COMPARATIVE GENOMICS AND EVOLUTION OF BACTERIAL REGULATORY SYSTEMS

M.S. Gelfand[1, 2, 3*], A.V. Gerasimova[2], E.A. Kotelnikova[2], O.N. Laikova[2], V.Y. Makeev[2], A.A. Mironov[2, 3], E.M. Panina[1], D.A. Ravcheev[1, 3], D.A. Rodionov[1], A.G. Vitreschak[1]

[1] Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoy Karetny pereulok 19, Moscow, 127994, Russia, e-mail: gelfand@iitp.ru; [2] State Scientific Center GosNIIGenetika, 1-j Dorozhny proezd 1, Moscow, 117545, Russia; [3] Department of Bioengineering and Bioinformatics, Moscow State University, Vorobievy Gory, 1-73, Moscow, 119992, Russia

* *Corresponding author*

Abstract: Recent advances in genome sequencing and development of comparative genomics techniques allow one to study evolution of regulation in prokaryotes at several different levels: microevolution of orthologous regulatory sites, changes in regulon content, evolution of interacting regulatory systems, and co-evolution of transcription factors and their binding signals. Regulatory interactions appear to be very dynamic in some cases and surprisingly stable in others. The review presents several examples where comparative analysis uncovered plausible scenarios of evolution of regulatory systems.

Key words: regulation; evolution; transcription; binding site; riboswitch

1. INTRODUCTION

Comparative analysis of regulatory signals is a powerful tool for functional annotation of genomes (Gelfand, 1999). Based on the assumption of conservation of regulatory interactions, it uses two related but somewhat different approaches.

Phylogenetic footprinting assumes that regulatory sites evolve slower than the surrounding non-coding sequences and thus are seen as conservation islands in alignments of intergenic regions. The term was introduced in analysis of eukaryotes (Gumucio et al., 1992), where it is a much-used technique (Frazer et al., 2004) that sometimes even provides motivation for sequencing of genomes (Boffelli et al., 2003; Cliften et al., 2003; Kellis et al., 2003; Thomas et al., 2003). At the same time, until lately it had not been applied to the analysis of bacterial genomes, as no genomes at the suitable evolutionary distances were available. For several years, the only group allowing for such analysis was enterobacteria (Florea et al., 2003).

A related approach is *consistency filtering* of candidate sites that has been successfully applied to many bacterial regulatory systems (Gelfand et al., 2000). Despite the fact that in most cases it is impossible to construct a reliable recognition rule for transcription factor binding sites, simultaneous analysis of multiple genomes allows one to retain only the sites occurring upstream of orthologous genes (and thus, likely to be true). The false positives are scattered at random and thus can be ignored. This approach was applied to the analysis of many diverse systems, and allowed us to make a number of functional predictions that were subsequently confirmed in experiment (Rodionov et al., 2000, 2002 a, b; Makarova et al., 2001; Panina et al., 2001, 2003a, b; etc.).

However, this approach allows one to find only the conserved regulon cores retained at relatively large evolutionary distances. Sequencing of numerous genomes uniformly spanning the evolutionary space made it possible to study the taxon-specific regulation and evolution of regulatory systems (Gelfand and Laikova, 2003). In particular, availability of many closely related genomes allowed for the use of *phylogenetic shadowing* (Boffelli et al., 2003) for identification of prokaryotic regulatory sites that look like conservation islands in multiple alignments (Figure 1).

Evolution of regulatory sites has several aspects:

1. Evolution of sites regulating expression of orthologous genes;
2. Co-evolution of transcription factors and their binding signals;
3. Evolution of *regulons*, that is, sets of co-regulated genes; and
4. Evolution of interacting systems.

We cannot yet suggest a uniform theory, or even drafts of a theory; however, there exist a number of non-trivial observations that can serve as a raw material for creating such a theory.

