

# Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria

Dmitry A. Rodionov\*

Burnham Institute for Medical Research, La Jolla, California 92037, and A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny per. 19, Moscow 127994, Russia

Received April 11, 2007

## Contents

1. Introduction	1	3.3. Analysis of Regulons to Support Metabolic Reconstruction and Functional Predictions	20
2. Computational Methods for Identification of Regulatory Motifs	3	3.3.1. L-Rhamnose Utilization	20
2.1. Structure, Function, and Representation of Transcription Factor Binding Sites	3	3.3.2. Other Catabolic Pathways	21
2.1.1. Position of TFBSs in Promoter Regions	3	3.3.3. Biosynthesis of Coenzyme B <sub>12</sub>	21
2.1.2. Structure of TFBSs	3	3.3.4. New Mechanisms for Alternative Cofactor Adaptation	23
2.1.3. TFBS Consensus and Logo Sequences	3	3.3.5. Prediction of Transporter Specificities	24
2.1.4. Positional Weight Matrices	4	4. Patterns and Mechanisms in Evolution of Transcriptional Regulatory Networks	24
2.1.5. Search for TFBSs in Microbial Genomes	5	4.1. Methionine Metabolism	24
2.2. Repertoire of Transcription Factors in Prokaryotic Genomes	5	4.2. Aromatic Amino Acid Metabolism	26
2.2.1. Distribution of TFs in Microbial Genomes	5	4.3. Fructose Regulon in $\gamma$ -Proteobacteria	26
2.2.2. Families of TFs	5	4.4. Iron and Manganese Regulatory Networks	26
2.2.3. Domain Architecture of TFs	7	5. Directions for Future Studies	27
2.2.4. Global and Local TFs	8	6. Abbreviations	27
2.2.5. Alternative Sigma Factors	8	7. Acknowledgments	28
2.3. Databases of Microbial TFs and TFBSs	8	8. Supporting Information	28
2.4. Computational Tools for Discovery of TFBSs in the Genomes	9	9. References	28
2.5. Comparative Genomic Approaches for Identification and Verification of TFBSs	10		
2.5.1. Consistency Check Approach	10		
2.5.2. Phylogenetic Footprinting Approach	10		
2.5.3. Genome-Wide Application of Comparative Approaches	11		
2.6. Interconnection of Transcription Factors and Their DNA Motifs	12		
2.7. Analysis of RNA Regulatory Elements	12		
3. Reconstruction and Comparison of Regulatory Networks That Control Central Metabolism in Bacteria	13		
3.1. Combining Experimental and Genomic Data to Predict TFBS Motifs	13		
3.2. Comparative Genomic Reconstruction of Regulatory and Metabolic Networks	14		
3.2.1. N-Acetylglucosamine and Chitin Utilization	15		
3.2.2. Sugar Acids Utilization	18		
3.2.3. Biotin Metabolism	18		
3.2.4. Nitrogen Metabolism	19		
3.2.5. NAD Metabolism	20		

## 1. Introduction

Microorganisms in most ecological niches are constantly exposed to variations of many environmental factors, including temperature; oxygen, nutrient, and water availability; presence of toxic compounds; and interaction with other organisms. Changing gene-expression patterns is a major adaptive response to these variations. Expression of genes in bacteria is controlled by a variety of mechanisms based on the level of transcription or translation. In most cases, the switch in gene expression is mediated by the specific regulatory proteins that receive an appropriate intra- or extracellular signal and trigger the specific transcriptional response.<sup>1</sup>

The key components of transcriptional regulatory machinery in prokaryotes are transcription factors (TFs) and transcription factor-binding sites (TFBSs), sigma factors and promoters, antiterminator proteins, and *cis*- and *trans*-acting regulatory RNAs. TFs are proteins that recognize specific *cis*-regulatory DNA sequences (TFBSs) to either stimulate or repress transcription of genes.<sup>2</sup> Sigma factors are prokaryotic transcription initiation factors that must be a part of RNA polymerase holoenzyme for specific binding to promoter sites encoded in the 5'-untranslated regions (UTRs) of genes.<sup>3</sup> Antiterminator protein factors bind to specific secondary structures in the leader region of mRNA to restart transcription of a gene.<sup>4</sup> Cellular signals that can modulate TFs

\* Address for correspondence: 10901 N. Torrey Pines Rd., La Jolla, CA 92037. Phone: (858) 646-3100 ext. 3082. Fax: (858) 795-5249. E-mail: rodionov@burnham.org.



Dmitry A. Rodionov was born in Moscow, Russia, in 1977. He studied biophysics at the Moscow Engineering Physics Institute and received his M.Sc. degree in 2000, analyzing transcriptional regulation of multidrug resistance genes in *Escherichia coli* under the supervision of Dr. Alexandra B. Rakhmaninova. His graduate research was conducted with Prof. Mikhail S. Gelfand and Prof. Andrei A. Mironov in the Laboratory of Bioinformatics of the State Research Institute for Genetics and Selection of Industrial Microorganisms in Moscow, Russia. At the same institution, he obtained his Ph.D. in Molecular Biology in 2003 with a thesis about comparative genomics reconstruction of transcriptional regulatory and metabolic networks in bacteria. He continued to study microbial regulatory networks using comparative genomics with Prof. Gelfand at the Institute for Information Transmission Problems of the Russian Academy of Sciences in Moscow. During this time, he participated in multiple collaborative projects as a visitor scientist at the National Center for Biotechnology Information in Bethesda, Maryland (spring 2000), the Unit of Microbiology and Genetics, INSA-Lyon, France (autumn 2003), Lawrence Berkeley National Laboratory in California (spring 2004), and the Institute for Microbiology at Humboldt University in Berlin, Germany (spring 2005). In autumn 2005, he moved to the laboratory of Prof. Andrei Osterman at the Burnham Institute for Biomedical Research in La Jolla, California, as a postdoctoral associate. His research interests are different aspects of comparative genomic reconstruction of metabolic pathways and regulatory networks in prokaryotes including identification of transcription factor-binding sites and RNA regulatory elements, functional annotation of genes and identification of missing genes in metabolic pathways, phylogenetic analysis of protein families, and evolution of metabolic pathways and regulatory systems.

include binding of a small molecule (effector), interaction with other proteins (e.g., phosphorylation), changing redox state, temperature, and other conditions.<sup>5</sup> Finally, various regulatory RNA structures including *cis*-acting metabolite-sensing riboswitches, T-boxes, and attenuators<sup>6,7</sup> and *trans*-acting small RNAs<sup>8</sup> control gene expression without involvement of specific proteins.

The operon is a set of adjacent genes that are transcribed as a single polycistronic mRNA. This organization of genes in operons achieves for bacteria a simple solution to the problem of coregulating genes that participate in the same metabolic process. However, the operon strategy has some limitations, e.g., the inability to combine both an independent regulation and a coordinated control. Regulon organization presents a level of control above the operons and permits coordinated control of operons that each have their own unique control. The regulon is a group of operons controlled by a common TF or regulatory RNA. The regulon usually includes genes that are implicated in a common cellular subsystem or a pathway. For example, in most species the regulons for arginine and thiamin biosynthesis genes are controlled by the ArgR repressor and the *THI* riboswitch, respectively.<sup>9,10</sup> However, most responses of bacterial cells to even a simple environmental stimulus are complex. Thus, the operational term stimulus was defined to refer to an

ensemble of genes (often involving multiple regulons and independent operons) that respond to a common environmental stimulus, although they may not share a common mechanism of regulation. For example, the heat-shock and phosphate-starvation stimulons include hundreds of genes in *Escherichia coli*, while only some of them are known members of the  $\sigma^{32}$  and PhoB regulons, respectively.<sup>11</sup> Finally, the term modulon was introduced to define a set of genes that are either directly or indirectly controlled by a certain regulatory system.

Fine-tuned environmental responses require efficient, flexible, and robust transcriptional regulatory networks (TRNs) that contain both internal checkpoints and feedback mechanisms to orchestrate the level of gene expression. To define a particular TRN, we need to specify which TFs bind to the promoter regions of which genes and what is the integrated effect of all these TFs on the expression of all these genes.<sup>12</sup> Reconstruction of TRNs helps us to better understand the metabolism and functions of prokaryotic organisms.<sup>13</sup> An accumulated amount of information about gene regulation networks was used to define the basic building blocks of complex TRNs, termed network motifs, and to understand their design principles.<sup>14</sup>

Traditional experimental methods for analysis of transcriptional gene regulation (such as gene cloning, knockout, reporter fusion, and *in vitro* transcription) and characterization of TFBSs (electrophoretic mobility shifts and nuclease protection assays) have been very powerful. However, they have certain limitations in terms of both productivity (the scale) and feasibility (e.g., for nonmodel organisms). Development of high-throughput transcriptome and proteome approaches allows thousands of genes and hundreds of proteins to be studied in a single experiment. DNA microarray technology has revealed the role of many regulatory factors in global regulatory networks in *Escherichia coli*, *Bacillus subtilis*, and other model bacteria.<sup>15</sup> However, in many cases, the complexity of the interactions between regulons makes it difficult to distinguish between direct and indirect effects on transcription. Another high-throughput experimental approach, the ChIP-on-chip technique (see section 3.1), is increasingly used for investigation of the genome-wide DNA binding of global TFs in bacteria.<sup>16–20</sup> The wide-ranging proteomic approach was used to assess phosphate-starvation response in *Vibrio cholerae*, the iron regulatory network in *Rhizobium leguminosarum*, and the bacteroid proteins network in *Bradyrhizobium japonicum*.<sup>21–23</sup> Finally, recent advances in tandem mass spectrometry and development of the powerful computational algorithms enable *de novo* shotgun sequencing of protein mixtures, thus providing another promising approach for high-throughput protein expression analysis.<sup>24–25</sup>

A constantly growing number of complete prokaryotic genomes allows computational biologists to extensively use comparative genomic approaches to predict *cis*-acting regulatory elements (TFBSs and RNA elements) and to reconstruct TRNs in bacteria.<sup>12,13,26–29</sup> The major directions of this analysis involve analysis and description of previously known regulons in uncharacterized organisms and *ab initio* prediction of novel regulons. Finally, the comparative analysis of regulons combined with other techniques of genome context analysis (see section 3.3) helps to significantly improve the quality and accuracy of functional gene annotations and predict novel genes in a variety of pathways.

The focus of this review is on novel approaches to the analysis of bacterial regulons, including the methods of identification of TFBSs and RNA regulatory elements based on comparative genomics (section 2). It provides a summary of several major studies on the computational reconstruction of certain TRNs in bacteria and an overview of Web-accessible databases of microbial TFs and their TFBSs (section 3). Finally, I discuss the likely evolutionary scenarios for bacterial regulons and the balance of conservation and flexibility in the composition of TRNs among species (section 4).

## 2. Computational Methods for Identification of Regulatory Motifs

### 2.1. Structure, Function, and Representation of Transcription Factor Binding Sites

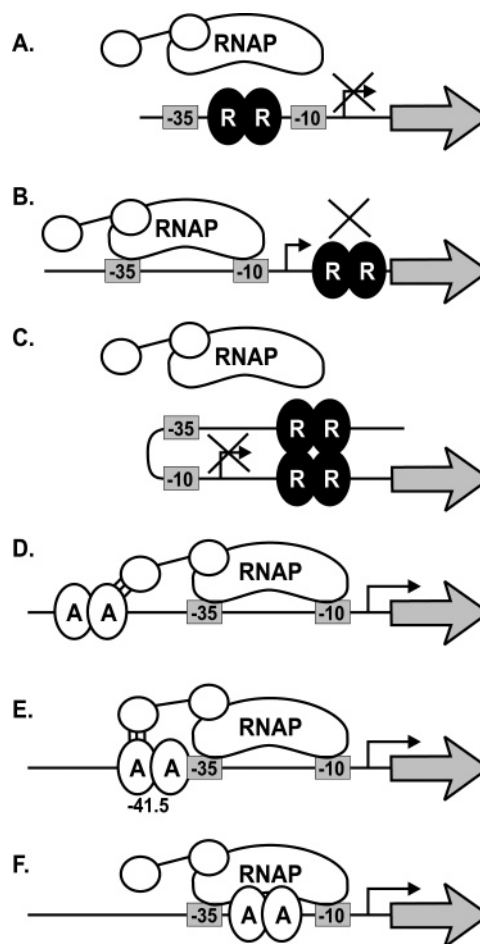
#### 2.1.1. Position of TFBSs in Promoter Regions

TFs regulate gene expression via specific binding to DNA sequences (or operators) located in promoter regions. The DNA-binding affinity and activity of TFs could be modulated by various signals including interaction with small ligands or covalent modification (e.g., phosphorylation by a specific sensor kinase). When a TF binds to an operator, it can either activate or repress transcription initiation.<sup>30</sup> In bacteria, there are TFs that act solely as repressors or as activators, whereas some other TFs have a dual regulatory role in gene expression. Positive or negative effects of such dual TFs depend on the position of the operator site within the target promoter region.

Most repressor sites are located between  $-60$  and  $+60$  relative to the transcriptional start site, suggesting that repression by steric hindrance of RNA polymerase binding to the promoter is the most common regulatory mechanism.<sup>31–33</sup> Alternatively, repressors may act by blocking transcription elongation or by looping DNA in the promoter region (Figure 1). The degree of repression depends significantly on the operator site position relative to the promoter.<sup>34</sup> Analysis of the data for various negatively acting regulators shows large variability in the relative positions of operators and promoters for each regulon. This variation in the repressor site position is in contrast to the relatively fixed positions of activator sites. Activators promote gene expression by binding to an operator that is located either upstream of or adjacent to the promoter  $-35$  element and by recruiting RNA polymerase to the promoter by direct protein–protein interaction (Figure 1). For example, the global catabolic activator Crp in *E. coli* binds operators, which have a preference to be centered at positions  $-62.5$ ,  $-72.5$ , or  $-92.5$  at Class I promoters or at position  $-41.5$  at Class II promoters.<sup>35</sup> Some activators (e.g., those from the MerR family) bind at or near the promoter elements and alter the conformation of the promoter to allow its interaction with RNA polymerase.<sup>30</sup>

#### 2.1.2. Structure of TFBSs

The size of a single TFBS usually varies between  $\sim 12$  and  $30$  nt, with the most common length being  $16$ – $20$  nt. Since TF proteins often recognize and bind to DNA as homodimers or homomultimeric protein complexes, the TFBSs usually possess an intrinsic symmetry. Cooperative binding of transcription factors to DNA plays an important role in regulating gene expression, ensuring a sigmoid response to the concentration of effector. Inverted repeats



**Figure 1.** Mechanisms of regulation by transcription factors in prokaryotes. A, repression by steric hindrance; B, repression by blocking of the transcription elongation; C, repression by DNA looping; D, class I activation; E, class II activation; F, activation by conformation change. RNAP, A, and R indicate RNA polymerase, activator, and repressor proteins, respectively. Promoter elements are shown by “ $-35$ ” and “ $-10$ ” boxes. Thin and thick arrows indicate transcription start sites and target genes, respectively. At class I promoters, the activator is bound to an upstream site and contacts the  $\alpha$  subunit of RNAP, thereby recruiting the polymerase to the promoter. At class II promoters, the activator binds to a target that is adjacent to the promoter (in most cases at position  $-41.5$  relative to the transcription start site), and the bound activator interacts with the  $\sigma 70$  subunit of RNAP.

(palindromes) and direct repeats are the most common structures of TFBSs. Some homomultimeric TFs cooperatively bind more complex TFBSs composed of both inverted and direct repeats (e.g., AraC in *E. coli*). However, in contrast to eukaryotes, complex regulatory cassettes containing heteromultimeric TFBSs are rare in prokaryotes. On average, TFBSs with dyad symmetry are predominant in bacteria. Although the spacing between two repeats may vary significantly, it is usually a specific value for a given TF. For instance, the distance between two half-sites in direct repeats is often a multiple of the length of one DNA helix turn ( $10.5$  nt). Examples of TFBS structures found in bacteria are given in Table 1.

#### 2.1.3. TFBS Consensus and Logo Sequences

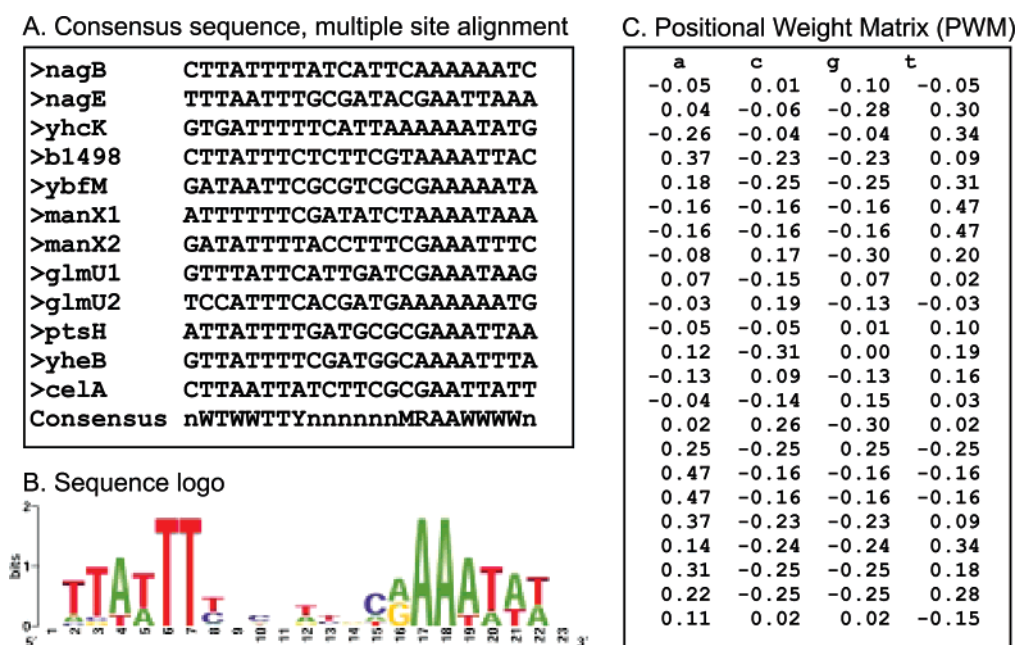
TFs bind to their DNA motifs in regulatory regions in a sequence-specific manner. However, the binding sites of a particular TF located upstream of different genes in the same genome could vary significantly, allowing for a more flexible



**Table 1. Symmetry of Some Bacterial Transcription Factor Binding Sites**

transcription factor	consensus half site <sup>a</sup>	type <sup>b</sup>	site size, nt	site structure <sup>c</sup>
biotin repressor BirA <sup>70</sup>	WTGTAAACC	IR	32–34	→ 14–16 nt ←
cAMP receptor protein Crp <sup>184</sup>	WWWGTGTA	IR	22	→ 6 ←
catabolic control protein CcpA <sup>186</sup>	WTGWAASC	IR	16	→ 0 ←
arabinose activator AraC <sup>120</sup>	YAGCNKNWNWRTCCATA	DR	38	→ 4 →
gluconate repressor GntR <sup>119</sup>	SWATGTTACC	IR	20	→ 0 ←
xylose repressor XylR <sup>121</sup>	GTTWGTTWWW	IR	21	→ 3 ←
heat shock repressor HrcA <sup>122</sup>	TTAGCACTC	IR	27	→ 9 ←
NAD repressor NadR <sup>69</sup>	TGTTTA	IR	18	→ 6 ←
nickel repressor NikR <sup>234</sup>	GTATGA	IR	27–28	→ 15–16 nt ←
methionine repressor MetJ <sup>262</sup>	RRACRTMY	DR	24	→ 0 → 0 →
iron repressor RirA <sup>240</sup>	SWTGA	IR	19	→ 9 ←

<sup>a</sup> Degenerate nucleotide designations are M (A or C), W (A or T), R (A or G), K (T or G), S (G or C), and Y (T or C), and N stands for any nucleotide. <sup>b</sup> The types of symmetry are inverted repeats (IR) and direct repeats (DR). <sup>c</sup> Arrows and numbers show the respective orientation of the half sites and the distance between them.



**Figure 2.** Representation of transcription factor binding sites. (A) Alignment of NagC binding sites in *E. coli*<sup>203</sup> and the derived consensus sequence. (B) Sequence logo representation generated by the WebLogo tool (<http://weblogo.berkeley.edu>). The relative height of the letters represents the frequencies of nucleotides at each position measured in bits of information. (C) PWM for the NagC binding motif, where the respective positional weights were calculated using the following formula:<sup>129</sup>  $W_{b,k} = \log(N_{b,k} + 0.5) - 0.25 \sum_{i=A,T,G,C} \log(N_{i,k} + 0.5)$ , where  $N_{b,k}$  is the count of nucleotide  $b$  in position  $k$ .

transcriptional control. This degenerate nature of most TFBSs is in contrast with much more strict conservation of recognition sequences for restriction enzymes. Different DNA-binding properties of TFs and restriction enzymes have an important biological meaning. Restriction enzymes need to have an all or none activity to protect the cell against phages and viruses, whereas TFs may have sites with different sequences and different affinities to regulate gene expression at different levels.<sup>36</sup>

The consensus sequence was commonly used to describe the DNA binding site specificity of TFs and generally refers to a sequence that matches all of the example sites closely but not necessarily exactly (Figure 2A). The number of mismatches allowed for the consensus sequence can be decreased by using the degenerate consensus sequence. This description of TFBSs uses an extended alphabet to show variable or degenerate nucleotides. For instance, Y stands for C or T (pYrimidine), R stands for A or G (puRine), W stands for A or T (Weak), and S stands for C or G (Strong). The sequence logo is a more precise graphic representation of the patterns within a multiple sequence alignment of

TFBSs (Figure 2B). The logo displays the frequencies of nucleotides at each position as the relative heights of letters (A, T, G, and C), along with the degree of sequence conservation as the total height of a stack of letters, measured in bits of information.<sup>37,38</sup>

#### 2.1.4. Positional Weight Matrices

A nucleotide frequency positional weight matrix (PWM) representation of the sites is an alternative to logo and consensus sequences (Figure 2C).<sup>39</sup> A PWM could be constructed by aligning known TFBS sequences, e.g., by using the program CONSENSUS.<sup>40</sup> The PWM-based approach is more sensitive and more precise for TFBS recognition than the consensus-based methods.<sup>41</sup>

Several methods have been proposed to determine the positional nucleotide weights for any particular collection of sites.<sup>42,43</sup> The method introduced by Staden<sup>44</sup> is very similar to current methods. In this method, the weights are calculated as the negative logarithms of the frequencies of each nucleotide at each position. Thus, the sum of weights for any particular site is the negative logarithm of the

probability of observing that particular sequence in the collection of known sites. Schneider defined and used the “information content” of binding sites on nucleotide sequences to calculate the amount of information that is required to locate the sites, given that they occur with some frequency in the genome.<sup>45</sup> Using statistical mechanics theory, it was shown that the information content is related to the average binding energy for the collection of sites.<sup>46</sup>

### 2.1.5. Search for TFBSs in Microbial Genomes

Computational algorithms for searching for potential TFBSs in genomic sequences most often use PWMs to evaluate the resemblance of any DNA sequence to a given TFBS pattern.<sup>39</sup> The score for a candidate TFBS sequence is calculated as the sum of the respective weights for each position. Any sequence with a score that is higher than the predefined cutoff is considered as a potential TFBS. One limitation of the PWM approach is the assumption that the positions in the site contribute additively to the total activity.

