Информационные процессы, Том 4, № 3, 2004, стр. 241-246.

Note on Cliques and Alignments

V.A. Lyubetsky, A.V. Seliverstov

*Institute for Information Transmission Problems RAS,
127994, Russia, Moscow, Bol'shoi Karetny per., 19,
e-mail: lyubetsk@iitp.ru, slvstv@iitp.ru*
Received July, 5, 2004

Abstract—We consider finding and counting small cliques in any graph. This algorithm is applied to detect conservative anchor motifs for use in multiple alignment. In this way the transcriptional termination control systems of some operons are investigated. Putative regulation structures in amino acid biosynthesis in actinobacteria are identified. Moreover, we present some statistic data for such structures in proteobacteria.

1. INTRODUCTION

A graph Γ is said to be *multipartite* if the vertex set is partitioned into *parts* such that there is no edge between any pair of vertices from the same part. A complete subgraph with exactly n vertices is called a n*-clique*.

Any algorithm for finding cliques allow to solve tasks of alignment of sequences, i.e. search of a set of similar words, one word in each sequence. Such set of words frequently name as a signal. The search of a signal is interesting to biology, for example, for alignment of nucleotide sequences. In this task for n given sequences in the alphabet $\{a, c, g, t\}$ is constructed an n-partite graph G, which vertices are labelled by the words from these sequences of fixed length. Two vertices are joined by an edge in the graph G, if they are words from different sequences and are similar against each other more some fixed threshold. Any system of similar word pairs (with one word in each of n of sequences) corresponds to a n-clique in the graph G.

It is to be kept in mind that the conventional attenuation regulation of amino acid biosynthesis implies presence of a leader peptide bearing regulatory codons, an antiterminator hairpin and a terminator hairpin that have a common half-stem. Refer to [**?**] as well as [**?**] for details.

Discovery and analysis of the regulation of gene expression in bacteria is currently an active field of research within the general framework of studying RNA-based regulation strategies. However, finding alternative regulation systems is still a long way even in bacteria. Further, understanding the attenuation regulation machinery is crucial for developing algorithmic tools of mass attenuation detection and deriving a descriptive attenuation model.

The main method for theoretical research is studying of multiple alignment of untranslated sequences before homologous genes. Conservative words or hairpins seems to be significant for regulation of gene expression.

Identifying relevant attenuation characteristics to study is by itself a separate task.

As an illustration we have considered four amino acids (triptophan, isoleucin, leucin, valin). There are not classical attenuation schema for $leuA$. At least we establish some conjecture about regulation of amino acid synthesis in actinobacteria. Furthermore, the conservative RNA structures are useful for verification of gene start.

2. FINDING AND COUNTING SMALL CLIQUES

We let Γ stand for an undirected graph with V vertices and E edges. Obviously

$$E \leq \frac{V(V-1)}{2}.$$

The graph Γ is defined by its adjacency matrix. That is a symmetrical order V 0-1 matrix A with zeroes on the principal diagonal.

The arithmetic complexity of matrix multiplication is $O(V^{\omega})$, where exponent $\omega < 2.376$. Recall that for the straightforward method $\omega = 3$ and for the Strassen algorithm $\omega = \log_2 7$.

For any two order V matrices A and B the Hadamard product $A * B$ is an order V matrix of item product

$$(A * B)_{ij} = A_{ij}B_{ij}.$$

Let us consider the Hadamard product $A * A^2$. For any $i \neq j$ its element $(A * A^2)_{ij}$ is equal to the number of triangles containing both i^{th} and j^{th} vertices.

The number of the triangles containing i^{th} vertex is equal to the matrix element $(A^3)_{ii}$. Moreover, the arithmetic complexity of counting triangles in the graph Γ is $O(V^{\omega})$, where $\omega < 2.376$ is the exponent of matrix multiplication. On the other hand the triangle enumeration can be done in $O(VE)$ time.

Any tetrahedron (i.e. 4-clique) is defined by a pair of disjoint edges. Thus, the tetrahedron enumeration can be done in $O(E^2)$ time.