Orthologous sites: unexpected conservation of non-consensus nucleotides. The traditional view on non-consensus nucleotides in transcription factor binding sites is that they represent random noise tolerated while the deviations from the consensus pass some threshold (Berg and von Hippel, 1988). A more complicated theory is that deviations from the consensus allow for activation or repression to occur at a fixed, gene-dependent level.

```
EC      AAA-GAGAAAAAAGCAGCAAACTTCGGTTGAAAAAGCCGCTATGATCGCCGGATAATCGTTTGCTTTTTTTA---
ST      AAA-GCATAAAAAGCGGCAAAGTTCAGTTGAAAAAGCGTTGATGATCGCTGGATAATCGTTTGCTTTTTTTTG--
YP      AAATGTATTAAATGTCGCATTCGGGTGTTGATTAGTCACCACTGATGGCTAGATAATCGTTTGCCTTAAATGACA
        *** *    *** *  ***       *****  *  *     **** **  ************* **   *

EC      -CCACCC--------GTTTTGT--------ATGCGCG----GAGCTAAACGTTTGCTTTTTTGCGACGCAGCA-A
ST      -CCACCC--------GTTTTGT--------ATACGTG----GAGCTAAACGTTTGCTTTTTTGCGGCGCCCCG-G
YP      TCTGCCCTAAACTTCGATTTTTTTTCAGTCATGCGTTCTCCCAGCTAATCGTTTGCTATTTTTCCCCGCTCTATG
         *  ***        * *** *        ** **       ****** ******** **** *  ***

EC      ATTGTCGCAAACCTGGA----------GCAGGAA-GATAACGTTTCGCTGGCAGGGGATTGTCCGCCACGCATCT
ST      -TTGTCAGTAATGTAGC----------ACAAGGA-GATAACGTTGCGCTGTTAGTGGATTACCTCCCACGTATAC
YP      AGTCAGGGAGAGTTAGTGAGTTCATCGACAGGAACGGAAACGATTACGTAGAGAAGGGCGCTTGGCTTGGCATGA
          *       *  * *            ** * * *  **** *    *      **        *   * **

EC      TGACGAAAATTAAACTCTCAGGGGATGTTTTCTTATGTCT------ACGCCATCAGCGCGTACCGGCGGTTCACT
ST      CGACGAATAATAAATTCTCAGGGGATGTTTTCT-ATGTCT------ACGCCTTCAGCGCGTACCGGCGGTTCACT
YP      CTATTTTAAATGA-CACACAGGGGACATCACC--ATGTCTAGCAGCAACCCTCAAGCACAGCCAAAGGGCACGCT
          *     * *   * *******  *   *  ******      *  **   *** *   *    **  * ** *
```

```
                                                            -35box
                         [------FNR-------]             [---===FNR----[===]-NrdR---
EC      CCGTACGCTCTGCTTTTTACTTTGAGCTACATCAAAAAAAGCTCAAACATCCTTGATGCAAAGCACTATATATAG
ST      CTGTACGCTCTGATTTTTACCTTGTTCTACATCAATAAAATTGCAAACATCCTTGATGCAAATCACTACATATAG
KP      CCGTACTCTCACCTTTTTACCTTGTTCTGGGTCAATAAAATCGCAAACATCTTTGATGCAAATCACTACATATAG
        * **** ***   ******* ***  **   **** ****   ******** ********** ***** ******
        --] -10box       >[-----NrdR-----]
EC      ACTTTAAAATGCGTCCCAACCCAATATGTTGTATTAATCGACTATAATTGCTACTACAGCTCCCCACG--AAAAA
ST      ACTTTAAAATGCACGCCGACCCAATATGTTGTATTAATTGACTACAATTGCTACAACACCTGTTCACT--CGACA
KP      AACTTAAAATGCGCCTCGGCCCAACATATTGTATTAATCGTCTATTAT-GTCACCATATCTTGTCGATGTCTGGC
        *  *********    *  ***** ** ********** * ***  ** *  ** * * **   *
                  [-DnaA--]
EC      GGTGCGGCGTTGTGGATAAGC-GGATGGCGATTGCGGA-AAGCACCGGAAAACGAAACGAAAAAACCGGAAAACG
ST      CAAGGTGAATTGTGGATAACCTGGGTCAGGATTGCGGG-AAGTCATTGGAAAAGAGATGAATAAACCTGTTA-TG
KP      GGTGATGAGATGTGGATAAAACGGGCCGGATCCGAAGGTAAACAGCACGAGCCGTAGCGTGCAGCGCCTTCG-GG
           *  *   *********   **    *    *  *  **        *   *    *   *   *       *
                      [-DnaA--]
EC      CCTTTCCCAATTTCTGTGGATAACCTGTTCTTAAAAATATGGAGCGATCATGACACCGCATGTGATGAAACGAGA
ST      GCTTCCCCGGCCTCTGTGGATAACCTGTTCTTACAAATATGGAGTGATCATGACACCGCATGTGATGAAACGAGA
KP      ATAACCTCCGCCTCTGTGGATAACCTGTTCT---ATATATGGAGTGATCATGACACCGCATGTGATGAAACGTGA
             * *    *******************   * ******** *************************** **
```

