Full text provided by www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Evolution of transcriptional regulatory networks in microbial genomes

Mikhail S Gelfand[1,2]

Advances in sequencing and generating high-throughput expression data have created a situation in which it is possible to integrate comparative analysis with genome-wide studies of the structure and function of regulatory systems in model organisms. Recent studies have focused on topological properties and the evolution of regulatory networks. This problem can be addressed on several levels: evolution of binding sites upstream of orthologous or duplicated genes; co-evolution of transcription factors and the DNA motifs that they recognize; expansion, contraction and replacement of regulatory systems; the relationship between co-regulation and co-expression; and, finally, construction of evolutionary models that generate networks with realistic properties. This should eventually lead to the creation of a theory of regulatory evolution with a similar level of detail and understanding to the theory of molecular evolution of protein and DNA sequences.

Addresses
[1] Institute for Information Transmission Problems, RAS, Bolshoi Karetny pereulok 19, Moscow, 127994, Russia
[2] Department of Bioengineering and Bioinformatics, MV Lomonosov Moscow State University, Vorobievy Gory 1-73, Moscow, 119992, Russia

Corresponding author: Gelfand, Mikhail S (gelfand@iitp.ru)

Current Opinion in Structural Biology 2006, **16**:420–429

This review comes from a themed issue on
Sequences and topology
Edited by Nick V Grishin and Sarah A Teichmann

Available online 2nd May 2006

0959-440X/$ – see front matter
© 2006 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.sbi.2006.04.001

Introduction

Regulation of gene expression is arguably one of the fastest developing areas of bioinformatics. There are two main reasons for this. Firstly, the number of sequenced genomes (hundreds of bacteria, dozens of archaea and yeasts) makes it possible to perform a wide variety of comparative studies. Secondly, advances in experimental techniques lead to the production of large amounts of non-genomic data that enable one to study whole networks, rather than individual systems, and place them in a cellular context. Here, I aim to review recent advances in the analysis of transcriptional regulation, and discuss the possible integration of functional and evolutionary approaches (Figure 1). I consider only microbial genomes, although developments concerning multicellular organisms are no less exciting. For example, large-scale studies of microRNA and their targets revealed a major, only recently appreciated, system of regulation. Large-scale mapping of promoters in conjunction with analysis of gene expression data obtained using microarrays and global transcriptome data obtained using tiling arrays pave the way for detailed studies of transcriptional regulation. In addition, the sequencing of many genomes from relatively narrow taxonomic groups, such as fruit flies and mammals, especially under the ENCODE project (http://genome.ucsc.edu/ENCODE/), enables large-scale comparative studies.

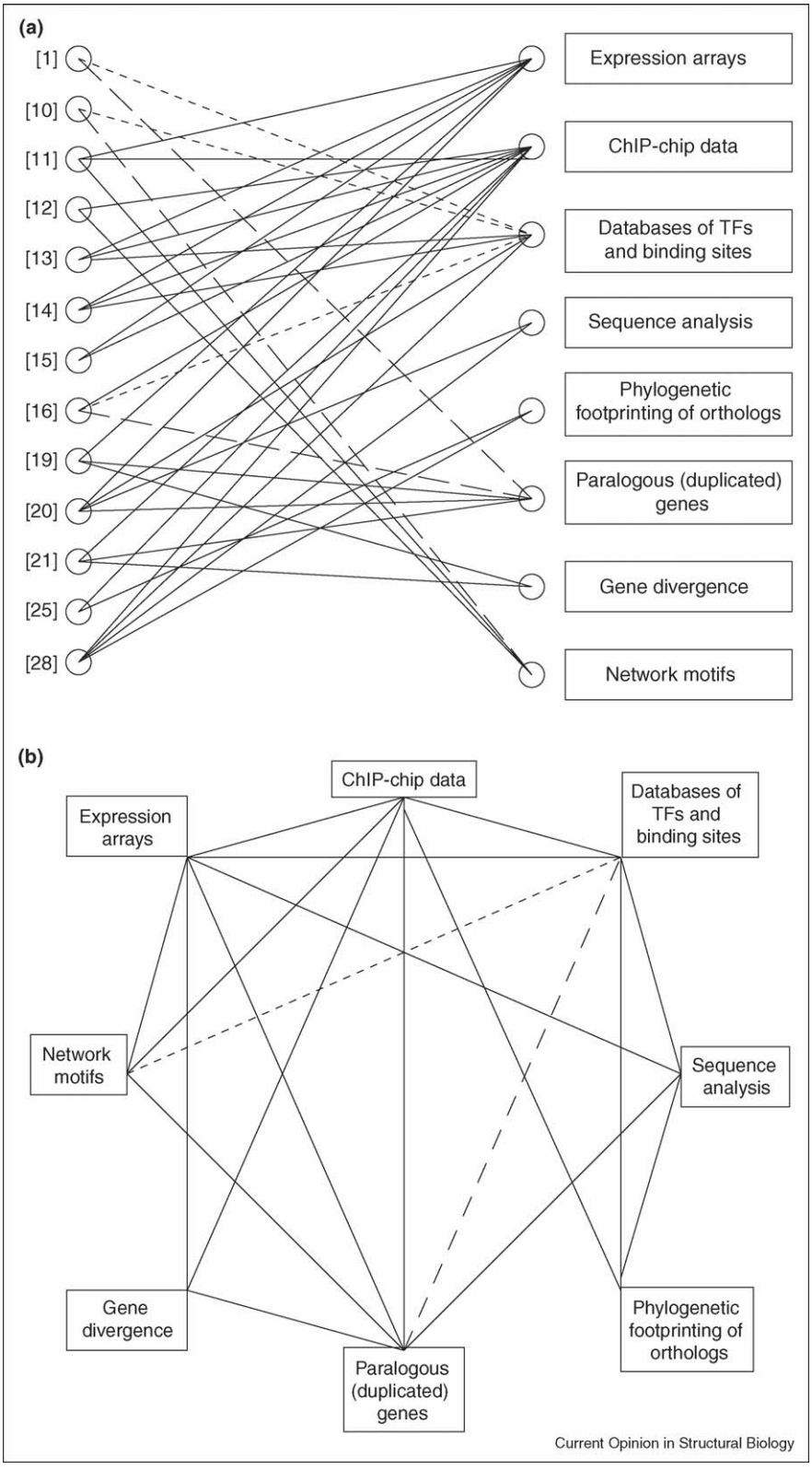
Structure and function of regulatory networks

