyk2 deletion mutants will also be a powerful tool to determine the structural requirements for Tyk2 binding to PAFR.

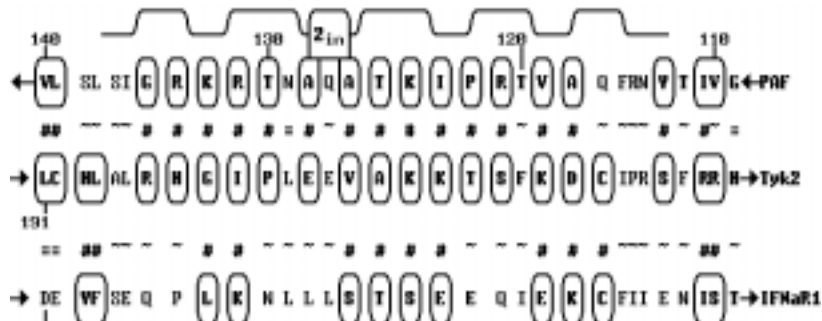


Figure 6. Schematic representation of the putative contact formed by the 2nd intracellular loop of the PAFR and Tyk2.

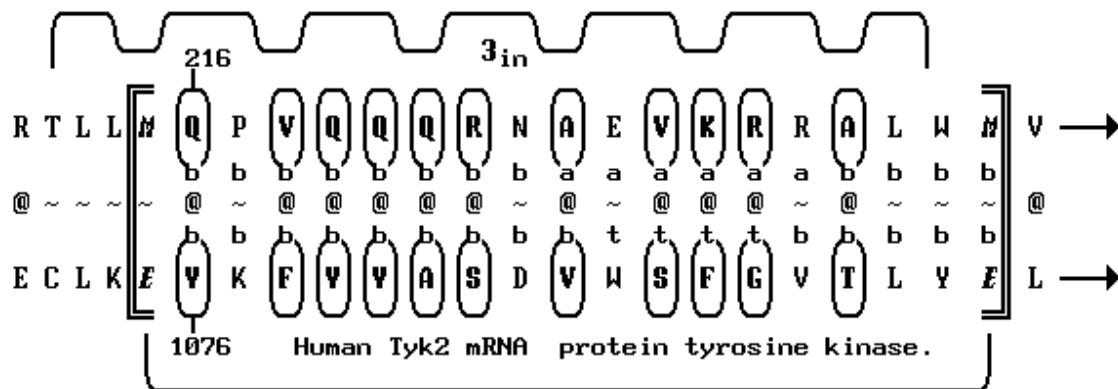


Figure 7. The probable binding site of the third intracellular loop "3-in" of the PAFR on human Tyk2.

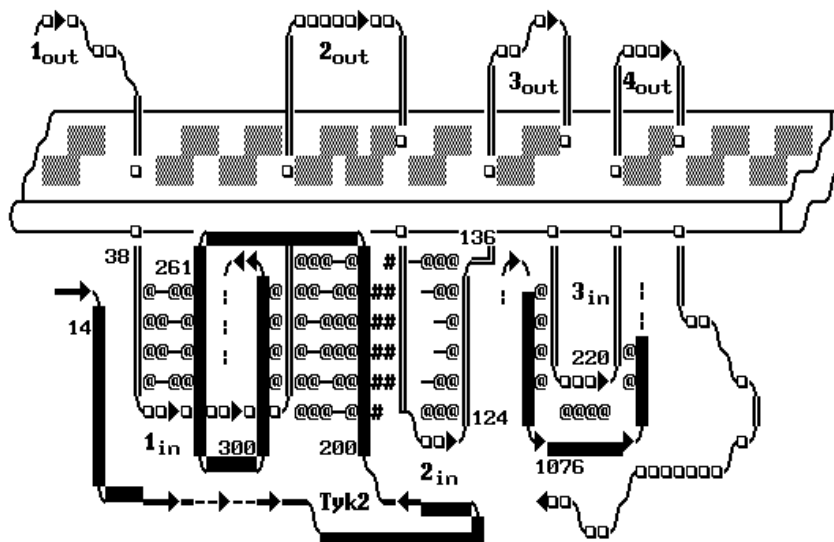


Figure 8. The model of interaction: the PAFR (Platelet Activating Factor Receptor) and tyrosine kinase Tyk2.

References

1. S. Pelegri, I. Dusanter-Fourt (1997) *Eourop. J. Biochem.* 248, 615-633
2. H. Yan, K. Krishnan et al. (1996) *Mol. Cell. Biol.* 16/5, 2074-2082
3. V. Lukashova, L. Bouchard, K. Asselin, M. Rola-Pleszczynski, J. Stankova (2000) in press
4. V.A. Lukashev, A.G. Bachinski, V.A. Kulichkov (1986) *Molecular Biology (Russian)* 20/5, 1192-1202
5. H. Yan, F. Piazza, K. Krishnan, J.J. Krolewski (1998) *J. Biol. Chem.* 273/7, 4046-4051
6. V.A. Lukashev, N.V. Bausk, L.N. Mazalov et al (1995) *Instr. and Meth. In Phys. Res.* A359, 259-262

STABILITY OF PARTIAL CORRELATION COEFFICIENT ESTIMATES FOR RESIDUE CHARACTERISTICS AT DIFFERENT POSITIONS OF AMINO ACID SEQUENCES

D.A. Afonnikov

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: ada@bionet.nsc.ru

*Corresponding author

Keywords: amino acid sequences, co-adaptive substitutions, partial correlation, numerical simulation

Introduction

Analysis of co-adaptive substitutions of amino acid residues is one of directions of investigating families of protein sequences. This analysis is based on the assumption that substitutions of amino acid residues displaying certain functionally important interactions are fixed evolutionary in a dependent manner [3,2]. Consequently, substitutions of these residues within a sample of protein family sequences are statistically correlated. Note, however, that a statistical correlation between amino acid substitutions does not necessarily indicate their interaction. One of the reasons is existence of mediated (long range) correlations [3], that is, the correlations of substitutions at residue positions i,j that are not interacting directly. These correlations stem from interactions of the residues at positions i,j with another or several other protein residues. This effect should be taken into account while analyzing and interpreting the relationships found. One of approaches to solving this problem is based on "maximum entropy" principle [3,4]. We have proposed the approach basing on partial correlation coefficients of physico-chemical characteristics at positions of protein sequences [5]. Partial correlation coefficients allow the interdependence of amino acid substitutions at a pair of positions within a protein to be estimated, provided that amino acids at the rest protein positions are fixed. Thus, they reflect the direct correlation of amino acid substitutions in a pair of positions. This work focuses on two questions—the behaviors of linear and partial correlation coefficients depending on (a) the size of sample analyzed and (b) the fraction of positions considered. The results obtained allowed us to assess the reliability of partial correlation coefficient estimates in case (a) a limited number of sequences is used and (b) certain fraction of positions with a protein are not involved in analysis (for example, due to their conservancy or presence of deletions). In our study, we use the data obtained through numerical simulation. The analysis performed has demonstrated that partial correlation coefficients allow the effect of remote correlations to be excluded while analyzing co-adaptive substitutions of amino acid residues.

Partial correlation coefficients

Let a physico-chemical characteristic f be chosen for a set of N protein sequences with the length L and the estimate of covariance matrix of this characteristic in the protein sample analyzed $\mathbf{S}=(s_{ij})$ be obtained (methods for assessing \mathbf{S} estimate are described, in particular in [5,6]). The linear correlation coefficient is the measure of linear dependence of f characteristics at the pair of positions i,j [7]:

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} \cdot s_{jj}}}, \quad (1)$$

where s_{ij} are the elements of covariance matrix for positions i,j . Partial correlation coefficients $r_{ij \cdot g}$ estimate the extent of dependence of f characteristics at a pair of positions i,j within a protein, provided that this characteristic at the rest g positions with the protein is fixed, and are calculated by the following expression [7]:

$$r_{ij \cdot g} = \frac{-M_{ij}}{\sqrt{M_{ii}M_{jj}}} \quad (2)$$

where M_{ij} is the minor of element ij of the covariance matrix \mathbf{S} . Equation (2) may be represented in another form, as $a_{ij} = M_{ij}/\det(\mathbf{S})$ is the element of matrix \mathbf{A} , which is the inverse covariance matrix (we assume \mathbf{S} not degenerate). Thus, equation (2) may be represented as

$$r_{ij \cdot g} = \frac{-a_{ij}}{\sqrt{a_{ii}a_{jj}}}. \quad (3)$$

To explicate the meaning of partial correlation coefficient, let us consider a protein of L amino acid residues in length. Let the total interaction energy of the protein residues E depend on characteristic f at its positions as

$$E = \frac{1}{2} \sum_{i,j} B_{ij} (f_i - \hat{f}_i)(f_j - \hat{f}_j) \quad (4)$$

where f_i is the characteristic f value at the protein position i ; \hat{f}_i , its optimal value at this position. Interactions of the residues are described by matrix \mathbf{B} ; if $b_{ij} = 0$, the residues at positions ij do not interact. Let us also assume that the probability to detect a mutant protein sequence with energy E has a Boltzmannian form:

$$p(E) = \frac{1}{Z} \exp(-\beta E), \quad (5)$$

where Z is a normalization constant; β , certain constant determining the selection intensity. This means that the smaller is the value of energy E of mutant sequences, the higher is their frequency. Consequently, a mutation decreasing the protein energy has a higher probability of being fixed. Substituting (4) into (5), we obtain the following equation:

$$p(E) = \frac{1}{Z} \exp\left(-\beta \frac{1}{2} \sum_{i,j} B_{ij} (f_i - \hat{f}_i)(f_j - \hat{f}_j)\right), \quad (6)$$

which has, accurate to the normalization constant, a form of L -meric normal distribution of the physico-chemical characteristic f value at the positions within the protein with the covariance matrix

$$\mathbf{C} = \mathbf{B}^{-1}/\beta. \quad (7)$$

Thus, the matrix \mathbf{A} , inverse to the covariation matrix of estimate \mathbf{S} , may serve in this model as an estimate of interaction matrix \mathbf{B} (accurate to the constant β). Consequently, it is reasonable to expect that partial correlation coefficients (3) for non-interacting position pairs will be close to 0, which is not necessarily met in case of linear correlation coefficients (1) due to (7).

Generation of amino acid sequences with interacting residues

To compare the efficiencies of linear and partial correlation coefficient estimates, we analyzed numerically simulated samples of N amino acid sequences with the length L . Serial number of an amino acid in the sequence was considered as a characteristic f . The interaction parameters were generated as follows:

$\hat{f}_i = 9.0$; $b_{ii} = 9.0$; each 15th non-diagonal matrix element equaled to +7 or -7 alternately; the rest non-diagonal matrix elements equaled to 0; and $\beta = 1$.

Sequences were generated independently according to the algorithm proposed by Metropolis *et al* [8]. First, a random sequence with equal residue frequencies was generated. Then, an amino acid at an arbitrary position of the sequence was randomly substituted. If the substitution decreased the protein energy, it was fixed. Otherwise, the mutation-caused change in the protein energy ΔE was calculated, and the mutation was fixed with a probability of $p \sim \exp(-\beta \Delta E)$. The number of steps corresponding to single substitutions at initial sequence was 10 000.

To perform the analysis, we generated the sample of $N_0 = 620$ sequences with the length of $L_0 = 40$ amino acid residues.

Study of stability of linear and partial correlation coefficient estimates

We have performed two types of analysis. First, we studied the dependence of linear and partial correlation coefficient estimates on the protein sample size. For this purpose, we randomly generated a sample of smaller size N' from the initial sample. Linear and partial correlation coefficients and their critical values r_c were estimated for each smaller sample to test the hypothesis on independence of positions (see Afonnikov *et al.*, 1997, and Kendall and Stuart, 1967, for r_c calculation). The values r_c were selected at 95% and 99% significance levels. Then, two parameters were calculated for each correlation type and significance level:

- n_{fneg} , the fraction of pairs displaying $b_{ij} \neq 0$ and $|r_{ij}| < r_c$ or significant correlation coefficients with the same sign as b_{ij} ("false negative" pairs) and
- n_{fpos} , the fraction of pairs with $b_{ij} = 0$ and $|r_{ij}| > r_c$ ("false positive" pairs).

Hereinafter, $n_{\text{fneg}}(l, 95)$ is the value n_{fneg} for linear correlation coefficient and r_c at 95% significance level; $n_{\text{fneg}}(p, 95)$, for linear correlation coefficient, etc.

The values n_{fneg} and n_{fpos} were averaged for each N' over 10 independent tests. The dependences of these parameters on the ratio N'/L were studied.

Second type numerical experiments were performed in a similar way, except for the fixed sample size of $N = 320$ and variable number of analyzed positions L' selected randomly from the initial alignment. The dependences of n_{fneg} and n_{fpos} on the ratio L/L_0 were studied.

Results and discussion

The results obtained analyzing the dependences of correlation estimates on the sample size are shown in Figs 1 and 2.

As is evident from Fig. 1, the values of parameter n_{fneg} for partial correlation coefficients are smaller than those for linear correlation coefficient and close to 0 in case of a large sample size ($N/L > 8$). In this case, partial correlation coefficient estimates allow virtually all the interacting position pairs to be detected with the less error probability compared to the linear correlation coefficients. However, the probability of partial correlation coefficients to miss residue interactions increases drastically with decreasing sample size and tends to 0 as $N/L \rightarrow 1$. It is likely that the accuracy of covariance matrix estimate decreases with the sample size (evident while analyzing linear correlation coefficients); however, these errors grow nonlinearly in case of inverting the matrix \mathbf{S}).

Analysis of n_{fpos} dependences has demonstrated that the error probability for partial correlation coefficients corresponds the level specified (95% and 99%) virtually across the entire range of N/L alterations. However, the value of this parameter for linear correlation coefficients is severalfold higher due to the long range correlations of amino acid substitutions. Thus, the data obtained suggest that partial correlation coefficients allow the overprediction error rate to be decreased essentially. In addition, partial correlation coefficients provide for a higher prediction accuracy in case of large samples.

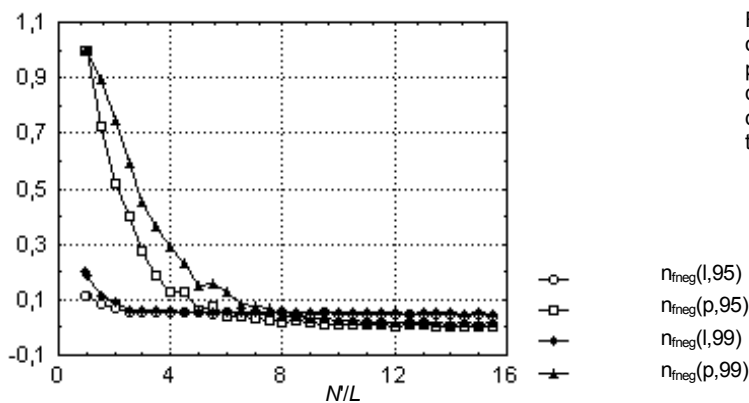


Figure 1. Dependences of n_{fneg} on the N/L ratio for linear and partial correlation coefficient at different r_c values. Curve designations are indicated to the right.