Figure -1. Phylogenetic shadowing. Binding sites are set in boldface; promoter boxes, Shine–Dalgarno boxes; and genes, in italics. EC: *E. coli*, ST: *Salmonella typhimurium*, KP: *Klebsiella pneumoniae*, YP: *Yersinia pestis*. (Top) PurR binding sites upstream of *yjcD* genes in enterobacteria look like conserved islands. (Bottom) Multiple overlapping FNR, DnaA, and NrdR binding sites in the regulatory region of *nrdR* gene of *E. coli* and close relatives. Overlapping sites are shown by '='. Transcription start is marked by '>'.

However, analysis of binding sites regulating expression of orthologous genes demonstrated unexpectedly high conservation of non-consensus nucleotides (examples are shown in Table 1). Note that deviations from the consensus occur at different positions and thus cannot be explained by erroneous assignment of consensus nucleotides.

The simplest explanation for this phenomenon could be that insufficient time has passed for mutations that would revert a non-consensus position to the consensus state or change a non-consensus nucleotide to another non-consensus one. Indeed, if one considers very close genomes, e.g., different strains of the same species, coincidence of non-consensus nucleotides would be absolutely natural. However, statistical analysis demonstrated that the observed degree of conservation is much higher than the one expected under a neutral model (Kotelnikova et al., 2005).

This phenomenon was analyzed in two ways. Firstly, the degree of conservation in non-consensus positions was shown to be much higher than conservation in synonymous codon positions assumed the best available approximation to the neutrally evolving DNA. Secondly, ANOVA analysis demonstrated that dependence of the non-consensus nucleotide on the orthologous row of genes is higher than the dependence on the genome.

Table -1. Orthologous sites with conserved non-consensus nucleotides

Genome	Binding site	
	PurR site upstream of purL	PurR site upstream of purM
Escherichia coli	ACGCAAACG**g**TT**t**CGT	**t**CGCAAACGTTTGC**t**T
Salmonella typhi	ACGCAAACG**g**TT**t**CGT	**t**CGCAAACGTTTGC**t**T
Yersinia pestis	ACGCAAACG**g**TT**t**CGT	**t**CGCAAACGTTTGC**c**T
Haemophilus influenzae	AtGCAAACGTTTGCtT	**t**CGCAAACGTTTGC**t**T
Pasteurella multocida	ACGCAAACGTTT**t**CGT	**t**CGCAAACGTTTGC**t**T
Vibrio cholerae	ACGCAAACG**g**TTGCtT	ACGCAAACGTTTtC**c**T

Non-consensus nucleotides are shown by lower case boldface symbols; conserved non-consensus positions are underlined.

One possible explanation for this phenomenon is the following. Recent experimental studies demonstrated that the speed of gene activation depends on gene position in a metabolic pathway (Zaslaver et al., 2004).

Thus, the binding sites perform a fine-tuning of the regulation level by maintaining the gene-specific binding constant of the transcription factor. The latter depends on the site sequence and, in particular, on nucleotides in non-consensus positions. Thus, these positions are not neutral, and evolution of each particular site follows a rather narrow path dependent on the gene position in the metabolic pathway.

Regulons: plasticity of content. Comparison of even very close genomes demonstrates that point mutations can destroy a site and thus release a gene from regulation (Figure 2). Analysis of bacterial regulatory systems demonstrates that, beside the conserved core, many regulons contain taxon-specific members. A regulated gene may be genome-specific and absent in related genomes, or be released from regulation by a given factor.

An example is provided by the NadR regulon in enterobacteria. It is well studied in *E. coli*, where it includes the main NAD-synthesis genes—*nadA*, *nadB*, and *pncB*. However, even in very close genomes of *Yersinia* and *Erwinia* spp., the candidate binding sites are observed only upstream of *nadA*, but not other genes. On the other hand, these genomes have a conserved NadR-binding site upstream of the *nadR* gene itself, so that the latter is autoregulated. Thus, even relatively simple regulons covering essential metabolic pathways may be quite flexible.