Repertoire

Transcription factors (TFs) form one of the largest functional protein groups in most genomes. As demonstrated by Stover *et al.* [1] and, for a larger data set, by van Nimwegen [2], the fraction of TFs in bacterial genomes scales approximately as the square of the total gene number of a genome: that is, doubling of the genome complement size leads to quadrupling of the TF repertoire. Although it is clear that this behavior theoretically cannot be maintained indefinitely, there are no signs of saturation for larger genome sizes, reaching, for the largest genomes, 9.4% in *Pseudomonas aeruginosa* [1] and 12.3% in *Streptomyces coelicolor* [3].

The contribution of particular TF families to this diversity is very uneven. In *Escherichia coli*, 271 TFs form 11 classes, as defined by the structure of their DNA-binding domain, which are further subdivided into 74 families of paralogs based on their domain architecture [4]. The largest TF family in *E. coli* is LacI, with up to 100 representatives (but see below). However, in other genomes, the prevalence of TF families is different. Just by looking at clusters of orthologous groups, one can roughly define several types of TF families. In many cases (NrdR, HrcA, ArgR), an almost ubiquitous TF family forms a perfect group of orthologs, with a single representative in each genome and no indication of duplications. Some such factors (BirA, ModE) are also present in archaea. As shown by comparative genomic analysis, the binding motifs of such factors tend to be conserved, although in some cases there is a complete change in the binding motif of orthologous factors (e.g. DinR in the *Bacillus/Clostridium* group and LexA in proteobacteria). One further strategy is to have a small number of TF family

Figure 1



Network of recent integrative studies in systems biology. Solid lines, yeast *Saccharomyces cerevisiae*; short dashed lines, bacterium *Escherichia coli*; long dashed lines, both. **(a)** Literature and data. Left column, references as cited in this review; right column, associated data and methods. **(b)** Integration of various types of data and resources.

representatives per genome, as exemplified by the CRP/FNR and FUR families. Finally, there are families that demonstrate a burst of activity in certain taxa, for example, the LacI family in many α- and γ-proteobacteria (O Laikova, personal communication), the LysR family in *E. coli* [5], two-component systems in *Desulfovibrio* spp (E Permina, personal communication) and sigma-factors in *Streptomyces* spp [3]. One reason for this may be that TFs from the same family tend to perform specific, related functions (e.g. regulation of sugar catabolism for LacI and GntR, redox state sensing for FNR, metal ion homeostasis for FUR) and therefore the prevalence of a family in a genome indirectly reflects its lifestyle.

Networks

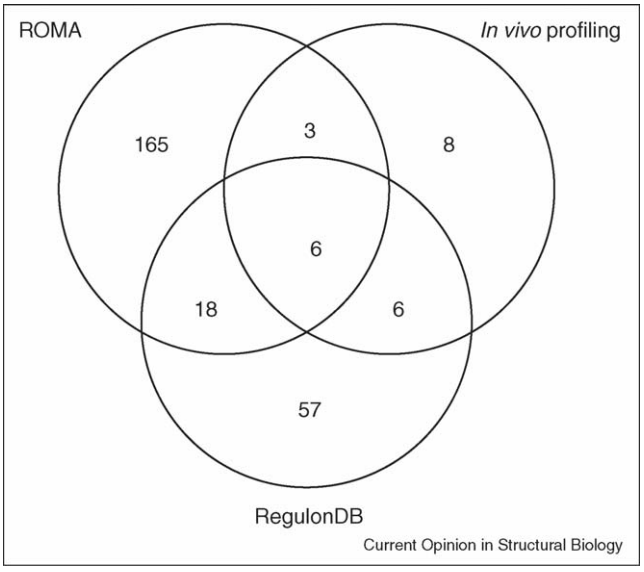
Advances in experimental techniques, specifically ChIP-chip analysis (an oligonucleotide-array-based technique for the large-scale identification of protein–DNA interactions), enabled a more or less complete elucidation of the yeast transcriptional regulatory network [6]. The number of promoters per regulator ranges from 0 to 181 (for the Abf1 TF) and the distribution follows a power law, with an average of 38 promoters per regulator. For comparison, analysis of the YPD database of yeast promoters (compiled from the literature) produced an average of 10 and a maximum of 72 promoters per regulator [7]. In contrast, the number of regulators per promoter seems to follow an exponential distribution, with the observed maximum of 21. Similar observations have been made for the *E. coli* regulatory network based on analysis of the literature data: the power law distribution with the maximum of 85 regulated genes per TF and the exponential distribution with up to 6 TF per promoter. About two-thirds of the *E. coli* and four-fifths of the yeast interactions are activations.

However, these numbers have to be considered as preliminary, as it is clear that further work will produce more data. A slightly more recent study of *E. coli* regulation using essentially the same type of data produced maxima of 197 sites per TF and 7 TFs per regulated operon [4], whereas a recent large-scale run-off transcription/microarray analysis (ROMA) identified 192 CRP-regulated operons, which contain about 300 genes [8]. The latter study also demonstrated that all methods tend to miss some sites and thus the actual number of CRP-regulated operons may be even larger (Figure 2).