With the constructed PWM, one can scan the whole genome and find additional genes that share the same DNA signal within their potential regulatory regions. Various programs, including PATSER<sup>40</sup> and MAST,<sup>47</sup> allow scanning sets of DNA sequences to identify potential TFBSs by using the generated PWMs. The Genome Explorer software for bacterial genome analysis provides tools for both genome-wide identification of TFBSs and comparison of gene sets in several genomes.<sup>48</sup>

Finally, it is instructive to mention one curious example of genomic identification of TFBSs by experimental biologists who have ignored available TFBS search tools. In this work, the authors used the Microsoft Word 2000 Find tool to identify putative binding sites of the ferric uptake regulator Fur in the genome of *Staphylococcus aureus* by using the consensus sequence and by inserting the “any character” function to enable mismatches.<sup>49</sup>

## 2.2. Repertoire of Transcription Factors in Prokaryotic Genomes

### 2.2.1. Distribution of TFs in Microbial Genomes

The estimated number of DNA-binding transcription factors varies in different microorganisms depending on their genome size, lifestyle, and habitat. Earlier literature analysis and similarity searches in *E. coli* K12 suggested that close to 7.5% of genes (around 300–350) encode TFs.<sup>50</sup> A collection of 237 candidate TFs was identified in another model microorganism, *B. subtilis*.<sup>33</sup> The smaller number of TF genes in *B. subtilis* could be explained by the higher number of RNA attenuators, in particular riboswitches, that contribute significantly to the regulation of numerous fundamental metabolic pathways in Gram-positive bacteria.<sup>51</sup>

Analyses of other bacterial genomes revealed a reasonable correlation between the number of TFs and the genome size.<sup>33,52–55</sup> The number of DNA-binding TFs was recently assessed in all sequenced organisms through homology-based prediction using profile hidden Markov models (HMMs) of domains and collected in the DBD database.<sup>56</sup> Thus, it is limited to factors that are homologous to those HMMs. The collection of HMMs was taken from two existing databases (PFAM<sup>57</sup> and SUPERFAMILY<sup>58</sup>) and is limited to models that include TFs that specifically recognize TFBSs.<sup>56</sup> Using

the DBD database, we chose 230 prokaryotic organisms and plotted the number of TFs per genome against the total number of open reading frames (ORFs), which serves as an indicator of genome size in prokaryotes (see Table 1 of the Supporting Information). Various taxonomic groups of bacteria and archaea showed the similar trend of a nearly linear increase in the number of TFs starting from ~1500 ORFs per genome (Figure 3A). Obligate pathogens and endosymbionts with a small genome size (less than ~1500 ORFs, see Table 1 of the Supporting Information and Figure 3B) have a much lower proportion of TFs (average 1%) compared to free-living and facultative pathogenic microorganisms (average 4.5%). The presence of only a few TFs in intracellular parasites (Chlamydia, Rickettsiales), symbiotic  $\gamma$ -proteobacteria (e.g., *Buchnera*), and some obligate pathogens (Mycoplasma and Spirochetes) is consistent with a reductive genome evolution that led to the loss of genes not essential for life within the host.<sup>59</sup>

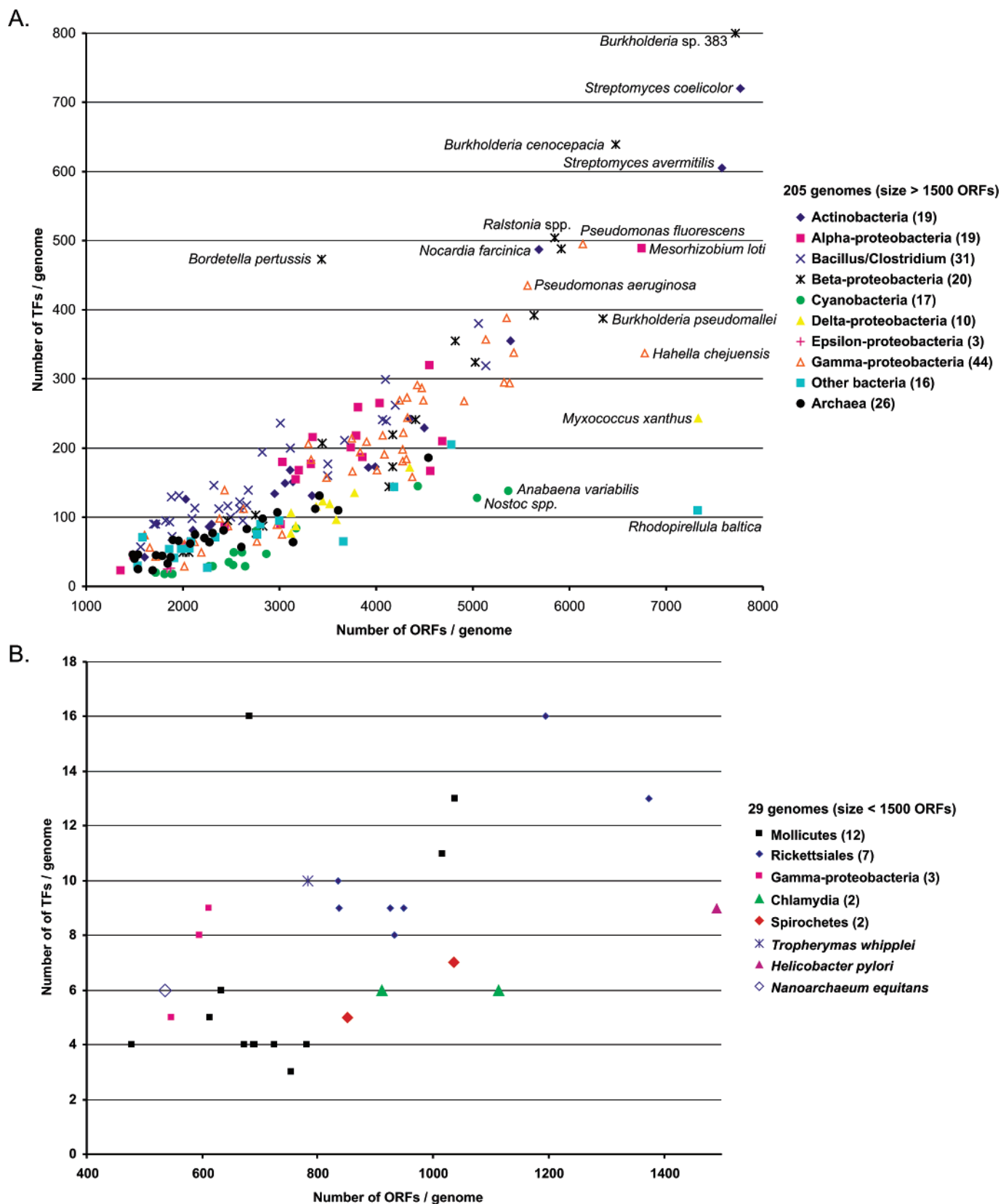
Environmental properties and metabolic capabilities of microorganisms strongly influence the proportion of TFs encoded in their genomes (Table 1 of the Supporting Information). Complex lifestyles and metabolic versatility require a higher number of TFs to better coordinate a response to changing conditions. Pathogenic and nonpathogenic bacteria from the same taxonomic group usually have similar proportions of TFs (e.g., see *Vibrio cholerae* vs *V. fischeri*, *Bacillus cereus* vs *B. subtilis*, and *Pseudomonas aeruginosa* vs *P. fluorescens*). Several free-living bacterial species with large genomes (*Myxococcus xanthus*, *Rhodospirillum rubrum*, *Anabaena*, and *Nostoc* spp.) have characteristically small proportions of TFs (Figure 3A). This discrepancy is compensated by the presence of complex regulatory pathways that involve serine–threonine protein kinases and sensor histidine kinases linked to  $\sigma^{54}$  activators, as well as many specialized  $\sigma$  factors.<sup>60–62</sup>

### 2.2.2. Families of TFs

Known microbial transcription factors are classified in at least 50 protein families based on the sequence similarity of their DNA-binding domains.<sup>33,52,54</sup> Major protein families that contain a large number of TF representatives with various regulatory roles are listed in Table 2. The largest known family of TFs is LysR, followed by AraC and TetR. The distribution of TFs by families varies among different species; e.g., the LysR protein family is the most abundant in  $\alpha$ - and  $\gamma$ -proteobacteria,<sup>50</sup> whereas the MarR family is the largest in *B. subtilis*,<sup>33</sup> and the IclR family is over-represented in *Bordetella* species.<sup>63</sup>

The number of TF families detected in archaea is significantly lower than that in bacteria: 19 families are shared by bacteria and archaea, whereas 33 families of bacterial TFs were not found in archaea.<sup>54</sup> Significant divergence of transcription regulatory systems in bacteria and archaea could be explained by emergence of novel TF families after divergence of two kingdoms. However, in contrast to a large number of bacteria-specific TF families, there is only one known family of archaea-specific TFs, HTH-10. One possible reason for this fact is the limiting number of experimental studies on transcriptional regulation in archaea.

Almost half of the characterized TF protein families contain regulators with only one characterized functional role/specificity.<sup>54</sup> Examples of TFs with unique functional roles include ArgR, MetJ, and TrpR for arginine, methionine, and



**Figure 3.** Number of TFs in prokaryotic genomes against the total number of ORFs per genome. Predicted TFs are from the DBD database.<sup>56</sup> The different taxonomic groups listed in the right inset are represented by dots of different form and color. The number of genomes in each taxonomic group is given in parentheses. (A) Plot for 205 prokaryotic genomes with size more than 1500 ORFs. (B) Plot for 29 genomes of obligate pathogens and symbionts with size less than 1500 ORFs.

tryptophan metabolism, respectively; BirA and NadR for biotin and NAD biosynthesis; HrcA and LexA for heat shock and SOS responses; ModE and NikR for molybdenum and

nickel homeostasis; and NrdR for deoxyribonucleotide synthesis. Some of these protein families are universally distributed among most bacterial species (e.g., NrdR, BirA,

**Table 2. Major Families of Transcription Factors in Prokaryotes**

family	TF examples	PFAM <sup>a</sup>	count <sup>b</sup>	functional roles of regulated genes	mode <sup>c</sup>	Pos. <sup>d</sup>
AraC	MelR, RhaS, XylR, MarA, SoxS, RhrA	PF00165	6954	carbohydrates utilization, stress response, iron siderophore uptake	A	C
ArsR	CadC, CzcA, NmtR, SmtB, ZiaR	PF01022	2064	homeostasis of transition metals (Cd, Co, Zn, Ni, Zn, As, Pb)	R	M
AsnC	Lrp, BkdR, PutR	PF01037	1527	amino acid metabolism	A, R	N
Cro	Cro, CI, CopR, Xre	PF01381	5258	bacterial plasmid copy number control	R	N
Crp	Fnr, Dnr, NtcA, PrfA, CooA, HcpR, ArcR	PF00325	891	anaerobic switch, catabolic repression, stress response, nitrogen metabolism	A (R)	C
DeoR	GlpR, AgaR, IolR	PF08220	915	carbohydrates utilization	R	N
Fis	NtrC, NifA, NorR, FhlA, TyrR, PrpR	PF02954	2843	nitrogen, amino acid, and secondary metabolism, flagella ( $\sigma$ 54-dependent)	A	C
Fur	Zur, Mur, Nur, Irr, PerR	PF01475	888	metal ion homeostasis (Fe, Zn, Mn, Ni), peroxide stress	R (A)	N
GntR	AraR, ExuR, DgoR, TreR, FadR, HutC, CitR, PdhR, BioR	PF00392	4293	carbohydrates, fatty acid and amino acid utilization, biotin metabolism	R	N
IclR	KdgR, PcaR, AilR, MhpR	PF01614	1122	sugar acids and aromatic compounds utilization, secondary metabolism	R (A)	N
LacI	GalR, CcpA, CytR, NagR, ScrR, PurR	PF00356	2000	carbohydrates utilization, catabolite repression, purine metabolism	R	N
LuxR	RhlR, TraR, ComA, NarP, NarL, FixJ	PF00196	3706	quorum sensing, competence, nitrogen metabolism, anaerobic switch	A (R)	C
LysR	IlvY, CysB, MetR, CynR, NodD, AmpR	PF00126	9421	secondary metabolism and amino acid biosynthesis	A (R)	N
MarR	SlyA, PecS, AdcR, BadR, HucR	PF01047	3280	antibiotic resistance, virulence, zinc uptake, aromatic compounds utilization	R	M
MerR	GlnR, TnrA, SoxR, BmrR, CueR, CadR, PbrR, ZntR	PF00376	2337	nitrogen metabolism, response to stress, multidrug efflux, heavy metal resistance (Hg, Cu, Cd, Pb, Zn)	A, R	N
OmpR	ArcA, PhoB, CiaR, ToxR, VirG	PF00486	5010	OM porins, respiration, phosphate metabolism, competence, virulence	A	C
ROK	NagC, XylR, Mlc	PF00480	1198 <sup>c</sup>	carbohydrates utilization	R	N
RpiR	HexR	PF01418	636	carbohydrates utilization	R	N
Rrf2	IscR, NsrR, RirA	PF02082	818	FeS cluster, iron, nitrogen metabolism	R	N
TetR	AcrR, QacR, FabR, RutR, BioQ	PF00440	6190	antibiotic resistance, fatty acids, pyrimidine, and biotin metabolism	R	N

<sup>a</sup> The PFAM database (<http://www.sanger.ac.uk/Software/Pfam/>)<sup>57</sup> identification number of the TF protein family is indicated. <sup>b</sup> The total number of proteins in the PFAM family across all prokaryotic genomes is indicated. The total number of proteins in the ROK family includes both TFs and sugar kinases that do not have a DNA-binding domain. <sup>c</sup> The following modes of regulation are indicated: A, activator; R, repressor; A (R), mostly activator; R (A), mostly repressor. <sup>d</sup> The position of the DNA-binding domain in the TF protein is indicated: C, C-terminal; N, N-terminal; M, central.

ArgR), whereas others are restricted to certain taxonomic groups (e.g., TrpR and MetJ in  $\gamma$ -proteobacteria, NadR in enterobacteria). Importantly, representatives of these unique TF families are usually present in one copy per genome.

Two-component signal transduction regulatory systems are widely used by prokaryotic cells to transmit and propagate a wide variety of environmental and intracellular signals. These regulatory systems typically comprise a sensory histidine kinase and the cognate response regulator.<sup>64</sup> Phosphorylation of the Asp residue in the N-terminal receiver domain of the latter regulatory component induces conformational changes, allowing it to form dimers and bind to DNA operators. Most DNA-binding domains in response regulators belong to the OmpR, LuxR/NarL, Fis/NtrC, and LytR families. The large number of paralogous subfamilies of histidine kinases and response regulators includes a repertoire of recently evolved signaling genes, which may reflect selective pressure to adapt new environmental conditions.<sup>65</sup> Both lineage-specific gene family expansion and horizontal gene transfer play major roles in the appearance of novel two-component regulatory systems.

### 2.2.3. Domain Architecture of TFs

Structural analysis revealed that the helix–turn–helix (HTH) signature is the most common DNA-binding motif present in all major prokaryotic TF families (Table 2). A

“recognition”  $\alpha$ -helix in the HTH motif forms specific contacts with DNA by fitting into the DNA major groove. Other structural DNA-binding motifs, including zinc-finger (e.g., in Ros<sup>66</sup>), zinc-ribbon (e.g., in NrdR), and antiparallel  $\beta$ -sheets (e.g., in MetJ), are much less abundant in prokaryotes. The position of the DNA-binding domain within the polypeptide (e.g., N-terminal, C-terminal, or central) is conserved within a TF family but may vary between families (Table 2). The position of the DNA-binding domain correlates with a positive or negative mode of regulation by TF: N-terminal DNA-binding domains are consistently present in repressors, whereas activators usually have C-terminal DNA-binding domains.<sup>67,68</sup>

In addition to DNA-binding domains, transcriptional regulators possess domain(s) involved in dimerization and/or sensing of particular environmental stimuli. Most of these non-DNA-binding TF domain families are found exclusively within TFs. However, some of them are shared with proteins of distinct cellular functions, i.e., periplasmic substrate-binding proteins of ABC transporters (LacI family of TFs) or sugar kinases (ROK family of TFs). Interestingly, regulators of the biotin and NAD metabolic pathways in *E. coli* (BirA and NadR, respectively) are bifunctional proteins with N-terminal HTH domains and C-terminal enzymatic domains, which allows them to contribute to both biochemical transformations and gene expression of the respective metabolic pathways.<sup>69,70</sup>



### 2.2.4. Global and Local TFs

Global transcription factors are defined as regulators that control more than 20 different genes in different transcriptional units and are usually involved in a number of distinct pathways.<sup>50</sup> Major global regulators in *E. coli* are the cAMP receptor protein Crp; the anaerobic regulators Fnr and ArcA, the leucine-responsive regulator Lrp, the histone-like DNA-binding proteins Ihf, Fis, and Hns, the iron-responsive regulator Fur, and the nitrite response regulator NarL.<sup>71</sup> Global regulators identified in *B. subtilis* include the growth phase transition factors AbrB and CodY, the carbon catabolic protein CcpA, the late competence regulator ComK, the regulator of initiation of sporulation Spo0A, and the nitrogen assimilation regulator TnrA.<sup>33</sup>

Local TFs usually regulate one or several transcriptional units encoding proteins from the same metabolic pathway. There is a tendency of genes encoding local TFs to cluster with TF-regulated genes on the chromosome (e.g., to form an operon or divergon).<sup>72,73</sup> It is quite common when a gene cluster involved in the utilization (catabolism) pathway for a specific compound also includes a local TF gene providing a specific transcriptional control of this gene cluster in response to this compound. For example, the sugar-specific repressors from the LacI family in *E. coli* (e.g., LacI, GalS, GntR, MalI, RbsR, and TreR) are encoded by the same cluster together with target genes involved in utilization of the specific sugar (lactose, galactose, gluconate, maltose, ribose, and trehalose, respectively).<sup>73</sup> Other examples of local TFs include specific regulators of biosynthetic pathways for cofactors (e.g., NadR, BirA in *E. coli*), amino acids (e.g., MetJ, ArgR, TrpR, TyrR), nucleotides (e.g., PurR), uptake transporters for essential metals (e.g., Zur, ModE, MntR, NikR), and specific stress or drug response pathways (LexA, OxyR, AcrR, MarR).

### 2.2.5. Alternative Sigma Factors

Bacterial sigma factors are an essential component of RNA polymerase and determine promoter selectivity.<sup>74</sup> The regulon of a single sigma factor can be comprised of hundreds of genes. The  $\sigma^{70}$  subunit of RNA polymerase in *E. coli* specifies transcription from promoters that are responsible for basal gene expression during vegetative growth. Sigma factors can be classified into two structurally unrelated families: the  $\sigma^{70}$  and the  $\sigma^{54}$  families.

The first family includes primary sigma factors (e.g., *E. coli*  $\sigma^{70}$ , *B. subtilis*  $\sigma^A$ ) and related alternative sigma factors that mediate transcription initiation of various sets of genes in bacteria. For instance, RpoH ( $\sigma^H$ ) transcribes the genes of the heat shock response regulon. The main regulatory role of FlhA ( $\sigma^{28}$ ) in many bacterial species is to transcribe the genes required for flagellar synthesis and bacterial motility. The sigma factors  $\sigma^B$  in Gram-positive bacteria and RpoS ( $\sigma^S$ ) in Gram-negative bacteria are functionally similar to each other in that they are responsible for stationary phase and stress response gene expression. Many alternative sigma factors also play an important role in bacterial pathogenesis by regulating the expression of virulence-associated genes (e.g.,  $\sigma^S$  and  $\sigma^{28}$  in *Salmonella*).

RpoN ( $\sigma^{54}$  or  $\sigma^N$ ) has several features that distinguish it from other sigma factors: it is not homologous to other sigma subunits,  $\sigma^{54}$ -dependent expression absolutely requires an activator, and the activator binding sites can be far from the transcription start site.<sup>75</sup> A physiological theme for  $\sigma^{54}$ -dependent genes has not yet emerged, as the regulated genes

described to date control a wide diversity of processes including nitrogen assimilation, uptake and catabolism of amino acids, secondary metabolism, and virulence.

Extracytoplasmic function (ECF) sigma factors are a phylogenetically distinct subfamily within the  $\sigma^{70}$  family. ECF sigma factors are small TFs that, upon receiving a stimulus from the environment, are released and can bind to RNA polymerase to stimulate transcription of a specific group of genes.<sup>76</sup> The number and functional roles of ECF sigma factors encoded in bacterial genomes are highly variable. For example, FecI in *E. coli* and PvdS in *P. aeruginosa* are involved in iron siderophore synthesis and uptake, whereas  $\sigma^W$  in *B. subtilis* and the orthologous sigma factor  $\sigma^E$  in *E. coli* control intrinsic resistance to antimicrobial compounds, heavy metals, and oxidative stress.

The number and diversity of sigma factor genes per genome are related to the environmental variation allowing growth for a given species. The distribution of three classes of sigma factors ( $\sigma^{70}$ ,  $\sigma^{54}$ , and ECF) in bacterial genomes was determined using HMM profiles based on experimentally verified sigma factors (see Table 2 in the Supporting Information).<sup>77</sup> Most bacteria species have either one or no  $\sigma^{54}$  genes. However, there is a larger divergence in the number of genes from the  $\sigma^{70}$  family. For example, the large genomes of *Streptomyces* species have 14  $\sigma^{70}$  genes, while the rest of the actinobacteria have far less. In cyanobacteria, there is quite a high number of  $\sigma^{70}$  genes (5 to 8); the same is true for the sporulating bacilli and clostridia. ECF sigma factors are generally far more numerous than the other two classes, but since they are not essential, they are missing in many organisms. *Streptomyces* species have 45 ECF sigma factors that are mainly involved in the control of secondary metabolism. *Bacteroides thetaiotaomicron* currently holds the record with 48 predicted ECF genes. Other genomes with a high amount of ECF sigma factors are *M. xanthus* (31), *R. baltica* (29), and *Pseudomonas fluorescens* (28). Interestingly, the first two species have a relatively small proportion of normal TFs in their large genomes.

## 2.3. Databases of Microbial TFs and TFBSs

With the increasing amount of information on transcriptional regulation in bacteria, many public databases specializing in microbial regulation are becoming available (Table 3). These include web resources specializing in transcriptional regulatory networks in model microorganisms, such as *E. coli* (RegulonDB<sup>78</sup>), *B. subtilis* (DBTBS<sup>79</sup>), Corynebacteria (CoryneRegNet<sup>80</sup>), and *Mycobacterium tuberculosis* (MtbRegList<sup>81</sup>). These databases compile an arsenal of TFs with their regulated genes as well as their recognition TFBS sequences, which were experimentally characterized. In addition to integration of the published experimental data on gene regulation, these resources provide genome-scale computational predictions of operons, promoters, TFBSs, and regulons. Use of these resources helps us to propose new regulatory hypotheses for wet-lab verification.

Several databases provide information about known and predicted TFs in multiple microbial genomes. The DBD database<sup>56</sup> classifies TFs by DNA-binding domain protein families. The ExtraTrain database<sup>82</sup> provides the distribution of the 16 largest families of TFs in microbial genomes. The AraC, TetR, and IclR families of TFs were reviewed and analyzed in detail<sup>83–85</sup> and integrated in the BacTregulators database.<sup>86</sup> The Sentra database<sup>87</sup> of signal transduction proteins lists manually curated two-component histidine



**Table 3. Databases for Microbial TFs and TFBSs**

name	URL	description	ref
RegulonDB	<a href="http://regulondb.ccg.unam.mx/">http://regulondb.ccg.unam.mx/</a>	DB of transcriptional regulation (TFs, TFBSs) in <i>E. coli</i> (literature data and predictions)	78
DBTBS	<a href="http://dbtbs.hgc.jp">http://dbtbs.hgc.jp</a>	DB of transcriptional regulation (TFs, TFBSs) in <i>B. subtilis</i> (literature data and predictions)	79
CoryneRegNet	<a href="http://www.cebitec.uni-bielefeld.de/groups/gi/software/coryneregnet/">http://www.cebitec.uni-bielefeld.de/groups/gi/software/coryneregnet/</a>	DB of TFs and TRNs in corynebacteria	80
MtbRegList	<a href="http://mtbreglist.dyndns.org/MtbRegList/">http://mtbreglist.dyndns.org/MtbRegList/</a>	DB for analysis of gene expression and regulation data in <i>Mycobacterium tuberculosis</i>	81
cTFbase	<a href="http://ceg wz.com/">http://ceg wz.com/</a>	DB for comparative genomics of TFs in cyanobacteria	88
DBD	<a href="http://transcriptionfactor.org">http://transcriptionfactor.org</a>	DB of TF and families prediction (all genomes)	56
ExtraTrain	<a href="http://www.era7.com/ExtraTrain">http://www.era7.com/ExtraTrain</a>	DB of extragenic regions and TFs in prokaryotes	82
BacTregulators	<a href="http://www.bactregulators.org/">http://www.bactregulators.org/</a>	DB of TFs in prokaryotes (specific TF families)	86
Sentra	<a href="http://compbio.mcs.anl.gov/sentra">http://compbio.mcs.anl.gov/sentra</a>	DB of sensory signal transduction proteins	87
PRODORIC	<a href="http://prodoric.tu-bs.de">http://prodoric.tu-bs.de</a>	DB of prokaryotic gene regulation (several specific organisms)	89
RegTransBase	<a href="http://regtransbase.lbl.gov">http://regtransbase.lbl.gov</a>	DB of TFBSs and regulatory interactions in prokaryotes (literature data and predictions)	90
TRACTOR	<a href="http://www.tractor.lncc.br/">http://www.tractor.lncc.br/</a>	DB of TRNs and TFBSs in $\gamma$ -proteobacteria	91

kinases and response regulators encoded in completely sequenced prokaryotic genomes. The cTFbase database classifies TFs identified in 21 cyanobacterial genomes and provides a resource for comparative analysis of putative TFs in cyanobacteria.<sup>88</sup>

A prokaryotic database of gene regulation, PRODORIC,<sup>89</sup> integrates different types of data including regulators and TFBSs, promoter structures, and operon and regulon organization by screening the original literature. Another manually curated database of gene regulation in prokaryotes, RegTransBase,<sup>90</sup> captures experimental knowledge on regulatory sequences and interactions published for a variety of microorganisms. In addition, these two web resources provide a set of tools to predict and compare TFBSs in multiple genomes. The TRACTOR database<sup>91</sup> contains comparative genomic predictions of new members of 74 known *E. coli* regulons in the genomes of  $\gamma$ -proteobacteria.