In taxonomic groups evenly covered by sequenced genomes, one can study evolution of regulons in detail. Other interesting examples are the fructose, ribose, and purine regulons.

```
Consensus                         ttGtACAagttaactaGTacaa
Escherichia coli          gtcgccgaATGTACTAGAGAACTAGTGCATtagcttat
Salmonella typhimurium    accgcaggATGTACTAGTAAACTAGTTTAAtggattgg
Yersinia pestis           gtcgtcggATGTTTTAACTAAATATTTTCAtgagtgat
Erwinia chrysanthemi      ctcgccgcATGTACTGATGGGTAACCGGCGctgaactg
Conserved positions        •++••+ ++++••+•   •• •+ •    • • •
```

Figure -2. Degeneration of TrpR binding site upstream of the trpH gene. The site region is set in capitals; functional sites, in boldface; and non-consensus nucleotides are underlined.

The fructose repressor FruR is a global regulator of the *E. coli* metabolism (Ramseier et al., 1995). However, in Vibrionaceae and Pasteurellaceae, it regulates only transport and metabolism of fructose. Preliminary analysis shows that expansion of the regulon occurred in the *E. coli* lineage.

A slightly more complicated story is that of purine and ribose repressors. The common ancestor of gamma-proteobacteria contained a ribose repressor that regulated the ribose catabolism operon; this state was retained in Pseudomonadaceae. Somewhere along the branch leading to Enterobacteriaceae, Vibrionaceae, and Pasteurellaceae, this repressor was duplicated. One copy (RbsR) retained the specificity towards ribose, but its DNA binding signal has changed. The other copy retained the signal, but changed the specificity, becoming the repressor of purine biosynthesis genes, PurR. Analysis of genomes from the latter three families demonstrated a gradual sliding of the regulon on the metabolic map (Ravcheev et al., 2002).

Interacting regulatory systems. Although regulation in bacteria is simpler than in eukaryotes, many genes are regulated by several factors and thus belong to several regulons simultaneously. In particular, this is a common feature of genes encoding enzymes belonging to several metabolic pathways (we do not discuss here a very non-trivial question of what set of reactions may constitute a pathway).

A somewhat more interesting situation occurs when one functional system is controlled by several regulators. Sometimes these regulators act independently, e.g., tryptophan attenuator and repressor TrpR of *E. coli*. In other cases, a complex functional system uses several regulators responding to different external stimuli.

One of examples of the latter kind is regulation of respiration in *E. coli* involving aerobic/anaerobic switch FNR, two-component regulator ArcAB, and nitrate/nitrite regulators NarPQ/NarLX. These regulators form different cascades in different genomes; orthologous and non-homologous isofunctional operons are regulated by different factors in different genomes (Table 2; Gerasimova et al., 2004). Similar observations were made for heat shock regulators in beta- and gamma-proteobacteria, regulated by specific sigma-factor σ^H and repressor HrcA (Permina and Gelfand, 2003b).

Table -2. Regulation of respiration in gamma-proteobacteria

Regulated gene	Regulator		
	FNR	ArcA	NarPQ/LX
fnr	E P V	- P V	- P -
arcAB	- E -	- E -	- - -
narL/narP	E P V	- - V	- - V
Escherichia coli (nuo)	FNR	ArcA	NarL
Yersinia pestis	FNR	ArcA	—
Yersinia entercolitica	FNR	—	—
Pasteurella multocida	FNR	ArcA	NarP
Actinobacillus actinomycetemcomitans	—	ArcA	NarP
Haemophilus influenzae	FNR	ArcA	—
Haemophilus ducreyi	FNR	ArcA	NarP
Vibrio vulnificus	—	ArcA	—
Vibrio parahaemolyticus	—	ArcA	—
Vibrio cholerae	FNR	ArcA	—
Vibrio fischeri	—	ArcA	—
Yersinia pestis	FNR	—	—
Yersinia entercolitica	FNR	ArcA	—
Pasteurella multocida	FNR	ArcA	—
Actinobacillus actinomycetemcomitans	FNR	—	NarP
Haemophilus influenzae	FNR	—	NarP
Haemophilus ducreyi	FNR	ArcA	NarP
Vibrio vulnificus	—	—	NarP
Vibrio parahaemolyticus	—	—	NarP
Vibrio cholerae	—	—	NarP
Vibrio fischeri	—	ArcA	NarP

(Top) Regulatory cascades. Notation: E, Enterobacteriaceae; P, Pasteurellaceae; and V, Vibrionaceae. (Middle) Respiratory chain operons *nuo* (*E. coli*) and *nqr* (other genomes). (Bottom) Molibdate cofactor biosynthesis operon *moa*.