Development of network motif analysis has taken two directions. One is the detailed analysis of transcriptional motifs, mainly the most over-represented one, the feed-forward loop (FFL): T → S → G, T → G, where T and S are transcription factors, and G is a gene. In particular, when the mode of regulation is taken into account, two types of FFLs (out of eight possible) turn out to be the most frequent both in *E. coli* and in yeast: an activation cascade, T → ⊕S → ⊕G, T → ⊕G, which delays the activating response of G to the stimulation of T if both factors (T and S) are needed for the expression of G; and T → ⊕S → ∅G, T → ⊕G, which accelerates the response of G to the stimulation of T in the absence of the stimulation of S, and represses the response if S is stimulated (⊕ and ∅ denote activation and repression, respectively, upon stimulation) [9]. The prevalence of certain types of motifs, in particular FFLs, has been linked to their dynamic properties: more prevalent motifs are those that lead to stable behavior under a larger set of repression/activation constants [10]. However, it is not immediately clear whether the latter study accounted for the following trivial explanation: among motifs formed by the same number of genes and interactions, relatively unstable (and infrequent) motifs are simply those that contain more TFs as vertices (as various loops are required for instability), whereas motifs with a relatively larger fraction of non-TF genes are more frequent simply because non-TF genes (terminal nodes in the transcription network) are more frequent than TF genes.

Analysis of the network context demonstrates that FFLs comprise just a few connected components of two types: both T and S regulate a large number of genes (e.g. the aerobic/anaerobic switch regulated by FNR → ArcA and the flagellar motor regulon); or T is a global regulator that influences a large number of local systems, each with its own local regulator (e.g. CRP and sugar catabolism repressors) [11].

Figure 2



Venn diagram of the number of CRP-regulated operons identified by ROMA, *in vivo* transcriptional profiling and diverse experiments, as collected in RegulonDB (data from [8]). The numbers in the intersection areas indicate the number of operons identified by more than one method.

Another direction is integration of diverse networks, putting transcriptional regulation into a functional framework. Simultaneous analysis of transcriptional regulation, protein–protein interactions, homology, co-expression and synthetic lethals in yeast revealed a number of over-represented motifs, most of which have a clear biological interpretation [12•]. Some of these motifs are FFLs; interacting TFs regulating the same gene; TFs regulating interacting and/or co-expressed genes; homologous TFs regulating same gene; and finally TFs regulating homologous genes. Again, it should be emphasized that the motifs do not exist in isolation, but rather form large 'themes', mainly by multiplication of the regulated genes. Some examples are multiple FFLs with the same pair of interacting TFs; interacting TFs and a large regulon; TF and its regulon, whose members are co-expressed and/or form a protein complex (Figure 3, top row).

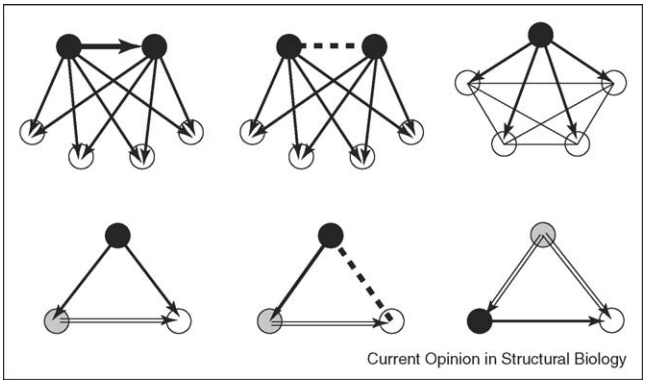
The next important step was to incorporate data on upstream signaling pathways. This became possible after the development of high-throughput methods for the identification of phosphorylation targets on proteome chips [13•]. Again, some motifs are over-represented in the resulting network (Figure 3, bottom row). How these simple motifs interact within a larger map remains to be analyzed.

Co-regulation and co-expression

Large sets of expression data and a surge in the development of clustering algorithms for the identification of co-expressed genes immediately raised the question of how well the constructed clusters approximate co-regulation. Comparison of co-expression clusters with a database of TF-binding sites and ChIP data on protein–DNA interactions demonstrated that the measure of overlap between these clusters and regulons stabilizes in the interval between 50 and 100 microarray experiments, but both false-negative and false-positive rates remain rather high [14]. Only about 20% of binding sites identified in the literature are recovered in ChIP experiments; on the other hand, at most 28% of genes in one cluster share at least one TF. However, this analysis did not take into account the nature of regulation; it had been demonstrated that only co-activated genes tend to be co-expressed in a large variety of conditions, whereas co-repressed genes may shut down simultaneously, but are uncorrelated otherwise [15]. It might be interesting to see whether it is possible to predict whether a TF is an activator or a repressor (given DNA-binding data) based on the difference between the modes of behavior of regulated genes.

Another way to integrate expression and regulation data is to identify regulatory interactions that are active in various conditions [16•]. For any given condition, a gene was assumed to be active if its expression changed in this condition; a TF regulating an active gene was assumed to be active if it was expressed at a significant level and this procedure was iterated, activating additional TFs, until convergence. At the end, only the links between active genes were retained, producing an active subnetwork. It turned out that the structure of the active subnetwork depends on the experimental conditions. In 'endogenous' conditions (cell-cycle phases, sporulation), complex TF combinations were revealed, with few targets per TF, relatively many TFs per target, long paths, high clustering coefficients, many connections between TFs and, in particular, many FFLs. In 'exogenous' conditions (diauxic shift, DNA damage, environmental stress), there were few connections between TFs, many targets per factor and most genes were regulated by a single TF.