## 2.4. Computational Tools for Discovery of TFBSs in the Genomes

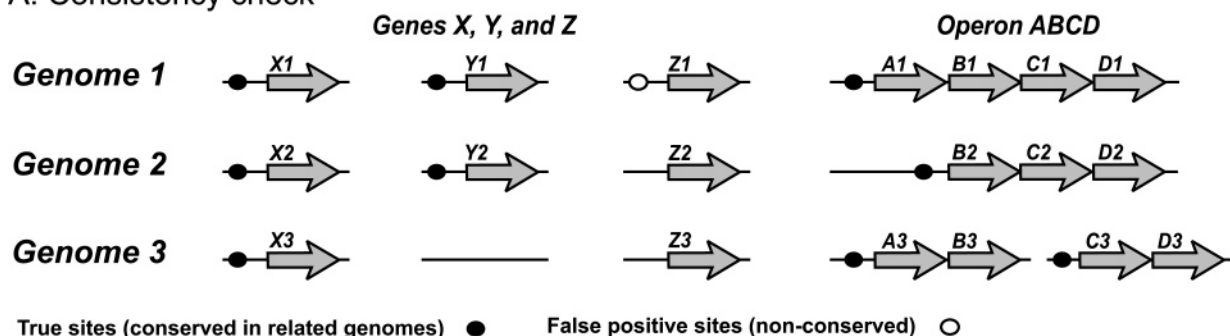
A number of computational methods have been developed for identification of candidate TFBSs.<sup>92,93</sup> These methods are subdivided into the consensus-building and PWM-based approaches. The consensus-based approaches, including the “word counting” and exhaustive enumeration algorithms, are more useful for motif finding in eukaryotic regulatory regions which, in contrast to bacteria, mostly include composite regulatory signals.<sup>94–97</sup> In PWM-based approaches, the specificity of the protein is represented as a matrix rather than the consensus sequence, allowing the binding site pattern to be identified. All these methods identify a common regulatory motif from multiple DNA fragments. The input training set of regulatory regions might be composed based on many sources for possibly coregulated genes, including microarray experiments, gene knockout experiments, and functional classes of genes that form a common metabolic pathway from the literature.<sup>98</sup>

There are many different implementations of PWM-based algorithms, and the most popular of them are outlined below. Detailed operation principles, technical data, and URLs of 13 different PWM-based tools were recently reviewed by Tompa and coauthors.<sup>99</sup> In this study, the authors conducted comparative assessment of these computational tools and estimated their accuracy and correctness for TFBS discovery in various input settings and data sets.<sup>99</sup>

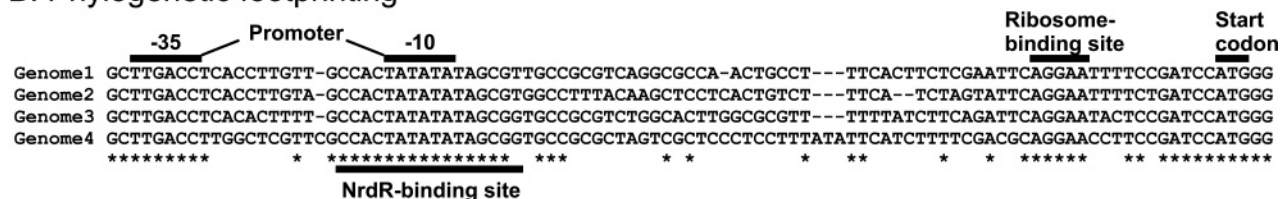
One of the first algorithms builds up a multiple alignment of the sites by adding new sites at each iteration and identifies the best alignment with the highest information content.<sup>100</sup> Expectation-maximization (EM) methods simultaneously optimize the PWM description of a motif and the binding probabilities for its associated sites.<sup>101</sup> One popular implementation of the EM algorithm, MEME, performs a single iteration for each site in the target sequence, selects the best motif from this set, and then iterates only that one to convergence.<sup>102</sup> SignalX is another EM-based program that uses an iterative procedure of clustering all weak palindromic sequences in the training set of DNA fragments to identify a palindromic signal of a given length with the highest information content.<sup>103</sup> A Gibbs-sampling algorithm is a stochastic implementation of the EM method that samples the space of all multiple alignments of small sequence segments in search of the one that is most likely to consist of samples from a common PWM.<sup>104</sup> The AlignACE program,<sup>105</sup> the Gibbs Recursive Sampler,<sup>106</sup> and the SeSimCMC program<sup>107</sup> are variants of the Gibbs sampling algorithm optimized for finding multiple distinct TFBS motifs within a single set of unaligned DNA fragments.

Another modification of the PWM-based approach uses the fact that many TFs in bacteria bind to a palindromic motif with intrinsic symmetry. This symmetry-based approach was applied to single bacterial genomes to predict novel regulatory DNA sequence motifs represented by PWMs.<sup>108,109</sup> The algorithm identifies all statistically significant palindromes in upstream intergenic regions and groups overrepresented sites into clusters of similar patterns. The set of PWMs constructed based on these patterns was used to scan the genome for additional candidate sites and to infer putative regulons. In the model species *E. coli* and *B. subtilis*, many derived clusters represent characterized regulatory motifs, whereas the large group of the remaining PWMs is likely to describe uncharacterized TFBSs.<sup>108,109</sup> A similar dimer-based approach was used by a different research group to predict 2497 regulatory motifs (PWMs) in the genome of *Streptomyces coelicolor*.<sup>110</sup> Functional analysis of genes located downstream of these DNA motifs identified several motifs that may be biologically significant as regulatory elements. These include a DNA motif found preferentially in UTRs immediately upstream of genes involved in polysaccharide degradation and sugar transport.<sup>110</sup>

### A. Consistency check



### B. Phylogenetic footprinting



**Figure 4.** Comparative genomic approaches for TFBSs identification. (A) Consistency check of the candidate TFBSs in a group of genomes. First, all UTRs in the genomes are scanned by the constructed PWM to identify candidate TFBSs. Then, the predicted TFBSs are differentiated based on their conservation in other genomes. False positive sites usually are not conserved in related genomes with orthologous TFs. Accounting for changes in operon structure in different genomes (gene loss and split and fusion of operons) increases the rate of predicted true positive sites. (B) Phylogenetic footprinting of orthologous UTRs on the example of the *nrdA* gene in *Pseudomonas* species. Highly conserved DNA regions that correspond to the NrdR-binding site, candidate  $-35$  and  $-10$  promoter elements, and the ribosomal binding site are shown by thick lines.

## 2.5. Comparative Genomic Approaches for Identification and Verification of TFBSs

Identification and recognition of TFBSs in genomic sequences is an old problem in computational molecular biology. Until this section we considered sets of coregulated genes from one genome and assumed that the 'UTRs of these genes contain a common TFBS motif. The problem of identification of additional TFBSs for known TFs was addressed in many studies, conducted mostly in model organisms such as *E. coli* and *B. subtilis*.<sup>111–118</sup> In these studies, either a consensus sequence or a PWM constructed on the experimentally known sites was used to scan the genome of interest in order to predict novel regulon members. However, even for relatively well-studied regulons, it is difficult to set thresholds reliably distinguishing between true and false sites. Besides, our ability to construct good recognition rules is severely impaired by the limited availability of experimental data on TFBSs. The availability of hundreds of bacterial genomes opens opportunities for using comparative genomic approaches to identify conserved functionally important sites (e.g., TFBSs, promoters, RNA regulatory sites) by genomic comparison of different species.

### 2.5.1. Consistency Check Approach

The presence of orthologous TFs in the analyzed microbial genomes is a prerequisite to the comparative analysis of their regulons. Furthermore, selection of genomes for the analysis depends on conservation of a TFBS signal between species. Very closely related genomes (e.g., different strains of the same species) usually have almost identical intergenic regions that do not permit getting rid of false positives. On the other hand, regulatory signals are usually poorly conserved or are at least highly divergent in distant taxonomic groups (e.g., between Gram-positive firmicutes and Gram-negative proteobacteria). Finally, possible changes in the operon struc-

tures of the coregulated genes need to be taken into account while comparing the sets of genes with a common regulatory motif. The consistency-check comparative approach was successfully applied for the prediction and verification of regulatory sites for many TFs in various taxonomic groups of bacteria and archaea.<sup>9,69,70,119–127</sup> An overview of these and other comparative studies of microbial regulons for reconstruction of regulatory networks is outlined in the next section of this review.

The consistency-check comparative approach is based on the assumption that regulons (sets of coregulated genes) have a tendency to be conserved between the genomes that contain orthologous TFs.<sup>128,129</sup> Therefore, the presence of the same TFBS upstream of orthologous genes is an indication that it is a true regulatory site, whereas TFBSs scattered at random in the genome are considered false positives (Figure 4A). Simultaneous analysis of multiple phylogenetically related genomes allows one to make reliable predictions of TFBSs even with weak recognition rules. The consistency check sharply increases the specificity of predictions, although it may lose species-specific members of regulons. This technique not only allows the transfer of data on regulatory interactions from well-studied genomes to newly sequenced ones, but it also makes it possible to find additional members of regulons and map novel regulons.

### 2.5.2. Phylogenetic Footprinting Approach

The phylogenetic footprinting approach identifies regulatory elements by finding highly conserved regions in a set of DNA sequences located upstream of orthologous genes from multiple species.<sup>130</sup> The term “phylogenetic footprint” was first introduced to describe several conserved *cis*-regulatory elements in primates.<sup>131</sup> The simple assumption of the method is that functional DNA sequences (such as TFBSs) diverge more slowly than nonfunctional ones (spac-

ers). Identification of conserved regulatory elements by this method requires a certain degree of phylogenetic relatedness of the analyzed upstream regions of orthologous genes (or orthologous UTRs). The standard approach uses a global multiple alignment of the candidate orthologous UTRs to identify a conserved region in the alignment (Figure 4B). It should be noted that the identification of a novel conserved regulatory element does not automatically reveal a TF that could recognize this site. This important problem of assignment of any identified TFBS to the corresponding TF is discussed in more detail in section 2.5.

The reliability of the phylogenetic footprinting method depends critically on the selection of species for the analysis. If the species are too closely related, the alignment is not informative. On the other hand, if they are too distant, it is difficult or impossible to construct an accurate alignment. The problem of species selection and the number of species required for the phylogenetic footprinting analysis of TFBSs was addressed using a set of 166 *E. coli* genes with experimentally identified TFBSs and genomic data from nine additional  $\gamma$ -proteobacteria.<sup>132</sup> It was found that just three species were sufficient for accurate motif predictions of TFBSs and that an appropriate phylogenetic distance between species is an important factor to consider.

In the case when many closely related genomes are available, one can use various computational tools for multiple alignments of bacterial genomes for mapping of potential TFBSs. The Menteric server (<http://globin.bx.psu.edu/enterix/>) is a visualization tool for bacterial genome alignments designed for *E. coli* and related enterobacteria.<sup>133</sup> The VISTA family of computational tools (<http://genome.lbl.gov/vista/>) provides precomputed full scaffold alignments for both microbial and eukaryotic genomes.<sup>134</sup> The phylogenetic shadowing approach was developed to compute and statistically evaluate conservation profiles of multiple sequence alignments from closely related species.<sup>135</sup> Micro-FooterPrinter is a phylogenetic footprinting program for discovering conserved *cis*-regulatory elements in prokaryotic genomes.<sup>136</sup>

### 2.5.3. Genome-Wide Application of Comparative Approaches

With the increasing number of sequenced genomes, phylogenetic footprinting approaches are becoming very popular tools of TFBS discovery. The quality of the TFBS prediction by phylogenetic footprinting can be substantially improved by combining this approach with the existing motif discovery tools (such as MEME, AlignACE, and Gibbs sampling). Several algorithms based on such a combination of phylogenetic footprinting and motif discovery tools have been developed for eukaryotic genomes, including PhyloGibbs<sup>137</sup> and PhyME.<sup>138</sup> Here, I will outline key studies that use this combined approach for identification of regulatory elements in bacterial genomes. In these studies, the motif discovery tools are applied to a training set of orthologous UTRs across species. I will illustrate these studies by multiple examples when the predicted TFBSs and regulons became validated in follow-up experiments.

The cross-species comparison of orthologous UTRs in *E. coli* and eight related  $\gamma$ -proteobacteria by the Gibbs-sampling algorithm revealed a large set of conserved DNA motifs (for almost 2000 *E. coli* genes), many of which coincide with documented TFBSs.<sup>139</sup> In the follow-up study, application of a Bayesian motif-clustering algorithm to the previously

predicted by Gibbs sampling  $\gamma$ -proteobacterial motifs led to accurate identification of many experimentally reported *E. coli* regulons (for example, PurR, LexA, MetJ, Crp, TrpR, NtrC, Mlc, and ModE), prediction of their additional members, and identification of novel regulons.<sup>140</sup> The Bayesian motif-clustering algorithm is based on an explicit statistical model that describes the relationship between the observed motifs and the putative regulons and a Markov chain Monte Carlo computational method.<sup>81</sup> Several novel regulons identified in *E. coli* by the combined comparative genomic approach<sup>80,81</sup> were later experimentally confirmed. These include fatty acid biosynthesis regulon FabR (previously YijC)<sup>141</sup> and novel ribonucleotide reductase regulon NrdR (YbaD).<sup>142</sup>

In genome-wide analysis of eight  $\alpha$ -proteobacteria,<sup>143</sup> the recursive Gibbs-sampling algorithm was applied to a set of orthologous upstream regions, and the resulting motifs were filtered and clustered into regulons by the Bayesian motif-clustering algorithm.<sup>140</sup> The phylogenetic footprinting approach allowed the authors to identify 101 putative regulons in *Rhodopseudomonas palustris*. Among them are several regulons of particular interest: the FixK, NnrR, NtrC, and RpoN regulons related to nitrogen metabolism; the hydroperoxide stress OhrR regulon; the DNA damage response LexA regulon; the flagellar synthesis FlhD regulon; and the photosynthetic PspR regulon.

Another comparative study of three *Bacillus* species using a local pairwise alignment program has detected nearly 1900 phylogenetically conserved elements in the upstream intergenic regions of  $\sim 1500$  *B. subtilis* genes.<sup>144</sup> Subsequent clustering of these genes according to the motif similarity allowed the authors to predict 154 different DNA motifs; each of those possibly coregulates a specific set of genes. Many of these motifs correspond to the previously described regulatory elements in *B. subtilis*, including various TFBSs (e.g., CtsR, CcpA) and RNA attenuators (e.g., S-box, T-box). The authors tentatively identified several new members of known regulons (e.g., *dnaJ* in CtsR) and many potential regulons that were not yet reported. One of these novel regulons, a hypothetical xanthine regulon for the *purE*, *xpt*, and *pbuG* genes, was later described to operate by a novel type of a metabolite-responsive riboswitch, the guanine-responsive G-box.<sup>51</sup>

In a comparative genomic study of two related groups of Gram-positive bacteria, lactobacilli and bacilli, clusters of orthologous transcriptional units were first identified, and the conserved DNA motifs were determined for two species sets using the MEME algorithm.<sup>145</sup> These motifs were subsequently used to scan the upstream regions using the MAST program, and nearly 200 conserved motifs in each set of species were selected. Many of the predicted motifs from bacilli and lactobacilli were very similar, including several well-described regulatory motifs (e.g., T-box, CIRCE, LexA-box). Interestingly, this method revealed 18 lactobacilli-specific candidate regulatory motifs including 13 that had not been described previously. The PhyloScan algorithm was developed to increase the flexibility and sensitivity of scanning for potential TFBSs and to decrease false-positive site predictions using cross-species evidence.<sup>146</sup>

The regulon detection by PhyloScan combines the evidence from matching sites found in orthologous data from several related species with the evidence from multiple sites within intergenic regions. The statistical significance of the TFBS predictions is calculated directly, without employing



training sets. Application of the PhyloScan algorithm to seven Enterobacteriales genomes allowed authors to identify several novel TFBSs for global transcription factors Crp and PurR in *E. coli*.

The Regulogger computational approach discriminates true regulon members from false-positive predictions on the basis of conservation of regulons across multiple genomes.<sup>147</sup> To quantify the degree of conservation of putative TFBSs, the Regulogger calculates for each predicted regulon member a relative conservation score using the fraction of orthologs that are preceded by the same candidate TFBS. Regulon members that have orthologs with conserved candidate TFBSs are considered true-positive predictions, and such a set is defined as a regulog. Application of Regulogger to the genome of *Staphylococcus aureus* and six related Gram-positive bacteria identified 125 high-scoring regulogs, many of which are consistent with previously characterized regulons (e.g., TnrA, Fnr, Fur, CtsR, LexA). Some of these regulogs correspond to the highly conserved regions within the known RNA regulatory elements (e.g., T-box, *THI* riboswitch). The regulog approach also predicted novel members of known regulons and revealed novel potential regulons. One of the predicted regulogs containing various ribonucleotide reductase genes was later investigated in detail and shown to operate by the novel transcription regulatory system NrdR for the ribonucleotide reductase genes in most bacterial lineages.<sup>87,142,148</sup>

## 2.6. Interconnection of Transcription Factors and Their DNA Motifs

Many putative TFs in prokaryotes have been identified only on the basis of their homologies and are still uncharacterized with regard to their cognate DNA-binding motifs, sets of target genes (regulons), and effectors. New candidate TFBSs discovered by computational approaches such as phylogenetic footprinting and clustering<sup>139,140</sup> may be connected to particular TFs using a combination of different types of evidence such as (i) positional clustering of TFBSs and TFs on the chromosome;<sup>72</sup> (ii) correlation in the phylogenetic pattern of co-occurrence of TFBSs (the presence or absence of a regulon) and TFs (the presence or absence of a candidate TF gene) in the genomes;<sup>148</sup> and (iii) binding specificity constraints for TFs having structurally similar DNA-binding domains.<sup>149,150</sup> Tan et al.<sup>151</sup> combined these types of information to calculate the probability of a given TF–TFBS pair and predicted many new connections between uncharacterized TFs and candidate DNA motifs in *E. coli*. Positional evidence of the first type provides the strongest impact on the assignment of a TF to its DNA sites. This is not surprising, since bacterial TFs are often auto-regulated<sup>71</sup> and the genes encoding TFs tend to colocalize on the chromosome with the genes they regulate.<sup>151</sup> For instance, in many local sugar utilization regulons, the target genes preceded by upstream TFBSs are located adjacent to the regulatory gene encoding the corresponding TF.

Conservation of the gene neighborhood is very useful not only for functional annotation of enzymes and transporters<sup>152</sup> but also to predict the cellular and biological processes that TFs potentially regulate.<sup>72</sup> However, some known TF genes, mostly those whose products have more than one target TFBS in the genome, are located remotely from their target genes (e.g., FruR and PurR in *E. coli*). Another limitation of the positional approach (especially if applied to a small group of species) is illustrated in the example of the NrdR

regulatory system for ribonucleotide reductase genes. Based on the conserved positional clustering with riboflavin biosynthesis genes in most proteobacteria, the hypothetical gene *ybaD* in *E. coli* was originally annotated as a regulator of riboflavin biosynthesis.<sup>72,153</sup> However, a subsequent comparative genomic study<sup>148</sup> using phylogenetic co-occurrence patterns of TFs and TFBSs in combination with the phylogenetic footprinting approach assigned a different role of a universal regulator of the deoxyribonucleotide metabolism (named NrdR) to the YbaD protein family. An extended positional analysis of NrdR sites allowed identifying several cases of colocalization of *nrdR* genes with target ribonucleotide reductase genes in other bacterial lineages, e.g., in Actinobacteria.<sup>148</sup> The predicted regulatory role of NrdR was finally confirmed in experiments conducted in *Streptomyces* species.<sup>154</sup>

## 2.7. Analysis of RNA Regulatory Elements

Various RNA regulatory systems including riboswitches, translational attenuators, T-boxes, and RNA-binding proteins have been described in bacteria.<sup>7</sup> The main mechanisms involved in regulation by *cis*-regulatory RNAs are based on the formation of alternative mRNA structures that either terminate transcription (terminators) or inhibit initiation of translation (sequesters). Different classes of RNA elements use different mechanisms to sense the concentration of a metabolite. Typically, an effector-responsive protein factor specifically binds the *cis*-regulatory RNA that is rather small and simple in structure (e.g., the tryptophan-responsive TRAP protein in *B. subtilis*). A unique class of RNA elements, T-boxes in Gram-positive bacteria, interacts directly with specific uncharged tRNAs to promote expression of target genes in response to amino acid concentrations. Riboswitches are widespread RNA elements with a complex structure that directly sense metabolites and control gene expression of related metabolic pathways.<sup>155,156</sup> Each riboswitch class is defined by a core of conserved base-paired elements and consensus nucleotides at specific positions and is highly specific to its cognate effector metabolite. Among various metabolites detected by known classes of riboswitches are vitamins (coenzyme B<sub>12</sub>, thiamin pyrophosphate, and flavin mononucleotide), amino acids (lysine, glycine, and *S*-adenosylmethionine), and nucleotides (adenine, guanine, and queuosine).

A high level of conservation of primary and secondary structures of riboswitches and T-boxes is remarkable and very useful for their identification by comparative genome analysis. Various classes of riboswitches that regulate the cobalamin, thiamin, riboflavin, lysine, methionine, and queuosine biosynthesis pathways were discovered by comparative genomic analysis<sup>10,157–162</sup> and experimentally characterized by in-line probing assays.<sup>162–168</sup> Representatives of 13 known classes of riboswitches identified in prokaryotic genomes are available within the Rfam database.<sup>169</sup> RibEx,<sup>170</sup> RegRNA,<sup>171</sup> and Riboswitch finder<sup>172</sup> Web tools were designed to search any input sequence for the presence of known regulatory RNA elements.

Discovery of new classes of RNA motifs and riboswitches in orthologous UTRs of genes is an interesting computational challenge. Comparison of UTRs between species resulted in identification of many novel RNA motifs with extensive sequence and secondary-structure conservation.<sup>173–176</sup> Some of these RNA motifs were experimentally validated (e.g., two novel *S*-adenosylmethionine riboswitches, queuosine

riboswitch).<sup>162,174–177</sup> Since the target genes for several other classes of new RNA motifs are mostly hypothetical, the effector molecules and the mechanism of regulation for these putative RNA regulatory elements remain unknown.

### 3. Reconstruction and Comparison of Regulatory Networks That Control Central Metabolism in Bacteria

During the past decade, the number of studies that use integrative genomic approaches for the analysis of regulons and metabolic pathways has substantially increased. Various techniques of genome context analysis, including chromosomal gene clustering, protein fusions, and occurrence profiles are extremely useful for metabolic reconstruction and functional gene annotation (see ref 152 for a review). In this section, the key principles and practical steps of genomic-based reconstruction of regulatory networks in bacteria are outlined. In the first part, single-microorganism studies of TF regulons that combine both high-throughput experimental approaches (such as expression profiling) and the genomic identification of TFBSs are outlined. The second part of this section summarizes comparative genomic studies describing TFBSs identification and reconstruction of TF regulons in complete microbial genomes. Finally, the last part illustrates the power of the comparative analysis of regulons for metabolic reconstruction and functional predictions, including novel functional roles in metabolic pathways, candidates for missing genes, and specificities of transporters.

#### 3.1. Combining Experimental and Genomic Data to Predict TFBS Motifs

DNA microarray technology detects changes in mRNA levels under different conditions and is extensively used for the analysis of transcriptional responses in bacteria.<sup>13</sup> Expression profiling allows thousands of genes in the cell to be studied simultaneously in a single experiment. By comparing gene expression under different conditions or between different genetic backgrounds (e.g., a gene knockout mutant vs a wild-type strain), one can identify a set of genes with the same pattern of expression, which could be potentially controlled by the same TF. However, because of experimental and biological variability, the interpretation of DNA microarray data is often ambiguous.<sup>178</sup> Technical imperfections of the method include random biological variations, sample handling errors, and measuring errors. Furthermore, covariation of expression level alone does not automatically imply that the corresponding genes form a single regulon (i.e., a set of genes directly controlled by a single TF). More accurately, such genes may be considered as a part of a so-called modulon (i.e., a set of genes either directly or indirectly controlled by a certain regulatory system).