Changes in regulatory systems. Even more radical path of the regulon evolution is a complete change in a regulatory system. For instance, zinc repressors in most genomes are homologous proteins ZUR, whereas they are absent in streptococci, and zinc repressor is the AdcR protein from a different family (Panina et al., 2003a).

One of the most remarkable examples of this kind is regulation of the methionine biosynthesis in firmicutes (Rodionov et al., 2004; Figure 3). The ancestral system, S-box riboswitch (Grundy and Henkin, 1998), exists not only in firmicutes, but also in some other taxa, in particular, Actinobacteriaceae, Thermotogales, and some proteobacteria (*Xanthomonas* and *Geobacter*). S-boxes bind S-adenosyl-methionine, and the resulting change in the arrangement of helices regulates premature termination of transcription. This system was retained in bacilli and clostridia and lost in the common ancestor of streptococci and lactobacilli: none of the extant genomes from these taxa contains S-boxes. The regulatory role in lactobacilli was assumed by a different RNA-based system, T-boxes, that normally regulate aminoacyl-tRNA-synthetase genes (Henkin, 1994). In streptococci, the methionine pathway is regulated by the transcription factor MtaR.

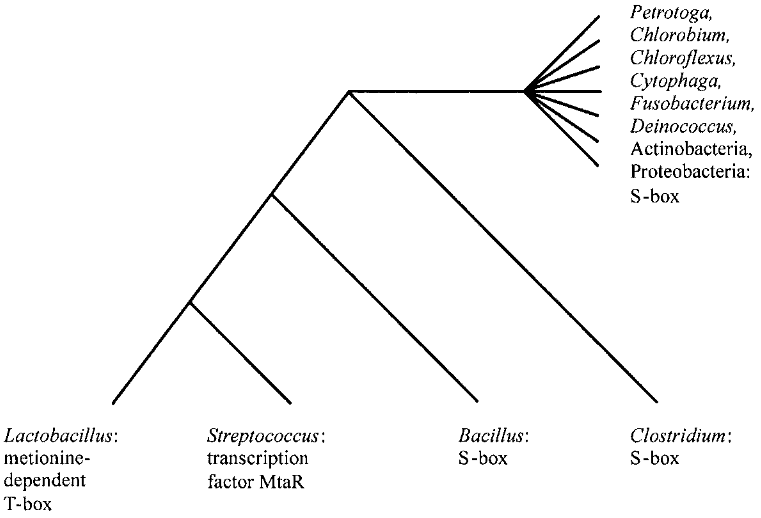
Petrotoga, Chlorobium, Chloroflexus, Cytophaga, Fusobacterium, Deinococcus, Actinobacteria, Proteobacteria: S-box
Lactobacillus: metionine-dependent T-box
Streptococcus: transcription factor MtaR
Bacillus: S-box
Clostridium: S-box

Figure -3. Regulation of methionine biosynthesis in firmicutes.

A similar situation occurs in the aromatic amino acid biosynthesis system, regulated by T-boxes, RNA-binding protein TRAP, and two unknown transcription factors whose signals have been identified by computational analysis (Terai et al., 2001; Panina et al., 2003b).

Co-evolution of regulators and signals. Analysis of protein–DNA interactions 'from the DNA point of view' demonstrated that the consensus positions, that is, the positions that clearly prefer one nucleotide, form significantly more contacts with transcription factors than non-consensus positions (Mirny and Gelfand, 2002). On the other hand, specificity-determining positions in transcription factor families cluster in three regions: the ligand-binding pocket, the subunit contact region, and DNA-binding helices (Kalinina et al., 2004).