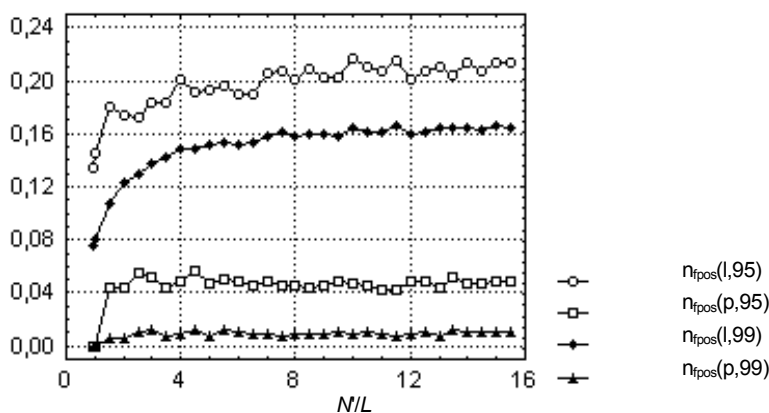


Figure 2. Dependences of n_{fpos} on the N/L ratio for linear and partial correlation coefficient at different r_c values. Curve designations are indicated to the right.

Analysis of n_{fneg} and n_{fpos} dependences on the number of positions analyzed (Fig. 3) has demonstrated that these errors display approximately similar rates up to $L'/L_0 \approx 0.7$. This allows us to hope that the prediction accuracy of residue interaction at the other positions within the protein will remain at least at the same level even if the information on substitutions at 30% positions is absent (for example, due to their conservancy of presence of deletions)

Summing up, the results obtained suggest that partial correlation coefficients allow the interaction between residues to be estimated more precisely, especially in case of large protein samples.

Acknowledgments

The work was supported by grants Nos. 98-07-91078 and 99-04-49879 of the Russian Foundation for Basic Research. The author is grateful to Galina Chirikova for assistance in translation.

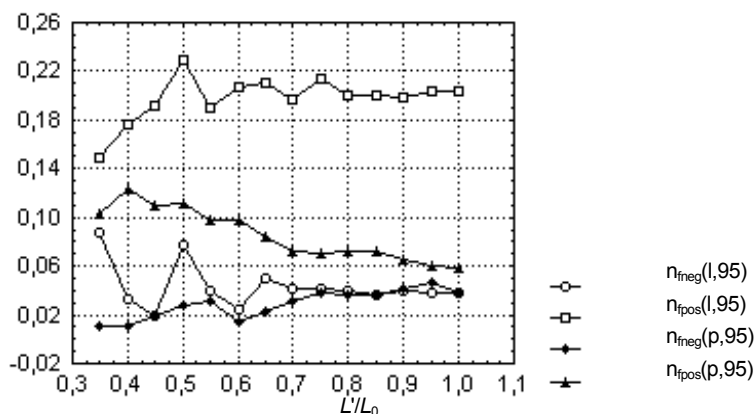


Figure 3. Dependences of n_{fpos} and n_{fneg} on L'/L_0 ratio for linear and partial correlation coefficients at r_c values corresponding to 95% significance level. Curve designations are indicated to the right.

References

1. Lim V.I. and Ptitsyn O.B. (1970) On the constancy of the hydrophobic nucleus volume in molecules of the myoglobins and hemoglobins, *Mol. Biol. (USSR)*, **4**, 372.
2. Göbel U., Sander C., Schneider R., and Valencia A. (1994) Correlated mutations and residue contacts in proteins. *Prot. Struct. Funct. Genet.*, **18**, 309–317.
3. Lapedes A.S., Giraud B.G., Liu L.C., and Stormo G.D. (1997) Correlated mutations in protein sequences: phylogenetic and structural effects. *Proc. AMS/SIAM Conference on Statistics in Molecular Biology*, Seattle, USA.
4. Giraud B.G., Lapedes A., and Liu L.C. (1998) Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E*, **58**, 6312–6322.
5. Afonnikov D.A., Kondrakhin Yu.V., and Titov I.I. (1997) Revealing of correlated positions of the DNA-binding region of the CREB and AP-1 transcription factor families. *Mol. Biol. (Mosk.)*, **31**, 741–748.
6. Afonnikov D.A., Oshchepkov D.Yu., and Kolchanov N.A. (2000) Estimation of variances and covariances of protein physico-chemical characteristics in families of homologous sequences. *Numerical Technologies* (in press).
7. Kendall M.G. and Stuart A. (1967) *The Advanced Theory of Statistics. Vol. 2. Inference and Relationship*. 2nd edition. London: Charles Griffin & Co Ltd.
8. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., and Teller E. (1953) Equation of state calculations for fast computing machines. *J. Chem. Phys.*, **21**, 219–236.

CONTEXT DEPENDENCIES IN AMINO ACID SEQUENCES OF PROTEIN DOMAINS

*Orlov Yu. L., Ivanisenko V.A., ¹Potapov V.N.

Institute of Cytology and Genetics of SB RAS, Novosibirsk, Russia

¹Sobolev Institute of Mathematics of SB RAS, Novosibirsk, Russia

e-mail: orlov@bionet.nsc.ru

*Corresponding author

Keywords: classification of protein topology, secondary protein structure, stochastic complexity, Markov models

Introduction

To determine spatial structure of globular proteins is of great significance in molecular biology. For comparison with the known sequences, the methods designed for rapid search of short sequences of coincident oligopeptides may be applied, including BLAST and PSI-BLAST packages [Altschul et al., 1997]. Statistically significant coincidence of short oligopeptide sequences enables to extract related sequences from the data banks. If similarity is not well pronounced, the method comparing the profiles of amino acid physico-chemical properties is suitable [Gribskov M. et al., 1987]. The more complete usage of information about related proteins gives better results of searching for remote homologs than pairwise comparison of sequences [Park J. et al., 1998]. The methods accounting dependencies between amino acids in different positions are being developed for analysis of protein families. Position-specific score matrices are coded as hidden Markov models (HMMs) [Durbin R. et al., 1998; Bateman A. et al., 1999]. In the last releases of PROSITE [Hofmann K. et al., 1999], hidden Markov models are used for description of protein functional sites.

If clear homology with the proteins with the known structures is absent then the methods detecting the structural class of these proteins become valuable. There exist a series of methods such as neurone networks [Valuev V., 2000], which make possible to detect the type of protein packaging by amino acid frequencies or dipeptides. In this work, the method is suggested for application of the Markov model for detection of the protein structural class.

Methods and algorithms

The algorithm considered is aimed for detection of statistical model according to amino acid sequence. It is based on the method of estimating stochastic complexity that was initially developed in data compression theory. It makes possible to construct a model by the sequence (generating source-tree with the contexts of interest) and to calculate data complexity in this model. Automated detection of statistically significant contexts is based on evaluation of complexity of data in a model and complexity due to adding novel parameters into the model [Orlov Yu. & Potapov V., 2000]. Thus, the problem of excessive parameters is being solved and the data are described more precisely.

A model of a sequence is considered such that probability of the next symbol occurrence in communication depends on preceding context. The sequence of residues is considered as the textual communication, from the left to the right, or from N-termini to the C-termini. The preceding context determines the state of Markov chain and distribution of probabilities for generating the next-in-turn symbol (i.e., the set of numbers determining the probabilities of occurrence of a symbol after the context given). The length of a context is not fixed. Thus, the sequences were analysed, of the form $X_n \dots X_2 X_1 Y$, where Y is a symbol considered, X_i are preceding symbols, $1 < i < n$, $n=7$ (maximal length of a context n does not exceed 7 in the present study).

The distinctive feature of our model of generating the text is the fact that the set of probabilities of generating the symbol is determined exactly by preceding from the left word (oligopeptide). Besides, the set of these words is a prefix set, that is, neither word is a prefix (or beginning) of another word. Hence, the state of Markov chain is determined uniquely. For example, note that for the two-lettered alphabet, each finite prefix set corresponds to the set of suspended vertexes of some binary tree. We simply add the symbol from the left to the context under analysis (a word from the set) or terminate the procedure.

The set of contexts determining the state of Markov chain is convenient to represent in a graphical form as a binary tree. For the sequences in a three-lettered alphabet, the tree will have 3 branches at each level corresponding to a preceding symbol (see Fig. 1-3). It is possible to "read" contexts in a tree according the routes connecting the leaves (suspended vertexes) with the root. For DNA sequences, the tree had 4 branches [Orlov Yu. & Potapov V., 2000].

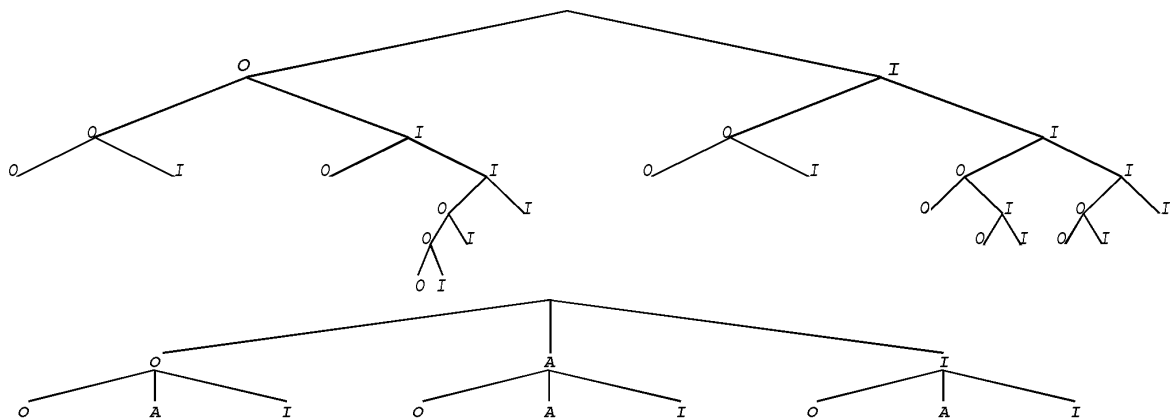
Initially, amino acid sequences are written in 20-lettered alphabet. However, to evaluate the frequency of oligopeptide occurrence in this alphabet, a large size of a sample is necessary. Amino acid residues were

separated into groups in accordance with their preferment to be located on the surface or inside the protein globula. The coding is given in denotation to the Figure 1.

Prediction of a protein class may be based on a Markov model constructed by the source-tree obtained. Estimated complexity of a sequence in this model is the value related to probability to obtain this sequence by an accident.

As the source of data, the database SCOP was used [Murzin A. et al., 1995]. The database SCOP gives detailed classification of proteins with the known structures described in Brookhaven National Laboratory's Protein Data Bank (PDB). We have used the sequences that have the level of pairwise homology at most 40%.

a)



b)

Figure 1. Context source-trees for the α -helical protein sequences (class 1): a) for two-lettered alphabet O (Outer) = {R, N, D, C, Q, E, G, H, K, S, T, Y}, I (Inner) = {A, I, L, M, F, P, W, V}; b) for three-lettered alphabet O (Outer) = {R, N, D, Q, E, H, K}, A - (Ambivalent) = {A, C, G, P, S, T, W, Y}, I (Inner) = {I, L, M, F, V}. The sample contained 262 sequences of different length.

Implementation and results

The results of analysis of the sample compiled of α -helical amino acid sequences in two- and three-lettered alphabets are given in Fig. 1 a, b.

Thus, the model of generating the letters in the helical protein sequences, which are written in two-lettered alphabet discriminating surface (outer, o) and inner (i) residues, is statistically described in our model by the following set of contexts: *ooo, ioo, oio, ooi, ioi; oiiio, ooii, iiii; ioiio, oioii, iioii, ooiii, ioiii; oooiio, iooiio*. In total, 15 contexts were analysed.

As can be seen from the figure, from the point of view of information theory, α -helical proteins are characterized by continuous enough repeated blocks of hydrophobic and hydrophilous residues. In α -helices, hydrogen bounds are formed between the residues in positions (i, i+4). Thus, the length of a context, on which the following symbol depends, should be at least 4, this rule is supported by the contexts given in Fig. 1a and in Table 1. It is well known that for producing a hydrophobic core of a globula, the hydrophobic residues in α -helice should alternate with the period of three-four residues. Such alternating patterns of residues could be detected separately. The method simultaneously detects all such patterns that influence the frequency of residues located to the right of them. Note that in three-lettered alphabet, the α -helical sequences are characterised only by dependencies of the second order. That is, in this case, the more is the alphabet, the less information can be obtained about the class of a protein.

Context trees for β -structure protein sequences are given in Fig. 2.

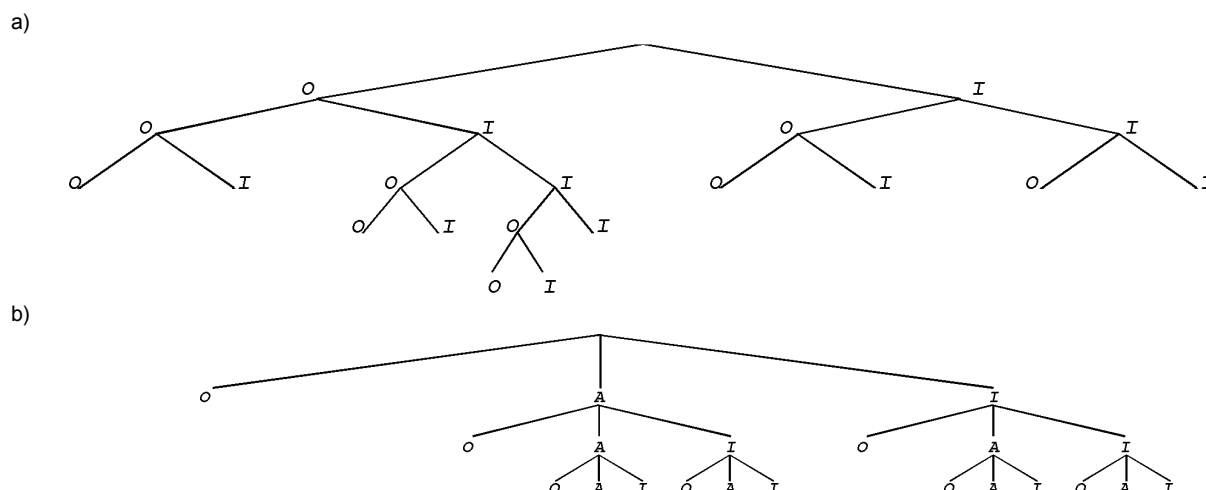


Figure 2. Context trees for β -structure proteins (class 2): a) for two-lettered alphabet; b) for three-lettered alphabet. The sample contained 316 sequences. (Denotations are as in the text and in Fig. 1).

In this case, for β -structure proteins, the source-tree of context dependencies in two-lettered alphabet (Fig. 2a) is much less than for α -helical proteins (Fig.1a). Nevertheless, the dependencies of symbols from preceding context of at least third order are noted.