Figure 3



Typical motifs involving transcriptional regulation and other types of interaction (following [12•,13•]). Arrows, transcriptional regulation (the thick arrow shows the interaction between TFs in generating the FFL motifs); dotted line, protein–protein interactions; double line, phosphorylation; thin solid lines, protein complexes and/or co-expression. Black circles, TFs; grey circles, kinases; white circles, other genes.

Evolution

Duplications within genomes

In several studies, it has been observed that paralogous TFs tend to regulate paralogous genes, both in *E. coli* [4,16•] and in yeast [12•,17]. This has been considered evidence that transcriptional regulatory networks tend to evolve by duplication and, indeed, a simple duplication-based model is sufficient to explain the observed scaling of TF number with genome size and the distribution of the node degrees in the transcriptional network [18]. However, although there are numerous obvious examples of recently duplicated TFs and their targets, another explanation might be massive horizontal transfer of fragments containing both regulated genes and their regulators. Indeed, recent horizontal transfers of bacterial enzymes outnumber duplications by about tenfold [19].

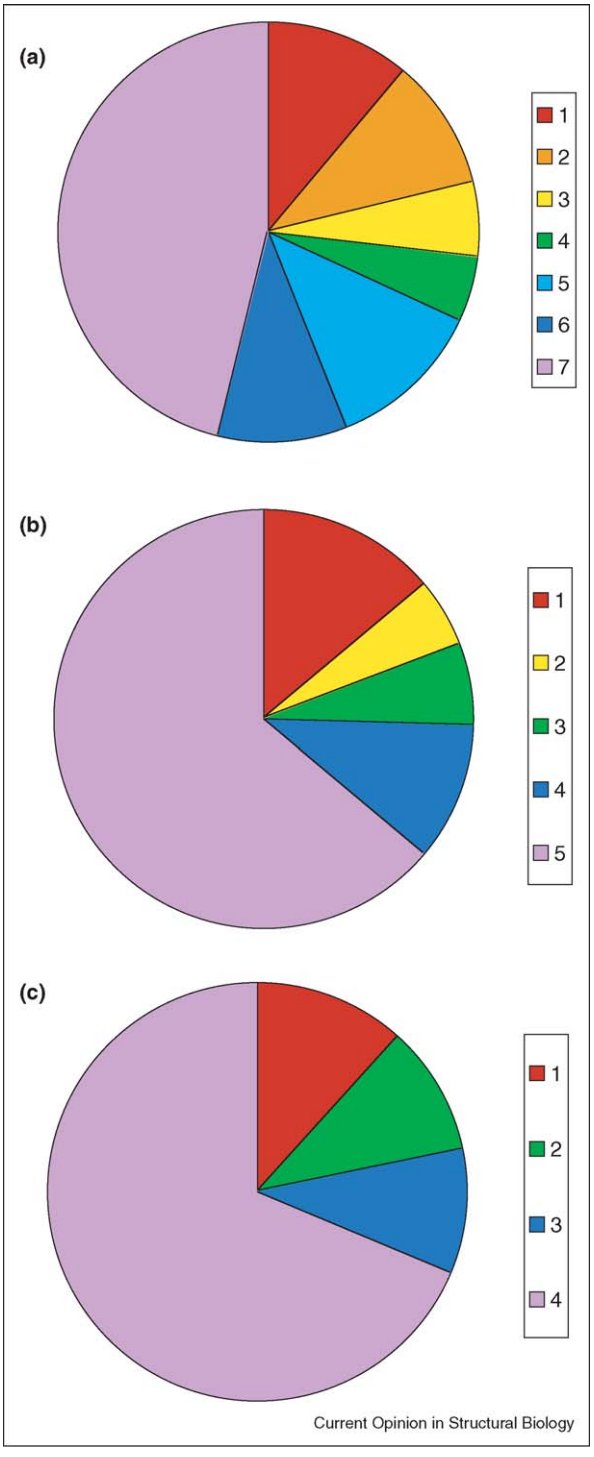
One of the main sources of examples of coordinated TF and target gene duplications in bacteria are the carbohydrate catabolism regulons. However, it looks like the main mode of introduction of new catabolic pathways is the LEGO-like construction of new operons from pre-existing components rather than simple duplication of entire operons (O Laikova, personal communication). As such, TFs, transporters and enzymes in functionally similar pathways may belong to different non-homologous families, creating numerous instances of non-orthologous gene displacement. Although this observation may be difficult to quantify, support comes from the fact that the TF content of close strains may be considerably different. For example, there are 100 groups of orthologous TFs from the LacI family in seven *Escherichia* and *Shigella* spp genomes, but only 46 of them occur in all seven genomes and a single genome contains 69–88 TFs; similarly, there are 72–80 representatives out of 91 possible in five *Salmonella* spp genomes and 36–43 out of 49 in four bacteria from the *Bacillus cereus* group (Figure 4) (O Laikova, personal communication).

In another series of studies, the divergence in regulation and/or expression of recently duplicated yeast genes was considered. The co-expression of paralogs was shown to deteriorate rather rapidly: 40% of paralogs show lack of co-expression even at a rather close evolutionary distance (synonymous substitution rate $K_s < 0.10$) and more than 80% of paralogs with $K_s > 1.5$ are not co-expressed [20]. A similar trend is seen when the non-synonymous substitution rate, K_a, is used as the measure of evolutionary distance, but only for the initial segment of K_a values ($K_a < 0.30$). Rapid loss of coincidence of predicted TF-binding sites was demonstrated in [21]. However, the loss of 3% of experimentally determined sites for common TFs per 1% amino acid divergence was observed in [22]; this coincides nicely with the above observations, as it means that, at $K_a \sim 0.3$, almost all common sites would be lost.