Above, we have mentioned the ribose repressor RbsR, whose signal has changed in several gamma-proteobacterial families. Analysis of the LacI family of transcription factors, to which RbsR belongs, demonstrates that it is a common situation: factors with the same specificity form different branches, whereas signals are similar within branches (Gelfand and Laikova, 2003).

Sometimes, one can observe simultaneous changes in factors and signals. Transcription factors from the FNR/CRP family have similar sequences that allow for a reliable alignment. There are two groups of positions in the protein–DNA contact zone that demonstrate universal correlations (Table 3). In one such group, arginine in the protein yields TG in the binding signal, whereas in the second group, glutamate and one more arginine situated at the same side of the alpha-helix recognize the GA dinucleotide.

Table -3. Correlation between amino acid sequences of transcription factors from the FNR/CRP family and their signals

Genome	Factor	Fragment of protein alignment	Binding signal
DD	CooA	altteqlslhmgat**R**QtvsTllnnlvr	n**TG**TCGGCnnGCCGA**CA**n
DV	CooA	eltmeqlaglvgtt**R**QtasTllndmir	
EC	CRP	kitrqeigqivgcs**RE**tvg**R**ilkmled	T**TG**T**GA**nnnnnn**TC**A**CA**A
YP	CRP	kxtrqeigqivgcs**RE**tvg**R**ilkmled	
VC	CRP	kitrqeigqivgcs**RE**tvg**R**ilkmlee	
DD	HcpR	dvsksllagvlgta**RE**tls**R**alaklve	T**TG**T**gA**nnnnnn**Tc**A**CA**A
DV	HcpR	dvtkgllagllgta**RE**tls**R**clsrmve	
EC	FNR	tmtrgdignylgltV**E**tis**R**llgrfqk	nnTT**GA**TnnnnA**TC**AAnn
YP	FNR	tmtrgdignylgltV**E**tis**R**llgrfqk	
VC	FNR	tmtrgdignylgltV**E**tis**R**llgrfqk	

Correlated positions are shown by single- and double- underlined symbols. Genome notation: DD and DV: epsilon-proteobacteria *Desulfovibrio desulfuricans* and *Desulfovibrio vulgaris*, respectively; EC, YP, and VC: gamma-proteobacteria *E. coli*, *Yersinia pestis*, and *Vibrio cholerae*, respectively.

Another process forming the transcription signals is changes in the spacer length between halves of palindromic signals bound by dimeric factors. In particular, binding signals of the biotin repressor BirA in gram-positive bacteria and archaea, and in proteobacteria are similar and differ mainly by the size of the non-conserved spacer between the complementary half-sites (Figure 4*a*; Rodionov et al., 2002b). Sometimes, these two processes occur simultaneously.

Thus, zinc repressors from the ZUR family have similar signals in different taxa. However, in alpha-proteobacteria, one can observe point differences from the common superconsensus, whereas in gamma-proteobacteria, the spacer length is different (Figure 4*b*; Panina et al., 2003a).

```
a BirA
  wwTGTtAAC   15-16   GTTaACAww    (gram-positive bacteria and archaea)
   ////////             \\\\\\\\
  tTGTaAACC   15-16   GGTTtACAa    (gram-negative bacteria)

b ZUR
  GaaATGTtA-----TAACATttC          (common superconsensus)
  GAAATGTTAtantaTAACATTTC          (gamma-proteobacteria)
  GAtATGTTA     TAACATaTC          (Rhodobacter spp.)
  GtAATGTAA     TAACATTaC          (other alpha-proteobacteria)
```

Figure -4. Evolution of binding signals. (*a*) BirA signals in bacteria and archaea; (*b*) ZUR signals in proteobacteria. Lower-case letters: weakly conserved positions (BirA) and deviations from the common superconsensus (ZUR).

2. CONCLUSIONS

Analysis of regulatory systems and their evolution is currently at the level where the protein comparison and evolution was about twenty years ago. We know a number of interesting examples and see the direction of further studies. However, we are very far from a comprehensive, or even a draft theory describing the evolution of regulation.

ACKNOWLEDGMENTS

This study was supported by grants from HHMI (No. 55000309, M. Gelfand), RFBR (No. 04-04-49361, D. Rodionov), and the Russian Academy of Sciences (programs 'Molecular and Cellular Biology' and 'Origin and Evolution of the Biosphere').