The tree of context dependencies for the β -structure protein sequences in the three-lettered alphabet (Fig. 2b) is bigger than for α -helical ones (Fig.1a). Dependencies for β -structure proteins in three-lettered alphabet have the third order, whereas for the α -helical proteins – only the second order. Notably, for β -structure proteins there is no dependencies from the contexts containing two and more surface (hydrophilous) residues.

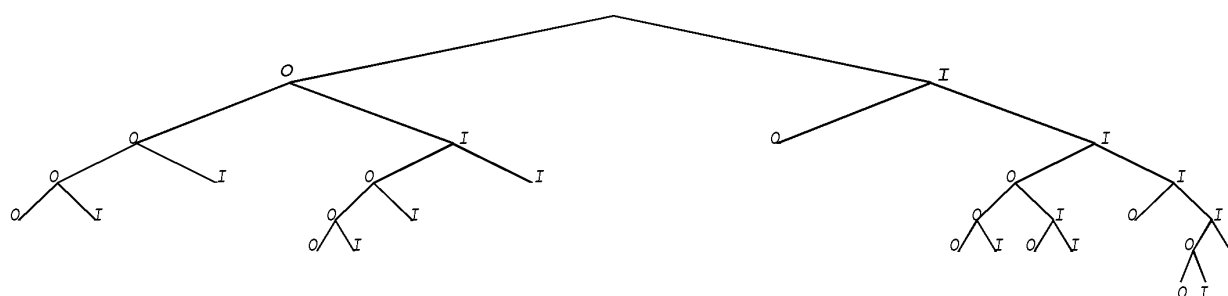


Figure 3. Context tree for the sequences of the class 3, alpha-helical and beta-structure proteins (a/b) in two-lettered alphabet. The sample contained 338 sequences.

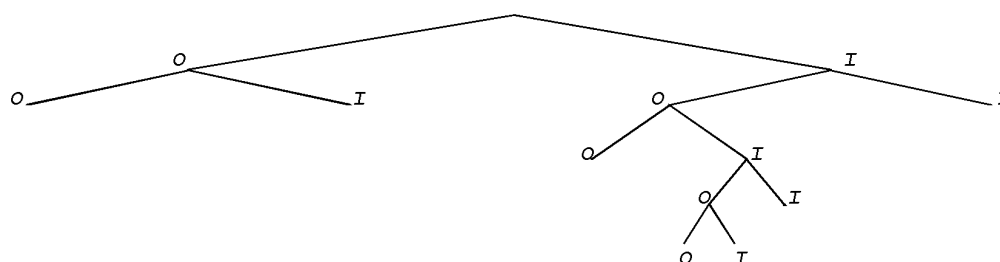


Figure 4. Context tree for the sequences of the class 4, alpha-helical and beta-structure proteins (a+b) (segregated alpha and beta regions) in two-lettered alphabet. The sample contained 269 sequences.

In Figures 3 and 4, the context-trees for the protein classes 3 (a/b) and 4 (a+b) corresponding to protein classes with mixed and segregated alpha- and beta-regions. It is seen that context tree for the proteins with mixed alpha- and beta-structures (Fig.3) contains the same contexts as the trees only for simply α -helical proteins and only for β -structure proteins (Fig.1a, 2a). Moreover, the tree for proteins with segregated alpha and beta regions (Fig. 4) has no clear coincidences with the trees only for classes of α -helical or only β -structure proteins.

Such result evidences about complexity of the protein classification. Protein groups may be considered in more details. Due to the short length, the separate amino acid sequence even in simplified two-lettered alphabet is insufficient for statistical analysis.

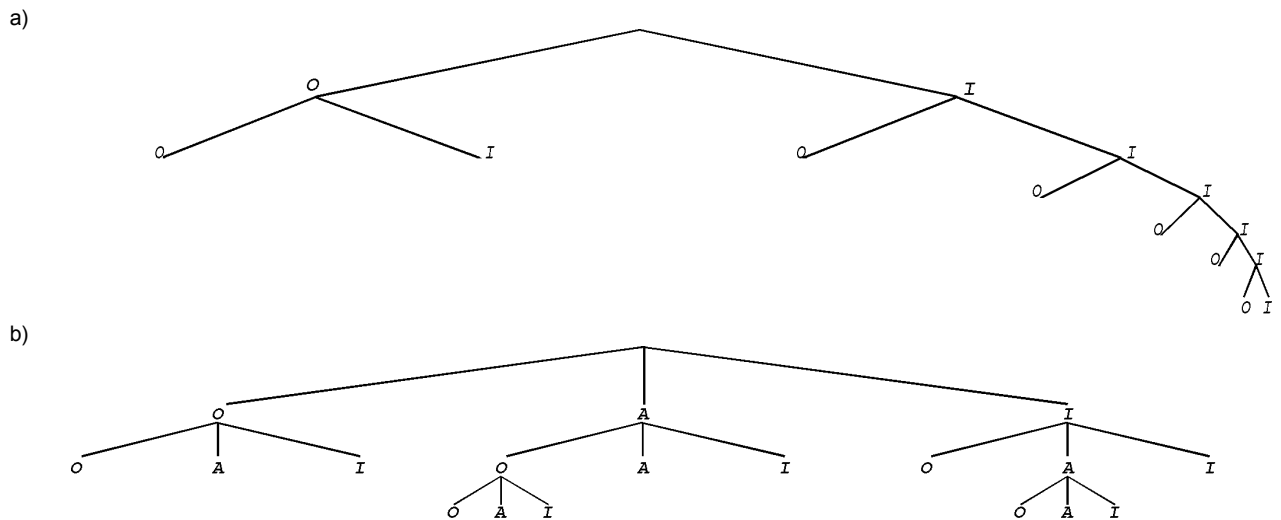


Figure 5. Context tree for sequences of the class 1 referring to the globin family: a) in two-lettered alphabet; b) in three-lettered alphabet. The sample contained 14 sequences.

Let us consider the protein class of more narrow structure, α -helical proteins referring to the globin family. Context trees for amino acid sequences of this class are given in Fig. 5.

This group belongs to the class 1. It is seen that context trees both in two-lettered alphabet and in three-lettered alphabet (Fig. 5a, 5b) are looking like context trees of α -helical protein class (Fig.1a,1b). Moreover, more long contexts occurred to be more significant. Thus, appearance of significant contexts depends upon the homogeneity level in a sample, this homogeneity being produced by structural similarity of proteins. Interestingly (see Fig.5) that for the significant context, there appears a seria of six consecutive hydrophobic amino acids (IIIIII, in denotations given above). Such context may be interpreted as repeatability of hydrophobic clusters on the surface of α -helical proteins, these clusters being important for protein hydrophobic core formation.

The analogous trees were constructed for all groups of proteins classified in SCOP. Each structure group of proteins is characterized by its own structure of context trees.

Discussion

It was demonstrated that by using generalized amino acid residues alphabets, the sequences coding globular protein domains are characterized by unique patterns of context dependencies generating the symbols. These dependencies may be interpreted in terms of correlation of residues in the elements of secondary structure, which are prevailing in the type of domains considered.

Protein sequences are surely non-homogeneous, because the elements of secondary structure, α -helices and β -threads, are supplied by different statistical properties. However, it is hard to predict with 100% accuracy the protein structure, basing only on statistical properties of the sequence regions. It is necessary to consider the protein in a whole integrity, because remote residues may provide mutual influence on the secondary structure. Hence, an important task is to reveal significant features in protein sequences of particular class. Probably, as a unique characteristic of a protein structure may serve not a pre-determined ability of a certain sequence region to form definite secondary structure, but particularly ordered alternation of hydrophilous and hydrophobic amino acid residues, definitely charged residues, or repeats of amino acids. Detection of such parameters becomes an alternative approach to prediction of protein structure. This approach uses neither prediction of secondary structure nor search for homology in databases. The advantage of the method is an automated detection of statistically significant parameters in a model.

Acknowledgements

The authors are grateful to G. Orlova for help in translation of the manuscript into English, to N.A. Kolchanov, D.A. Afonnikov, V.P. Valuev for valuable comments and scientific discussion. The work was supported by Russian Foundation for Basic Research and Integration Project of SB RAS No 66.

References

1. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389-3402.
2. Bateman A., Birney E., Durbin R., Eddy S.R., Finn R.D., Sonnhammer E.L.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, 27, no.1, 260-262.
3. Durbin R., Eddy S.R., Krogh A. and Mitchison G. (1998) *Biological sequence analysis: probabilistic models of protein and nucleic acids*. Cambridge University Press, 1-347.
4. Gribskov M., McLachlan A.D. and Eisenberg D. (1987) Profile analysis: detection of distantly related proteins. *Proc.Natl Acad. Sci. USA*, 84, 4355-4358.
5. Hofmann K., Bucher P., Falquet L. and Bairoch A. (1999) The PROSITE database, its status in 1999, *Nucleic Acids Res.*, 27, no.1, 215-219.
6. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
7. Orlov Yu.L. and Potapov V.N. (2000) Estimation of stochastic complexity of genetical texts. *Computational technologies (Novosibirsk)*, 5, spec.issue, 5-15.
8. Park J., Karplus K., Barrett C., Hughey R., Haussler D., Hubbard T. and Chothia C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise method. *J. Mol. Biol.*, 198, 567-577.
9. Valuev V.P. (2000) 3D protein structural class recognition, *Proceedings of BGRS'2000 (this issue)*, Novosibirsk.

HIERARCHICAL FEATURE DECOMPOSITION IN FUNCTIONAL DOMAINS

*Murray D., Honig B.H., *Califano A.¹*

Columbia Presbyterian Medical Center, Columbia University, New York, NY, USA

¹IBM Computational Biology Center, T.J.Watson Research Center, Yorktown Heights, NY, USA

e-mail: dmurray@columbia.edu, bh6@columbia.edu, acal@us.ibm.com

*Corresponding author

Keywords: protein models, C2 domains, membrane targeting, SPLASH, structural patterns, motif discovery

Resume

Subcellular targeting of peripheral membrane proteins is crucial to many processes such as signal transduction, vesicle trafficking, and viral assembly. Localization of proteins in cell is often accomplished through protein domains that target specific membranes.

If key functional residues in these domains are mutated, function is either lost or significantly modified. As a consequence, it is reasonable to expect that residue clusters, which are preserved by evolutionary processes, would tend to correlate well with the biophysical properties of a protein. These clusters or motifs can be exhaustively identified using pattern discovery algorithm.

This paper analyzes families of homologous membrane targeting sequences from the C2 domain family, which are significantly diverse from a functional perspective. Conserved motifs are exhaustively and hierarchically identified using the SPLASH pattern discovery algorithm. Statistically significant motifs are then correlated with the underlying biological function to identify the protein regions that are responsible for specific functions. It is shown that SPLASH generated motifs correlate very well with functional or structural characteristics of the subset of sequences that support them. By studying the physicochemical properties of the regions corresponding to key motifs, based on structural protein models, it is possible to gain insights into the mechanisms involved in membrane binding activity.

Introduction

We have found that the biophysical factors that coordinate membrane localization of many membrane targeting domains, e. g. electrostatics and hydrophobicity, are manifested as sequence and structural patterns. Several of these amino acid residue signatures, or motifs, can be detected by the deterministic pattern discovery program SPLASH [4]. These are shown to correlate well with the relevant biophysical properties.

By focusing on C2 domains, we illustrate how detailed examination of a protein family can provide insight into the molecular basis of membrane targeting. About 200 C2 domain sequences are known and recent structural studies reveal the family has a common Ig-like β -sandwich fold [10, 13, 14]. Although the family members share the same fold and significant sequence similarity, they perform a variety of functions. Most proteins with C2 domains function in interfacial signal transduction pathways and synaptic vesicle exocytosis. Many C2 domains exhibit calcium-dependent membrane association, while others bind membranes constitutively, mediate protein-protein interactions, or have no currently defined function [10, 13, 14]. Among the C2 domains that target membranes in response to calcium, different C2 domains target different membranes. For example, upon binding calcium, the C2 domains from protein kinase C associates with membranes containing acidic lipids, while the C2 domain from cytosolic phospholipase A2 penetrates into the hydrocarbon core of electrically neutral, zwitterionic membranes [5]. Experimental studies suggest that electrostatic interactions are involved, and detailed calculations with atomic models of proteins and membranes [9, 12] have established a strong correlation between electrostatic potential on the surfaces of C2 domain structures and the type of membrane targeted. By combining the computational analyses of C2 sequences, structures, and biophysical properties, we have discovered signature motifs that contribute to electrostatic-mediated membrane association and can differentiate C2 domain into functional sub-classes.

The unique combination of computational approaches, for the detection of biologically significant sparse amino acid residue signatures, and quantitative analysis of biophysical properties allows us to understand the membrane targeting function of proteins of known structure at the molecular level. These results can be generalized to make predictions for protein families of unknown structure or function. The synthesis of these computational approaches is being used to analyze and exploit genomic data by examining and comparing a number of protein families involved in membrane association.

Method

In [6] it has been shown that deterministic pattern discovery using SPLASH algorithm [4], combined with a statistical framework to assess the significance of discovered patterns [16], can successfully identify biologically relevant regions in protein families. In particular, in 70% of the almost 1000 families considered in the study, the single most statistically significant pattern overlaps with key functional motifs identified by biological assays and reported in the PROSITE database [8]. Patterns are regular expressions of the form $\Sigma(\Sigma\cup\cdot)^*\Sigma$, where a token Σ is either an amino acid or set of amino acids that have a high probability of inter-mutation based on a probability of mutation matrix, such as PAM or BLOSUM [7]. For instance, C.C.[ILMV].[KRD] could be one such pattern.

Based on that result, and exhaustive, top-down, hierarchical model for the unsupervised classification of large protein superfamilies has been proposed [11]. This has been applied to the G-Protein coupled receptor superfamily and it has been shown to correlate extremely well with the underlying biological function.

In this approach, pattern discovery is performed on a set of protein sequences s_1, \dots, s_m , at increasingly lower values of minimum pattern support, until a number of patterns greater than or equal to a predefined threshold, N_{min} , is reported by the algorithm. Typically, $N_{min}=100$. By minimum pattern support, we indicate the minimum number of sequences that must contain a given pattern before it is reported.

Among all reported patterns, the single most statistically significant one, π_1 , is selected. The latter is used to build a position specific scoring matrix, or profile M_1 [1], by measuring the frequency of individual amino acids at each position relative to the pattern, based on the locations where it occurs on the sequence set. Pseudo counts are computed using the minimum-risk procedure in [18].

Based on π_1 , the sequences $\{s\}$ are divided in two subset. One for which $M_1(s) \leq t$, the matching set, and one for which $M_1(s) > t$, the not-matching set. Here $M(s)$ is the lowest p-value computed for the profile M on the sequence s . The two thresholds, t and m_0 , can be selected such that $t > m_0$, leading to disjoint or partially overlapping sets. We shall call these sets respectively $\{s\}_1$ (matching) and $\{s\}_0$ (not-matching). The procedure can be repeated independently for these two sets, after masking the amino acids that support the pattern, leading potentially to two new profiles M_{11} and M_{01} . These can be used to further divide $\{s\}_1$ and $\{s\}_0$ into $\{s\}_{11}$, $\{s\}_{10}$, $\{s\}_{01}$ and $\{s\}_{00}$. The procedure can be repeated iteratively to produce a binary tree until each leaf contains fewer than m_0 sequences.