The combination of chromatin immunoprecipitation (ChIP) and high-density microarrays, also known as the ChIP-on-chip technique, has been widely exploited to investigate interactions between eukaryotic proteins and their DNA targets *in vivo*.<sup>179,180</sup> The method is based on capturing of protein–DNA interactions by chemical cross-linking and filtering them out using antibodies specific to the protein of interest. The enriched DNA population is then labeled and applied to DNA microarrays to detect enriched signals. In bacteria, ChIP-on-chip was successfully used for whole-genome identification TFBSs for global TFs, such as CtrA in *Caulobacter crescentus*, Crp and Fnr in *E. coli*, and Spo0A

in *B. subtilis*.<sup>16–20</sup> In comparison with the DNA microarray approach, ChIP-on-chip avoids complications due to genes indirectly controlled by a TF or genes that are regulated by multiple TFs. However, it also has an important limitation due to its inability to detect all TF–DNA interactions, which may be caused by inefficient cross-linking at some locations. For example, ChIP-on-chip analysis of the Fnr regulon in *E. coli* identified 63 binding target sites, including several novel targets and missing many previously validated targets.<sup>17</sup>

The development of high-throughput experimental techniques has allowed the generation of vast amounts of data related to TRNs. These data combined with the information on known TRN structures from databases and the literature have opened the way for genome-scale reconstruction of microbial TRNs.<sup>12</sup> The matrix formalism was introduced to represent a series of regulatory rules for the individual genes of a TRN in a matrix form.<sup>181</sup> In this form, the state of a gene is represented as either transcribed or not transcribed in response to regulatory signals. The matrix formalism allows for the systematic characterization of functional states of transcriptional regulatory systems and facilitates the computation of the transcriptional state of the genome under given environmental condition.<sup>181</sup> The consistency between known TRNs and gene-expression data in *E. coli* is influenced by both the structural features of the network and the functional classes of genes involved in TRNs.<sup>182</sup>

The increased availability of high-throughput data will further improve the prospects of TRN reconstruction, and additional data types can be used to resolve inconsistencies. For instance, a large-scale mapping of *E. coli* TRNs inferred from a compendium of 445 *E. coli* Affimetrix expression arrays and 3216 known *E. coli* regulatory interactions from RegulonDB<sup>78</sup> was performed by the context likelihood of relatedness algorithm, allowing prediction of 1079 regulatory interactions (with a 60% true positive rate), of which one-third were in the previously known TRN and two-thirds were novel predictions.<sup>183</sup>

Computational identification of TFBSs in the genomes, combined with the gene expression data, improves the determination of bacterial regulons and allows one to distinguish between direct and indirect effects of a certain TF on the gene regulation. Many specific regulons were analyzed using high-throughput transcriptome comparisons between wild-type and TF-knockout strains of a single bacterial species and were supported by the genomic identification of candidate binding sites for the respective TFs. These include many global regulatory systems, such as Crp, ArcA, NarL, Fnr, and Fur in *E. coli*,<sup>113,114,184,185</sup> CcpA, Fnr, and TnrA in *B. subtilis*,<sup>115,117,186</sup> and Fur in *Shewanella oneidensis*<sup>187</sup> and *Yersinia pestis*,<sup>188</sup> as well as some specific regulons, such as the SOS response system LexA in *B. subtilis*,<sup>189</sup> the iron-responsive systems DtxR in *Corynebacterium glutamicum*,<sup>146,190</sup> and the system Irr in *Bradyrhizobium japonicum*.<sup>191</sup>

Comparison of gene expression between TF knockout mutant and wild-type strains, subsequent selection of differentially regulated genes, and comparative analysis of the corresponding upstream gene regions help to accurately predict candidate TFBSs. For example, analysis of the CodY regulon in *Lactococcus lactis* revealed a novel overrepresented motif in the upstream regions of genes derepressed in the *codY* mutant strain. This motif was confirmed to function as a high-affinity CodY-binding site using electrophoretic mobility shift and nuclease protection assays.<sup>192</sup> In

another example, the *C. glutamicum* sulfur metabolism regulon McbR was analyzed by the same approach, resulting in the identification and experimental verification of a consensus McbR binding site.<sup>193</sup> Whereas the DNA microarray detected 86 genes with enhanced transcription in the *mcbR* mutant strain, the genomic analysis identified candidate McbR-binding sites upstream of 22 genes and operons, suggesting that the transcription of at least 45 genes involved in the sulfur metabolism is directly controlled by the McbR repressor. The remaining genes, which showed an enhanced expression in the *mcbR* mutant but which are not part of the McbR regulon, are likely the subject of an indirect coregulation.

Similar conclusions were obtained by comparing the ArcA and Fnr modulons and regulons in *E. coli* that are involved in global anaerobic respiration control.<sup>194</sup> The data about modulon composition were taken from two microarray studies of *arcA* and *fnr* mutants of *E. coli*,<sup>195,196</sup> whereas regulons were predicted by TFBS search and comparison between *E. coli* and related enterobacteria.<sup>194,197</sup> The Fnr and ArcA modulons were defined as sets of genes with at least a 2-fold change in expression and included 151 and 135 operons, respectively. However, in these groups of *E. coli* operons, candidate Fnr- and ArcA-binding sites were determined in the regulatory regions of 38 and 23 operons, respectively. It was concluded that the Fnr–ArcA regulatory cascade and additional regulatory systems significantly expand the respiratory modulons in comparison with the respective regulons.<sup>194</sup>

Another novel technique, which combines *in vitro* runoff transcription with microarray analysis (ROMA), was used to analyze the  $\sigma^W$  regulon in *B. subtilis*.<sup>198</sup> Comparison of *in vivo* transcriptional profiling, ROMA, and consensus search approaches showed that these methods are complementary to each other and that each tends to miss some sites. Maximal coverage in the definition of a bacterial regulon was obtained by combining all three approaches. In a similar study of the *E. coli* Crp regulon,<sup>35</sup> the *in vivo* and *in vitro* transcription profiling methods were combined with Crp-binding site determination. Comparison of the ability of each of these methods to identify known members of the Crp regulon demonstrates that the site-search approach prevails over *in vivo* transcription profiling. The main reason of the failure to identify many Crp-regulated genes using microarrays is the complexity of the Crp regulon, where the Crp-activated promoters are dependent on the presence of additional regulators in response to a specific substrate.

In contrast to *E. coli* and *B. subtilis*, specific regulatory mutants are rarely available for many other species. Nevertheless, the combination of both hierarchical clustering of microarray data under different conditions and TFBS-finding approaches is an efficient approach for describing novel regulons. Mao et al.<sup>199</sup> investigated the photosynthetic regulons PrrA and PpsR and the anaerobic regulon FnrL in *Rhodobacter sphaeroides* by detection of genes that share similar expression patterns under photosynthetic and/or anaerobic conditions and by identification of possible TFBS motifs that may be involved in their coregulation. This approach allowed the authors to find and improve FnrL- and PpsR-binding motifs and to predict a candidate TFBS motif for the photosynthetic response regulator PrrA.

Finally, integration of *in silico* genomic approaches with *in vitro* and *in vivo* experimental methods helps identify novel regulatory systems in poorly characterized microorgan-

isms. For instance, transcriptional regulation of the glycolytic genes in the hyperthermophilic archaea *Pyrococcus* and *Thermococcus* was elucidated by experimental determination of the transcription initiation sites and computational comparison of the promoter regions.<sup>200</sup> This analysis of thermococcal archaea revealed a potential TFBS motif within 20 glycolytic promoters and a candidate regulator from the TrmB family, which is likely involved in recognition of this DNA motif. Only a limited number of regulons have been characterized experimentally in Archaea.<sup>201</sup> The above-described and other *in silico* genomic studies<sup>103</sup> demonstrated an importance of the genomic approaches for analysis of archaeal regulons.

### 3.2. Comparative Genomic Reconstruction of Regulatory and Metabolic Networks

A general strategy to analyze known regulons consists of the following steps: (i) search for orthologous TFs to reveal phylogenetic distribution of the regulon, (ii) obtain binding-site models from known sites in a model genome(s), (iii) obtain sets of orthologous upstream gene sequences from genomes at the appropriate phylogenetic distance, (iv) apply pattern recognition programs, (v) construct PWMs and search for additional sites in the genomes of interest, and (vi) perform consistency check or cross-species comparison of the predicted members of the regulon (see Figure 5A). The last step schematically represented in Figure 4A is very important for the comparative approach, which is based on the assumption that regulatory events tend to be conserved in closely related species with orthologous regulators. Thus, conservation of a candidate regulatory site upstream of orthologous genes in a group of genomes is used to eliminate false-positive site predictions. The consistency-check stage requires special attention to the selection of a group of genomes for comparison and to a threshold for the TFBS search. Also, to account for possible differences in the operon structures of orthologous genes, it needs an accurate operon prediction for the candidate regulon members.

Depending on the availability of experimental data, the training set for signal identification may be obtained in different ways. In the simplest situation, the training set is composed of experimentally known TFBSs that have been defined in model species, such as *E. coli* or *B. subtilis* (strategy Ia). In the absence of such knowledge, the training set may be composed of candidate regulatory regions of genes that are known to be controlled by a given TF in model species (strategy Ib). The accuracy of the *de novo* identification of a regulatory signal depends on the number of sequences in the training set and may be improved by inclusion of orthologous upstream regions from related species.

To identify novel regulons in the absence of any experimental data about regulation of specific genes, two alternative comparative genomic strategies could be used (see Figure 5B). The subsystem-oriented approach (strategy IIa) is based on the assumption that the genes from the same metabolic pathway may be coregulated by one TF. This approach starts with the identification of a set of functionally linked genes within the taxonomic group of interest (e.g., genes from the same metabolic pathway). First, all possible operons including the genes of interest are defined and the corresponding upstream UTRs are collected. Then, the collection of candidate regulatory regions is used by signal-recognition programs to predict a common DNA pattern allowing a limited number of input sequences to be excluded from the



**A. Strategy I: Known TFs****Ia. Known TFBS model**

Collection of known TFBSs from model species

Pattern construction  
(CONSENSUS,  
SignalX)Positional  
Weight  
MatrixApplication of pattern  
recognition program  
(MEME, AlignACE,  
SignalX, SeSiMCMC)Scanning the analyzed genomes for TFBSs  
(PATSER, MAST, Genome Explorer)**Ib. Unknown TFBS model**

Collection of genes that are known to be controlled by a TF in model species

Search for orthologous genes  
Collection of candidate TF-regulated orthologs in the analyzed genomes

Training set of orthologous upstream gene regions in the analyzed genomes

Sets of candidate TFBSs scored above the cutoff in the analyzed genomes

Known TF in model species

Search for orthologous TFs  
Phylogenetic distribution of the analyzed TF regulonConsistency check  
accounting for operons  
(see Fig. 3A)

Conserved members of TF regulon

**B. Strategy II: Putative TFs****IIa. Subsystem-oriented approach**

Defining of a taxonomic group of species, and a group of genes involved in the metabolic pathway (subsystem) of interest

Search for orthologous genes

Collection of orthologous operons involved in the same metabolic pathway (subsystem)

Training set of orthologous upstream gene regions in the analyzed genomes

Pattern recognition program

Positional Weight Matrix

Scanning the analyzed genomes for additional TFBSs + consistency check

Attribution of a candidate TF to the novel regulatory motif

**IIb. Phylogenetic footprinting**

Collection of orthologous upstream regions of genes of interest in the taxonomic group of closely related species

Analysis of individual genes:  
Multiple alignment of upstream regions (see Fig. 3B)

Selection of highly conserved regions with intrinsic symmetry

Pattern construction

Positional Weight Matrix

Scanning the analyzed genomes for additional TFBSs + consistency check

Attribution of a candidate TF to the novel regulatory motif

Genome-wide analysis:

Pattern recognition program (Gibbs Sampler, MEME, AlignACE)

Grouping of cross-species motifs into regulons (Bayesian motif clustering)

Reconstruction of novel regulon (a candidate TF, a DNA motif, a set of genes)

**Figure 5.** Schematic representation of two strategies for comparative genomic reconstruction of regulons. (A) Strategy I for analysis of known regulons with experimentally determined TFs. Known TFBSs are collected to construct a PWM, which is used to scan the genomes for additional sites. If the TFBS model is unknown, the set of upstream regions of known TF-regulated genes and their orthologs in other genomes is collected and used as an input for TFBS pattern recognition programs and a PWM construction. (B) Strategy II for discovery of novel regulons operating by previously unknown TFs. In the subsystem-oriented approach, the training set for the TFBS recognition program includes upstream regions of genes from the same metabolic pathway in the defined taxonomic group of bacteria. Phylogenetic footprinting identifies highly conserved regions in multiple alignments of upstream gene regions across the closely related species that are used to construct a PWM to search for additional TFBSs in the genomes.

pattern. In the next step, the genomes of interest are scanned with the constructed DNA pattern to reveal the distribution of similar sites, which are verified by the consistency-check procedure.

An alternative approach for discovery of novel regulons is based on the phylogenetic footprinting method (strategy IIb). Orthologous upstream UTRs of a gene of interest are collected from a group of closely related genomes and used to construct a multiple alignment. The group of genomes is selected based on the presence of orthologous target genes and on the extent of conservation of UTRs. In an ideal case, the multiple sequence alignments contain several highly conserved regions that are broken by relatively unconserved regions. These islands of conservation in UTRs are obvious candidates to serve as promoters or *cis*-regulatory sites. Since most prokaryotic TFs bind DNA as homodimers recognizing symmetrical sites, the analyzed conserved regions might be inspected for the presence of either inverted or direct repeats with allowable mismatches. Candidate TFBS regions are used to construct search profiles. In the next stage, these TFBS regions are verified by genome-wide searches for similar sites in intergenic regions of analyzed species. The combination of the phylogenetic footprinting approach with clustering of

predicted TFBSs can help to identify novel regulons. For example, the fatty acid biosynthesis regulon FabR in *E. coli* was first identified by this combined *in silico* approach<sup>139</sup> and then experimentally validated.<sup>141</sup> Finally, a novel predicted TFBS motif could be connected to a specific TF using positional genomic evidence, phylogenetic co-occurrence profiles, and binding specificity constraints (see section 2.6).

The above strategies of identification of TFBSs were successfully applied to analyze many regulons involved in the central metabolism of sugars, amino acids, nucleotides, metals, and cofactors as well as important regulons controlling respiration, nitrogen metabolism, and stress response (Table 4). The combination of metabolic maps and regulatory networks shows many species- and taxon-specific differences in the structure of metabolic pathways and regulons in bacteria. Several representative examples that illustrate the power of these comparative genomic approaches for discovery and characterization of microbial regulons are outlined in Table 5 and are briefly discussed below.

### 3.2.1. *N*-Acetylglucosamine and Chitin Utilization

The NagC regulon for *N*-acetylglucosamine and chitin utilization was initially characterized in *E. coli*<sup>202</sup> and further

**Table 4. Transcription Factors and Regulons Analyzed by Comparative Genomics Methods**

regulated metabolic pathway	regulon <sup>m</sup>	phylogenetic distribution <sup>n</sup>	strategy <sup>o</sup>
<b>Sugar Utilization Pathways</b>			
pectin <sup>119, 206</sup>	KdgR	γ (Ent, Vib)	Ia
chitin <sup>203</sup>	NagC	γ (Ent, Vib)	Ia
	NagR*	γ (Alt, Xan)	IIa
	NagQ*	γ (Pse), β, α	IIa
glucuronate <sup>119</sup>	UxuR	γ (Ent, Pas)	Ib
gluconate <sup>119</sup>	GntR	γ (Ent, Vib)	Ib
arabinose <sup>120, 121</sup>	AraC, AraR	γ (Ent), BCl	Ia
xylose <sup>120, 121</sup>	XylR	γ (Ent, Pas), BCl	Ib
ribose <sup>120, 121</sup>	RbsR	γ, BCl	Ib
rhamnose <sup>a</sup>	RhaS, R1*, R2*, R3*	γ, α, BCl, Act	Ia, IIa
glycerol <sup>b</sup>	GlpR	γ (Ent, Vib, Pse)	Ib
<b>Metal Homeostasis</b>			
iron <sup>123, 240, 241, c</sup>	Fur	γ (Ent, Vib, Pse), δ, α	Ia, Ib
	Irr, RirA	α	Ib
	IdeR	Act	Ia
zinc <sup>124</sup>	Zur	γ (Ent, Vib), α	Ib
manganese <sup>240</sup>	Mur, MntR	α	Ib
nickel <sup>123, 241</sup>	NikR	γ, β, α, δ, ε, BCl, Arc	Ia, Ib
molybdenum <sup>241, d</sup>	ModE	γ, δ, CFB, Arc	Ia
heavy metal resistance <sup>260</sup>	CueR, CadR, HmrR, PbrR	γ, β, α, BCl	Ia
<b>Cofactors and Amino Acid Metabolism</b>			
NAD <sup>69, e</sup>	NadR	γ (Ent)	Ia
	YrxA	BCl	IIa
	NadQ*	γ, β	IIa
	NrtR**	Cya, Act	IIa
biotin <sup>70, 208, a</sup>	BirA	γ, β, ε, BCl, Arc	Ia, Ib
	BioR*	α	IIa
	BioQ*	Act	IIa
aromatic amino acids <sup>125, 129</sup>	TyrR, TrpR	γ (Ent, Vib, Pas)	Ia
arginine <sup>9, 129</sup>	ArgR	γ, BCl, TM	Ia
<b>Nitrogen Metabolism</b>			
nitrogen assimilation <sup>127, f, g</sup>	NtcA	Cya	Ia
	TnrA, GlnR	BCl	Ia
	NtrC	γ (Ent, Vib, Pse), α	Ia
nitrogen fixation <sup>103, 215, g</sup>	NifA	α	Ia
	NrpR	Arc	IIa
denitrification <sup>216</sup>	Dnr, NnrR	γ, β, α	Ib
nitrogen oxides respiration <sup>126</sup>	NarP	γ (Ent, Pas, Vib)	Ib
nitrogen oxides detoxification <sup>216</sup>	NsrR	γ, β, α, BCl, Act	IIb
	HcpR*	δ, BCl, CFB, Cya, TM	IIb
	NorR	γ, β	Ib
<b>Other Metabolic Pathways</b>			
heat shock <sup>122</sup>	HrcA, σ <sup>32</sup>	γ, β, ε	Ia
DNA damage (SOS system) <sup>h, i</sup>	LexA	γ, β, α, Cya, BCl	Ia
ribonucleotides metabolism <sup>147, 148</sup>	NrdR**	bacteria	IIa, IIb
purine biosynthesis <sup>129, j</sup>	PurR	γ (Ent, Vib, Pas)	Ia
anaerobic respiration <sup>197</sup>	Fnr	γ (Ent, Vib, Pas)	Ia
global catabolic regulation <sup>27</sup>	Crp	γ (Ent, Vib, Pas)	Ia
fatty acid biosynthesis <sup>132</sup>	FabR**	γ (Ent)	IIb
phosphate metabolism <sup>k</sup>	PhoB	γ, α	Ia
sporulation <sup>l</sup>	Spo0A	BCl	Ib

<sup>a</sup> Analyzed in this study regulons. <sup>b</sup> Reference 277. <sup>c</sup> Reference 278. <sup>d</sup> Reference 279. <sup>e</sup> D.A.R., manuscript in preparation. <sup>f</sup> Reference 280. <sup>g</sup> D.A.R. and Natalia Doroshchuk, unpublished observation. <sup>h</sup> Reference 281. <sup>i</sup> Reference 282. <sup>j</sup> Reference 283. <sup>k</sup> Reference 284. <sup>l</sup> Reference 285. <sup>m</sup> Novel regulons tentatively predicted by comparative genome analysis and those of them that were experimentally confirmed are marked by one and two asterisks, respectively. <sup>n</sup> Abbreviations of taxonomic groups of microorganisms: α, β, γ, δ, and ε correspond to α-, β-, γ-, δ-, and ε-proteobacteria; Ent, Enterobacteriales; Vib, Vibrionales; Alt, Altermonadales; Xan, Xanthomonadales; Pse, Pseudomonadales; Pas, Pasteurellales; BCl, *Bacillus/Clostridium* group; Act, Actinobacteria; Arc, Archaea; CFB, *Chlorobium/Bacteroides* group; Cya, Cyanobacteria; TM, Thermotogales. <sup>o</sup> Strategies for regulon analysis are described in Figure 5.

identified by comparative genomics in other species from two taxonomic groups, the Enterobacteriales and Vibrionales.<sup>203</sup> The NagC-binding motif was constructed using upstream regions of known NagC-controlled genes in *E. coli* and their orthologs in related genomes. Scanning of the *nagC*-containing genomes using the constructed motif identified additional candidate NagC-regulated genes that are involved in the degradation and uptake of chitin and its *N*-acetylglucosamine derivatives. In *Vibrio cholerae*, the predicted NagC regulon was in agreement with microarray

data on the induction of gene expression by *N*-acetylglucosamine.<sup>204</sup>

In contrast to Enterobacteriales and Vibrionales, many species from other taxonomic groups contain genes for *N*-acetylglucosamine utilization but lack orthologs of NagC. Analysis of conserved *nag* gene clusters in other groups of proteobacteria (Altermonadales, Pseudomonadales, Xanthomonadales, β- and α-proteobacteria) identified two previously uncharacterized regulators from the LacI and GntR protein families. These two TFs, called respectively NagR

Table 5. Binding Motif Details for Microbial TFs Analyzed by Comparative Genomics

Metabolic pathway	Regulon	Phylogenetic distribution	Binding site consensus logo <sup>a</sup>	TF family
Chitin and N-acetylglucosamine utilization	NagC	γ-proteobacteria (Enterobacteriales, Vibrionales)		ROK
	NagR	γ (Alteromonadales, Xanthomonadales)		LacI
	NagQ	γ (Pseudomonadales), β (Burkholderiales), α (Rhizobiales, <i>Caulobacter</i> )		GntR
Pectin utilization	KdgR	γ (Enterobacteriales, Vibrionales)		IclR
Glucuronate utilization	UxuR	γ (Enterobacteriales, Pasteurellales)		GntR
Gluconate utilization	GntR	γ (Enterobacteriales, Vibrionales)		LacI
Nitrogen assimilation	NtcA	Cyanobacteria		Fnr
	TnrA	<i>Bacillus/Clostridium</i> group		MerR
Nitrogen fixation	NifA	α-proteobacteria		Fis
	NrpR	Methanogenic archaea		COG1693
Biotin metabolism	BirA	γ- and β-proteobacteria		BirA
	BirA	<i>Bacillus/Clostridium</i> group, Archaea		BirA
	BioR	α-proteobacteria		GntR
	BioQ	Actinobacteria		TetR
NAD metabolism	NadR	γ (Enterobacteriales)		NadR
	YrxA	<i>Bacillus/Clostridium</i> group, TM		COG1654
	NrtR	Cyanobacteria, Actinobacteria		COG1051
	NadQ	α- and β-proteobacteria		COG4111

<sup>a</sup> Sequence logos were generated by the WebLogo tool (<http://weblogo.berkeley.edu>).

and NagQ, were tentatively predicted to control the *nag* genes in a subset of species based on positional genomic evidence and phylogenetic co-occurrence profiles.<sup>203</sup> For each group of species containing one of these TFs, a conserved DNA binding motif was identified in the training set of poten-

tially coregulated genes from the chitin and *N*-acetylglucosamine pathways. The constructed recognition profiles were then used to scan against a subset of genomes of proteobacteria having a respective Nag regulator and to identify additional conserved regulon members. The results



of this study suggested that at least three nonorthologous types of TFs control expression of the *N*-acetylglucosamine and chitin utilization genes in various groups of proteobacteria (Table 5).

### 3.2.2. Sugar Acids Utilization

*E. coli* is capable of using various sugar acids (e.g., gluconate, hexuronates) as a source of carbon and energy. The respective sugar acid catabolic pathways converge to the common Entner–Doudoroff glycolytic pathway and are controlled by three different sugar acid-responsive TFs, GntR, UxuR, and KdgR. Comparative analysis of these sugar acid regulons in  $\gamma$ -proteobacteria predicted novel regulon members and TFBS motifs.<sup>119</sup> Combination of metabolic maps with regulatory networks showed the differences in the structure of the sugar acid catabolic pathways and regulons in different species.

The *E. coli* gluconate repressor GntR controls operons involved in the gluconate catabolism (*gntT*, *gntKU*) and the Entner–Doudoroff pathway (*edd*–*eda*). A GntR-binding site search profile was constructed by application of the signal detection procedure SignalX to the training set of upstream regions of the GntR-regulated genes and their orthologs in enterobacteria.<sup>119</sup> The GntR consensus site obtained by this procedure coincides with the experimentally mapped GntR sites at *gntT* (Table 5). Reconstruction of the GntR regulons by a genomic search with the GntR motif profile revealed some differences in the regulon content of various  $\gamma$ -proteobacteria. For instance, the GntR regulon in *Yersinia pestis* consists of only *gntK* and *gntU* genes, whereas *edd* and *eda* are in different operons that have no candidate GntR sites. In many cases, the candidate GntR sites occur in pairs, suggesting possible cooperative interactions of GntR dimer pairs.