Let us construct the graph Γ' by a following way. Vertices of the graph Γ' correspond to edges of the graph Γ. Two vertices of the graph Γ' are adjacent ones if related edges of the graph Γ are two disjoint edges of a tetrahedron. Any triangle of the graph Γ' corresponds to a 6-clique of the graph Γ. Thus, for counting 6-cliques, we get the bound of $O(E^{\omega})$ arithmetic operations, where $\omega < 2.376$ is the exponent of matrix multiplication.

Furthermore, for $n > 6$ we have the fast algorithm for dense subgraph finding [**?**]. See also [**?**].

3. APPLICATION: PUTATIVE REGULATION STRUCTURES IN ACTINOBACTERIA

These 12 samples are examined: *Corynebacterium diphtheriae* NCTC 13129, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum* ATCC 13032, *Kineococcus radiotolerans* SRS30216 Krad_60, *Mycobacterium avium* subsp. paratuberculosis str. k10, *Mycobacterium bovis* subsp. bovis AF2122/97, *Mycobacterium leprae* strain TN, *Mycobacterium marinum* M (unfinished genom), *Mycobacterium tuberculosis* H37Rv, *Streptomyces avermitilis* MA-4680, *Streptomyces coelicolor* A3(2) and *Thermobifida fusca* Tfus_36. All nucleotide sequences are obtained from NCBI
http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi.

Tryptophan. Let us consider genes involved in tryptophan biosynthesis or utilization.

- *trpA* tryptophan synthase alpha chain;
- *trpB* tryptophan synthase beta chain;
- *trpC* indole-3-glycerol phosphate synthase;
- *trpD* anthranilate phosphoribosyltransferase;
- *trpE* anthranilate synthase component I;
- *trpG* anthranilate synthase component II;
- *trpS* tryptophanyl-tRNA synthetase.

In *Corynebacterium* many genes comprise one operon. Let us consider $5'$-untranslated leaders. We found 7 examples for attenuation regulation. Each putative leader peptide have either the codon gtg (Valin) or the codon atg (Methionin) as the start codon.

1. *C. diphtheriae* $trpB_1EGDC_1$ MetAsnAlaHisAsn**TrpTrpTrp**ArgAla
2. *C. diphtheriae* $trpB_2A$ MetAsnAlaAlaPheLysPhe**TrpTrp**ArgAla
3. *C. efficiens* $trpEGDCBA$ ValAsnAsnPheCysGlnSerGlnGlyThrGln**TrpTrpTrp**ArgAlaArg
4. *C. glutamicum* $trpEGDCBA$ ValAsnAsnSerCysLeuSerGlnSerThrGln**TrpTrpTrp**ArgAlaAsn
5. *S. avermitilis* $trpE_1$ MetPheAlaHisSerIleGlnAsn**TrpTrpTrp**ThrAlaHisProAlaAlaHis
6. *S. avermitilis* $trpS_2$ MetThrThrArgThrCysThrGlnGln**TrpTrp**AlaAla
7. *S. coelicolor* $trpE_2$ MetPheAlaHisSerThrArgAsn**TrpTrpTrp**ThrAlaHisProAlaAlaHis

By means finding cliques conservative fragments are computed. These fragments approximately correspond to putative terminator hairpins with poly-U and antiterminator hairpins. Our alignment reveals that both the antiterminator and terminator half-stems are highly conservative in the organisms studied.

The distance between the Trp-codons and the antiterminator left half-stem varies between 9 and 12 bases. For γ-proteobacteria *trp* operons its value varies between 14 and 17. For α-proteobacteria — between 9 and 17. The terminators in all operons have rich-G right half-stem. This is true for γ-proteobacteria too.

There are **antiterminators** for the *trp* operons. The half-stems are underlined.