The following parameters have been chosen: patterns must contain at least 3 tokens in any substring of 12 characters starting with a token and minimum of 5 tokens globally. The latter constraint guarantees that patterns are specific enough so that random matches are unlikely in the training set. Amino acids are considered equivalent if the corresponding entry in BLOSUM50 is greater than 0. Patterns must have a statistical significance of z-score $\geq 10^3$ [16]. Here, $z=(n-n^*)/\sigma$, where n is the number of discovered patterns equivalent to π , n^* is the average number of patterns equivalent to π expected in a random set of sequence of same length and composition and σ is the standard deviation for n^* . Finally, $t=t_0=10^{-3}$ and $m_0=4$.

Results

We have studied 65 sequences for known C2 domains from a recent review [13], using the unsupervised clustering scheme described in the previous section. The method organizes the 65 sequences into a binary tree with 16 nodes and 9 leafs, using 10 independent patterns. This is shown in Fig.3. By analyzing the sequences in each node, we have been able to relate most of the patterns to important and distinguishing function/structural motifs. Descending the hierarchical tree, SPLASH discovered patterns progressively discriminate the sequence set into more and more specific functional sub-classes.

The first three motifs (M1, M11, M111) define a subclass of C2 domains which bind acidic phospholipids [5, 13, 14] in a calcium-dependent manner, this group encompasses the C2 domains of synaptic vesicle proteins (SVPs, which contain two C2 domains, C2A and C2B) and the classical protein kinases C (cPKCs). There are three loops at one end of the C2 domain structure which contain the residues responsible for calcium coordination in the calcium-dependent family members, each of the first three motifs corresponds to one of these loops. Fig.1 locates the calcium-binding motifs [14] on a GRASP surface representation of the structure of the synaptotagmin I C2A domain. In response to calcium binding, the electrostatic properties of this region of the domain change dramatically from negative to positive, resulting in targeting to membranes containing acidic lipids through electrostatic interactions. The next branch of the tree (M1111) separates the SVPs from the cPKCs by a motif which represents the first two beta strands of the domain structure. Given that the SVPs and cPKCs share structural homology in this region, this portion of the structure may play a role in mediating protein-protein intramolecular interactions. Following the M1111 branch further, the SVPs are separated into C2A and C2B domains. Splash detects a sequence motif (M11111, illustrated in Fig.2) that corresponds to a C-terminal alpha helix that had previously gone undetected by traditional sequence search algorithms and was not discovered until the three-dimensional structure of the rabphilin 3A C2B domain was solved [17].

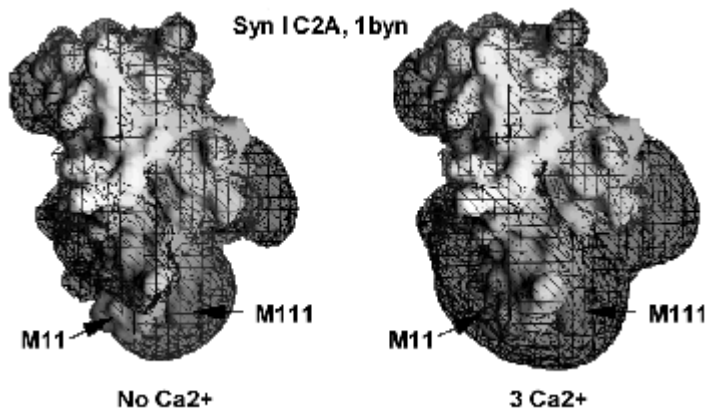


Figure 1. Electrostatic properties of the C2A domain from synaptotagmin I. Blue (red) meshes represent the +25 mV (-25 mV) equipotential contours. In the absence of calcium (left image), the calcium-binding loops (denoted M11 and M111) are negatively charged. When calcium binds (right image), this region becomes highly positively charged and targets the domain to the plasma membrane which is rich in acidic phospholipids. The Splash motifs M11 and M111 define the class of C2 domains that bind calcium and acidic lipids.

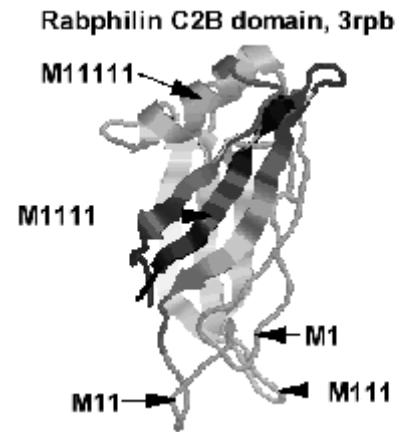


Figure 2. Four Splash motifs that define for the C2B domains of synaptic vesicle proteins. C2A and C2B domains are separated on motif M11111 which corresponds to a C-terminal alpha helix that remained undiscovered until the structure of a representative domain was solved.

Conclusions

This paper shows how the combination of pattern discovery and statistics can lead to a top-down unsupervised functional/structural classification algorithm. By applying this procedure to C2 membrane targeting domains, we have shown that a biologically plausible sub-classifications of the general protein set can be efficiently and accurately performed. Furthermore, all major functional subgroups of the C2 domain have been successfully identified and related to specific motifs that appear to play an important electrostatic role on the surface of the proteins. One of the motifs that were automatically discovered by this method could potentially play an important role in protein-protein interaction, based on correspondence to biochemical and structural studies. Also the method was able to identify a structural motif, M11111, that corresponds to an additional alpha-helix structural element in the structural core of C2 domains which had been previously gone undetected and has only recently been reported in the literature [17].

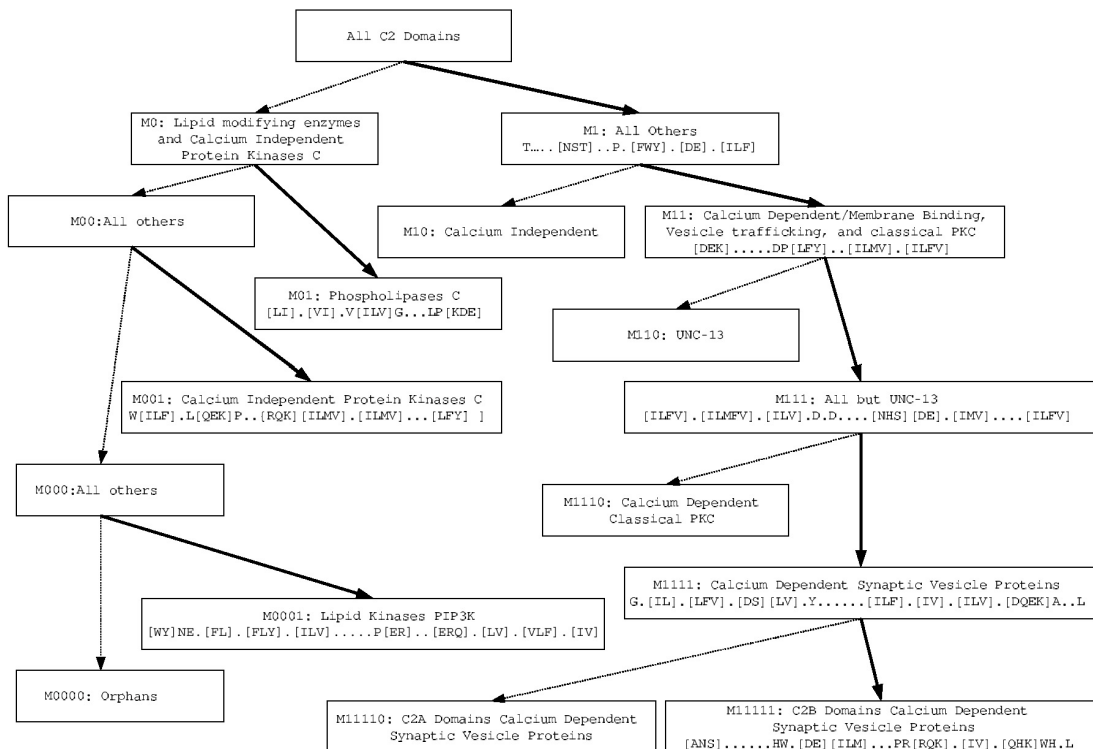


Figure 3. Functional feature decomposition tree for C2 membrane targeting domain

This indicates that the methodology could play an important role in the identification of protein motifs that play a critical functional or structural role in families of related sequences. We are planning to extend this study to include other membrane-targeting domains as well as other, previously uncharacterized domains that play a critical biological role, such as pleckstrin homology, C1, and FYVE domains [10].

References

1. Bailey T.L. and Gribskov M. "Methods and statistics for combining motif match scores", *J.Comp.Biol.*, **5**, 211-221, (1998).
2. Bork P. and Koonin E.V. "Protein sequence motifs", *Curr.Opin.Struct.Biol.*, **6**, 3666-376, (1996).
3. Brazma A. et al. "Approaches to the automatic discovery of patterns in biosequences", *J.Comp.Biol.*, **5**(2), 279-305, (1998).
4. Califano A. "SPLASH: structural pattern localization analysis by sequential histograms", *Bioinformatics*, **16**(4), 341-357, (2000). (Software available at www.research.ibm.com/splash).
5. Davletov B.A., Peristic O. and Williams R.L. "Calcium-dependent membrane penetration is a hallmark of the C2 domain of cytosolic phospholipase A2 whereas the C2a domain of synaptotagmin binds membranes electrostatically", *J.Biol.Chem.*, **273**, 19093-19096, (1998).
6. Hart R.K., Royyuru A, Stolovitzky G. and Califano A. "Systematic and automated discovery of patterns in PROSITE families" in *Proc.4th annual ACM Intl.Conf. on Comp.Mol.Biol., RECOMB'00*, S.Istrail, P.Pevzner and M.S.Waterman editors. (Apr.2000).
7. Henikoff S. and Henikoff J.G. "Amino acid substitution matrices from protein blocks", *Proc.Natl.Acad.Sci.USA*, **89**, 10915-10919, (1992).
8. Hofman K., Bucher P., Falquet L. and Bairoch A. "The PROSITE database, its status in 1999", *Nucleic Acids Research*, **27**, 215-219, (1999).
9. Honig B.H. and Nicholls A. "Classical electrostatics in biology and chemistry", *Science*, **268**, 1144-1149, (1995).
10. Hurley J.H. and Misra S. "Lipid protein interactions in subcellular targeting", *Annu.Rev.Biophys.Biophys.Chem*, In press, (1999).
11. Liu A. and Califano A. "Hierarchical classification of G-Coupled protein receptors by pattern discovery", in *3rd TIGR Conference on Genome Analysis*, Baltimore, (Nov.1999).
12. Murray D., Ben-Tal N., Honig B. and McLaughlin S. "Electrostatic interaction of myristoylated proteins with membranes: simple physics, complicated biology", *Structure*, **5**, 985-989, (1997).
13. Nalefski E.A. and Falke J.J. "The C2 domain calcium-binding motif: Structural and functional diversity", *Protein Science*, **5**, 2375-2390, (1996).
14. Rizo J. and Sudhof T.C. "C2-domains, structure and function of a universal Ca²⁺-binding domain", *J.Biol.Chem.*, **273**, 15879-15882, (1998).
15. Schwartz R.M. and Dayhoff M.O. "Matrices for detecting distant relationships", *Atlas of Protein Sequence and Structure*, ed. Dayhoff M.O., 353-358, (1978).
16. Stolovitzky G. and Califano A. "Pattern statistics in biological datasets", *IBM RC*, (1998). Available at www.research.ibm.com/splash.
17. Ubach J., Garcia J., Paige Nitter M., Sudhof T.C. and Rizo J. "Structure of the Janus-faced C2B domain of rabphilin", *Nature Cell Bio*, **1**, 106-112, (1999).
18. Wu T.D., Nevil-Manning C.G. and Brutlag D.L. "Minimum-risk profiles of protein families based on statistical decision theory", *J.Comp.Biol.*, **6**(2), 219-235, (1999).



**SECTION 5.
BIOINFORMATICS OF GENOME
STRUCTURE AND EVOLUTION**

MONO- AND BIVARIATE FLUORIMETRIC FLOW SORTING OF HUMAN CHROMOSOMES: QUANTITATIVE DATA ANALYSIS

**Kravatsky Yu.V., Poletaev A.I.*

Cytometry Group, Engelhardt Institute of Molecular Biology, Moscow, Russia

e-mail: jiri@mx.eimb.relarn.ru

*Corresponding author

Keywords: flow cytofluorimetry, chromosome sorting, human chromosomes, spectral decomposition, flow karyotype

Resume

Motivation:

Fluorimetric flow analysis of human chromosomes is not only an efficient means of obtaining cytogenetical information, but also permits preparative isolation of individual chromosome fractions. Statistical distributions obtained from fluorimetric analysis of a suspension of stained metaphase chromosomes convey information on the chromosome sets of the cells under study. Information can be directly obtained from experimental distributions only partially, so there is no wonder that system of valid quantitative analysis that can fully reveal all hidden information in the distribution is in high need. A procedure is proposed for quantitative processing of the data obtained by fluorimetric flow analysis of human chromosomes in the mono- and bivariate mode with one or two fluorochromes. In the bidirectional case one fluorochrome is specific toward GC and the other toward AT pairs in DNA.

Results:

The method was implemented as a software system for IBM-compatible PC AT 486 up, permitting comprehensive analysis of mono- and bivariate flow data and detection of chromosomal aberrations as well as quantitative comparison of homologous chromosomes and entire chromosome sets. System that is built consists of:

1. The program for quantitative analysis of monovariate statistical distributions that are obtained in the process of single-fluorochrome flow analysis of human chromosomes;
2. The program for quantitative analysis of bivariate statistical distributions that are obtained in the process of two-fluorochrome flow analysis of human chromosomes;
3. Programs for visualizing results of analysis; programs for converting obtained data to the common formats for report building convenience.
4. The results of analysis with our software system not only yields objective cytogenetic information at the chromosome level, but also provides information for preparing chromosomal material with controlled characteristics for genomic studies at the molecular level.

Availability:

Available on request from the authors.

Introduction

Univariate distributions, obtained in the flow cytometers, provide information on the relative DNA content in the chromosomes, whereas the bivariate system records two kinds of signal, each corresponding to the intensity of the fluorescence of an AT- or GC-selective fluorochrome specifically adsorbed on the DNA of a particular chromosome. Thus, the bivariate distribution patterns reflect not only the DNA content in the chromosomes but also the differences in the relative content of the AT and GC pairs accessible for dye binding. For this reason the bivariate distributions are more informative, and a larger number of chromosome fractions with an allowed extent of contamination can be obtained in this mode. The required cytogenetic information can be partly derived from the experimental statistical distributions, known as flow karyotypes of the cells under study, by qualitative analysis involving the available cytogenetic data on the cell line karyotype. However, even for the simpler case of univariate distribution it has been shown that only correct quantitative analysis can make the conclusions really convincing. Here quantitative processing means deriving the characteristics of the individual components of the distribution, i.e., groups of recorded signals corresponding to chromosomes of particular types.