The UxuR repressor in *E. coli* regulates the glucuronate utilization genes, but its DNA binding site was unknown. Using of signal-detection procedure and a sample of upstream regions of UxuR-regulated genes and their orthologs, a candidate UxuR-binding motif was identified and used to locate additional UxuR target genes (e.g., *gntP*) in the genomes of enterobacteria.<sup>119</sup> These comparative genomic predictions were later confirmed in experiments where the UxuR repressor was proved to bind its candidate operator sites in *E. coli* and control the expression of *gntP* in response to fructuronate concentrations.<sup>205</sup>

Utilization of pectin and its derivatives, oligogalacturonates and 2-keto-3-deoxygluconate (KDG), is controlled by the KDG-responsive repressor KdgR. All previously characterized KdgR-binding sites in the plant pathogen *Erwinia chrysanthemi* were collected from the literature and used to construct a search profile.<sup>119,206</sup> Comparative genomic analysis of the KdgR regulon in other enterobacteria and *Vibrio* species helped identify many new KdgR-regulated genes. For example, the predicted oligogalacturonide transporter OgtABCD<sup>119</sup> was confirmed in an independent study to have the proposed function (renamed TogMNAB) and to be regulated by KdgR in *E. chrysanthemi*.<sup>207</sup> Regulation of most other regulon members predicted in *E. chrysanthemi* was experimentally validated using *in vivo* transcriptional fusions, and for the first time a new phenomenon of positive regulation by KdgR was described.<sup>206</sup> Complete reconstruction of the KdgR regulons in various  $\gamma$ -proteobacteria yielded a metabolic map reflecting a globally conserved pathway for the catabolism of pectin and its derivatives, but with

significant variability in transport and enzymatic capabilities among species.<sup>206</sup>

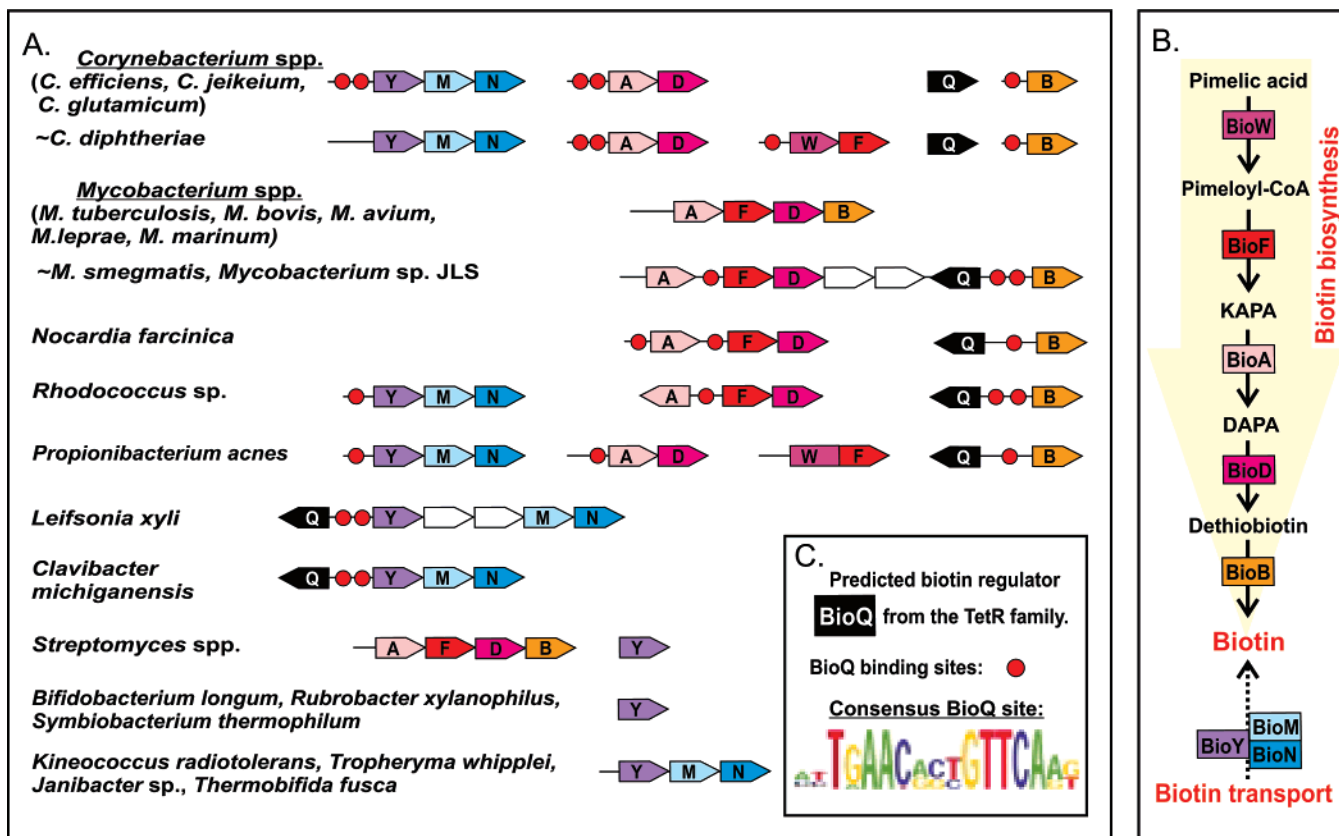
### 3.2.3. Biotin Metabolism

Biotin is an obligate cofactor of numerous biotin-dependent carboxylases in a variety of microorganisms. The strict control of biotin biosynthesis in *E. coli* is mediated by the bifunctional BirA protein, which acts both as a biotin-protein ligase and as a transcriptional repressor of the *bio* operon. A comparative genomic approach was used to reconstruct the biotin biosynthesis pathways and regulatory networks in a wide range of prokaryotic organisms.<sup>70</sup> Although *birA* is a widely distributed gene, only a fraction of BirA orthologs possess the N-terminal DNA-binding domain with the HTH motif (D-b-BirA). Based on phylogenetic analysis of DNA-binding domains, all D-b-BirA proteins were divided into two major groups, proteobacterial and nonproteobacterial. Accordingly, two partially similar recognition profiles for the BirA binding sites were constructed using the sets of upstream regions of the *bio* genes from various genomes (Table 5). The constructed profiles successfully detected new candidate members of the biotin regulon bacteria that contain D-b-BirA. In particular, the previously uncharacterized hypothetical transmembrane protein BioY was predicted to encode a transporter for biotin. Additional scanning of microbial genomes showed that the occurrence of potential BirA-binding sites upstream of biotin-related genes coincides with the presence of D-b-BirA in a genome.<sup>70</sup>

BirA represents a rare example of a TF in which the binding signal is conserved in various bacteria and archaea. However, the mode(s) of biotin-dependent regulation in the bacteria that lack D-b-BirA is still not known. This gap in our knowledge was partially filled by comparative genomic analysis using strategy IIa, which allowed us to identify a novel GntR-type TF for the *bio* genes (named BioR) and its binding signal in 8 out of 19 species of  $\alpha$ -proteobacteria.<sup>208</sup>

Here, I report, for the first time, the application of a similar approach (strategy IIa) to a set of the biotin biosynthesis and transport *bio* genes in Actinobacteria, another lineage that lacks D-b-BirA. In 11 out of 27 genomes of actinobacteria, there is a novel palindromic DNA motif associated with the *bio* genes (Figure 6). In many cases, these novel candidate sites occur in tandem, suggesting cooperative binding of an unknown TF to DNA. A candidate regulatory gene that encodes a TetR-type TF (named BioQ) is colocalized with the biotin synthase gene *bioB* in five genomes (*Nocardia*, *Rhodococcus* and *Propionibacterium* and two *Mycobacteria* species) and with the biotin transport operon *bioYMN* in two genomes (*Leifsonia* and *Clavibacter* species). Orthologs of the *bioQ* gene are also present in four *Corynebacterium* species but not in other actinobacteria. In *Corynebacterium* species, the *bioQ* and *bio* genes are not clustered on the chromosome. The phyletic distribution and genomic localization of novel candidate TFBS motifs and *bioQ* genes strongly suggest that BioQ mediates the biotin-dependent transcriptional regulation of the *bio* genes in the 11 species of actinobacteria. However, the mode of control of *bio* genes in other actinobacteria (including *Streptomyces* and pathogenic *Mycobacterium* species) is yet unknown.

These observations demonstrate that the biotin metabolism in bacteria is regulated by at least three distinct systems, including the bifunctional enzyme/repressor D-b-BirA, and two specialized TFs from the GntR and TetR protein families, BioR and BioQ (Table 5).



**Figure 6.** Novel biotin regulon BioQ in actinobacteria tentatively predicted by strategy IIa. (A) Chromosomal clusters of biotin synthesis and transport genes (shown by arrows) and localization of candidate BioQ-binding sites (red circles). Homologous genes are marked by matching colors. (B) Biotin biosynthesis and uptake pathway. (C) Consensus sequence logo for the predicted BioQ-binding sites.

### 3.2.4. Nitrogen Metabolism

Expression of nitrogen assimilation genes in bacteria is under the control of many regulatory systems, including the RpoN sigma factor and a set of lineage-specific TFs. In proteobacteria, this metabolic pathway is regulated by the two-component Ntr system, whose response regulator belongs to the Fis family of TFs.<sup>209</sup> Regulators from different protein families mediate the control of nitrogen assimilation genes in other bacterial lineages: MerR-type regulators TnrA and GlnR in the *Bacillus/Clostridium* group,<sup>210</sup> the Fnr-type regulator NtcA in cyanobacteria,<sup>211</sup> and the TetR-type regulator AmtR in actinobacteria.<sup>212</sup> These and other regulons were analyzed by various comparative genomic techniques.

The NtcA regulon in cyanobacteria was analyzed using the comparative genomic algorithm that combines information about cooccurrence of candidate NtcA and sigma-factor binding sites and the presence of similar motifs in the regulatory regions of orthologous genes in other related genomes.<sup>213</sup> Using the phylogenetic footprinting approach, the authors were able to predict new members of the NtcA regulons in the genomes of nine cyanobacteria. In addition to multiple nitrogen assimilation genes, high-scoring NtcA sites were found for many genes involved in the various stages of the photosynthesis process, suggesting tight coordination of these metabolic processes in cyanobacteria.<sup>213</sup>

Comparative analysis of the homologous TnrA and GlnR regulons in the *Bacillus/Clostridium* group revealed their significant plasticity in different bacteria.<sup>127</sup> The TnrA and GlnR orthologs were distinguished using the constructed phylogenetic tree for the MerR family of transcription factors. *Streptococcus*, *Listeria*, and *Staphylococcus* species lack

TnrA but have the highly conserved GlnR regulon, which mainly contains genes of glutamine transport and utilization. In *Bacillus* species, the duplicated regulators TnrA and GlnR control many genes for utilization of glutamine and other nitrogen-containing compounds.

Genes involved in nitrogen fixation are under control of the  $\sigma^N$ -dependent transcriptional activator NifA in bacteria,<sup>214</sup> whereas, in the nitrogen-fixing species of archaea, these genes are regulated by the transcriptional repressor NrpR, which represents a new family of regulators unique to the euryarchaeota.<sup>215</sup> Accordingly, these two different regulatory systems operate by different binding motifs (Table 5). The NifA regulon in  $\alpha$ -proteobacteria was analyzed in conjunction with identification of RpoN ( $\sigma^{54}$ ) binding sites using the training sets of experimentally characterized sites of both factors (Natalia A. Doroshchuk and D.A.R., unpublished observation). Simultaneous comparative analysis of upstream NifA binding sites and downstream  $\sigma^{54}$ -dependent promoters decreases the rate of false-positive site predictions, allowing for more-accurate reconstruction of the nitrogen fixation regulons in the sequenced genomes of  $\alpha$ -proteobacteria. Finally, the archaeal nitrogen fixation regulon NrpR was analyzed using the consistency-check approach and the training set of nitrogen fixation genes.<sup>103</sup>

Two dissimilatory processes in the bacterial inorganic nitrogen cycle, denitrification and detoxification of nitrogen oxides, are controlled by an evolutionary variable transcriptional network that involves the Fnr-like transcription factors HcpR, Dnr, and NnrR; the two-component systems NarXL and NarPQ; the nitric oxide-responsive  $\sigma^{54}$ -dependent activator NorR, and the nitrite-sensitive repressor NsrR.<sup>126,216</sup> Comparative reconstruction of the nitrogen oxides regulatory

network has revealed multiple interconnections between the regulons, conservation of some regulatory interactions, and changing of other regulatory interactions, as well as extensions, reductions, or even loss of some regulons.<sup>216</sup> For instance, the nitrogen oxides detoxification genes *hcp* and *hmp* are regulated by various TFs (NsrR, NorR, Dnr, and HcpR) in various bacterial species.

### 3.2.5. NAD Metabolism

Transcriptional regulation of NAD biosynthesis genes has been extensively studied in enterobacteria, where at high NAD levels the multifunctional protein NadR represses the *de novo* NAD synthesis and salvage genes.<sup>217</sup> In addition to the N-terminal DNA-binding domain, NadR has two enzymatic domains involved in the salvage of nicotinamide riboside.<sup>218</sup> The application of the comparative genomic approach to the analysis of the NadR regulon in the enterobacteriales revealed similar patterns of NadR binding sites in this lineage and predicted the autoregulation of the *nadR* gene.<sup>69</sup> In contrast to enterobacteria, an N-terminal DNA-binding domain of NadR is absent in other bacterial lineages and the mechanism of regulation of the NAD metabolism in these species is unclear. Different taxonomic groups of bacteria may have a variety of regulatory strategies for control of the same pathway. Application of the signal-detection procedure and the subsystem-oriented strategy IIa of comparative genomics allowed us to identify and reconstruct three other novel NAD regulons in different bacterial lineages (D.A.R., Nadia Raffaelli, and Andrei Osterman, unpublished observations).

A different nicotinate-responsive transcriptional repressor encoded by the *yxrA* gene was later identified in *Bacillus subtilis*, where it controls the NAD biosynthesis operon; however, its DNA-binding site was unknown.<sup>219</sup> We applied the comparative approach to the genomes of other firmicutes that have *yxrA* orthologs and identified a conserved palindromic motif in upstream regions of NAD biosynthesis and salvage operons from the *Bacillus/Clostridium* group (Table 5). Based on co-occurrence and colocalization with *yxrA* genes in the genomes, this novel DNA motif was tentatively attributed to the YrxA-like NAD regulator. A search for additional YrxA sites, complemented by genome context analysis and cross-species comparisons, led to identification of new candidate members of the YrxA regulon, in particular, different types of candidate transporters for NAD metabolic precursors (D.A.R. and Andrei Osterman, manuscript in preparation).

Comparative analysis of potential regulatory regions of NAD biosynthesis operons in  $\alpha$ - and  $\beta$ -proteobacteria revealed a conserved DNA motif (Table 5) and a connected hypothetical TF, named NadQ, which is encoded by an adjacent gene immediately upstream of the *nad* operon (D.A.R., unpublished observation). Similar analysis of the NAD metabolic genes in the genomes of cyanobacteria and actinobacteria identified another DNA motif and a hypothetical TF, named NrtR, which is encoded in close proximity to the NAD biosynthesis and salvage genes and has this candidate motif upstream. Although the predicted NAD regulators NrtR and NadQ belong to different protein families, they share similar HTH domains on their C-terminal parts. The candidate binding sites for NadQ and NrtR have some resemblance to each other, consistent with the similarity of their DNA-binding domains. Recently, the novel predicted NAD regulator NrtR was purified from *Synechocystis* sp.

and confirmed in electrophoretic mobility shift assays to bind specifically the candidate NrtR sites upstream of the *nadE*, *nadMV*, and *nadA* genes (Nadia Raffaelli and D.A.R., manuscript in preparation).

Apart from these taxonomic groups, the mode of regulation of NAD metabolism in other prokaryotic lineages remains unclear and requires further study.

### 3.3. Analysis of Regulons to Support Metabolic Reconstruction and Functional Predictions

Comparative analysis of regulons based on the identification and cross-genome comparison of shared regulatory sites (e.g., TFBSs, RNA regulatory elements) is an important technique for functional annotation of hypothetical genes. It predicts coregulation of a set of genes, providing evidence that these genes may be functionally coupled. First, the identification of novel members of metabolic regulons helps to locate candidates for so-called missing genes in metabolic pathways, attempting to connect known functional roles to genes that have not yet been characterized.<sup>152</sup> On the other hand, the metabolic regulon reconstruction allows one to identify novel metabolic enzymes and to predict novel enzymatic reactions that were not known before. Finally, the analysis of bacterial regulons promotes substantial progress in functional annotation of hypothetical transporter genes that could be tentatively attributed to the regulated metabolic pathway (D.A.R. and Mikhail Gelfand, in preparation).

Integration of the comparative genomic analysis of microbial regulons with traditional approaches of genome context analysis is an efficient method for functional gene annotation and metabolic pathway reconstruction. The traditional approaches of genome context analysis largely fall into one of the following three categories:<sup>152</sup>

(1) Clustering of genes on the chromosome (or gene neighborhood) approaches are based on the known tendency that proteins, whose corresponding genes are located “close” to each other in multiple genomes, are expected to be “functionally coupled” and form the same metabolic pathway.<sup>220,221</sup>

(2) Gene fusion-based approaches attempt to discover pairs or sets of genes in one genome that are merged to form a single gene in another genome, providing further evidence of potential functional coupling.<sup>222,223</sup>

(3) The phylogenetic profiling approach is based on the assumption that functionally associated proteins are expected to have very similar occurrence profiles across various organisms.<sup>224</sup>

Several examples below illustrate how the comparative analysis of regulons helps in metabolic reconstruction and, in particular, how useful is it to predict novel functional roles, missing genes, and transporters in microbial metabolic pathways.

#### 3.3.1. L-Rhamnose Utilization

The first example, presented here for the first time, describes in detail the general strategy for reconstruction of a metabolic pathway and associated regulatory mechanisms. To reconstruct the L-rhamnose utilization system in bacteria, we used a subsystem-based approach combining a number of comparative genomic techniques as implemented in the SEED platform.<sup>203,225,226</sup> The utilization of L-rhamnose in *E. coli* is catalyzed via L-rhamnose mutarotase (RhaM), L-rhamnose isomerase (RhaA), L-rhamnulose kinase (RhaB), and L-rhamnulose-1-phosphate aldolase (RhaD). The detailed



results of this analysis are captured in the SEED subsystem available online (<http://theseed.uchicago.edu/FIG/subsys.cgi>) and in Figure 7.

The transcriptional activator RhaS in *E. coli* belongs to the AraC family and controls the L-rhamnose transporter *rhaT* and the catabolic operon *rhaBAD*.<sup>227,228</sup> Orthologs of *rhaS* and these L-rhamnose catabolic genes are present in some other  $\gamma$ -proteobacterial genomes. The analysis of upstream regions of *rha* genes in this taxonomic group results in construction of the RhaS search profile and identification of additional RhaS targets (Figure 7). For example, *Salmonella typhimurium* and *Erwinia carotovora* are predicted to possess a RhaS-regulated hypothetical transport system (named *rhiABC*), which is similar to the C4-dicarboxylate transport system Dcu. Candidate RhaS regulons in two *Erwinia* species and *Klebsiella pneumoniae* also include the *rhiT-rhiN* operons involved in the uptake and catabolism of rhamnogalacturonides, L-rhamnose-containing oligosaccharides.<sup>229</sup> Based on the gene-occurrence pattern and candidate coregulation, *RhiABC* is tentatively predicted to encode an alternative transporter for rhamnogalacturonides, which replaces *RhiT* in *S. typhimurium*.

The RhaS regulon in  $\gamma$ -proteobacteria is also predicted to include various genes that are likely involved in utilization of L-lactaldehyde, a final product of the L-rhamnose catabolism. The rhamnose operons in *K. pneumoniae* and *S. typhimurium* include an additional gene (named *rhaZ*) encoding the hypothetical iron-containing alcohol dehydrogenase (PF00465 protein family in PFAM<sup>57</sup>). *E. carotovora* has a single RhaS-regulated gene *aldA* encoding alcohol dehydrogenase from another protein family (PF00171). In contrast, the RhaS regulons in *Erwinia chrysanthemi* and *Mannheimia succiniproducens* include the lactaldehyde reductase *fucO*, whereas *aldA* and *rhaZ* are absent from their genomes. These observations suggest that  $\gamma$ -proteobacteria use three different enzymes and two different pathways for the final stage of the L-rhamnose pathway.

Analysis of other taxonomic groups outside the  $\gamma$ -proteobacteria tentatively identifies previously uncharacterized members of the LacI, DeoR, and AraC families as alternative transcriptional regulators of the L-rhamnose pathway (Figure 7). In actinobacteria, a LacI-type regulator (named here R2) is identified in a chromosomal cluster with the *rha* genes. The predicted palindromic R2-binding signal is characteristic of DNA-binding sites of LacI family regulators. Two TFs from the DeoR family (R3 and R3', 28% similar to each other) are inferred based on chromosomal clustering with *rha* genes in the Bacillaceae and  $\alpha$ -proteobacteria groups, respectively. The deduced binding motifs consist of two or three imperfect direct repeats (AACAAAA for R3 and TGATTGA for R3') separated by three base pairs. Finally, another potential regulator from the AraC family with a very weak similarity to RhaS (named R1) is identified in some species from the *Bacillus/Clostridium* group. Thus, at least five nonorthologous types of TFs appear to regulate the L-rhamnose utilization genes in bacteria.

In addition to TFs, a high level of variation is also observed for the components of transport machinery. The L-rhamnose-specific transporter RhaT is a conserved member of *rha* operons and RhaS regulons in  $\gamma$ -proteobacteria. An alternative system of L-rhamnose transport via a committed ABC cassette (named RhaFGHJ) is predicted to substitute RhaT in  $\alpha$ -proteobacteria and *Streptomyces* spp., whereas *K. pneumoniae* has both of them encoded in the *rha* gene

cluster. Another novel transporter for L-rhamnose (named RhaY) is tentatively identified in actinobacteria and in the *Bacillus/Clostridium* group (Figure 7). RhaY has no similarity to RhaT and belongs to the PF00083 family of sugar transporters from the MFS superfamily.

The reconstruction of bacterial L-rhamnose utilization pathways reveals that all but one enzymatic pathway component occur in many alternative forms, with the L-rhamnulose kinase RhaB being the only invariant component of the pathway. A nonorthologous isomerase (named RhaI) is inferred by the genome context analysis in actinobacteria,  $\alpha$ -proteobacteria, and *B. licheniformis*. Instead of the canonical form of aldolase (RhaD), the *rha* clusters in actinobacteria, bacilli, and  $\alpha$ -proteobacteria contain a chimeric gene (e.g., *yuxG* in *B. subtilis*), which encodes a two-domain protein with an N-terminal class II aldolase domain and a C-terminal short chain dehydrogenase domain (named RhaE and RhaW, respectively). The phylogenetic occurrence profile suggests that RhaW may encode the missing L-lactaldehyde dehydrogenase. Thus, this bifunctional protein is tentatively predicted to catalyze two final reactions in the L-rhamnose utilization pathway.