1. tggtggtggcgcgct**taa**cc.gcgggcc.gtttt...cacgcattcatttc...........aac..aggctcgcc
2. ttctggtggcgcgcc**tag**caggcgggccccttttgtgtgagcattcaccacacaactttggaaacacaagcccgcg
3. tggtggtggcgcgctaga**taa**gcgggcccacggatcaccaagttgttttcacactgaagattt...caaggctcgtg
4. tggtggtggcgcgctaac**taa**gcgagcctgacacctcaagttgttttcactt..tgatgaatttttaaggctcgta
5. tggtggtggaccgctcatccggcg.gcccac**tga**.........ctgcgcgt.acgcaagacttcgcgaaggc.cgccc
6. cagtggtgggccgcc**tga**.cggcg.gccgtacacacgtatgtactc...................aacggc.cgccgc
7. tggtggtggaccgctcacccggcg.gcccac**tga**.........ctgcgcgcgactcaagac.tcgcgaaggc.cgccc

There are **terminators** for the *trp* operons. The terminator half-stems are set in uppercase and the right-hand antiternimator half-stems are underlined.

1. ....aac..AGGCTCGCCTTGTcca....AC.AAGcaGCGGGCCTttttgttagc
2. ttttggaaacacAAGCCCGCGtat..............C.GCGGGC.TTtttcgtatat
3. aagattt...cAAGGCTCGTgtaCTTCGTtcgACGAAGcaGCGGGCCTTtgtggttca
4. atgaatttttAAGGCTCGT..aCTTCGTtcgACGAAGaaGCGGGCCTTtgtggttttt
5. aagacttcgcgAAGGCCGCCC...............gagGGGCGGCCTTtcgtgtttccg
6. ....AACGGCCGCCGcct.............CGGCGGCCGTTctcgtttctc
7. aagac.tcgCGAAGGCCGCCC...............gagGGGCGGCCTTCGgtgttttcg

ilv operon. In many actinobacteria genes $ilvB$ (acetolactate synthase large subunit), $ilvH$ (acetolactate synthase small subunit) and $ilvC$ (ketol-acid reductoisomerase) comprise a single operon. There are leader peptides and conservative terminators with poly-U at right end. The terminators in all operons have rich-G right half-stem. There are **leader peptides** for the ilv operon.

1. *C. diphtheriae* MetAsn**IleIle**Arg**LeuValValIle**ThrThrArgArg**Leu**Pro
2. *C. efficiens* MetThrSer**Ile**ArgPro**ValValIleVal**AlaAlaArgArg**Leu**Pro
3. *C. glutamicum* MetThr**IleIle**Arg**LeuValValVal**ThrAlaArgArg**Leu**Pro
4. *M. tuberculosis* Met**LeuValValIle**GlyArgArg**Leu**GlyAla
5. *M. bovis* Met**LeuValValIle**GlyArgArg**Leu**GlyAla
6. *M. leprae* Met**LeuValValIle**CysGlnArg**Leu**GlyGly
7. *M. avium* Met**LeuValValIle**ArgArg**Leu**GlyAla
8. *M. marinum* MetAspThrAlaGlyThrProGlyLys**LeuValValLeu**GlyArgArg**LeuVal**Ala
9. *S. avermitilis* MetArgThrArg**IleLeuValLeu**GlyLysArg**Leu**Gly

10. *S. coelicolor* MetArgThrArg**IleLeuValLeu**GlyLysArg**Leu**Gly

Terminators for the *ilv* operon. The terminator half-stems are set in uppercase. Also distance SU between the initial position of the leader peptide stop codon and the beginning of the poly-U, loop length L of terminator hairpins and the ratio $\frac{G}{G+C}100$ are computed.

Bacteria	Terminator hairpin and poly-U	SU	L	$\frac{G100}{G+C}$
C. diphtheriae	`cgaaaagcGCCCTCGaCAGCAccacacaTGCTGag.CGGGGGCtttccttat`	62	7	88
C. efficiens	`caagcGCCCTCGACAGTACccaccacaGTGCTGttTCGAGGGCtttgttgt.`	69	8	70
C. glutamicum	`caagcGCCCTCGaCAACACTcaccacAGTGTTGgaaCGAGGGCtttcttgtt`	67	6	80
M. tuberculosis	`ccaacgcgACCCTCGtgCAGCagctgaGCTGg..CGAGGG.Tttttctt`	57	6	77
M. bovis	`ccaacgcgACCCTCGtgCAGCagctgaGCTGg..CGAGGG.Tttttctt`	57	6	77
M. leprae	`ccaacgcgcAACCCTCGtgCAGCTagtcAGCTG..tCGAGGG.TTttttgtt`	74	4	75
M. avium	`ccaacgcgcAACCCTCGtgCAGCacaaGCTG..tCGGGGG.TTttttgtt`	72	4	77
M. marinum	`ccaacgcgcAACCCTCGTgCAGCagctgaGCTG..ACGGGGG.TTttttgtt`	59	6	77
S. avermitilis	`cccggcgcgctCCCCTCGctTGCCtcacGGCACGAGGGGtttttgtt`	84	4	77
S. coelicolor	`ccgacgcgctCCCCTCGctTGCCttacGGCACGAGGGGtttttgtt`	84	4	77