The analytical methods for solving the problem above had not been adequately systematized when this work was initiated. Furthermore, the available realizations ignored the complex nature of the problem, which requires accounting for diverse factors. Analysis of the problem reveals that it is mathematically incorrect to solve of the

problem directly, because, first, there is no unique solution, and second, the more elements are taken in decomposition the closer one can approximate the experimental peak, thereby depriving the decomposition of physical sense. A correctly performed analytical procedure must not provide smooth fitting to the experimental data but should yield a set of parameters describing the physical object under study.

The cornerstone of a correct decomposition is a concrete physical model of the object (a chromosome set in our case), which describes and interrelates its real physical characteristics. The availability of a model is actualized as certain constraints on the decomposition process, which not only reduces the number of variants to be exhausted but also keeps the process physically realistic. The vast number of degrees of freedom makes a formal decomposition into initial peaks not only laborious (especially in bivariate analysis) but also meaningless, since its results do not reflect the actual parameters of the object. Even for the simpler univariate case it has been demonstrated that the availability of a model and the nature of information included therein are decisive for the precision and reasonability of the decomposition.

Furthermore, the real sets of experimental data arise from limited sampling and hence have a certain statistical scatter of the number of events registered in a particular experiment. Hence, one of the tasks of preliminary analysis is to smoothen or interpolate the dataset in one or another way, otherwise the precision of the approximation would be lowered because of the presence of a "high-frequency" noise component.

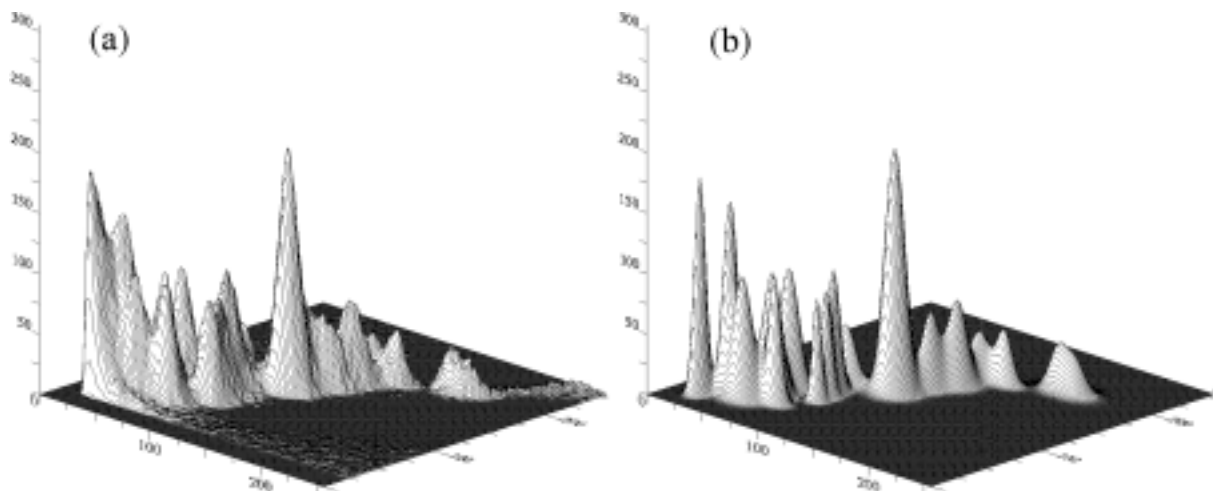


Figure 1. (a) Distribution of the fluorescence intensity of human chromosomes cleansed of contaminating signals and chromosome aggregates and then smoothed. The artifacts were filtered off by approximation with a two-sheeted hyperboloid of rotation, and chromosome aggregates were removed upon approximation with a 3D Gaussian. Filtered distribution was smoothed with B-spline interpolation. (b) Distribution of the fluorescence intensity of human chromosomes reconstructed using the results of analysis.

Methods and algorithms

The methods in practical use for analysis of flow karyotypes are: simple regional, least squares, and least squares in partial derivatives.

Comparative assessment of the methods for flow karyotype analysis allows the following conclusions. First, in the cases when the peaks overlap (large human chromosomes) or the background contamination is pronounced the simple regional method leads to incorrect estimates. Second, in the cases when the peak is superimposed on slowly changing background contaminations, the only method yielding correct values of the approximated parameters is the least squares in partial derivatives. Thus direct use of the least squares method as such gives an erroneous volume for human chromosome 21 and falsely suggests trisomy in this chromosome for the assessed karyotype. Third, the mean values of the distributions can be estimated with high precision by any of the above methods. Thus, the simple regional method is quite suitable for obtaining the initial approximations to be further processed with least squares and least squares in partial derivatives. Moreover, it is unreasonable to apply these methods as such without approximating (or removing) background contaminations, as this yields inadequate results of the target approximation. It should also be noted that approximation by any method is meaningless without building a corresponding physical model; if the background contaminations have been approximated (or removed), both least squares versions would yield the same results. Hence it must be admitted that in the complex system of flow karyotype analysis the use of the standard least squares method is more expedient, as its computing costs are less and its implementation is simpler, usually requiring only modification of the already implemented standard algorithms. Our procedure combines the sequential truncation technique with subsequent approximation of the chosen region in the least

squares mode basing on the physical model of the process built for this purpose. This also includes an original approach to filtering off the contaminating signals.

Implementation and results

We can illustrate the procedure developed by the following Fig.1 where you can see that parameters has been correctly obtained for all chromosomal groups, including the most complex cases like 9-12, 15-17 and 19-22 chromosoma.

Discussion

We have created the procedure for adequate quantitative flow karyotype analysis. Our procedure proved stable enough against arbitrary variation of the initial parameters of the approximation process within separate groups of peaks. Procedure has been implemented as software system on contemporary PCs and is therefore feasible for any laboratory equipped with a modern cytofluorimetric flow sorter.

EDUCATION ON THE BASIS OF THE GENEEXPRESS SYSTEM: BUSINESS GAME "REGULATORY SIGNALS"

^{*}*Ponomarenko M.P., Ponomarenko J.V., Lavryushev S.V., Vorobiev D.G.,* [§]*Minina A.V.,*
[§]*Ivashin S.A.,* [§]*Mikhailov Yu.I.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

[§]Siberian Trade University, Novosibirsk, Russia

e-mail: pon@bionet.nsc.ru

*Corresponding author

Keywords: education, DNA, regulatory signals, recognition

Resume

Motivation:

According to the Russian state standard "higher education", this standard includes obligatory discipline "Modern concepts in natural sciences", which incorporates the basic knowledge in bioinformatics and introduction to the Russian National Project "Human Genome".

Results:

The scenarios are designed and implemented for application of the GeneExpress system for education of future specialists in economy and law to the bases of bioinformatics, which produces them with valuable informational-analytical acquirments.

Availability:

The system GeneExpress, <http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/>. The textbooks for application of the GeneExpress system with educational purposes [2] are at hand of Prof. Yu.I. Mikhailov (natural@sibupk.nsk.su).

Introduction

Representation about existence of common features in organization and functioning of self-reproducing systems originated naturally becomes of common knowledge over the recent years. To these systems may be referred living organisms, languages of human communication, social processes, law systems, and economical markets. With respect to this representation, informational–analytical methods, intensively developed within the frames of the Russian National Project "Human Genome", are of vital interest for education of economists and lawyers.

The goal of the class hours is to formulate a representation about composition, organization, functioning, and regulatory capacities of inherited information in living organisms. For the training example of genetic information, a regulatory signal was chosen. The practical class hours with the usage of the GeneExpress system permit to fix the theoretical course materials by the following directions:

- exposing students to the bases of bioinformatics and the concepts of the "Human Genome" Project;
- mastering of scenarios for genetical information analysis;
- providing of economists and lawyers with informational-analytical attainments.

Scenario № 1: "DETECT THE SIGNAL !" (Figure 1)

You are a bioinformatisist detecting the signals of transcription (reading) of genes.

1. Insert into the Web-browser an ADDRESS <http://wwwmgs.bionet.nsc.ru/mgs/>.
2. The proper result is given in fig.1A: you visit the Molecular-genetic server of IC&G SB RAS (Fig.1a). Click the script "SAMPLES" (arrow 1).
3. The proper result is seen in fig.1B: this is the "SAMPLES" database. It accumulates regulatory signals within DNA. Here one can find the examples of the signals under analysis. Click the script "ENTER SAMPLE" (arrow 2).
4. The proper result is seen in fig.1C. In order to identify transcription regulatory signals, choose the option "Eukaryotic Transcription Factor Binding Site Compilation" (arrow 3) and click the option "Group Content" (arrow 4).
5. The proper result is given in fig.1D. Choose the name of the signal that you analyze from the list (arrow 5) and then choose the option "format Fasta"(arrow 6) in order the data were compatible to the program for analysis. Click "GET SAMPLE" (arrow 7).

6. The proper result is given in fig.1E. These are the DNA fragments with the length of 120 bp, which are characterized by the presence of the signal under study. In menu "EDIT", choose the options "SELECT ALL" (arrow 8) and "COPY" (arrow 9). Insert URL="http://www.mgs.bionet.nsc.ru/programs/gibbs_nuc/" (arrow 10).
7. The proper result is seen in fig.1F. This is a program tool GIBBS-Sampler. It enables to find the regions of the best textual similarity between all DNA fragments inserted. Set the cursor into the window "INPUT" (arrow 11) and from menu "EDIT" choose the option "INSERT" (arrow 12). Set the boundaries (length) of the regulatory signal under study (arrow 13). Click the script "SUBMIT" (arrow 14).
8. The proper result is seen in figures 1G and 1H. Fig.1G demonstrates the part of the screen with the revealed regions of the best similarity between DNA fragments (in the frame). Since each DNA fragment that you have inserted contains the signal analyzed, then, you may expect that more likely these regions are exactly the signals you have detected. In fig. 1H, the part of the screen is shown, with nucleotide frequencies in positions of the signals detected together with their consensus (i.e., significantly frequent nucleotides in positions of the signal, the level of significance $\alpha < 0.01$).

Scenario № 2: "RECOGNIZE A SIGNAL!" (Figure 2)

You are a bioinformaticist detecting the presence of a definite regulatory signal in the annotated DNA fragment.

1. Insert via the URL=<http://www.mgs.bionet.nsc.ru/mgs/systems/geneexpress/>.
2. The right result of your action is seen in fig. 2A: This is the GeneExpress System of the Institute of Cytology and Genetics of SB RAS. Click the script "SITE RECOGNITION" or the icon shown by the arrow 1.
3. The right result is given in fig. 2B. This is the section "Site Recognition". It accumulates the programs for recognition of regulatory signals. Here one can find the programs for analysis of the signal under study. Click one of the scripts "BDNAvideo", "Nsamples", "Selex_DB" or "ConsFreq" (below the choice of "ConsFreq" is exemplified and shown by the icon, at which the arrow 2 points out).
4. The right result is shown in fig. 2C. It leads to the section "Consensuses and Weight matrices". Here are accumulated the programs for recognition of regulatory signals according to their textual similarity, as it was described above, in the game "Detect the signal!" Click one of the scripts, "MATRIX" or "CONSENSUS" (further the choice of "MATRIX" visualised by its icon (arrow 3), is described).
5. The right choice is shown in fig. 2D. This is the knowledge database MATRIX. It accumulates the programs for regulatory signals recognition according nucleotide frequencies in each position of these signals. In order to find such program for the signal, which you are analyzing, click "Site Name" (arrow 4).
6. The right result is shown in the figure 2E. This is a vocabulary of signals. Click the script "List values" (arrow 5), find the signal you are analyzing from the list of names appeared and click its name (arrow 6).
7. The right result is demonstrated in fig. 2F. This is the entrance to the document, where the programs for recognition of the signal you study are described. Click the script "MATRIX: NAME" (arrow 7), find in the text of the document appeared the script "WWW ... Programs" and click it (arrow 8).
8. The right result is shown in the figure 2G. This is the program for recognition of the signal you study. Input into the window "INPUT..." the DNA sequence that you analyze (fig.2F, arrow 9) and click the script "EXECUTE" (fig.2F, arrow 9).
9. The proper result is shown in fig. 2H. This is a similarity profile (Y axis) of the DNA analyzed with the signal under study (at the X axis, the numbers of positions are given). The ring marks the maximum of such similarity. This maximum with the high probability corresponds to the signal recognized.

In the Table, the models describing how different factors act on identification and recognition of signals are listed. Analysis of such models enables economists and lawyers to obtain informational-analytical acquirments necessary for their professional activities. Thus, the business game "Regulatory signals" is a novel approach to such education. Its novelty is a simultaneous learning to the basic concepts of bioinformatics and informational-analytical methods by the example of solving the task of deciphering of genetical information by means of computer systems developed within the frames of the Project "Human Genome". This game responds to demands of contemporary education of the specialists in economy and law.

The work was supported by Russian Foundation for Basic Research, Russian National Human Genome Project, DOE and TACIS.

References

1. Kolchanov, N.A., Ponomarenko, M., et al. (1999) Integrated databases and computer systems for studying eukaryotic gene expression. *Bioinformatics*, **15**, 669-686.
2. Minina A.V., Mikhailov Yu.I. et al. "Concepts in the modern natural science. III. Genetic information: regulatory signals (manual)". 1999. Novosibirsk: Ed. by Sib.UPK. 27 p.