### 3.3.2. Other Catabolic Pathways

A similar approach was applied for the comparative genomic analysis of other sugar catabolic pathways in bacteria.<sup>119–121,203,206</sup> In the *N*-acetylglucosamine utilization subsystem, a similarly high level of variations and non-orthologous gene displacements was observed for specific TFs and transport systems. Most notably, the PTS-mediated transport of *N*-acetylglucosamine in enterobacteriales and vibionales appears to be functionally replaced by a specific MFS-type permease in alteromonadales and xanthomonadales or an ABC cassette in  $\alpha$ -proteobacteria in conjunction with a novel bacterial *N*-acetylglucosamine kinase enzyme. In addition to that, two nonorthologous versions of the *N*-acetylglucosamine-6-phosphate deaminase NagB were found and experimentally verified.<sup>203</sup>

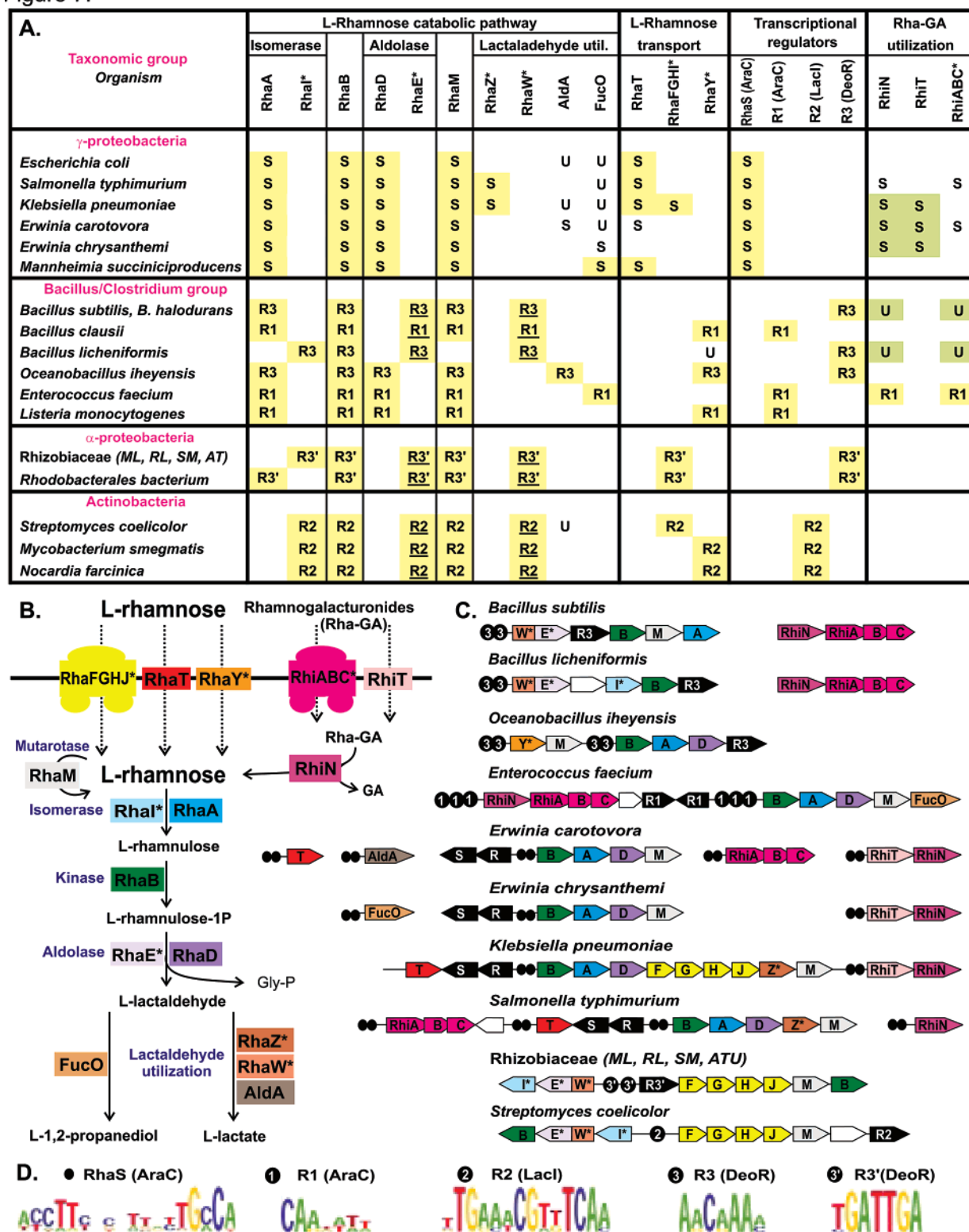
Analysis of the arabinose utilization subsystem identified a novel nonorthologous variant of L-ribulokinase in a number of Gram-positive bacteria.<sup>121</sup> Reconstruction of the xylose regulon XylR in enterobacteriales resulted in identification of operons comprising putative transporters and hydrolases for utilization of xylose oligosaccharides.<sup>120</sup> Analysis of the KdgR regulon revealed several novel transport systems and enzymes (e.g., sugar isomerase SpiX) involved in the utilization of products of pectin degradation, such as galacturonate, glucuronate, and KDG.<sup>119,206</sup>

The comparative analysis of the fatty acid degradation regulon FadR revealed new members of this regulon in the *E. coli* genome (*fadIJ*, formerly *b2342-41*) and demonstrated that the candidate FadR-regulated gene *yafH* encoding acyl-CoA dehydrogenase is identical to the gene *fadE* previously identified by genetic techniques.<sup>230</sup> The identity of *yafH* and *fadE* in *E. coli* was then experimentally confirmed by targeted gene disruption, and the FadR-dependent regulation of its transcription was further confirmed.<sup>231</sup>

### 3.3.3. Biosynthesis of Coenzyme B<sub>12</sub>

Biosynthesis of adenosylcobalamin (coenzyme B<sub>12</sub>) requires about 25 enzymes encoded by *cbi* and *cob* genes that catalyze the *de novo* synthesis of a tetrapyrrole-derived corrin ring, insertion of a cobalt ion, adenosylation and attachment of an aminopropanol arm to the corrin ring, and assembly

Figure 7.



**Figure 7.** Reconstruction of the L-rhamnose utilization system in bacteria. (A) Occurrence and features of genes involved in L-rhamnose utilization. Species in several taxonomic groups of bacteria are shown as rows. The presence of genes for the respective functional roles (columns) is shown by capital letters corresponding to the four identified rhamnose regulons: S, RhaS regulon (as in *E. coli*); R1, R2, R3, and R3' correspond to the novel regulons of the same names. Other genes that were not identified within the above rhamnose regulons are marked by "U". Genes clustered on the chromosome (operons) are outlined by matching background colors. Tentatively predicted functional roles are marked by asterisks. Functional roles corresponding to the predicted bifunctional enzymes RhaE–RhaW are underlined. The four Rhizobiaceae genomes that have the same set of genes and genome context are *Mesorhizobium loti* (ML), *Agrobacterium tumefaciens* (AT), *Rhizobium leguminosarum* (RL), and *Sinorhizobium meliloti* (SM). (B) The reconstructed L-rhamnose utilization pathway. (C) Chromosomal clusters of L-rhamnose utilization genes (arrows) and localization of candidate binding sites (circles) for rhamnose-specific TFs. The genes corresponding to the rhamnose-specific regulators RhaS, RhaR, R1, R2, and R3 are shown by black arrows with S, R, R1, R2, and R3 letters, respectively. Other homologous genes are marked by matching colors. (D) Consensus sequence logos for predicted binding sites of rhamnose-specific TFs. The corresponding TF protein family name is given in parentheses.

of the nucleotide loop that bridges the lower ligand dimethylbenzimidazole and the corrin ring.<sup>232</sup> Most *cbi* and *cob* genes in bacteria are organized in extended operons and controlled by B<sub>12</sub> riboswitch elements, conserved mRNA leader sequences that directly bind an effector molecule, adenosylcobalamin.<sup>159,165</sup> The genomic identification and comparative analysis of B<sub>12</sub> riboswitches combined with other genome context techniques identified a large number of new candidate B<sub>12</sub>-regulated genes with tentatively assigned functional roles in the B<sub>12</sub> biosynthesis pathway.<sup>233</sup> For example, nine different types of candidate cobalt transporters were identified within the bacterial B<sub>12</sub> regulons in different lineages, emphasizing the importance of cobalt uptake for the *de novo* coenzyme B<sub>12</sub> biosynthesis.<sup>198</sup> Experimental analysis confirmed cobalt transport activity for several representatives of two families of metal uptake transporters, CbiMNQO and NiCoT.<sup>234,235</sup>

Metabolic reconstruction of the B<sub>12</sub> biosynthesis pathway revealed a large number of missing genes, most of which were identified as non-orthologous displacements.<sup>233</sup> Most remarkably, various nonorthologous gene displacements for the *cobC* gene involved in the nucleotide loop assembly were identified in archaea,  $\alpha$ -proteobacteria, and Actinobacteria (named *cobZ*, *cblXY*, and *cblZ*, respectively). Later, the *cobZ* gene of *Methanosarcina mazei* was confirmed to encode a nonorthologous replacement of the  $\alpha$ -ribazole-5-phosphate phosphatase (CobC) enzyme of enterobacteria.<sup>236</sup>

A novel functional role of the L-threonine kinase PduX has been proposed for the pathway of synthesis of the lower ligand of coenzyme B<sub>12</sub>. In some Gram-positive bacteria, the *pduX* gene of unknown function was found within the B<sub>12</sub> biosynthesis gene clusters adjacent to the *cobD* gene. In *Streptomyces coelicolor*, the single *pduX* gene is predicted to be regulated by a B<sub>12</sub> riboswitch. The PduX proteins belong to the GHMP kinase superfamily and are weakly similar to L-homoserine and mevalonate kinases. The lower ligand of B<sub>12</sub> is synthesized by the CobD aminotransferase, which requires L-threonine-3-phosphate as a substrate. Based on these facts, the novel B<sub>12</sub>-regulated gene *pduX* was proposed to encode L-threonine kinase involved in B<sub>12</sub> biosynthesis,<sup>233</sup> and experimental verification of the PduX function is currently underway (Aaron Best, personal communication).

### 3.3.4. New Mechanisms for Alternative Cofactor Adaptation

Comparative analyses of several cofactor-specific regulons revealed several cases where distinct isofunctional genes appear to be regulated according to the availability of respective cofactors (Table 6).

B<sub>12</sub> riboswitches were detected upstream of the *metE*, *nrdAB*, and *nrdDG* genes encoding the B<sub>12</sub>-independent isozymes of methionine synthase and ribonucleotide reductase in various genomes from diverse taxonomic groups of bacteria (e.g.,  $\alpha$ -proteobacteria and actinobacteria). These microbial genomes also encode the MetH and NrdJ isozymes that perform the same functional roles but require B<sub>12</sub> as a cofactor. Thus, it was proposed that when vitamin B<sub>12</sub> is present in the cell, expression of B<sub>12</sub>-independent isozymes is inhibited, and only relatively more-efficient B<sub>12</sub>-dependent isozymes are used.<sup>233</sup> Although the repression of B<sub>12</sub>-independent isozymes by the excess of coenzyme B<sub>12</sub> looks rational, this regulatory strategy was not known before the comparative genomic identification of B<sub>12</sub> riboswitches

**Table 6. Cross-talk in Transcriptional Regulation of Isozymes with Different Cofactor Requirements**

isozyme <sup>a</sup>	cofactor	regulon and its effector <sup>b</sup>
Ribonucleotide Reductase ( $\alpha$ , BCl, Act, CFB)		
NrdJ	[B <sub>12</sub> ]	
NrdAB or NrdDG		[B <sub>12</sub> ]-riboswitch (represses)
Methionine Synthase( $\alpha$ , BCl, Act, CFB)		
MetH	[B <sub>12</sub> ]	
MetE		[B <sub>12</sub> ]-riboswitch (represses)
Fumarate Hydratase ( $\gamma$ )		
FumA	[Fe]	[Fe]-Fur (activates)
FumC		[Fe]-Fur (represses)
Fumarate Hydratase ( $\alpha$ )		
FumA	[Fe]	
FumC		[Fe]-RirA (represses)
Superoxide Dismutases ( $\gamma$ )		
SodB	[Fe]	[Fe]-Fur (activates)
SodA	[Mn]	[Fe]-Fur (represses)
Electron-Transfer Proteins ( $\delta$ )		
ferredoxin	[Fe]	
flavodoxin		[Fe]-Fur (represses)
Hydrogenases ( $\delta$ )		
[Ni-Fe] Hyd	[Ni-Fe]	
[Fe] Hyd	[Fe]	[Ni]-NikR (represses)
GTP Cyclohydrolase I (BCl, $\gamma$ , $\beta$ )		
FolE	[Zn]	
YciA	[?]	[Zn]-Zur (represses)

<sup>a</sup> The taxonomic distribution of the observed transcriptional regulatory cross-talk is indicated, where the abbreviations of the taxonomic groups are the same as those in Table 4. <sup>b</sup> Regulatory effector molecules are shown in square brackets. Positive or negative mechanism of regulation is indicated in parentheses.

upstream of the *nrd* and *metE* genes. Recently, this hypothesis about regulation by B<sub>12</sub> riboswitches was experimentally confirmed for *metE* in *Bacillus clausii*<sup>237</sup> and for *nrdAB* in *Streptomyces coelicolor*.<sup>238</sup> Interestingly, the methionine synthetase *metE* in *B. clausii* is subject to dual regulation by tandem riboswitches that respond to *S*-adenosylmethionine and coenzyme B<sub>12</sub>.<sup>237</sup>

Several genes encoding iron-containing enzymes (e.g., *sdh*, *acnA*, *fumA*, *sodB*) are positively regulated in high iron concentrations by Fur in *E. coli* through repression of synthesis of a small antisense RNA.<sup>239</sup> Another regulatory strategy for iron metabolism, where an alternative iron-independent enzyme is negatively regulated by high iron concentrations, was reported for the non-iron fumarate hydratase FumC and [Mn] superoxide dismutase SodA in  $\gamma$ - and  $\alpha$ -proteobacteria.<sup>123,240</sup> In addition, [Fe]-Fur was predicted to repress a flavodoxin gene in *Desulfovibrio* species, which may be used in an electron-transfer chain as an alternative to ferredoxins present in the genomes.<sup>241</sup> Finally, the nickel repressor NikR was predicted to regulate the *hyd* operon encoding [Fe] hydrogenase in *Desulfovibrio desulfuricans*, whose genome also encodes [NiFe] hydrogenase.<sup>241</sup>

A similar regulatory strategy has been proposed for ribosomal proteins in the comparative genomic study of bacterial zinc regulons.<sup>124</sup> Repression by the zinc repressor Zur was predicted for genes encoding paralogs of L36, L33, L31, and S14 ribosomal proteins. The original copies of these proteins contain zinc-ribbon motifs and thus likely bind Zn, whereas these motifs are not present in zinc-regulated paralogs that substitute the main proteins during zinc starvation. Since ribosomes are highly abundant in the cell,



this alternation may lead to an increased concentration of zinc ions available for other zinc-binding proteins in the cell. Therefore, this regulatory system would contribute to the zinc homeostasis in the cell under zinc starvation. This regulatory model of zinc-dependent regulation of ribosomal proteins by Zur was experimentally confirmed in *B. subtilis*<sup>242,243</sup> in *S. coelicolor*.<sup>244,245</sup>

Taken together, these data suggest that a flexible strategy of transcriptional regulation of isozymes and other isofunctional proteins with different cofactor requirements may represent a common theme in the environmental adaptation of bacteria.

### 3.3.5. Prediction of Transporter Specificities

Transport systems are essential components of the cell.<sup>246</sup> They are involved in uptake of all nutrients into the cytoplasm, supporting the utilization of exogenous sources of carbon and nitrogen and also providing the source of essential microelements (e.g., vitamins, metal ions). Classification of enzymes and reconstruction of metabolic pathways from genomic data have led to the development of metabolic databases such as MetaCyc<sup>247</sup> and KEGG.<sup>248</sup> In contrast to metabolic pathways, much less effort has been expended on genomic reconstruction of transport systems. The Transport DB collects known and predicted transport systems encoded in complete microbial genomes and annotated based on a series of experimental and bioinformatic evidence.<sup>249</sup> However, most potential transport systems are still annotated as hypothetical and need to be characterized.

Projection of transporter annotations by homology only is not reliable in many cases, as, in comparison with enzymes, the substrate specificity of transporters is much more changeable during evolution. The use of reporter gene fusions in a high-throughput platform offers the possibility of screening hundreds of compounds against all candidate transporter operons to identify specific inducers for transport systems and predict their solute specificity.<sup>250</sup> Particular genome context evidence (chromosomal clustering, coregulation, and co-occurrence profiles) and careful phylogenetic analysis of transport protein families contribute significantly to functional annotation of hypothetical transporters (D.A.R. and Mikhail Gelfand, in preparation).

Comparative genomic analysis of specific metabolic regulons has led to substantial progress in functional annotation of hypothetical transporter genes. For instance, candidate uptake transporters for the amino acids arginine, lysine, methionine, and glycine in *Shewanella oneidensis* (ArgP, LysW, MetT, and GlyP, respectively) and for the vitamins riboflavin, biotin, and thiamin in *B. subtilis* (YpaA, BioY, and YuaJ, respectively) were predicted based on coregulation with the respective amino acid/vitamin biosynthetic genes by a specific metabolite-responsive riboswitch or TF regulon.<sup>10,70,157,158,160,174,208,251</sup> The predicted specificities of YpaA (re-named RibU) and BioY transporters were later confirmed by direct measurements of riboflavin and biotin uptake, respectively.<sup>252–255</sup>

## 4. Patterns and Mechanisms in Evolution of Transcriptional Regulatory Networks

Although the comparative genomics of microbial regulons is an emerging field of research, a substantial amount of data have already been accumulated for the description of the most common and important types of events associated with the

evolution of TRNs in bacteria.<sup>26,73</sup> Duplications and losses of TFs and their TFBSs result in regulon expansions, shrinkages, mergers, and split-ups. New regulons could be introduced by duplication and specialization of a TF paralog. Similar to metabolic enzymes, microbial TFs are subject to horizontal gene transfer and nonorthologous gene displacement events, leading to considerable rewiring of TRNs. The inference of these evolutionary events is strongly supported by the observation of multiple cases when nonorthologous TFs control equivalent pathways or, vice versa, orthologous regulators control distinct pathways in different species.

In this section, several approaches for analysis of evolutionary dynamics of TRNs are described and illustrated by examples of how the inferred evolutionary events could contribute to the flexibility and interchangeability of regulons in bacteria.

The best-characterized TRN currently available, that of the model bacterium *E. coli* (documented in RegulonDB<sup>78</sup>), was used in a number of studies to analyze the conservation patterns of this network across completely sequenced prokaryotic genomes. A high level of conservation of coregulation between two well-characterized model bacteria was first reported by the comparison of the operon map of *B. subtilis* with the regulon map of *E. coli*.<sup>256</sup> In three other studies, the conservation of individual components of TRNs in *E. coli* was analyzed by identification of orthologs of TFs and their target genes.<sup>53,257,258</sup> All three investigations reported an extreme flexibility of TRNs in bacteria. TFs are typically less conserved than the target genes and appear to evolve independently. The majority of *E. coli* regulons get rapidly lost over the increase of phylogenetic distance, as other microorganisms tend to have their own sets of TFs.<sup>53,257</sup> Despite a generally poor conservation of the regulatory interactions across genomes, certain regulons (e.g., ArgR, Fur, BirA, LexA) have been conserved across different taxonomic groups.<sup>257</sup>

However, the above approach does not take into account the presence and distribution of TFBSs in the genomes, limiting its ability to predict the loss and gain of regulatory interactions, novel regulon members, and the rewiring of regulons.

Combining identification of orthologous TFs with the genome-scale search for their cognate TFBSs is a powerful approach to the analysis of coevolution of TFs and TFBSs. This integrated approach allows us to describe the divergence and adaptation of regulons in conjunction with duplication, birth, or loss of TFBSs.<sup>70,124,215,240,259–261</sup> Several examples below illustrate a remarkable variability of TRNs associated with a particular metabolic pathway, that allows us to make first steps toward the reconstruction of possible evolutionary scenarios for these TRNs.

### 4.1. Methionine Metabolism

Methionine metabolism in bacteria is regulated by a variety of RNA and DNA regulatory systems (Table 7A). Analysis of the distribution of these regulatory systems in bacterial species helps to elucidate possible evolutionary scenario(s) for regulation of this metabolic pathway.

In  $\gamma$ -proteobacteria, two TFs, MetJ and MetR, are implicated in the control of methionine metabolism. The *S*-adenosylmethionine repressor MetJ in *E. coli* controls all methionine biosynthesis and transport genes by binding to operators that contain two to five tandem repeats of an 8-bp sequence.<sup>262</sup> The homocysteine-responsive activator MetR

**Table 7. Regulatory Systems for Methionine and Aromatic Amino Acid Metabolism in Bacteria**

system	type <sup>a</sup>	effector <sup>b</sup>	phylogenetic distribution <sup>c</sup>	regulated genes <sup>d</sup>
<b>A. Methionine</b>				
MetJ	TF (MetJ)	SAM	$\gamma$ (Ent, Pas, Vib, Alt)	<i>metK</i> (SAM synthesis); <i>metABCFEHY</i> (Met synthesis); <i>metNPQ</i> , <i>metT</i> (Met transport); <i>metJ</i> (autoregulation)
MetR	TF (LysR)	homocysteine	$\gamma$ (Ent, Vib, Alt, Pse), $\beta$ (Bor, Bur, Ral)	In Ent: <i>metAEFH</i> (Met synthesis); <i>metR</i> (autoregulation)
McbR	TF (TetR)	SAH	Act (corynebacteria)	<i>metK</i> (SAM synthesis); <i>metBFEHXY</i> (Met synthesis); <i>cysNDHIJEK</i> (sulfur assimilation)
SAM-I (S-box)	riboswitch	SAM	BCI (Bac, Clost), $\gamma$ (Xan), $\delta$ (Geo), TM, DR, FN, CT	<i>metK</i> (SAM synthesis); <i>metBCFEHIXY</i> (Met synthesis); <i>metNPQ</i> , <i>metT</i> (Met transport)
SAM-II	riboswitch	SAM	$\alpha$ , $\beta$ (Bor), CFB	<i>metK</i> (SAM synthesis); <i>metACHXY</i> (Met synthesis)
MtaR/MET-box	TF (LysR)	Met	BCI (Strep, LL)	In Strep: <i>metNPQ</i> (Met transport); <i>metBEFIY</i> (Met synthesis). In LL: <i>metEF</i> only
CmbR/CYS-box	TF (LysR)	<i>O</i> -acetylserine	BCI (LL, Strep)	Experimental data in LL only: <i>cysM</i> , <i>tcy</i> (Cys synthesis, transport); <i>yrhBA</i> (Met to Cys synthesis); <i>metNPQ</i> (Met transport); <i>metBIY</i> (Met synthesis)
Met-T-box	antiterminator	Met-tRNA	BCI (LB, Bac, Clost)	In LB: <i>metBCFEIY</i> (Met synthesis); <i>metNPQ</i> (Met transport). In Bac, Clost: <i>metS</i> (tRNA synthesis)
SAM-III	riboswitch	SAM	BCI (LB, Strep, LL)	<i>metK</i> (SAM synthesis)
<b>B. Aromatic Amino Acids</b>				
TrpR	TF (TrpR)	Trp	$\gamma$ (Ent, Pas, Vib, Alt), chlamydia	In Ent, Pas, Vib: <i>aro</i> , <i>trp</i> (Trp synthesis); <i>mtr</i> (Trp transport); <i>trpR</i> (autoregulation). In Alt: <i>aro</i> , <i>tyr</i> (Tyr synthesis); <i>tyrP</i> (Tyr transport); <i>trpR</i> . In Chlamydia: <i>trp</i> (Trp synthesis)
TyrR	TF (TyrR)	Tyr, Phe	$\gamma$ (Ent, Pas, Vib, Alt, Pse)	In Ent, Pas, Vib: <i>aro</i> , <i>tyr</i> , <i>tyrP</i> , <i>aroP</i> (Tyr, Phe synthesis and transport); <i>tyrR</i> (autoregulation). In Pse: Phe and Tyr catabolism. In Alt: amino acid catabolism; <i>tyrR</i>
Phe-atten	attenuator	Phe-tRNA	$\gamma$ (Ent, Vib, Alt)	<i>pheA</i> (Phe synthesis)
Trp-atten	attenuator	Trp-tRNA	$\gamma$ (Ent, Vib, Alt, Pse), $\alpha$	<i>trp</i> (Trp synthesis)
TRAP	RNA-binding protein	Trp	BCI (Bac - except Bcer)	<i>trp</i> (Trp synthesis); <i>trpP</i> (Trp transport)
Trp-T-box	antiterminator	Trp-tRNA	BCI (Bcer, LB, LL, Strep, Clost)	<i>trp</i> (Trp synthesis); <i>trpP</i> , <i>trpXYZ</i> (Trp transport)
Tyr-T-box	antiterminator	Tyr-tRNA	BCI (LB, Bcer)	<i>aro</i> , <i>tyr</i> (Tyr synthesis); <i>tyrT</i> (Tyr transport)
PCE-box	uncertain TF	uncertain	BCI (Bac - except Bcer)	<i>aro</i> (Tyr, Phe synthesis)
ARO-box	uncertain TF	uncertain	BCI (LL, Strep)	<i>aro</i> (Tyr, Phe synthesis)

<sup>a</sup> TF protein families are indicated in parentheses. <sup>b</sup> Abbreviations for effectors: Met, methionine; Trp, tryptophan; Tyr, tyrosine; Phe, phenylalanine; SAM, *S*-adenosylmethionine; SAH, *S*-adenosylhomocysteine. <sup>c</sup> Abbreviations for taxonomic groups are the same as those in Table 3. Additional abbreviations are as follows: Bac, Bacillales; Bcer, *Bacillus cereus* group; LB, Lactobacilli; Strep, Streptococci; LL, *Lactococcus lactis*; Clost, Clostridiales; Bor, *Bordetella* spp.; Bur, *Burkholderia* spp.; Ral, *Ralstonia* spp.; Geo, *Geobacter* spp.; DR, *Deinococcus radiodurans*; FN, *Fusobacterium nucleatum*; CT, *Chlorobium tepidum*. <sup>d</sup> Functional roles of genes and operons from the amino acid regulons are indicated in parentheses.

controls the expression of *metE*, *metH*, *metA*, and *metF* genes, which are under the dual control of MetJ and MetR.<sup>263</sup> Computational analysis of the distribution of MetJ-binding sites in bacteria whose genomes have a *metJ* ortholog revealed significant conservation of the MetJ regulon in  $\gamma$ -proteobacteria (D.A.R., unpublished observation). In a limited number of species (e.g., in 3 out of 11 *Shewanella* species), the MetJ regulon is extended to include the methionine degradation and salvage genes (*mdeA*, *mht*). The MetR regulon possibly has emerged earlier than MetJ, since it is present also in various  $\beta$ -proteobacteria.