Earlier results for proteobacteria. These data for proteobacteria are originally from [?]. The third column contains distances SU between the initial position of the leader peptide stop codon and the beginning of the poly-U for terminator. The next column contains loop length L of terminator hairpins. The fifth and sixth columns contain the amount of G and C bases in the right half-stem of the terminator. The distance between the antiterminator left half-stem and the stop codon varies between -8 (stop codon to the left of the antiterminator) and 33 (stop codon in the middle of the antiterminator loop).

Subdivision	Operon	SU	L	G	C	$\frac{G}{G+C}100$
α	*ilvIH*	51-55	4-7	2-5	1-3	40-66
α	*trp(E/G)*	52-72	3-10	4-6	2-5	44-66
γ	*ilvBN*	53-57	4-6	6-7	3	66-70
γ	*ilvGMEDA*	37-64	4-6	5-8	0-3	66-100
γ	*leuABCD*	42-69	3-7	4-5	1-3	64-83
γ	*thrABC*	46-62	3-8	3-7	1-3	50-88
γ	*his*	90-113	3-7	2-5	1-4	50-83
γ	*trp*	44-73	4-8	3-4	1-2	60-80
γ	*pheA*	61-72	3-5	4-6	1-2	71-86
γ	*pheST*	68-69	3-6	4-5	1	80-83

Gene $leuA$ (2-isopropylmalate synthase). We suppose that the start codon of the gene $leuA$ in *Corynebacterium efficiens* is located on 35 codons more to the left, than is specified in the base NCBI. Besides the start codon of a gene $leuA$ in *Thermobifida fusca* is located on 199 codons more to the right, than is specified in NCBI. The appropriate sequences for *M. bovis*, *M. tuberculosis* H37Rv and *M. tuberculosis* CDC1551 completely coincide.

There are putative **leader peptides**.

1. *C. diphtheriae* MetAsnArgAlaAsn**LeuLeuLeuLeu**ArgArgGlyGlySerGlnAla
2. *C. efficiens* MetPheSerSerHisGluArgSerAla**LeuLeuLeu**ArgArgGlyGlySerGlnArgSer
3. *C. glutamicum* MetThrSerArgAlaAsn**LeuLeuLeuLeu**ArgArgGlyGlySerGlnArgSer
4. *K. radiotolerans* ValAlaArg**Leu**GluAsn**LeuLeuLeu**Arg ArgArgGlyGlyAlaSer
5. *M. avium* ValGlnArgVal**LeuLeuLeu**GlyArgArgAspGlyVal
6. *M. bovis* Val**Leu**HisValGlnArgVal**LeuLeuLeu**GlyArgArgAspGlyVal
7. *M. leprae* ValGlnArgVal**LeuLeuLeu**GlyArgArgAspGlyAla
8. *M. marinum* ValGlnArgVal**LeuLeuLeu**GlyArgArgAspGlyAla
9. *S. avermitilis* MetArgPheGly**LeuLeuLeuLeu**SerCysArgGlyGluGly**Leu**
10. *S. coelicolor* MetArgPheGly**LeuLeuLeuLeu**SerCysArgGlyGluGly**Leu**