As it has been recently demonstrated that a large fraction of the yeast paralogs arose from whole-genome duplication [23], it would be interesting to compare the pattern of site loss and expression change in surviving descendants of this duplication (in this case, the post-duplication time is the same for all pairs, and it is possible to check whether there is a correlation between the rate of sequence and regulatory evolution) and in the paralogs arising from local duplications. One could expect that the reasons for retention of duplicated genes following global duplication and local duplication might be different.

No systematic studies of this kind were done in bacteria, although some observations have been made in the course of comparative analyses. One such observation is the rapid divergence after duplication of the ribose repressor in the common ancestor of *Enterobacteriaceae*, *Vibrionaceae*

Figure 4



Representation of groups of orthologous TFs from the LacI family in three groups of closely related strains. The chart reflects the number of groups that are represented exactly in a given number of genomes. **(a)** Seven *Escherichia* and *Shigella* spp, **(b)** five *Salmonella* spp and **(c)** four bacteria from the *B. cereus* group (*B. cereus* ATCC10987 and E33L, *B. anthracis* and *B. thuringiensis*) (O Laikova, personal communication).

and *Pasteurellaceae* (M Gelfand, unpublished). One lineage, the purine repressors (PurR), retained the DNA motif, but changed the cofactor specificity and the regulated pathway; the other lineage, the ribose repressors (RbsR), changed the DNA motif.

Genome comparisons

Comparative analyses of regulatory sites in related genomes were initially used to identify conserved predicted sites and thus to filter out false positives. It was successfully applied to bacteria and was one of the main reasons prompting the sequencing of multiple yeast genomes [24,25]. The results were somewhat mixed.

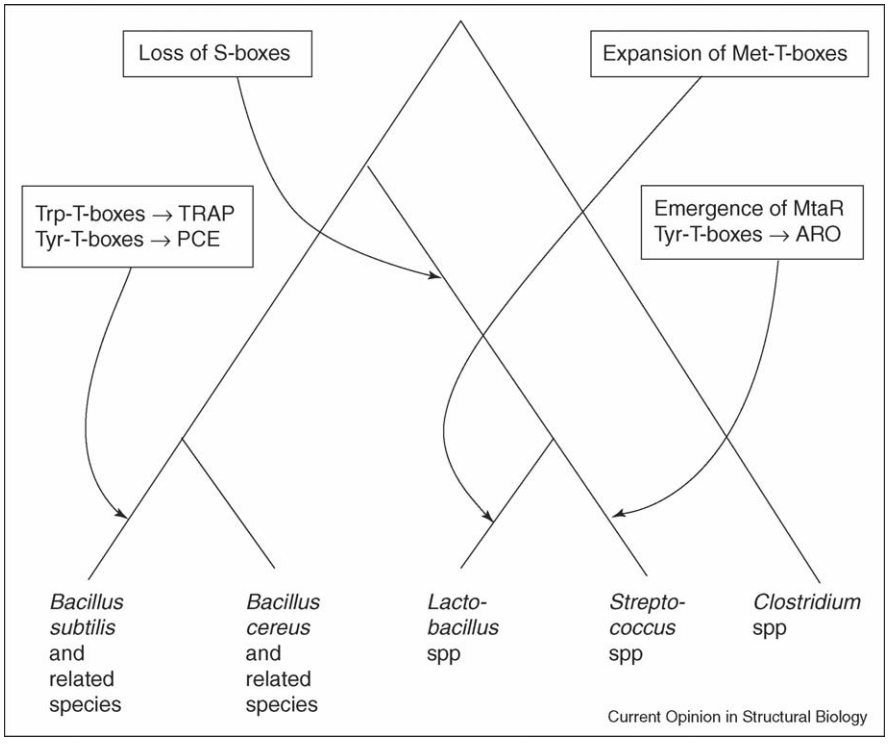
The lists of candidate motifs obtained in the two large-scale comparative studies [24,25] contained few common motifs, although more robust results were obtained subsequently when phylogenetic footprints were overlaid with genome-wide location data [26••]; in particular, it demonstrated that regulons of many TFs strongly depend on cellular conditions.

Similarly, automated comparative analyses of bacterial genomes produced a considerable amount of false positives. One reason could be frequent shuffling of genes in operons. Taking this possibility into account led to the notion of a regulog, a set of genes regulated by orthologous factors in several genomes [27•].

On the other hand, analysis of individual systems was quite successful and produced numerous experimentally validated predictions. One of the candidate motifs from [26••] was analyzed in detail, leading to the discovery of the NrdR system, which regulates ribonucleotide reductase genes in various bacteria [28•]. This is an example of a completely new system that has been characterized in minute detail solely by comparative analysis. This description includes: the conserved motif; the corresponding TF (identified using phylogenetic co-occurrence patterns); the mode of regulation (repression following cooperative binding to tandem regulatory sites); and the link between this regulon, deoxyribonucleotide salvage and replication (based on the occurrence of candidate sites upstream of DNA ligases, topoisomerases, helicases, replication initiator *dnaA* and genes involved in chromosome partitioning).