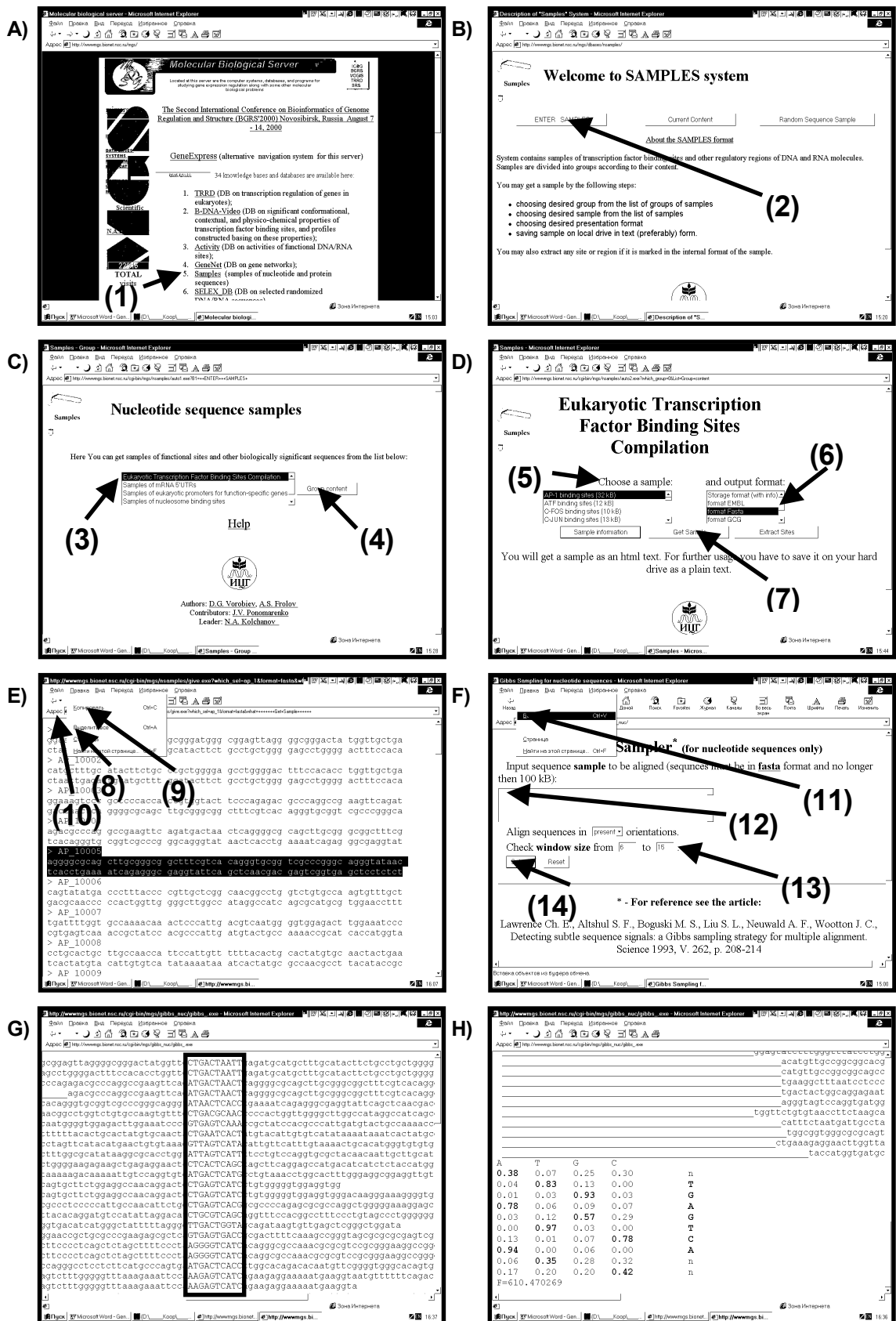
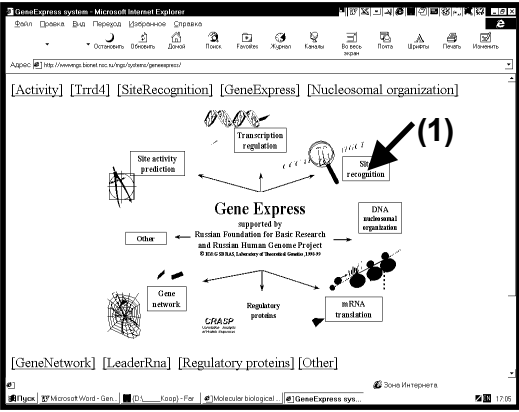
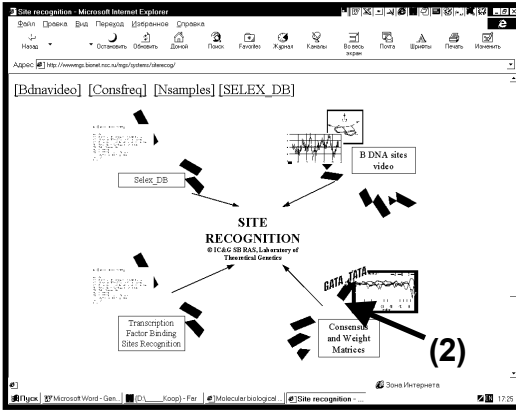
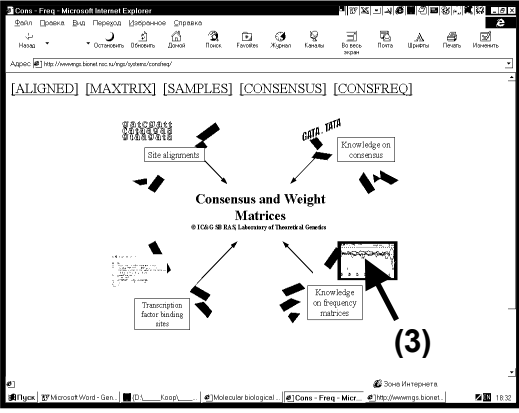
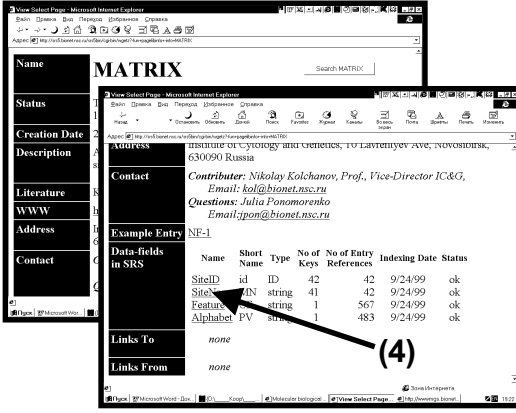


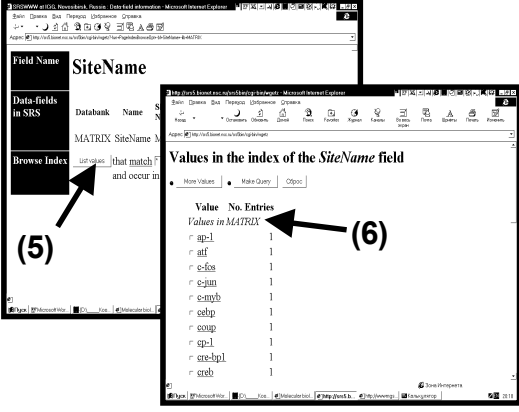
Figure 1. An example of scenario №1 of the business game "DETECT THE SIGNAL!".

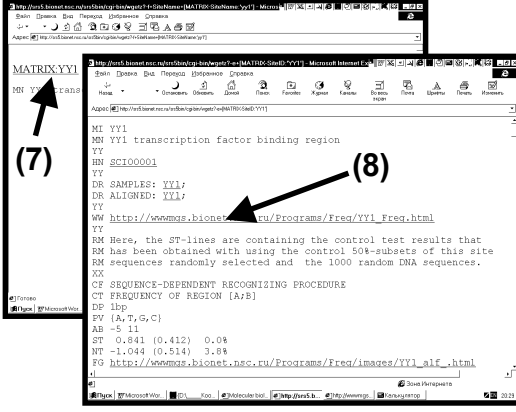
A) 

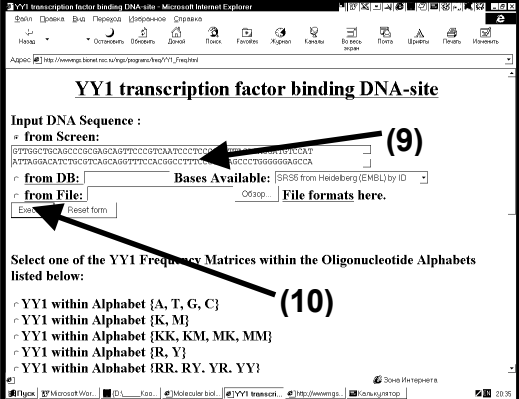
B) 

C) 

D) 

E) 

F) 

G) 

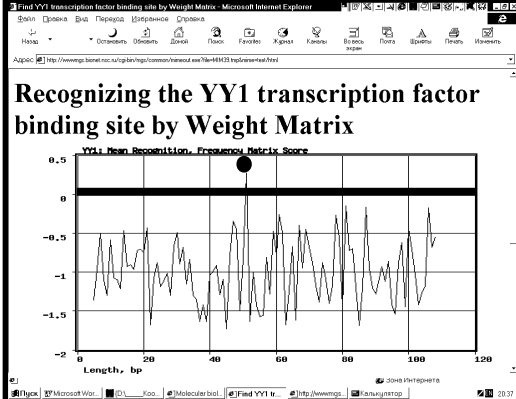
H) 

Figure 2. An example of scenario №2 of the business game "RECOGNIZE A SIGNAL!".

Table. Models of the action of different factors on the signal analysis.

Factor	Model of the factor's action	→
Scenario No. 1 "DETECT THE SIGNAL!" (Figure 1)		
<i>A priori</i> knowledge	Vary the expected value of the signal	13
Data incompleteness	Delete some DNA fragments with the signal	12
Data excess	Repeat some DNA fragments with the signal	12
Erroneous data	Add arbitrary DNA (for example, AAAAAA.....AAA)	12
Data misrepresentation	Insert A, T, G, C into the fragments with the signal	12
Data loss	Remove "short parts" of DNA fragments with the signal	12
Noise in data	Add to the set of DNA fragments with the signal analyzed some DNA fragments with the other signals	7, 12
Invent a factor	Model the action of YOUR factor	all
Scenario No. 2 "RECOGNIZE A SIGNAL!" (Figure 2)		
Coding	Vary the alphabets of nucleotides and oligomers	10
Method for analysis	Recognize the signal on the base of its consensus	3
Nature of signals	Recognize the signal by similarity to its analogs (SELEX_DB)	3
Signal is unknown	Vary the "name" of the regulatory signal	6
Data misrepresentation	Change some nucleotides in the DNA sequence under analysis into some other nucleotides	9
Data loss	Remove "short parts" from DNA analyzed	9
Invent a factor	Model the action of YOUR factor	all

COURSE “INTRODUCTION TO BIOINFORMATICS”

**Valuev V.P., Afonnikov D.A.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

e-mail: valuev@bionet.nsc.ru

*Corresponding author

Keywords: bioinformatics teaching

Resume

Motivation:

There is no widely known courses on bioinformatics in Russian, though this field is of key importance in today's science.

Results:

The course “Introduction to bioinformatics” is given to the students of the 4th year of the department of Cytology and Genetics of the Natural Science Faculty of Novosibirsk State University. It comprises theoretical matter, description of Internet resources with links to them and individual tasks for the students.

Availability:

<http://www.test/mgs/mix/bioinformatics/>

Introduction

Bioinformatics is one of the most dynamically evolving areas of knowledge. Having emerged in the late 1970s, by the mid-90s it appeared as a vast independent field. By now the whole biological community realized that progress in understanding life, and first of all its molecular basis, is determined by application of computer methods (Wiley, 2000). So basic knowledge of bioinformatics, familiarity with Internet resources and ability to duly apply them grew vitally important for every biologist. On the other hand, just as each science that evolves drastically, bioinformatics needs an inflow of new people, well educated and competent. In this connection the problem of teaching bioinformatics acquires a paramount importance. But its short history and fast development not only determine great profit from its teaching, but make it difficult also. First, some manuals and review books come out of date soon. Second, there is a gap between the developers of algorithms and software and their users. Books written by mathematicians and programmers often pay not enough attention to the biological aspect of the matter (e.g. see Baldi and Brunak, 1998). On the contrary, in the works of biologists the balance is shifted towards concrete biological examples, and computer methods are only slightly mentioned. Besides, by virtue of dynamism of the field, there are a great variety of methods and approaches, and because the authors of manuals are active researchers and participants of scientific dispute, not all the approaches and methods are equally represented and given an impartial judgement. Finally, the third thing specific for bioinformatics teaching is the fact that virtually all the bioinformatics is plunged into the Internet, so the databases, analysis tools, interactive maps of metabolism, navigation systems and other resources require certain practical skills (<http://www.nsu.ru/biology/courses/internet/main.html>). The said above ensues the need for development of a bioinformatics course in Russian, which hopefully would be free of mentioned flaws. We are developing such course for students of the Cytology and Genetics department of Novosibirsk State University.

Structure of the course

The course is realized as an Internet-resource, split into separate HTML-documents, corresponding to the theoretical lessons and practical tasks. The navigation in the course is possible by means of hyperlinks. During the study students solve real biological tasks with the help of true scientific resources of the web, to which they come through hyperlinks in the body of lessons and tasks. For educational purposes there are examples for many tasks that simulate working with remote resources (examples of query and query result pages, results of analysis etc.) The goals of the course are: giving students some skills for work with the Internet; giving them a review of problems that theoretical genetics of today faces; making them familiar with the biological resources of the World Wide Web; showing them mathematical principles underlying computer methods for genetic information analysis.

Content of the course

The course includes the following sections:

1. General notion on the Internet. Protocols HTTP, FTP, Gopher. Search engines.

2. General biological resources of the web. Notion on databases. Bibliographic databases MEDLINE and AGRICOLA.
3. SRS system. Main databases on molecular biology: EMBL, GenBank, SwissProt, PIR, PDB.
4. Search and analysis of genetic information. Alignment. FASTA, Blast, Clustal.
5. Primer design for PCR.
6. Gene networks.
7. Completed genomes and model organisms.
8. Useful referential Internet resources for biologists.

The material within each section is tied with a common subject. The necessary theoretical matter is presented, review of main web resources is given and individual tasks are developed.

1. Notion on the Internet.

The first lesson is devoted to the basis of working with Internet. The material is organised in such a way that one who has no experience in working in the web can get basic skills of navigating with the help of hyperlinks, get familiar with principal search engines, learn about different protocols (http, ftp, gopher). In the same time, a rather experienced user also can take profit from a glance at the first lesson. Individual tasks include principal operations with the Internet (search of information with the help of search engines, saving files, dealing with ftp-server).

2. Bibliographic databases.

Every scientific research starts with a look at literary references on the topic. In biology there are two search engines for reviewed literary information: MEDLINE, dealing with journals on biology and medicine, and AGRICOLA, which specialises on biological and agricultural subjects. These engines have different format, but both allow complex queries, with the help of which one can get great number of references covering given topic as well as find given paper starting from incomplete data. The lesson comprises a short description of content of these databases and their query forms and tasks for various kinds of searching these databases.

3. Main molecular-biological databases.

Operations with main molecular-biological databases, such as GenBank, EMBL, Swiss-Prot et al., are inevitable parts of each biologist's activity. Besides practical skills of dealing with SRS (Sequence Retrieval System), which is the uniform tool for searching molecular-biological databases, the material of the lesson comprises the general review, namely short descriptions of the main databases, their content, statistics on sources, organisms represented, cross-links.

4. Genetic information analysis.

The main operations with the genetic sequences are homology search and alignment. The problems of the most effective alignment algorithms, statistical significance, similarity matrices are from the oldest in bioinformatics, but their significance has not decreased with years. The description of basic algorithms (FASTA, BLAST, CLUSTAL), introduction of the notion of statistical significance, description of main types of similarity matrices, supplied with examples and individual tasks, allow students to grasp the mathematical basis and to apply these methods correctly.

5. Primer design for PCR.

Polymerase chain reaction (PCR) is one of the most powerful and necessary tools of modern molecular biology. Thus, though primer design does not belong to the range of fundamental genetic problems, the course includes a lesson that describes the primer design algorithm starting from the set of sequences on the example of GeneFisher software.