In contrast, in the actinobacterium *C. glutamicum*, the McbR repressor responds to *S*-adenosylhomocysteine and coregulates the methionine biosynthesis, sulfur assimilation, and cysteine biosynthesis genes.<sup>193</sup> Comparative analysis confirmed major conservation of the McbR regulon in three other *Corynebacterium* species.<sup>264</sup> Although two other species from the Corynebacteriaceae group, *Nocardia farcinica* and *Mycobacterium smegmatis*, have a *mcbR* ortholog preceded by a candidate McbR binding site, the McbR regulon was not identified in other cctinobacteria. Thus, it is tempting to speculate that the global sulfur metabolism regulon McbR was only recently evolved in the common ancestor of corynebacteria.

Three different classes of *S*-adenosylmethionine-responsive RNA regulatory elements regulate the methionine metabolism in various taxonomic groups. The SAM-I riboswitch (or the S-box system) is widely distributed in the *Bacillus*/

*Clostridium* group and is also present in some additional diverse bacterial lineages.<sup>251</sup> Most SAM-II riboswitches were found in  $\alpha$ -proteobacteria and the CFB group.<sup>175</sup> Thus, it is likely that SAM-I and SAM-II were already present in the last common ancestors of firmicutes and  $\alpha$ -proteobacteria, respectively. However, among firmicutes, SAM-I riboswitches were only identified in the Bacillales and Clostridiales lineages but not in the Lactobacillales and Streptococcaceae, where they were likely substituted by other methionine-specific regulatory systems.<sup>251</sup>

The loss of the SAM-I regulatory system in the Streptococcaceae group is correlated with the emergence of two novel LysR-type TFs that control the methionine and cysteine metabolism in *Streptococcus* and *Lactococcus* species (Table 7A). MtaR was first identified as a regulator of methionine transport in the group B streptococci, but its binding site was unknown.<sup>265</sup> Comparative genomic analysis of the *Streptococcus* genomes allowed us to identify a potential binding motif (MET-box) in the regulatory regions of methionine biosynthesis and transport genes.<sup>251</sup> The co-occurrence of MET-boxes and *mtaR* orthologs suggests that MET-boxes are likely MtaR-binding sites. The *O*-acetylserine-responsive TF CmbR in *L. lactis* was characterized as a master regulator of the sulfur amino acid metabolism that controls all genes likely involved in methionine and cysteine synthesis and transport except *cysE* and *metEF*.<sup>266</sup> Interestingly, the latter operon is the only potential target of MtaR in *L. lactis*.<sup>251</sup> Further comparative analysis of the

CmbR regulon suggests that it mostly controls the cysteine metabolism in *Streptococcus* species and also has some overlap with the MtaR regulon (Galina Kovaleva, personal communication). The consensus binding sites of CmbR (CYS-box) and MtaR (MET-box) differ from each other but follow the general symmetry for LysR-type regulators.

In the Clostridiales and Bacillales groups, the methionine-specific T-box RNA elements regulate expression of only one gene, the methionyl-tRNA synthetase *metS*. In contrast, the Met-T-box regulation is extensively used only in the Lactobacillales group, where it exclusively controls methionine genes in the absence of the SAM-I riboswitch regulon. This suggests that the family of Met-T-boxes initially associated with the *metS* genes has been likely expanded in the Lactobacillales lineage to include most of the methionine metabolism genes.<sup>251</sup> Indeed, the phylogenetic analysis of T-box families suggests that these RNA regulatory elements are subject to frequent duplications, deletions, and horizontal transfer between species (Alexei G. Vitreschak, personal communication).

The *S*-adenosylmethionine synthetase gene *metK* is regulated by the SAM-I riboswitch in Bacillales and Clostridiales; however, this gene is not a member of the MtaR and Met-T-box regulons in the Lactobacillales (including the Streptococcaceae family), probably because they do not use *S*-adenosylmethionine as an effector.<sup>251</sup> This gap was recently filled by identification of a novel *S*-adenosylmethionine-responsive riboswitch (SAM-III) for translational regulation of *metK* in the Lactobacillales.<sup>176</sup> Limited phylogenetic distribution of the SAM-III riboswitch and its limited appearance in the genomes (it was found only upstream of *metK*) suggest that this regulatory element has emerged relatively recent in the common ancestor of the Lactobacillales.

## 4.2. Aromatic Amino Acid Metabolism

A similar variability in regulatory mechanisms was identified for the aromatic amino acid (ARO) biosynthesis pathway (Table 7B). Biosynthesis of tyrosine, phenylalanine, and tryptophan starts from the common chorismate biosynthesis pathway encoded by the *aro* genes and then divides into the terminal pathways that are specific for each aromatic amino acid.

In  $\gamma$ -proteobacteria, the control of this pathway is mediated by two aromatic amino acid-responsive TFs, the tyrosine/phenylalanine-specific regulator TyrR and the tryptophan repressor TrpR.<sup>125</sup> In addition, the phenylalanine and tryptophan operons are controlled by Phe- and Trp-specific transcriptional attenuators, respectively.<sup>267</sup> Although the TFs and their cognate DNA signals are conserved in  $\gamma$ -proteobacteria, the content of TrpR and TyrR regulons varies widely. In the Enterobacteriales, Pasteurellales, and Vibrionales lineages, TrpR and TyrR control the biosynthesis and transport of tryptophan and tyrosine/phenylalanine, respectively. Some genes are under dual control of two different regulatory systems; for instance, *aroL* and *mtr* are regulated by both TyrR and TrpR, whereas the *trp* operon is controlled by the TrpR repressor and tryptophan attenuator.<sup>125</sup> An ortholog of TyrR in the Pseudomonadales was characterized as PhhR, an activator of the phenylalanine degradation operon, which binds to a TyrR-box-like motif in the presence of phenylalanine or tyrosine.<sup>268</sup> The comparative genomic analysis of TyrR- and TrpR-like regulons in *Shewanella* species that belong to the Alteromonadales group of  $\gamma$ -pro-

teobacteria revealed large-scale shifts in the metabolic content of regulons (D.A.R., unpublished observation). In *Shewanella*, TyrR is predicted to regulate degradation and transport of various amino acids (e.g., branch chain amino acids, proline, phenylalanine), whereas TrpR likely controls the tyrosine biosynthesis and transport. Finally, the tryptophan-responsive TrpR repressor was experimentally characterized in *Chlamydia* species, where it regulates the tryptophan synthase operon.<sup>269</sup>

Tryptophan biosynthesis and transport genes in the *Bacillus/Clostridium* group are regulated at the RNA level by two different mechanisms, the Trp-specific T-box RNA elements and the RNA-binding TRAP protein.<sup>269,270</sup> The TRAP-mediated regulation is used in all *Bacillus* species except the *B. cereus* group. In contrast, other lineages in the *Bacillus/Clostridium* group (including *B. cereus*) use Trp-T-boxes for tryptophan control. The common pathway of aromatic amino acid biosynthesis (encoded by *aro* genes) is likely regulated by two different conserved DNA elements, termed PCE in the *Bacillus* species and ARO-box in the Streptococcales.<sup>144,259</sup> Note, though, that the *B. cereus* group does not have PCE elements and uses tyrosine-specific T-boxes to control *aro* genes. In other firmicutes, Tyr-T-boxes mostly regulate tyrosine-specific aminoacyl-tRNA synthetase (Alexei G. Vitreschak, personal communication). These observations suggest that T-boxes in some Gram-positive species have undergone multiple duplications, leading to the expansion of the respective amino acid regulons from aminoacyl-tRNA synthetases to the biosynthesis and transport.

## 4.3. Fructose Regulon in $\gamma$ -Proteobacteria

The fructose repressor FruR, which belongs to the LacI family of TFs, demonstrates a noteworthy example of regulon expansion. This TF has a pleiotropic regulatory role in *E. coli* and closely related *Salmonella* species.<sup>271</sup> It responds to the level of fructose-6-phosphate (Fru-6P) repressing the fructose utilization operon *fruBKA*. Therefore, it was initially named FruR. Later, it was also implicated in global regulation of more than 20 operons involved in the central carbohydrate pathways (e.g., glycolysis, gluconeogenesis, the Entner–Doudoroff pathway, and the TCA cycle), which led to an alternative name, Cra, for catabolite repressor–activator protein.<sup>272</sup> Comparative analysis of the FruR regulons in other species of the Enterobacteriales revealed an intermediate situation, with a smaller number of genes being controlled by FruR compared to *E. coli*.<sup>73</sup> For example, the FruR regulon in *Erwinia* and *Yersinia* species does not include the *mtlADR*, *pckA*, *fbp*, and *aceBAK* operons that are FurR regulated in *Escherichia* and *Salmonella* spp. In other groups of  $\gamma$ -proteobacteria (e.g., Vibrionales and Pseudomonadales), FruR appears to be just a local regulator of the *fruBKA* operon.<sup>73</sup> These observations suggest a possible evolutionary scenario for FruR. The initially local fructose uptake regulon was expanded in various species of the Enterobacteriales at different extents to become a global TF mediating Fru-6P-dependent catabolic regulation of the central carbon metabolism genes.

## 4.4. Iron and Manganese Regulatory Networks

Global control of iron and manganese homeostasis, including uptake, storage, and usage of these metals, is mediated by TFs from at least three major protein families, Fur, DtxR,



**Table 8. Major TF Regulons for Iron and Manganese Homeostasis in Bacteria**

taxonomic group	iron regulons <sup>a</sup>	manganese regulons <sup>a</sup>
Cyanobacteria	Fur	
Actinobacteria	IdeR (DtxR family)	
firmicutes	Fur	MntR (DtxR family)
$\gamma$ (Enterobacteriales, Xanthomonadales)	Fur	MntR
$\alpha$ (Rhizobiales)	Irr (Fur family), RirA (Rrf2 family)	Mur (Fur family)
exception: Bradyrhizobiaceae group	Irr	Mur
exception: <i>Mesorhizobium loti</i>	Irr, RirA	MntR
$\alpha$ (Rhodobacterales)	Irr, iron-rhodo-box (uncertain TF)	Mur
exception: <i>Rhodobacter capsulatus</i>	Irr, iron-rhodo-box	MntR
$\alpha$ (other groups), $\beta$ , $\delta$ , $\epsilon$ , $\gamma$ (other groups)	Fur	

<sup>a</sup> TF protein families are indicated in parentheses.

and Rrf2.<sup>273</sup> The DtxR family of metalloregulators includes the manganese repressor MntR in proteobacteria and firmicutes and the iron-responsive regulators IdeR and DtxR in actinobacteria. TFs of the Rrf2 family are widespread in bacteria, where they regulate diverse metabolic processes, such as metabolism of nitrogen oxides (NsrR), Fe–S cluster biogenesis (IscR), and iron homeostasis (RirA). Metalloregulators from the Fur superfamily respond to specific metal ions (iron, zinc, manganese, nickel) and regulate respective metabolic pathways.

Fur, the global iron-responsive TF, is the most widely distributed regulator of iron homeostasis, since it is present in Gram-negative proteobacteria ( $\gamma$ -,  $\beta$ -,  $\epsilon$ -, and  $\delta$ -subdivisions), Gram-positive bacteria, and cyanobacteria. Some lineages of  $\alpha$ -proteobacteria (e.g., *Caulobacter*, *Magnetospirillum*) are predicted to have a similar regulon, suggesting that the last common ancestor of  $\alpha$ -proteobacteria used a Fur-like protein to control iron metabolism.<sup>240</sup> However, recent experimental<sup>274,275</sup> and comparative genomic<sup>240</sup> analyses demonstrated that the iron and manganese regulons in the Rhizobiales and Rhodobacterales groups of  $\alpha$ -proteobacteria are significantly different from other microbial lineages (Table 8).

An evolutionary scenario suggested for the Rhizobiales and Rhodobacterales lineages includes the following events (Table 8):<sup>240</sup> (i) change of the effector molecule (from Fe<sup>2+</sup> to Mn<sup>2+</sup>) and the regulon content (from the iron metabolism genes to the manganese uptake genes) for the Fur proteins, which were therefore renamed “Mur”; (ii) recruitment of two novel TFs to control of the iron metabolism, the [Fe–S]-responsive repressor RirA and the heme-responsive regulator Irr, that sense the physiological consequence of the iron availability rather than iron concentration *per se*; and (iii) the secondary loss of Mur and its substitution by MntR in at least two species, *Mesorhizobium loti* and *Rhodobacter capsulatus* (possibly achieved by horizontal gene transfer).

Interestingly, the candidate consensus DNA-binding sites of Fur and Mur in  $\alpha$ -proteobacteria still resemble each other and show a similarity to the classical Fur-box consensus from  $\gamma$ -proteobacteria and firmicutes. Furthermore, the Fur/Mur sites show faint similarity to RirA-binding sites (consensus TG-N11-CA), suggesting that iron-regulatory signals in  $\alpha$ -proteobacteria may have evolved from canonical Fur sites. Indeed, theoretical calculations of bacterial TFBSs demonstrate that TFBSs even weakly conforming to the requirements of cognate TFs may provide a selective advantage for the regulon to function and further positive selection may perfect a TFBS to a higher-affinity state.<sup>276</sup>

Comparative genomic reconstruction of the iron and manganese regulatory networks based on the identification of several classes of TFBS motifs in  $\alpha$ -proteobacteria

revealed the significant variability and cross-connectivity of these TRNs as follows: (i) the proposed mechanisms of regulation are different between various lineages and species (Table 8); (ii) the functional content of regulons is variable due to lineage-specific regulon extensions and reductions; (iii) there is an overlap between regulons and potential regulatory cascades involving the two different iron-responsive TFs, Irr and RirA.<sup>240</sup>

## 5. Directions for Future Studies

This review illustrates major advances in comparative genomic reconstruction of regulons associated with metabolic pathways in microorganisms. This area is still very young, and many unresolved questions and open problems listed below have to be addressed in the coming years.

(1) Development of new powerful comparative genomics tools for *in silico* analysis, annotation, and computational prediction of TRNs in the multitude of sequenced microbial genomes. Key components of prokaryotic TRNs, TFs, and their TFBSs need to be systematically classified and captured in specialized databases.

(2) Further accumulation of high-quality expression data generated by transcriptomics and proteomics techniques, and protein–DNA interaction (ChIP-on-chip) in a broad range of species and experimental conditions.

(3) Capture of experimental and computational data about TRNs within a framework of genomic integrations supporting reconstruction and comparative analysis of regulatory and metabolic networks. Such a broad integration will strongly impact annotation and reconstruction of both TRNs and metabolic pathways including prediction of previously uncharacterized genes (regulators, enzymes, transporters).

(4) Systematic comparison and cross-evaluation of high-throughput experimental data and *in silico* reconstructed microbial regulons. Assessment of advantages and limitations for each of these techniques.

(5) Development of theoretical models for the evolution of TRNs in prokaryotes. Incorporation of horizontal transfer and duplication into evolutionary models. Systematic analysis of coevolution of TFs and their TFBSs.

## 6. Abbreviations

ABC	ATP binding cassette
ChIP	chromatine immunoprecipitation
ECF	extracytoplasmic function sigma factors
EM	expectation-maximization method
HMM	hidden Markov model
HTH	helix-turn helix
NAD	nicotinamide adenine dinucleotide
ORF	open reading frame

PWM	positional weight matrix
TF	transcription factor
TFBS	transcription factor binding site
TRN	transcription regulatory network
UTR	untranslated region

## 7. Acknowledgments

I thank Andrei Osterman and Mikhail Gelfand for useful discussions, careful reading, and comments during the preparation of this manuscript. I am grateful to Aaron Best, Nadia Raffaelli, Galina Kovaleva, Alexei Vitreschak, Dmitry Ravcheev, and Natalia Doroshchuk for sharing unpublished data. I am indebted to Andrey Mironov, who designed and wrote the Genome Explorer program that was used in many comparative genomic studies of microbial regulation, and Ross Overbeek and the SEED development team for providing access to integrated genomes and tools for their comparative analysis. This work has been partially supported by the National Institutes of Health (1R01AI066244-01A2), the Department of Energy (DE-FG02-ER64384), the Howard Hughes Medical Institute (Grant 55005610, “Comparative genomics and evolution of regulatory systems”), and the Russian Academy of Sciences (Program “Molecular and Cellular Biology”).

## 8. Supporting Information

Tables showing the distribution of TFs in sequenced prokaryotic genomes and the distribution of sigma factors in sequenced bacterial genomes. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## 9. References

- Baumberg, S. *Prokaryotic Gene Expression*; Oxford University Press: Oxford, 1999.
- Lloyd, G.; Landini, P.; Busby, S. *Essays Biochem.* **2001**, *37*, 17.
- Kazmierczak, M. J.; Wiedmann, M.; Boor, K. J. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 527.
- Gollnick, P.; Babitzke, P. *Biochim. Biophys. Acta* **2002**, *1577*, 240.
- Hodgson, D. A. *Signals, Switches, Regulons, and Cascades*; Cambridge University Press: Cambridge, 2002.
- Winkler, W. C.; Breaker, R. R. *Annu. Rev. Microbiol.* **2005**, *59*, 487.
- Gelfand, M. S. *Mol. Biol.* **2006**, *40*, 609.
- Gottesman, S. *Annu. Rev. Microbiol.* **2004**, *58*, 303.
- Makarova, K. S.; Mironov, A. A.; Gelfand, M. S. *Genome Biol.* **2001**, *2*, Research0013.
- Rodionov, D. A.; Vitreschak, A. G.; Mironov, A. A.; Gelfand, M. S. *J. Biol. Chem.* **2002**, *277*, 48949.
- Neidhardt, F. C.; Savageau, M. F. In *Escherichia coli and Salmonella: cellular and molecular biology*; Neidhardt, F. C., Ed.; ASM Press: Washington, DC, 1996; Vol. 2, p 1310.
- Herrgard, M. J.; Covert, M. W.; Palsson, B. O. *Curr. Opin. Biotechnol.* **2004**, *15*, 70.
- Zhou, D.; Yang, R. *Cell. Mol. Life Sci.* **2006**, *63*, 2260.
- Shen-Orr, S. S.; Milo, R.; Mangan, S.; Alon, U. *Nat. Genet.* **2002**, *31*, 64.
- Dharmadi, Y.; Gonzalez, R. *Biotechnol. Prog.* **2004**, *20*, 1309.
- Grainger, D. C.; Aiba, H.; Hurd, D.; Browning, D. F.; Busby, S. J. *Nucleic Acids Res.* **2007**, *35*, 269.
- Grainger, D. C.; Hurd, D.; Harrison, M.; Holdstock, J.; Busby, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17693.
- Laub, M. T.; Chen, S. L.; Shapiro, L.; McAdams, H. H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4632.
- Breier, A. M.; Grossman, A. D. *Mol. Microbiol.* **2007**, *64*, 703.
- Molle, V.; Fujita, M.; Jensen, S. T.; Eichenberger, P.; Gonzalez-Pastor, J. E.; Liu, J. S.; Losick, R. *Mol. Microbiol.* **2003**, *50*, 1683.
- von Kruger, W. M.; Lery, L. M.; Soares, M. R.; de Neves-Manta, F. S.; Batista e Silva, C. M.; Neves-Ferreira, A. G.; Perales, J.; Bisch, P. M. *Proteomics* **2006**, *6*, 1495.
- Sarma, A. D.; Emerich, D. W. *Proteomics* **2005**, *5*, 4170.
- Todd, J. D.; Sawers, G.; Johnston, A. W. *Mol. Genet. Genomics* **2005**, *273*, 197.
- Sadygov, R. G.; Cociorva, D.; Yates, J. R., III. *Nat. Methods* **2004**, *1*, 195.
- Yates, J. R., III. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 297.
- Gelfand, M. S. *Curr. Opin. Struct. Biol.* **2006**, *16*, 420.
- Gelfand, M. S.; Novichkov, P. S.; Novichkova, E. S.; Mironov, A. A. *Briefings Bioinf.* **2000**, *1*, 357.
- Gelfand, M. S. *Res. Microbiol.* **1999**, *150*, 755.
- Stormo, G. D.; Tan, K. *Curr. Opin. Microbiol.* **2002**, *5*, 149.
- Browning, D. F.; Busby, S. J. *Nat. Rev. Microbiol.* **2004**, *2*, 57.
- Collado-Vides, J.; Magasanik, B.; Gralla, J. D. *Microbiol. Rev.* **1991**, *55*, 371.
- Espinosa, V.; Gonzalez, A. D.; Vasconcelos, A. T.; Huerta, A. M.; Collado-Vides, J. *J. Mol. Biol.* **2005**, *354*, 184.
- Moreno-Campuzano, S.; Janga, S. C.; Perez-Rueda, E. *BMC Genomics* **2006**, *7*, 147.
- Lanzer, M.; Bujard, H. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 8973.
- Zheng, D.; Constantinidou, C.; Hobman, J. L.; Minchin, S. D. *Nucleic Acids Res.* **2004**, *32*, 5874.
- von Hippel, P. H. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 79.
- Schneider, T. D.; Stephens, R. M. *Nucleic Acids Res.* **1990**, *18*, 6097.
- Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. E. *Genome Res.* **2004**, *14*, 1188.
- Stormo, G. D. *Bioinformatics* **2000**, *16*, 16.
- Hertz, G. Z.; Stormo, G. D. *Bioinformatics* **1999**, *15*, 563.
- Stormo, G. D.; Schneider, T. D.; Gold, L.; Ehrenfeucht, A. *Nucleic Acids Res.* **1982**, *10*, 2997.
- Staden, R. *Comput. Appl. Biosci.* **1989**, *5*, 89.
- Claverie, J. M.; Audic, S. *Comput. Appl. Biosci.* **1996**, *12*, 431.
- Staden, R. *Nucleic Acids Res.* **1984**, *12*, 505.
- Schneider, T. D.; Stormo, G. D.; Gold, L.; Ehrenfeucht, A. *J. Mol. Biol.* **1986**, *188*, 415.
- Berg, O. G.; von Hippel, P. H. *J. Mol. Biol.* **1987**, *193*, 723.
- Bailey, T. L.; Gribskov, M. *Bioinformatics* **1998**, *14*, 48.
- Mironov, A. A.; Vinokurova, N. P.; Gelfand, M. S. *Mol. Biol.* **2000**, *34*, 253.
- Horsburgh, M. J.; Ingham, E.; Foster, S. J. *J. Bacteriol.* **2001**, *183*, 468.
- Perez-Rueda, E.; Collado-Vides, J. *Nucleic Acids Res.* **2000**, *28*, 1838.
- Mandal, M.; Boese, B.; Barrick, J. E.; Winkler, W. C.; Breaker, R. R. *Cell* **2003**, *113*, 577.
- Perez-Rueda, E.; Collado-Vides, J.; Segovia, L. *Comput. Biol. Chem.* **2004**, *28*, 341.
- Madan Babu, M.; Teichmann, S. A.; Aravind, L. *J. Mol. Biol.* **2006**, *358*, 614.
- Minezaki, Y.; Homma, K.; Nishikawa, K. *DNA Res.* **2005**, *12*, 269.
- van Nimwegen, E. *Trends Genet.* **2003**, *19*, 479.
- Kummerfeld, S. K.; Teichmann, S. A. *Nucleic Acids Res.* **2006**, *34*, D74.
- Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonhammer, E. L.; Bateman, A. *Nucleic Acids Res.* **2006**, *34*, D247.
- Wilson, D.; Madera, M.; Vogel, C.; Chothia, C.; Gough, J. *Nucleic Acids Res.* **2007**, *35*, D308.
- Dobrindt, U.; Hacker, J. *Curr. Opin. Microbiol.* **2001**, *4*, 550.
- Studholme, D. J.; Dixon, R. *FEMS Microbiol. Lett.* **2004**, *230*, 215.
- Wang, L.; Sun, Y. P.; Chen, W. L.; Li, J. H.; Zhang, C. C. *FEMS Microbiol. Lett.* **2002**, *217*, 155.
- Goldman, B. S.; Nierman, W. C.; Kaiser, D.; Slater, S. C.; Durkin, A. S.; Eisen, J. A.; Ronning, C. M.; Barbazuk, W. B.; Blanchard, M.; Field, C.; Halling, C.; Hinkle, G.; Iartchuk, O.; Kim, H. S.; Mackenzie, C.; Madupu, R.; Miller, N.; Shvartsbeyn, A.; Sullivan, S. A.; Vaudin, M.; Wiegand, R.; Kaplan, H. B. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15200.
- Molina-Henares, A. J.; Krell, T.; Eugenia Guazzaroni, M.; Segura, A.; Ramos, J. L. *FEMS Microbiol. Rev.* **2006**, *30*, 157.
- Galperin, M. Y. *J. Bacteriol.* **2006**, *188*, 4169.
- Alm, E.; Huang, K.; Arkin, A. *PLoS Comput. Biol.* **2006**, *2*, e143.
- Bouhouche, N.; Syvanen, M.; Kado, C. I. *Trends Microbiol.* **2000**, *8*, 77.
- Perez-Rueda, E.; Gralla, J. D.; Collado-Vides, J. *J. Mol. Biol.* **1998**, *275*, 165.
- Madan Babu, M.; Teichmann, S. A. *Trends Genet.* **2003**, *19*, 75.
- Gerasimova, A. V.; Gelfand, M. S. *J. Bioinf. Comput. Biol.* **2005**, *3*, 1007.
- Rodionov, D. A.; Mironov, A. A.; Gelfand, M. S. *Genome Res.* **2002**, *12*, 1507.
- Martinez-Antonio, A.; Collado-Vides, J. *Curr. Opin. Microbiol.* **2003**, *6*, 482.
- Doerks, T.; Andrade, M. A.; Lathe, W., III; von Mering, C.; Bork, P. *Trends Genet.* **2004**, *20*, 126.