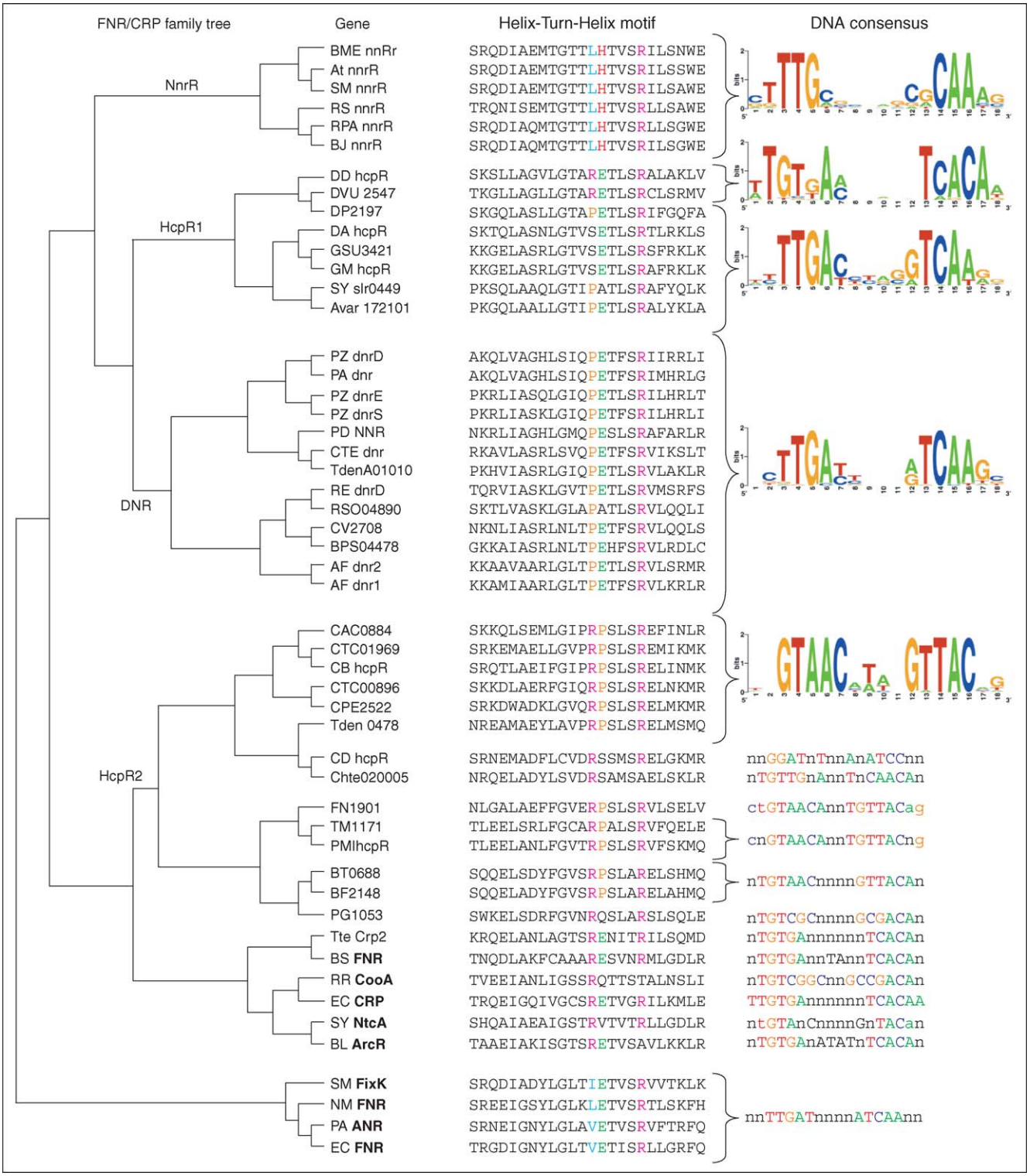
Analyses of particular regulatory systems demonstrated considerable flexibility of both motifs and systems. For example, although the Rpn4p factors that regulate

Figure 5



Changes in the regulation of amino acid biosynthesis operons in the *Bacillus/Clostridium* group of Gram-positive bacteria. The tree reflects the phylogenetic relationships. S-boxes, S-adenosylmethionine-dependent riboswitches; T-boxes, tRNA-binding regulatory RNA structures; TRAP, RNA-binding protein that regulates tryptophan biosynthesis genes; MtaR, transcriptional repressor of methionine biosynthesis genes; PCE and ARO: candidate signals of unknown TFs upstream of tyrosine and phenylalanine biosynthesis genes, respectively [43–45].

Figure 6



Correlation between residues in the DNA-binding helix-turn-helix domain of the FNR/CRP family of TFs and their DNA recognition motifs [32]. Two specificity-determining positions in the helix-turn-helix domain correlated with DNA motifs are colored (R180 and E181 in the TF correlate with G3 and A6 in the DNA, respectively). The fourth column shows sequence logos for presumably homogeneous and large sets of DNA sites, and sequence consensus for small sets of DNA sites and for well-established motifs of other factors (FNR, CRP, CooA, NtcA, ArcR).

proteasome genes in *Saccharomyces cerevisiae* and *Candida albicans* are similar, the nucleotide frequency distributions are significantly different in several positions, as are the binding affinities of the factors for various motifs [29]. On the other hand, the motif is the same in *C. albicans* and the distant *Neurospora crassa* genome, whereas the regulons are clearly different.

Analysis of the ribosomal protein (RP) genes in various yeasts demonstrated two types of behavior [30••]. One was the gradual change in the binding motif of the IFHL factor. The other was cooptation of the Rap1p factor after it acquired a new transactivation domain in the last common ancestor of *S. cerevisiae* and *Ashbya gossypii*, after branching of *C. albicans*. The situation is even more complicated because, in *S. cerevisiae*, genes encoding mitochondrial RPs are co-expressed with stress response genes, whereas in *C. albicans*, they are co-expressed and co-regulated with cytoplasmic RP and rRNA processing genes [31••]. The regulatory motif responsible for this was lost in mitochondrial RP genes of the *S. cerevisiae* lineage after the whole-genome duplication event (in particular, it exists in *A. gossypii*).