6. Gene networks.

In recent years due to the increasing rate of accumulating primary molecular-genetic information the accent of the research is being shifted to the study of metabolism and signalling pathways. Resources have been developed, that make an effort towards adequate description of hierarchical organisation of living systems on the levels from genes to separate organs and tissues, description of metabolism of model organisms etc. Within the lesson are reviewed main principles underlying gene networks, and examples of various Internet-resources on this subject are given.

7. Completed genomes and model organisms.

The advent of completely sequenced genome for several tens of organisms allowed to pass to qualitatively novel stage of analyzing genetic information. Became possible determination of orthologs and paralogs, reconstruction of metabolism, recovery of species determinants (unique genes and proteins specific for the organism), study of evolution character of gene and protein families etc. The lesson shows main stages of genome sequence annotation and up-to-date state of completed genome studies.

8. Useful referential Internet resources for biologists.

The last lesson represents a selection of links to the useful sites for biologist with their short description. It includes links to free digital libraries, newsgroups connected with biology, web-pages of various scientific societies and organizations etc.

Conclusion

In last decades the interest in molecular biology and genetics has increased strongly. The analysis of molecular-genetic information became a field of joint effort of biologists, mathematicians, physicists. But their collaboration is hindered substantially by difference in their background. In this connection arises acute problem of training people familiar with all these fields. Hopefully the course that we develop for students in biology, besides merely giving them practical skills, will help bridge the gap between biologists, on one side, and programmers and mathematicians, on the other side. The course being developed is one the first of the kind in Russian and has already proved important in the education of biologists.

Acknowledgement

The work was supported with Integration Project of SB RAS No 66.

References

1. Wiley HS (2000) Are Computers Evolving in Biology? HMSBeagle, 75
http://www.biomednet.com/hmsbeagle/75/viewpts/op_ed
2. Baldi,P. and Brunak,S.(1998) *Bioinformatics: a machine learning approach*. A Bradford Book. The MIT Press.
3. <http://www.nsu.ru/biology/courses/internet/main.html>

BIOINFORMATICS: NOVEL PROFILE OF VOCATIONAL EDUCATION

^{1*}Valishev A.I., ²Kolchanov N.A., ²Podkolodny N.L., ¹Melnikov V.N., ¹Alsymbayeva L.G.,
¹Yaroslavtseva R.G., ³Haans W.J.A.

¹Novosibirsk State University High College of Informatics, Novosibirsk, Russia
e-mail: valishev@ci.nsu.ru

²Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

³FONTYS University, Eindhoven, The Netherlands

*Corresponding author

Keywords: bioinformatics, middle vocational education, curriculum

Resume

Motivation:

The need in preparing of specialists in the field of bioinformatics is produced by (1) intensive development of biotechnology, gene engineering, molecular medicine, and drug design technology in the modern biology, (2) the lack of programming and information technology assistance in biological, ecological, medical, and agricultural research laboratories and practical institutions, together with medical insurance companies in Siberia.

Results:

A 4-year education programme is worked out in the Novosibirsk State University High College of Informatics (HCI) in order to train students for systematic and practical application of computer science, computer technology in combination with general natural disciplines, i.e., chemistry, general biology, ecology, human and animal physiology, genetics, biochemistry, botany, and fundamentals in anatomy and pathology. The project described is the first educational project in Russia aimed to train specialists in bioinformatics.

Availability:

The HCI is open to the international cooperation in the field. The development of Curriculum in Bioinformatics is a part of DELPHI project, with HCI being a pilot institution.

Introduction

Modern research laboratories, particularly in the field of genetics, deal with processing of tremendous numerical data sets and are instrumented by complex computer-assisted equipment. The support and service management of such systems need to engage a large number of auxiliary personnel who should be able to use up-to-date information and computer technologies and possess by acquirements to utilize them in order to solve concrete problems appearing in practice.

Content

The educational programme is divided into two phases, the preliminary phase (during the first two years of education) and the main phase (the subsequent two years). The goals of the preliminary phase are (1) to orient the students towards new speciality; (2) to study general humanitarian, social-economical, mathematical, general professional disciplines oriented to natural sciences and general computer science disciplines, and (3) to determine the student's intellectual capacity for acquirement of new skills that are necessary for successful completion of the study and to select the students for further stages of education.

During the main phase, the basic training necessary for the software and hardware assistance will be provided. A set of special disciplines and optional subjects will be studied. Students will have the practice in the research laboratories of the main social partner of the HCI, the Institute of Cytology and Genetics of SB RAS.

After graduation from the HCI, the students obtain the specialisation as bioinformation technology and programming assistant. They will be able to:

- set up a definition of the requirements for a software system with the help of software engineer;
- model the data and design the databases;
- define the requirements, design and build a robust, maintainable, efficient software product under the guidance of software engineer using the third and higher generation of programming languages;
- test, integrate and install software systems;

- set up a definition of requirements, specify simple hardware interfaces and design, and construct the machine necessary for biological experiments;
- consult in purchase of hardware and software for computer systems and networks, to look for and choose the distributing companies;
- give advice on information and technical nature to the manager implementing communication networks, which programmable resources should be used for the benefit of data, text, sound, image, and control signals quality;
- design and edit web-sites.

Principle discriminative feature of education at this specialization is an early involvement of students into the practical work in research laboratories equipped with modern devices, which enables to master the whole technological cycle of scientific experiment and production based on progressive technologies. In the process of training, the students will get acquirements in servicing and adjustment of specialized computational complexes, connection of computer with various devices and sensors, development of local networks, experimental data treatment and statistical analysis, usage of computer packages for data analysis, recognition of visual data, etc.

Among the basic educational topics are construction and development of databases and knowledge bases, digital libraries, mastering of business education and editing systems together with the usage of the Internet-technologies.

We suppose that the graduates from the HCI specialized in bioinformatics will get the job in private firms and state organizations engaged in application of modern biotechnology, medical service, testing for food and drug quality, design of medicines and food additives, ecological monitoring, etc.

MINK ENTERITIS VIRUS VP2 GENE FRAGMENTS ANALYSIS

**Tkachev S.E.*

Institute of Bioorganic Chemistry of SB RAS, Novosibirsk, Russia

e-mail: tkachev@niboch.nsc.ru

*Corresponding author

Keywords: mink enteritis virus (MEV), VP2 gene, Rodniki strain, Cherepanovo-98 virus isolate

Resume

Motivation:

Mink enteritis virus (MEV) causes acute enteric disease of minks resulting in high mortality (Christensen *et al*, 1994). Previously nucleotide sequences of some mink enteritis virus strains isolated in Japan (Higashihara *et al*, 1981), in Europe and USA (Christensen *et al*, 1994; Parrish *et al*, 1984) have been determined. Genome structure of MEV strains isolated in Russia have not been studied earlier. The virion protein VP2 is the MEV most abundant capsid protein and, consequently, the main viral immunogen. In baculovirus expression system VP2 protein is shown to form virus-like particles (Christensen *et al*, 1994) that might be used as effective DNA delivery system into eukariotic cells.

Results:

Mink enteritis virus (Rodniki strain and Cherepanovo-98 isolate) VP2 gene 5' and 3' - fragment nucleotide sequences were determined. The comparative analysis shown 100% homology of VP2 gene sequence between Rodniki strain with MEV antigenic type 2. In contrast, Cherepanovo-98 isolate had nucleotides typical for MEV antigenic type 1 and 2 as well as Abashiri strain.

Availability:

The presented sequence data have been submitted to the GenBank database and assigned accession numbers AF201476, AF201477, AF201478 and AF201479.

Introduction

Mink enteritis is an acute disease with high mortality of infected animals. This disease is caused by mink enteritis virus, which is the member of *Parvovirus* genera in *Parvoviridae* family. Viral particles are icosahedral about 26 nm in diameter and consist only of DNA and proteins. The MEV genome is linear, non-segmented single-stranded (-)sense DNA, 5094 nucleotides (n.) in length. There are two large open reading frames (ORFs), one in the 5'-half and the other in the 3'-half of the genome coding polypeptides of 722 and 668 amino acid residues long, respectively (Kariatsumari *et al*, 1991). The first ORF encodes nonstructural protein(s), the second one - three structural proteins. The mature virions are shown to contain 3 proteins - VP1, VP2 и VP3. The main viral immunogen is VP2 protein.

VP2 gene structure of feline panleukopenia virus (FPLV), canine parvovirus (CPV) and MEV were shown to have more than 98% homology (Horiuchi *et al*, 1994). The data are consistent with the hypothesis that MEV and CPV are host-range variants of FPLV. MEV isolates were subdivided into the three antigenic types (Parrish *et al*, 1984) based on their interaction with monoclonal antibodies panel.

MEV Rodniki strain was isolated from the infected mink in 1976 near Moscow. Now it is used for vaccine production. Virus Cherepanovo-98 isolate was found in 1998 in Novosibirsk region.

The aim of this work was to determine mink enteritis virus (Rodniki strain and Cherepanovo-98 isolate) VP2 gene sequences and compare them with known MEV sequences.

Methods

Amplification of the capsid protein gene VP2 by polymerase chain reaction (PCR).

The capsid protein VP2 gene was amplified by the PCR method. Two primers, M29 (5' – AAGAGGTCGACTTGACCAATGGGTGATGG - 3') and M30 (5' – AACATCTCGAGTATATAATAAATTTT TAGG - 3') were chosen from highly conservative sequences of 5' and 3'- regions of MEV antigenic types 1 and 2 VP2 gene (Parrish *et al*, 1984). MEV DNA isolated by method described earlier (Bloom *et al*, 1980) was used as a template. The reaction contained 20 pmol of each primer, buffer for *Taq* polymerase (16,6 mM (NH₄)₂SO₄; 67 mM tris-HCl (pH 8,9); 1,5 mM MgCl₂; 0,05% Tween 20), 0,4 mmol of each deoxynucleoside triphosphate. After the incubation of mixture at 94°C for 5 min 1 a.u. of *Taq* polymerase was added. Amplification was performed in the following conditions: denaturation at 94°C for 1 min, primer annealing at 50°C for 1 min and

extension at 72°C for 2 min (5 cycles); and then 30 cycles with the same conditions but with primer annealing at 60°C.

Nucleotide sequence analysis

Nucleotide sequence analysis was carried out by the dideoxynucleotide sequencing method (Sanger *et al*, 1977) with some modifications. Computer analysis of the nucleotide sequences was performed using program ClustalW (<http://www2.ebi.ac.uk/clustalw/>) (Higgins *et al*, 1994).

ABASH	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	GAACGGGTCT
MEV-E	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	GAACGGGTCT
MEV-D	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	GAACGGGTCT
WOLF	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	GAACGGGTCT
JOHNS	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	AAACGGGTCT
DK90	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	GAACGGGTCT
CH98	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	AAACGGGTCT
RODN	2836	AGCAGTTCAA	CCAGACGGTG	GTCAACCTGC	TGTCAGAAAT	GAAAGAGCTA	CAGGATCTGG	AAACGGGTCT
		*****	*****	*****	*****	*****	*****	*****
ABASH	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACGGGTAC	TTTCAATAAT	CAGACGGAAT
MEV-E	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACGGGTAC	TTTCAATAAT	CAGACGGAAT
MEV-D	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACGGGTAC	TTTCAATAAT	CAGACGGAAT
WOLF	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACAGGTAC	TTTCAATAAT	CAGACGGAAT
JOHNS	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACAGGTAC	TTTCAATAAT	CAGACGGAAT
DK90	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACAGGTAC	TTTCAATAAT	CAGACGGAAT
CH98	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACAGGTAC	TTTCAATAAT	CAGACGGAAT
RODN	2906	GGAGGCGGGG	GTGGTGGTGG	TTCTGGGGGT	GTGGGGATTT	CTACAGGTAC	TTTCAATAAT	CAGACGGAAT
		*****	**** *	*****	*****	**** *	*** *	**** *
ABASH	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
MEV-E	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
MEV-D	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
WOLF	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
JOHNS	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
DK90	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
CH98	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
RODN	2976	TTAAATTTTT	GGAAAACGGA	TGGGTGGAAA	TCACAGCAA	CTCAAGCAGA	CTTGATACA	
		*** *	*****	*****	*****	*****	*****	

Figure 1. MEV VP2 gene 5'-region sequences analysis. List of abbreviations: ABASH – Abashiri strain; MEV-E – MEV-e isolate; MEV-D – MEV-d isolate; WOLF – Wolf isolate (antigenic type 1); JOHN – Johnson isolate (antigenic type 2); DK90 – DK90 isolate; CH98 – Cherepanovo-98 isolate; RODN – Rodniki strain.

ABASH	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
MEV-E	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
MEV-D	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
WOLF	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
JOHNS	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
DK90	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
CH98	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
RODN	4331	GCGCCTAATT	TAACAAATGA	ATATGATCCT	GATGCATCTG	CTAATATGTC	AAGAATTGTA	ACTTACTCAG
		*****	**** *	*****	*****	*****	*****	*****
ABASH	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
MEV-E	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
MEV-D	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
WOLF	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
JOHNS	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
DK90	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
CH98	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
RODN	4401	ATTTTTGGTG	GAAAGGTAAA	TTAGTATTTA	AAGCTAAACT	AAGAGCATCT	CATACTTGA	ATCCAATTCA
		*****	*****	*****	*****	*****	*****	*****
ABASH	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
MEV-E	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
MEV-D	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
WOLF	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
JOHNS	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
DK90	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
CH98	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
RODN	4471	ACAAATGAGT	ATTAATGTAG	ATAACCAATT	TAACATATCTA	CCAAATAATA	TTGGAGCTAT	GAAAATTGTA
		*****	*****	*****	***** *	*****	*****	*****

Figure 2. MEV VP2 gene 3'-region sequences analysis.

Abbreviations are the same as for Figure 1.

Implementation and results

Mink enteritis virus (isolate Cherepanovo-98) was isolated from faeces of infected minks. MEV Rodniki strain was received from State Research Institute of Veterinary Drugs (Moscow, Russia). Viral DNA was isolated and then the amplification of MEV VP2 gene fragment 1363 n. in length was carried out. Obtained PCR fragments were used as a template for sequencing. Then the alignment of MEV VP2 gene 5'- and 3'-regions sequences with MEV DNA sequences contained in GeneBank database (<http://www.ncbi.nlm.nih.gov/Genbank/index.html>) was performed with ClustalW program. The results are shown in Fig.1 and Fig.2.

Discussion

Earlier nucleotide sequences of some MEV strains isolated in Japan, Europe and USA have been determined. In this study VP2 gene fragments of MEV strains isolated in Russia were investigated for the first time.

5'- and 3' – regions of MEV Rodniki strain VP2 gene were shown to have 100% homology with the corresponding sequences of MEV Johnson isolate which is belong to antigenic type 2.

MEV Cherepanovo-98 isolate was found on mink farm in Siberia in 1998 during parvoviral enteritis outbreak. Nucleotides 2920 and 2971 of the isolate appeared to differ from those of all the known MEV sequences. Interestingly, nucleotides at positions 2896, 2959, 4345 and 4396 of MEV Cherepanovo-98 genome sequence were identical to MEV antigenic type 2, nucleotide 4381 was the same as MEV antigenic type 1 and nucleotide 2950 was typical for Abashiri strain. So these data suggest that Cherepanovo-98 isolate could be the new antigenic type of MEV. Further investigations of MEV genome structure are necessary.