11. *T. fusca* Met**Leu**ArgGlu**LeuLeuLeuLeu**SerGlyArgGlyGlyGlyArg

All examined sequences are directly ahead of start codon (gtg for *Mycobacterium* and atg in other cases) of the gene $leuA$.

```
 1. cttcTCCTTCtt......cgccgcggcgggtcaca..ggcttaacgtcccttac
 2. GCTCtTCTTct......TCGccgcggcgggtcccagaggtcataa.........
 3. ctactTCTTCTT......cgccgcggcgggtcccagaggtcttaa.........
 4. aacCTCCTCcTTC...gtcgccgcggcggg........gccag...........
 5. cgggTGCTCCtcctcggacgccgcgacggg........gtctgatt........
 6. cgggTGCTCCtcctcggacgccgcgacggg........gtctgat.........
 7. caggtaCTCCtcctcgaacgccgcgacggg........gtctgat.........
 8. cgggTGCTCCtcctcggacgccgcgacggg........gcctgat.........
 9. gggctGCTCCTCcttagctgccgcggcgag.......ggcctgtaag.......
10. GGGCTgCTTCtccttagctgccgcggcgag.......ggcctgtag........
11. gagctGCTCCTGcttagcggccgcggcggg.......ggccgataa........
```

```
 1. acacagccggctc.cccgtcgcggagttct......agtgtagccggctg....
 2. ..gcgaccggcac.cccgtcgcggagttt........gtgttgccggtcgtgaa
 3. .cacgaccggcat.cccgtcgcggagttt.......ggtgttgccggtcgtg..
 4. .ctaggccggtctcccgtcgcgggacctcgtcgt..gcg.cgccggcc.....
 5. .ccagaccggctt.cccgtcgcgggt.gttcgc...gatg.cgccggtctg...
 6. .ccagaccggctt.cccgtcgcgggacgttcgc...gatg.cgccggtctg...
 7. cccagaccggctg.cccgttgtggaa.gttcact...atg.cgccggtctg...
 8. .ccagaccggctt.cccgtcgcggg.tgttcg...cgatg.cgccggtctgaag
 9. cagaggccgaccccctccccgcgg..agtctggcgttgcgccgtcggccg....
10. ...aggccgactccctccccgcgg..agcttggt..ggtgccgtcggccgtcct
11. ...gggccggctccctcgccgcggaggttcgacctgtctgctgtcggccg....
```

```
 1. .................................caacaagaacccacgtGAAGGAaactacca
 2. cccgcaacagcgctagagtttgattccagaaaacaagcgcacactccaCGAAAGAtGAGCacccatc
 3. ............................gacccacccaaactttttAAGAAGGttgaacaca
 4. ......................gccgcaccagccgctgaagaccgcGAAcGAGGAGaacgaa
 5. ................................aggttccttctgatatccccGGAGCAatcacc
 6. ................................aggttccttctcaccatcccGGAGCAactacc
 7. ................................aggttccttctcacatc.ccGGAGcaattatt
 8. ..................................ttccttctcgccacccccGGAGCAactacc
 9. ..................................tccttccggacaccacGAGGAGCccacgcatc
10. ..tccggacacgcggacgacgcggacaccgccgagatccgcggacatcacGAGGAGCCCacgccatc
11. ...........................cacgaccgcaagaaaaagtctcaCGGGAGCgtattcac
```

Conclusion. There is a pseudoknot with conservative half-stems. Moreover, the right half-stem overlap the ribosome-binding site. This structure can repress translation. See for example [**?**].

Acknowledgements. We wish to thank M.S. Gelfand, K.Yu. Gorbunov and A.A. Mironov for critical comments on this article.

REFERENCES

1. Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A., Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis, (2004), *FEMS Microbiology Letters*, (2004), 234, 2, pp. 357-370.

2. Singer M., Berg P. *Genes and genomes*, (1991), University Science Books, Mill Valley, California.

3. Lyubetsky V.A., Seliverstov A.V., Selected algorithms related with finite groups. *Information processes*, (2003), 3, 1, pp. 39-46.

4. Gorbunov K.Yu., Mironov A.A., Lyubetsky V.A. Search for Conserved Secondary Structures of RNA, *Molecular Biology*, (2003), 37, 5, pp. 723-732.

5. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. Riboswitches: the oldest mechanism for the regulation of gene expression? *TRENDS in Genetics* (2004), 20, 1 January.