Similar striking examples of flexibility in regulatory interactions were observed in analyses of bacterial regulatory systems. Some examples are multiple changes in the regulatory cascades of the aerobic/anaerobic switch and respiration in γ-proteobacteria involving the FNR, ArcA and NarL/NarP TFs (D Ravcheev, A Gerasimova, personal communication); shuffling of regulators of nitrogen oxide metabolism [32]; taxon-specific σ^E promoters in enterobacteria [33]; and loss and expansion of regulatory systems of amino acid metabolism genes in Gram-positive bacteria (Figure 5). The last system is of particular interest as it involves both TFs and RNA elements; the latter are sufficiently large to reconstruct the history of duplications and horizontal transfers (A Vitreschak, personal communication).

The evolution of TF-binding sites can be studied by considering sites upstream of orthologous genes. As expected, the positional mutation rate negatively correlates with information content, both in yeasts [34] and in bacteria [35]; indeed, the latter is positively correlated with the number of protein–DNA contacts [36] and thus the evolution of such positions is constrained. More surprisingly, non-consensus nucleotides in orthologous sites still tend to be more conserved than expected under a neutral model [37].

Analysis of the co-evolution of TFs and their binding motifs also demonstrates a mosaic of conservation and change. The architecture of binding signals tends to be the same for members of a TF family (palindromes for many families, short direct repeats with a helical pitch periodicity of 1011 base pairs for two-component systems, overlapping direct and inverted repeats for the FUR family), although changes are possible. For example, the GntR family TFs have both palindromic and direct repeat motifs [38,39]. The binding motif is conserved for most families that have a single representative per genome, a notable exception being LexA/DinR. For families with numerous representatives, such as LacI and GntR, common features of the binding motifs can be discerned (O Laikova, personal communication and [38], respectively); this may be compared with analogous observations for eukaryotic TF families [40,41•]. Moreover, analysis of a set of binding sites may allow one to predict the structural class of the TF, at least at the coarse level [41•,42•]. On the other hand, the analysis of correlations between TFs from one family and their motifs may lead to the identification of structural determinants of specific recognition (Figure 6).

Some open problems

The above discussion shows that analysis of regulatory networks is a very young area developing in a variety of directions. Thus, instead of a conclusion, I would like to list some open problems.

1. Development of a language for an intermediate level of network topology (interacting motifs, motif complexes).
2. Further incorporation of post-genomic data (kinase cascades, nucleosome positioning, histone modifications) for yeasts.
3. Systematic large-scale experimental analysis of protein–DNA interactions in prokaryotes.
4. Producing models of the evolution of TF families, linking TF family size, phylogenetic distributions and regulon sizes.
5. Systematic analysis of the evolution of protein–DNA interactions, and the co-evolution of TFs and their DNA motifs.
6. Incorporation of horizontal transfer into evolutionary models and characterizing metagenomic (common to several strains) TF pools.
7. Study of the consequences of whole-genome duplications in the yeast genome on the regulatory network; comparison of paralogs arising from this duplication and paralogs stemming from local duplications.
8. Systematic comparative analysis of complex regulatory systems, collection and generalization of observations, and their use in the construction of realistic models of regulatory network evolution.

Acknowledgements

I am grateful to Iaroslav Ispolatov, Vsevolod Makeev and Sergei Maslov for useful discussions, and to Anna Gerasimova, Olga Laikova, Elizabeth Permina, Dmitry Ravcheev, Dmitry Rodionov and Alexei Vitreschak for sharing unpublished data. This study was supported by grants from the Howard Hughes Medical Institute (55005610) and the Russian Academy of Science (programs Molecular and Cellular Biology, and Origin and Evolution of the Biosphere).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, Hickey MJ, Brinkman FS, Hufnagle WO, Kowalik DJ, Lagrou M *et al.*: **Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen**. *Nature* 2000, **406**:959-964.

2. van Nimwegen E: **Scaling laws in the functional contents of genomes**. *Trends Genet* 2003, **19**:479-484.

3. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D *et al.*: **Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)**. *Nature* 2002, **417**:141-147.

4. Babu MM, Teichmann SA: **Evolution of transcription factors and the gene regulatory network in *Escherichia coli***. *Nucleic Acids Res* 2003, **31**:1234-1244.

5. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV: **Lineage-specific gene expansions in bacterial and archaeal genomes**. *Genome Res* 2001, **11**:555-565.

6. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799-804.

7. Milo R, Shen-Orr SS, Itzkovitz S, Kashtan A, Alon U: **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**:824-827.

8. Zheng D, Constantinidou C, Hobman JL, Minchin SD: **Identification of the CRP regulon using *in vitro* and *in vivo* transcriptional profiling**. *Nucleic Acids Res* 2004, **32**:5874-5893.

9. Mangan S, Alon U: **Structure and function of the feed-forward loop network motif**. *Proc Natl Acad Sci USA* 2003, **100**:11980-11985.

10. Prill RJ, Iglesias PA, Levchenko A: **Dynamic properties of network motifs contribute to biological network organisation**. *PLoS Biol* 2005, **3**:e343.

11. Dobrin R, Beg QK, Barabasi AL, Oltvai ZN: **Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network**. *BMC Bioinformatics* 2004, **5**:10.

12. • Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP: **Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network**. *J Biol* 2005, **4**:6.

The authors identify and analyze structural motifs in a graph that represents five types of interactions: co-expression, transcriptional regulation, protein–protein interactions, synthetic lethals and homology.

13. • Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R *et al.*: **Global analysis of protein phosphorylation in yeast**. *Nature* 2005, **438**:679-684.

This large-scale experimental study of phosphorylation in yeast enabled reconstruction of the signaling pathways.

14. Yeung KY, Medvedovic M, Bumgarner RE: **From co-expression to co-regulation: how many microarray experiments do we need?** *Genome Biol* 2004, **5**:R48.