References

1. Bloom M.E., Race R.E., Wolfenbarger J.B. «Characterization of Aleutian disease virus as a parvovirus» *J Virol.* **35**, 836-843 (1980).
2. Christensen J., Alexandersen S., Bloch B., Aasted B., Uttenthal A. «Production of mink enteritis parvovirus empty capsids by expression in a baculovirus vector system: a recombinant vaccine for mink enteritis parvovirus in mink» *J Gen. Virol.* **75**, 149-155 (1994).
3. Higashihara T., Izawa H., Onuma M., Kodama H., Mikami T., Noda H. «Mink enteritis in Japan. I. Isolation and characterization of causative virus and its pathogenicity in cat.» *Jpn. J. Vet. Sci.* **43**, 841-851 (1981).
4. Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J. «CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.» *Nucleic Acids Res.* **22**, 4673-4680 (1994).
5. Horiuchi M., Goto H., Ishiguro N., Shinagawa M. «Mapping of determinants of the host range for canine cells in the genome of canine parvovirus using canine parvovirus/mink enteritis virus chimeric viruses» *J Gen. Virol.* **75**, 1319-1328 (1994).
6. Kariatsumari T., Horiuchi M., Hama E., Yaguchi K., Ishiguro N., Goto H., Shinagawa M. «Construction and nucleotide sequence analysis of an infectious DNA clone of the autonomous parvovirus, mink enteritis virus» *J Gen. Virol.* **72**, 867-875 (1991).
7. Parrish C.R., Gorham J.R., Schwartz T.M., Carmichael L.E. «Characterization of antigenic variation among mink enteritis virus isolates» *Am. J. Vet. Res.* **45**, 2591-2599 (1984).
8. Sanger F., Nicklen S., Coulson A.R. «DNA sequencing with chain-terminating inhibitors» *Proc Natl Acad Sci U S A.* **74**, 5463-5467 (1977).

AUTHOR INDEX

A

ABRIL J., 44
AFONNIKOV D.A., 145, 149, 160, 207, 229
ALIKINA T.YU., 46
ALSYNBAYEVA L.G., 232
AMIKISHIEV V.G., 85
AN J., 157
ANANKO G.G., 167
ANTONENKO O.V., 109
APWEILER R., 35
ARCHAKOV A.I., 198
ARRIGO P., 155
ASSMUS H., 93

B

BABENKO V.N., 59, 62
BACHINSKY A.G., 181, 188, 194
BAITALYUK M.V., 26
BAZHAN S.I., 125, 188
BELOVA O.E., 125
BERIKOV V.B., 82
BEYLINA A.G., 160
BONDAR A.A., 46
BORK P., 159, 197
BROWN C.M., 142
BUBENSHCHIKOVA E.V., 109

C

CALIFANO A., 27, 216
CARDO P.P., 155
CEBRAT S., 38
CHALEY M.B., 100
CHUZHANOVA N., 79
COLLADO-VIDES J., 104
COOPER D.N., 79

D

DIEMAND A., 180
DUDEK M.R., 38
DZHEL'YADIN T.R., 128

E

EGOROVA A.V., 106
ELLOUMI M., 103
ESIPOVA N.G., 74

F

FLEISCHMANN W., 35
FOKIN O.N., 160

G

GEBAUER-JUNG S., 44
GELFAND M.S., 12, 14, 16, 18, 22, 24, 26, 32, 42, 49, 51,
55, 57
GIERLIK A., 38
GIRAUD M., 61

GLAZKO G.V., 82, 89
GOEBEL U., 113
GRIGOROVICH D.A., 119, 160, 171, 181
GROMIHA M.M., 157
GUEX N., 180
GUIGÒ R., 44
GUSEV V.D., 79

H

HAANS W.J.A., 232
HERZEL H., 93
HIDE W., 59, 72
HONIG B.H., 216

I

IVALDI G., 155
IVANISENKO V.A., 119, 160, 171, 211
IVANOVA N.N., 128

J

JACOBS G.H., 142
JIANGHONG A.N., 152

K

KAMINUMA T., 175
KAMZOLOVA S.G., 128
KATOKHIN A.V., 62
KIELBASA SZ.M., 93
KOCHETOV A.V., 89, 135
KOLCHANOV N.A., 119, 135, 138, 171, 232
KOLESANOVA E.F., 198
KOLPAKOV R., 61
KONDRASHOV A.S., 105
KONO H., 131, 157
KOONIN E.V., 32
KORBEL J.O., 93
KOROTKOV E.V., 100
KOWALCZUK M., 38
KOZHINA E.M., 194
KRAVATSKAYA G.I., 74
KRAVATSKY YU.V., 221
KRAWCZAK M., 79
KUCHEROV G., 61

L

LAVRYUSHEV S.V., 224
LIKHOSHVAI V.A., 65
LOKHOVA I.V., 160
LUKASHEV V.A., 191, 204
LUKASHEV V.V., 204
LUKASHOVA V.V., 191, 204

M

MACKIEWICZ P., 38
MAKEEV V.JU., 96
MAKSYUTOV A.Z., 188
MATUSHKIN YU.G., 65
MATVEEV I.V., 198
MELNIKOV V.N., 232

MILANESI L., 77
MIRONOV A.A., 12, 14, 16, 18, 22, 24, 26, 32, 42, 49, 51,
185
MITCHELL-OLDS T., 113
MITRA, CHANCHAL K, 178
MULDER N.J., 35
MURRAY D., 216

N

NAKANO T., 175
NAKATA K., 175
NEMYTIKOVA L.A., 79
NIZOLENKO L.PH., 181, 194
NOVICHKOV P.S., 18, 26, 42
NOVICHKOVA E.S., 18
NOWICKA A., 38

O

OLENINA L.V., 198
ORLOV YU.L., 69, 211

P

PANINA E.M., 51, 57
PARK S.-J., 201
PEITSCH MANUEL C., 180
PEREZ-RUEDA E., 104
PERMINA E.A., 22
PETRENKO O., 160
PODKOLODNY N.L., 232
POLETAEV A.I., 221
POLOZOV R.V., 128
PONOMARENKO J.V., 224
PONOMARENKO M.P., 224
POROIKOV V.V., 198
POTAPOV V.N., 69, 211
PRABAKARAN P., 131, 157
PTITSYN A., 72

R

RAKHMANINOVA A.B., 49, 185
RAMENSKY V., 159, 197
RAMENSKY V.E., 96
RATNER V.A., 85, 106, 109
RODIONOV D.A., 49
ROGOZIN I.B., 77, 82, 89
ROLA-PLESZCZYNSKI M., 191, 204
ROYTBERG M.A., 96

S

SARAI A., 131, 152, 157
SCHWEDE TORSTEN, 180
SELVARAJ S., 131, 157
SEN, ARUSHARKA, 178
SHABALINA S.A., 133
SIVOZHELEZOV V.S., 128
SOBOLEV B.N., 198
SOROKIN A.A., 128
STANKOVA J., 204
STOCKWELL P.A., 142
SUNYAEV S., 159, 197
SZCZEPANIK D., 38

T

TAKAI T., 175
TITOV I.I., 122, 135, 138
TKACHEV S.E., 234
TRIFONOV E.N., 111
TUMANYAN V.G., 96

U

UEDAIRA H., 157

V

VALISHEV A.I., 232
VALUEV V.P., 160, 164, 229
VASILYEVA L.A., 109
VINOKUROVA N.P., 16
VISHNEVSKY O.V., 62
VITRESCHAK A.G., 55
VOROBIEV D.G., 135, 138, 224

W

WAKO H., 152
WESTON P.S., 98
WIEHE T., 44, 113

Y

YAMAMURA M., 201
YARIGIN A.A., 181, 194
YAROSLAVTSEVA R.G., 232
YUDANIN A.YA., 106

Z

ZELENIN S.M., 46
ZVEREV A.A., 115

KEYWORDS INDEX

A

adaptive evolution, 167
algorithm, 14, 61, 201
algorithm development, 72
alignment, 44, 77, 145
alternative processing, 72
alternative splicing, 59
amino acid characteristics, 145
amino acid sequences, 145, 149, 194, 207
antibodies, 160
antisense, 38
approximation algorithms, 103
AQP4, 46
aquaporin 4, 46
Archaea, 32
ARS elements, 93
attenuators, 55

B

Bacillus subtilis, 22
bacteria, 24, 104
bacterial genome, 16, 26
bacterial genomes, 26
bioinformatics, 180, 229, 232
bioinformatics teaching, 229
biologically active sites of proteins, 171
biosequences, 155
blood coagulation, 167

C

catabolite repression, 18
chromosome sorting, 221
classification of protein topology, 211
clustering, 72
clustering approaches, 103
co-adaptive substitutions, 145, 149, 207
coding probability, 38
codon frequencies, 65
colocalization., 89
comparative analysis, 22, 24, 26, 32
comparative genomics, 51
complement, 167
complete genome, 49, 57, 194
complete genome analysis, 57
complexities, 103
complexity, 61, 69, 79, 211
complexity analysis, 79
composition, 96
compositional asymmetries, 93
computer analysis, 65, 133
computer database, 142
computer model, 106
computer modeling, 204
computer program, 115
computer simulation, 122
computer tool, 35, 142
context, 82
contigs overlaps, 103
curriculum, 232

D

DAHP-synthase, 51
data banks, 194
database, 35, 77, 100, 142, 157, 160, 181
database searches, 77
databases, 119, 171
Delaunay tessellation, 152
dimerization, 191
disease, 155, 159
diseases, 105
DNA, 96, 224
DNA asymmetry, 38
DNA sequence assembly, 103
DNA unwinding, 74
DNA walk, 38
DNA-binding, 149
domain, 35
domains, 96, 216

E

E.coli 16S rRNA, 133
education, 224
efficiency, 65, 135
electrostatic potential, 128
energy landscape, 155
entropy, 185
epitopes, 188
errors correction, 62
EST, 59, 62, 72, 77
estimate, 105
eukaryotes, 14
eukaryotic genes, 42
eukaryotic mRNAs, 135
evolution, 44, 79, 104, 111, 167
evolution of noncoding regions, 113
evolution of patterns, 109
exact algorithms, 103
exon localization, 46
exon-intron structure, 42
exon-intron structure, 42

F

families, 89, 104, 194
family, 35
flow cytofluorimetry, 221
flow karyotype, 221
function, 35
functional sites, 69, 85
functional systems, 167

G

gamma-proteobacteria, 18
GenBank analysis, 100
gene, 46, 82, 234
gene expression, 27, 65, 155
gene number, 38
gene prediction, 44, 77
gene recognition, 14
gene regulation, 24
genetic algorithm, 201

genetic algorithms, 98
genetic and phenotypic variations, 159
genetical texts, 69
genomic DNA, 77
genomic research, 98
genomics, 27

H

heat shock response, 24
helix-turn-helix motif, 104
HIV-1, 125, 188
homeodomain, 149
Homology, 44
hotspot, 82
human chromosome 22, 59
human chromosomes, 221
hypertext, 125

I

immunoglobulin gene, 82
inbreeding, 106, 109
induction of transpositions, 109
information system, 125
internet, 175
Internet, 125
isoforms, 46

K

kinetics, 122

L

large search spaces, 98
latent periodicity regions, 100
leucine zipper motif, 191
ligand binding, 175
local complementarity, 65
local similarity, 188
logical modeling, 115
long range correlation, 38
loop, 185
LPD database, 100

M

Markov chain, 178
Markov models, 69
mathematical model, 65
matrix Fourier analysis, 74
maximal repetitions, 61
membrane targeting, 216
method SYNAP, 115
MGE pattern, 106
microarray data, 27
middle vocational education, 232
mobile genetic elements, 109
mobile genetics elements, 85
modifier, 106
molecular functions, 85
molecular screening, 155
motif, 104, 152, 191
motif discovery, 216
motifs, 181
motifs of functional sites, 85

mRNA, 46, 142
multidrug systems, 49
multiple sequence alignment, 145
mutation, 82
mutations, 105, 155, 159

N

neural networks, 164
new version, 115
nucleotide sequences, 61, 100
numerical simulation, 207

O

object, 115
operons, 55
ORF reconstruction, 62
ORFan, 38
origin of chromosome replication, 74
origin of replication, 93
orthologous genes, 26

P

p53 gene, 82
palindromes, 57
partial correlation, 207
pattern, 106, 109
pattern analysis, 27
pattern of complementarity, 133
patterns, 181, 194
PDB, 119
periodicity, 74
phage display, 160
phylogenetic fingerprinting, 49
phylogenetic footprint, 44, 113
phylogeny, 115
polygenes, 109
polygenic system, 106
polymerase chain reaction, 122
prediction, 42, 44, 77
primary sequence, 178
prokaryotic genomes, 57
profile search, 16, 18, 32
prokaryotic genomes, 57
promoter mutation, 79
promoter shuffling, 79
protein, 35, 152, 178
protein comparison, 194
protein families, 181, 194
protein folding, 185
protein folding & structure, 180
protein models, 216
protein sequence, 175
protein similarity, 16
protein structure, 164, 171, 178, 197, 211
protein structure classes, 164
protein-DNA complex, 131
protein-nucleic acid interaction, 157
proteins, 111, 119, 159, 171

R

random walk, 38
receptor, 175, 191, 204
recognition, 14, 62, 128, 224

regression analysis, 82
regularity, 178
regulation, 24, 32, 51, 142
regulatory regions, 113
regulatory signal, 16, 18, 22
regulatory signals, 224
regulons, 32
renaturation, 122
Rebase Update, 89
repeat, 35
repeated elements, 77
repetitive DNA families, 89
replication, 74, 93
repressor, 22
restriction-modification systems, 57
RNA secondary structure, 55, 138
RNA-polymerase, 128
RNA-RNA interactions, 133

S

S/MARs, 89
Saccharomyces cerevisiae, 38, 93
secondary protein structure, 211
secondary structure, 135, 138
selection, 106, 109
sequence, 72, 82, 96, 103, 145, 175, 178
sequence analysis, 178
sequence similarity, 72, 175
sequence-property-mapping, 113
sequencing errors, 14
similarity search, 181
single nucleotide polymorphisms, 197
SNP, 159
SOS repair, 22
spectral decomposition, 221
splice detection methods, 59
spliced alignment algorithm, 14
SRS, 119
statistical analysis, 135, 138
stochastic complexity, 69, 211

stress, 109
structural alignment, 201
structural motif, 152
structural patterns, 216
supercomputing, 72
synthetic peptides, 160

T

T4 phage DNA promoter, 128
tandem repeats, 61
target prediction, 131
tertiary protein structure, 171
thermodynamic data, 157
three-dimensional structure, 152
transcription factor, 104
transcription factors, 131
transcription regulation, 32, 104
translation, 138
translation efficiency, 135
translation initiation, 138
translational regulation, 142
tree of similarity, 106

U

untranslated regions, 142

V

vaccine, 125
vector, 115
vertebrate growth hormone genes, 79

W

world wide web, 175
WWW computer tool, 142