15. Yu H, Luscombe NM, Qian J, Gerstein M: **Genomic analysis of gene expression relationships in transcriptional regulatory network**. *Trends Genet* 2003, **19**:422-427.

16. • Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 2004, **431**:308-312.

Transcriptional subnetworks in yeast, activated in response to external and internal stimuli, were shown to possess different topological properties.

17. Teichmann SA, Babu MM: **Gene regulatory network growth by duplication**. *Nat Genet* 2004, **36**:492-496.

18. Foster DV, Kauffman SA, Socolar JES: **Network growth models and genetic regulatory networks**. *Phys Rev E* 2006, **73**:031912.

19. Pál C, Papp B, Lercher M: **Adaptive evolution of bacterial metabolic networks by horizontal gene transfer**. *Nat Genet* 2005, **37**:1372-1375.

20. Gu Z, Nicolae D, Lu HH, Li WH: **Rapid divergence in expression between duplicate genes inferred from microarray data**. *Trends Genet* 2002, **18**:609-613.

21. Papp B, Pal C, Hurst LD: **Evolution of cis-regulatory elements in duplicated genes of yeast**. *Trends Genet* 2003, **19**:417-422.

22. Maslov S, Sneppen K, Eriksen KA, Yan KK: **Upstream plasticity and downstream robustness in evolution of molecular networks**. *BMC Evol Biol* 2004, **4**:9.

23. Kellis M, Birren B, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae***. *Nature* 2004, **428**:617-624.

24. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements**. *Nature* 2003, **423**:241-254.

25. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting**. *Science* 2003, **301**:71-76.

26. •• Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al.*: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**:99-104.

The authors characterized the yeast transcriptional network using the integration of large-scale binding data in a variety of conditions with phylogenetic footprinting.

27. • Alkema WB, Lenhard B, Wasserman WW: **Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus***. *Genome Res* 2004, **14**:1362-1373.

The authors present a method of phylogenetic footprinting that takes into account changes in the operon structure of bacterial genomes.

28. • Rodionov DA, Gelfand MS: **Identification of a bacterial regulatory system for ribonucleotide reductases by phylogenetic profiling**. *Trends Genet* 2005, **21**:385-389.

The complete characterization of a regulatory system by bioinformatics analysis, including identification of an almost universal TF, its mode of action, DNA motifs and links to other systems in a variety of genomes.

29. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB: **Conservation and evolution of cis-regulatory systems in ascomycete fungi**. *PLoS Biol* 2004, **2**:e398.

30. •• Tanay A, Regev A, Shamir R: **Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast**. *Proc Natl Acad Sci USA* 2005, **102**:7203-7208.

See annotation to [31••].

31. •• Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N: **Rewiring of the yeast transcriptional network through the evolution of motif usage**. *Science* 2005, **309**:938-940.

Two papers [30••,31••] present detailed computational and experimental analyses of the regulation of genes that encode RPs in fungal genomes evenly spanning the taxonomic space from *S. cerevisiae* to *Schizosaccharomyces pombe*. This enabled the reconstruction of detailed evolutionary scenarios involving changes in regulatory motifs, cooptation of new factors and loss of existing sites.

32. Rodionov DA, Dubchak IL, Arkin AP, Alm EJ, Gelfand MS: **Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks**. *PLoS Comput Biol* 2005, **1**:e55.

33. Rhodius VA, Suh WC, Nonaka G, West J, Gross CA: **Conserved and variable functions of the σ^E stress response in related genomes**. *PLoS Biol* 2005, **4**:e2.

34. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB: **Position specific variation in the rate of evolution in transcription factor binding sites**. *BMC Evol Biol* 2003, **3**:19.

35. Brown CT, Calan CG Jr: **Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli***. *Proc Natl Acad Sci USA* 2004, **101**:2404-2409.

36. Mirny L, Gelfand MS: **Structural analysis of conserved base-pairs in protein-DNA complexes**. *Nucleic Acids Res* 2002, **30**:1704-1711.

37. Kotelnikova EA, Makeev VY, Gelfand MS: **Evolution of transcription factor DNA binding sites**. *Gene* 2005, **347**:255-263.

38. Rigali S, Schlicht M, Hoskisson P, Nothaft H, Merzbacher M, Joris B, Titgemeyer F: **Extending the classification of bacterial transcription factors beyond the helix-turn-helix motif as an alternative approach to discover new *cis*/*trans* relationships**. *Nucleic Acids Res* 2004, **32**:3418-3426.

39. Danilova LV, Gelfand MS, Lyubetsky VA, Laikova ON: **Computer-assisted analysis of regulation of the glycerol-3-phosphate metabolism in genomes of proteobacteria**. *Mol Biol* 2003, **37**:716-722.

40. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics**. *J Mol Biol* 2004, **338**:207-215.

41. • Xing EP, Karp RM: **MotifPrototyper: a Bayesian profile model for motif families**. *Proc Natl Acad Sci USA* 2004, **101**:10523-10528.
See annotation to [42•]

42. • Narlikar L, Hartemink AJ: **Sequence features of DNA binding sites reveal structural class of associated transcription factor**. *Bioinformatics* 2006, **22**:157-163.
Two studies [41•,42•] show that it is possible to predict the structural class of a TF from a set of its binding sites.

43. Panina EM, Vitreschak AG, Mironov AA, Gelfand MS: **Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria**. *FEMS Microbiol Lett* 2003, **222**:211-220.

44. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of the regulation of methionine metabolism in Gram-positive bacteria**. *Nucleic Acids Res* 2004, **32**:3340-3353.

45. Gutierrez-Preciado A, Jensen RA, Yanofsky C, Merino E: **New insights into regulation of the tryptophan biosynthetic operon in Gram-positive bacteria**. *Trends Genet* 2005, **21**:432-436.