- (73) Gelfand, M. S.; Laikova, O. N. In *Frontiers in Computational Genomics*; Galperin, M. Y., Koonin, E. V., Eds.; Caister Academic Press: Wymondham, 2003; p 195.
- (74) Kazmierczak, M. J.; Wiedmann, M.; Boor, K. J. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 527.
- (75) Reitzer, L.; Schneider, B. L. *Microbiol. Mol. Biol. Rev.* **2001**, *65*, 422.
- (76) Helmann, J. D. *Adv. Microb. Physiol.* **2002**, *46*, 47.
- (77) Kiil, K.; Ferchaud, J. B.; David, C.; Binnewies, T. T.; Wu, H.; Sicheritz-Ponten, T.; Willenbrock, H.; Ussery, D. W. *Microbiology* **2005**, *151*, 3447.
- (78) Salgado, H.; Gama-Castro, S.; Peralta-Gil, M.; Diaz-Peredo, E.; Sanchez-Solano, F.; Santos-Zavaleta, A.; Martinez-Flores, I.; Jimenez-Jacinto, V.; Bonavides-Martinez, C.; Segura-Salazar, J.; Martinez-Antonio, A.; Collado-Vides, J. *Nucleic Acids Res.* **2006**, *34*, D394.
- (79) Makita, Y.; Nakao, M.; Ogasawara, N.; Nakai, K. *Nucleic Acids Res.* **2004**, *32*, D75.
- (80) Baumbach, J.; Brinkrolf, K.; Czaja, L. F.; Rahmann, S.; Tauch, A. *BMC Genomics* **2006**, *7*, 24.
- (81) Jacques, P. E.; Gervais, A. L.; Cantin, M.; Lucier, J. F.; Dallaire, G.; Drouin, G.; Gaudreau, L.; Goule, J.; Brzezinski, R. *Bioinformatics* **2005**, *21*, 2563.
- (82) Pareja, E.; Pareja-Tobes, P.; Manrique, M.; Pareja-Tobes, E.; Bonal, J.; Tobes, R. *BMC Microbiol.* **2006**, *6*, 29.
- (83) Tobes, R.; Ramos, J. L. *Nucleic Acids Res.* **2002**, *30*, 318.
- (84) Krell, T.; Molina-Henares, A. J.; Ramos, J. L. *Protein Sci.* **2006**, *15*, 1207.
- (85) Ramos, J. L.; Martinez-Bueno, M.; Molina-Henares, A. J.; Teran, W.; Watanabe, K.; Zhang, X.; Gallegos, M. T.; Brennan, R.; Tobes, R. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 326.
- (86) Martinez-Bueno, M.; Molina-Henares, A. J.; Pareja, E.; Ramos, J. L.; Tobes, R. *Bioinformatics* **2004**, *20*, 2787.
- (87) D'Souza, M.; Glass, E. M.; Syed, M. H.; Zhang, Y.; Rodriguez, A.; Maltsev, N.; Galperin, M. Y. *Nucleic Acids Res.* **2007**, *35*, D271.
- (88) Wu, J.; Zhao, F.; Wang, S.; Deng, G.; Wang, J.; Bai, J.; Lu, J.; Qu, J.; Bao, Q. *BMC Genomics* **2007**, *8*, 104.
- (89) Munch, R.; Hiller, K.; Grote, A.; Scheer, M.; Klein, J.; Schobert, M.; Jahn, D. *Bioinformatics* **2005**, *21*, 4187.
- (90) Kazakov, A. E.; Cipriano, M. J.; Novichkov, P. S.; Minovitsky, S.; Vinogradov, D. V.; Arkin, A.; Mironov, A. A.; Gelfand, M. S.; Dubchak, I. *Nucleic Acids Res.* **2007**, *35*, D407.
- (91) Perez, A. G.; Angarica, V. E.; Vasconcelos, A. T.; Collado-Vides, J. *Nucleic Acids Res.* **2007**, *35*, D132.
- (92) Brazma, A.; Jonassen, I.; Eidhammer, I.; Gilbert, D. J. *Comput. Biol.* **1998**, *5*, 279.
- (93) D'Haeseleer, P. *Nat. Biotechnol.* **2006**, *24*, 959.
- (94) Sinha, S.; Tompa, M. *Nucleic Acids Res.* **2002**, *30*, 5549.
- (95) Eskin, E.; Pevzner, P. A. *Bioinformatics* **2002**, *18*, S354.
- (96) Pavesi, G.; Mauri, G.; Pesole, G. *Bioinformatics* **2001**, *17*, S207.
- (97) Marsan, L.; Sagot, M. F. *J. Comput. Biol.* **2000**, *7*, 345.
- (98) McGuire, A. M.; Hughes, J. D.; Church, G. M. *Genome Res.* **2000**, *10*, 744.
- (99) Tompa, M.; Li, N.; Bailey, T. L.; Church, G. M.; De Moor, B.; Eskin, E.; Favorov, A. V.; Frith, M. C.; Fu, Y.; Kent, W. J.; Makeev, V. J.; Mironov, A. A.; Noble, W. S.; Pavesi, G.; Pesole, G.; Regnier, M.; Simonis, N.; Sinha, S.; Thijs, G.; van Helden, J.; Vandenbogaert, M.; Weng, Z.; Workman, C.; Ye, C.; Zhu, Z. *Nat. Biotechnol.* **2005**, *23*, 137.
- (100) Stormo, G. D.; Hartzell, G. W. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 1183.
- (101) Lawrence, C. E.; Reilly, A. A. *Proteins* **1990**, *7*, 41.
- (102) Bailey, T. L.; Elkan, C. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 28.
- (103) Gelfand, M. S.; Koonin, E. V.; Mironov, A. A. *Nucleic Acids Res.* **2000**, *28*, 695.
- (104) Lawrence, C. E.; Altschul, S. F.; Boguski, M. S.; Liu, J. S.; Neuwald, A. F.; Wootton, J. C. *Science* **1993**, *262*, 208.
- (105) Roth, F. P.; Hughes, J. D.; Estep, P. W.; Church, G. M. *Nat. Biotechnol.* **1998**, *16*, 939.
- (106) Thompson, W.; Rouchka, E. C.; Lawrence, C. E. *Nucleic Acids Res.* **2003**, *31*, 3580.
- (107) Favorov, A. V.; Gelfand, M. S.; Gerasimova, A. V.; Ravcheev, D. A.; Mironov, A. A.; Makeev, V. J. *Bioinformatics* **2005**, *21*, 2240.
- (108) Mwangi, M. M.; Siggia, E. D. *BMC Bioinf.* **2003**, *4*, 18.
- (109) Li, H.; Rhodius, V.; Gross, C.; Siggia, E. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11772.
- (110) Studholme, D. J.; Bentley, S. D.; Kormanec, J. *BMC Microbiol.* **2004**, *4*, 14.
- (111) Djordjevic, M.; Sengupta, A. M.; Shraiman, B. I. *Genome Res.* **2003**, *13*, 2381.
- (112) Thieffry, D.; Salgado, H.; Huerta, A. M.; Collado-Vides, J. *Bioinformatics* **1998**, *14*, 391.
- (113) Constantinidou, C.; Hobman, J. L.; Griffiths, L.; Patel, M. D.; Penn, C. W.; Cole, J. A.; Overton, T. W. *J. Biol. Chem.* **2006**, *281*, 4802.
- (114) Liu, X.; De Wulf, P. *J. Biol. Chem.* **2004**, *279*, 12588.
- (115) Yoshida, K.; Yamaguchi, H.; Kinehara, M.; Ohki, Y. H.; Nakaura, Y.; Fujita, Y. *Mol. Microbiol.* **2003**, *49*, 157.
- (116) Brune, I.; Werner, H.; Huser, A. T.; Kalinowski, J.; Puhler, A.; Tauch, A. *BMC Genomics* **2006**, *7*, 21.
- (117) Reents, H.; Munch, R.; Dammeyer, T.; Jahn, D.; Hartig, E. *J. Bacteriol.* **2006**, *188*, 1103.
- (118) Zheng, M.; Wang, X.; Doan, B.; Lewis, K. A.; Schneider, T. D.; Storz, G. *J. Bacteriol.* **2001**, *183*, 4571.
- (119) Rodionov, D. A.; Mironov, A. A.; Rakhmaninova, A. B.; Gelfand, M. S. *Mol. Microbiol.* **2000**, *38*, 673.
- (120) Laikova, O. N.; Mironov, A. A.; Gelfand, M. S. *FEMS Microbiol. Lett.* **2001**, *205*, 315.
- (121) Rodionov, D. A.; Mironov, A. A.; Gelfand, M. S. *FEMS Microbiol. Lett.* **2001**, *205*, 305.
- (122) Permina, E. A.; Gelfand, M. S. *J. Mol. Microbiol. Biotechnol.* **2003**, *6*, 174.
- (123) Panina, E. M.; Mironov, A. A.; Gelfand, M. S. *Nucleic Acids Res.* **2001**, *29*, 5195.
- (124) Panina, E. M.; Mironov, A. A.; Gelfand, M. S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9912.
- (125) Panina, E. M.; Vitreschak, A. G.; Mironov, A. A.; Gelfand, M. S. *J. Mol. Microbiol. Biotechnol.* **2001**, *3*, 529.
- (126) Ravcheev, D. A.; Rakhmaninova, A. B.; Mironov, A. A.; Gelfand, M. S. *Mol. Biol.* **2005**, *39*, 832.
- (127) Doroshchuk, N. A.; Gelfand, M. S.; Rodionov, D. A. *Mol. Biol.* **2006**, *40*, 919.
- (128) Tan, K.; Moreno-Hagelsieb, G.; Collado-Vides, J.; Stormo, G. D. *Genome Res.* **2001**, *11*, 566.
- (129) Mironov, A. A.; Koonin, E. V.; Roytberg, M. A.; Gelfand, M. S. *Nucleic Acids Res.* **1999**, *27*, 2981.
- (130) Blanchette, M.; Tompa, M. *Genome Res.* **2002**, *12*, 739.
- (131) Tagle, D. A.; Koop, B. F.; Goodman, M.; Slightom, J. L.; Hess, D. L.; Jones, R. T. *J. Mol. Biol.* **1988**, *203*, 439.
- (132) McCue, L. A.; Thompson, W.; Carmack, C. S.; Lawrence, C. E. *Genome Res.* **2002**, *12*, 1523.
- (133) Florea, L.; McClelland, M.; Riemer, C.; Schwartz, S.; Miller, W. *Nucleic Acids Res.* **2003**, *31*, 3527.
- (134) Dubchak, I.; Ryaboy, D. V. *Methods Mol. Biol.* **2006**, *338*, 69.
- (135) Boffelli, D.; McAuliffe, J.; Ovcharenko, D.; Lewis, K. D.; Ovcharenko, I.; Pachter, L.; Rubin, E. M. *Science* **2003**, *299*, 1391.
- (136) Neph, S.; Tompa, M. *Nucleic Acids Res.* **2006**, *34*, W366.
- (137) Siddharthan, R.; Siggia, E. D.; van Nimwegen, E. *PLoS Comput. Biol.* **2005**, *1*, e67.
- (138) Sinha, S.; Blanchette, M.; Tompa, M. *BMC Bioinf.* **2004**, *5*, 170.
- (139) McCue, L.; Thompson, W.; Carmack, C.; Ryan, M. P.; Liu, J. S.; Derbyshire, V.; Lawrence, C. E. *Nucleic Acids Res.* **2001**, *29*, 774.
- (140) Qin, Z. S.; McCue, L. A.; Thompson, W.; Mayerhofer, L.; Lawrence, C. E.; Liu, J. S. *Nat. Biotechnol.* **2003**, *21*, 435.
- (141) Zhang, Y. M.; Marrakchi, H.; Rock, C. O. *J. Biol. Chem.* **2002**, *277*, 15558.
- (142) Torrents, E.; Grinberg, I.; Gorovitz-Harris, B.; Lundström, H.; Borovok, I.; Aharonowitz, Y.; Sjöberg, B.; Cohen, G. *J. Bacteriol.* **2007**, *189*, May 11.
- (143) Conlan, S.; Lawrence, C.; McCue, L. A. *Appl. Environ. Microbiol.* **2005**, *71*, 7442.
- (144) Terai, G.; Takagi, T.; Nakai, K. *Genome Biol.* **2001**, *2*, Research0048.
- (145) Wels, M.; Francke, C.; Kerkhoven, R.; Kleerebezem, M.; Siezen, R. J. *Nucleic Acids Res.* **2006**, *34*, 1947.
- (146) Carmack, C. S.; McCue, L. A.; Newberg, L. A.; Lawrence, C. E. *Algorithms Mol. Biol.* **2007**, *2*, 1.
- (147) Alkema, W. B.; Lenhard, B.; Wasserman, W. W. *Genome Res.* **2004**, *14*, 1362.
- (148) Rodionov, D. A.; Gelfand, M. S. *Trends Genet.* **2005**, *21*, 385.
- (149) Morozov, A. V.; Siggia, E. D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7068.
- (150) Sandelin, A.; Wasserman, W. W. *J. Mol. Biol.* **2004**, *338*, 207.
- (151) Tan, K.; McCue, L. A.; Stormo, G. D. *Genome Res.* **2005**, *15*, 312.
- (152) Osterman, A.; Overbeek, R. *Curr. Opin. Chem. Biol.* **2003**, *7*, 238.
- (153) Wolf, Y. I.; Rogozin, I. B.; Kondrashov, A. S.; Koonin, E. V. *Genome Res.* **2001**, *11*, 356.
- (154) Grinberg, I.; Shteinberg, T.; Gorovitz, B.; Aharonowitz, Y.; Cohen, G.; Borovok, I. *J. Bacteriol.* **2006**, *188*, 7635.
- (155) Winkler, W. C.; Breaker, R. R. *Annu. Rev. Microbiol.* **2005**, *59*, 487.
- (156) Vitreschak, A. G.; Rodionov, D. A.; Mironov, A. A.; Gelfand, M. S. *Trends Genet.* **2004**, *20*, 44.
- (157) Gelfand, M. S.; Mironov, A. A.; Jomantas, J.; Kozlov, Y. I.; Perumov, D. A. *Trends Genet.* **1999**, *15*, 439.
- (158) Vitreschak, A. G.; Rodionov, D. A.; Mironov, A. A.; Gelfand, M. S. *Nucleic Acids Res.* **2002**, *30*, 3141.



- (159) Vitreschak, A. G.; Rodionov, D. A.; Mironov, A. A.; Gelfand, M. S. *RNA* **2003**, *9*, 1084.
- (160) Rodionov, D. A.; Vitreschak, A. G.; Mironov, A. A.; Gelfand, M. S. *Nucleic Acids Res.* **2003**, *31*, 6748.
- (161) Grundy, F. J.; Henkin, T. M. *Mol. Microbiol.* **1998**, *30*, 737.
- (162) Roth, A.; Winkler, W. C.; Regulski, E. E.; Lee, B. W.; Lim, J.; Jona, I.; Barrick, J. E.; Ritwik, A.; Kim, J. N.; Welz, R.; Iwata-Reuyl, D.; Breaker, R. R. *Nat. Struct. Mol. Biol.* **2007**, *14*, 308.
- (163) Winkler, W. C.; Cohen-Chalamish, S.; Breaker, R. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15908.
- (164) Mironov, A. S.; Gusarov, I.; Rafikov, R.; Lopez, L. E.; Shatalin, K.; Kreneva, R. A.; Perumov, D. A.; Nudler, E. *Cell* **2002**, *111*, 747.
- (165) Nahvi, A.; Barrick, J. E.; Breaker, R. R. *Nucleic Acids Res.* **2004**, *32*, 143.
- (166) Winkler, W.; Nahvi, A.; Breaker, R. R. *Nature* **2002**, *419*, 952.
- (167) Sudarsan, N.; Wickiser, J. K.; Nakamura, S.; Ebert, M. S.; Breaker, R. R. *Genes Dev.* **2003**, *17*, 2688.
- (168) McDaniel, B. A.; Grundy, F. J.; Artsimovitch, I.; Henkin, T. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3083.
- (169) Griffiths-Jones, S.; Moxon, S.; Marshall, M.; Khanna, A.; Eddy, S. R.; Bateman, A. *Nucleic Acids Res.* **2005**, *33*, D121.
- (170) Abreu-Goodger, C.; Merino, E. *Nucleic Acids Res.* **2005**, *33*, W690.
- (171) Huang, H. Y.; Chien, C. H.; Jen, K. H.; Huang, H. D. *Nucleic Acids Res.* **2006**, *34*, W429.
- (172) Bengert, P.; Dandekar, T. *Nucleic Acids Res.* **2004**, *32*, W154.
- (173) Abreu-Goodger, C.; Ontiveros-Palacios, N.; Ciria, R.; Merino, E. *Trends Genet.* **2004**, *20*, 475.
- (174) Barrick, J. E.; Corbino, K. A.; Winkler, W. C.; Nahvi, A.; Mandal, M.; Collins, J.; Lee, M.; Roth, A.; Sudarsan, N.; Jona, I.; Wickiser, J. K.; Breaker, R. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6421.
- (175) Corbino, K. A.; Barrick, J. E.; Lim, J.; Welz, R.; Tucker, B. J.; Puskarz, I.; Mandal, M.; Rudnick, N. D.; Breaker, R. R. *Genome Biol.* **2005**, *6*, R70.
- (176) Fuchs, R. T.; Grundy, F. J.; Henkin, T. M. *Nat. Struct. Mol. Biol.* **2006**, *13*, 226.
- (177) Mandal, M.; Lee, M.; Barrick, J. E.; Weinberg, Z.; Emilsson, G. M.; Ruzzo, W. L.; Breaker, R. R. *Science* **2004**, *306*, 275.
- (178) Hatfield, G. W.; Hung, S. P.; Baldi, P. *Mol. Microbiol.* **2003**, *47*, 871.
- (179) Hudson, M. E.; Snyder, M. *Biotechniques* **2006**, *41*, 673.
- (180) Lee, T. I.; Johnstone, S. E.; Young, R. A. *Nat. Protoc.* **2006**, *1*, 729.
- (181) Gianchandani, E. P.; Papin, J. A.; Price, N. D.; Joyce, A. R.; Palsson, B. O. *PLoS Comput. Biol.* **2006**, *2*, e101.
- (182) Herrgard, M. J.; Covert, M. W.; Palsson, B. O. *Genome Res.* **2003**, *13*, 2423.
- (183) Faith, J. J.; Hayete, B.; Thaden, J. T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J. J.; Gardner, T. S. *PLoS Biol.* **2007**, *5*, e8.
- (184) Gosset, G.; Zhang, Z.; Nayyar, S.; Cuevas, W. A.; Saier, M. H., Jr. *J. Bacteriol.* **2004**, *186*, 3516.
- (185) McHugh, J. P.; Rodriguez-Quinones, F.; Abdul-Tehrani, H.; Svis-tunenko, D. A.; Poole, R. K.; Cooper, C. E.; Andrews, S. C. *J. Biol. Chem.* **2003**, *278*, 29478.
- (186) Lorca, G. L.; Chung, Y. J.; Barabote, R. D.; Weyler, W.; Schilling, C. H.; Saier, M. H., Jr. *J. Bacteriol.* **2005**, *187*, 7826.
- (187) Wan, X. F.; Verberkmoes, N. C.; McCue, L. A.; Stanek, D.; Connelly, H.; Hauser, L. J.; Wu, L.; Liu, X.; Yan, T.; Leapheart, A.; Hettich, R. L.; Zhou, J.; Thompson, D. K. *J. Bacteriol.* **2004**, *186*, 8385.
- (188) Zhou, D.; Qin, L.; Han, Y.; Qiu, J.; Chen, Z.; Li, B.; Song, Y.; Wang, J.; Guo, Z.; Zhai, J.; Du, Z.; Wang, X.; Yang, R. *FEMS Microbiol. Lett.* **2006**, *258*, 9.
- (189) Au, N.; Kuester-Schoeck, E.; Mandava, V.; Bothwell, L. E.; Canny, S. P.; Chachu, K.; Colavito, S. A.; Fuller, S. N.; Groban, E. S.; Hensley, L. A.; O'Brien, T. C.; Shah, A.; Tierney, J. T.; Tamm, L. L.; O'Gara, T. M.; Goranov, A. I.; Grossman, A. D.; Lovett, C. M. *J. Bacteriol.* **2005**, *187*, 7655.
- (190) Wennerhold, J.; Bott, M. *J. Bacteriol.* **2006**, *188*, 2907.
- (191) Rudolph, G.; Semini, G.; Hauser, F.; Lindemann, A.; Friberg, M.; Hennecke, H.; Fischer, H. M. *J. Bacteriol.* **2006**, *188*, 733.
- (192) den Hengst, C. D.; van Hijum, S. A.; Geurts, J. M.; Nauta, A.; Kok, J.; Kuipers, O. P. *J. Biol. Chem.* **2005**, *280*, 34332.
- (193) Rey, D. A.; Nentwich, S. S.; Koch, D. J.; Ruckert, C.; Puhler, A.; Tauch, A.; Kalinowski, J. *Mol. Microbiol.* **2005**, *56*, 871.
- (194) Tsyganova, M. O.; Gelfand, M. S.; Ravcheev, D. A. *Mol. Biol.*, in press.
- (195) Salmon, K. A.; Hung, S. P.; Steffen, N. R.; Krupp, R.; Baldi, P.; Hatfield, G. W.; Gunsalus, R. P. *J. Biol. Chem.* **2005**, *280*, 15084.
- (196) Salmon, K.; Hung, S. P.; Mekjian, K.; Baldi, P.; Hatfield, G. W.; Gunsalus, R. P. *J. Biol. Chem.* **2003**, *278*, 29837.
- (197) Gerasimova, A. V.; Rodionov, D. A.; Mironov, A. A.; Gelfand, M. S. *Mol. Biol.* **2001**, *35*, 853.
- (198) Cao, M.; Kobel, P. A.; Morshedi, M. M.; Wu, M. F.; Paddon, C.; Helmman, J. D. *J. Mol. Biol.* **2002**, *316*, 443.
- (199) Mao, L.; Mackenzie, C.; Roh, J. H.; Eraso, J. M.; Kaplan, S.; Resat, H. *Microbiology* **2005**, *151*, 3197.
- (200) van de Werken, H. J.; Verhees, C. H.; Akerboom, J.; de Vos, W. M.; van der Oost, J. *FEMS Microbiol. Lett.* **2006**, *260*, 69.
- (201) Geiduschek, E. P.; Ouhammouch, M. *Mol. Microbiol.* **2005**, *56*, 1397.
- (202) Plumbidge, J. *Nucleic Acids Res.* **2001**, *29*, 506.
- (203) Yang, C.; Rodionov, D. A.; Li, X.; Laikova, O. N.; Gelfand, M. S.; Zagnitko, O. P.; Romine, M. F.; Obratsova, A. Y.; Nealon, K. H.; Osterman, A. L. *J. Biol. Chem.* **2006**, *281*, 29872.
- (204) Meibom, K. L.; Li, X. B.; Nielsen, A. T.; Wu, C. Y.; Roseman, S.; Schoolnik, G. K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2524.
- (205) Bates Utz, C.; Nguyen, A. B.; Smalley, D. J.; Anderson, A. B.; Conway, T. J. *Bacteriol.* **2004**, *186*, 7690.
- (206) Rodionov, D. A.; Gelfand, M. S.; Hugouvieux-Cotte-Pattat, N. *Microbiology* **2004**, *150*, 3571.
- (207) Hugouvieux-Cotte-Pattat, N.; Blot, N.; Reverchon, S. *Mol. Microbiol.* **2001**, *41*, 1113.
- (208) Rodionov, D. A.; Gelfand, M. S. *FEMS Microbiol. Lett.* **2006**, *255*, 102.
- (209) Reitzer, L. *Annu. Rev. Microbiol.* **2003**, *57*, 155.
- (210) Yoshida, K.; Yamaguchi, H.; Kinehara, M.; Ohki, Y. H.; Nakaura, Y.; Fujita, Y. *Mol. Microbiol.* **2003**, *49*, 157.
- (211) Su, Z.; Mao, F.; Dam, P.; Wu, H.; Olman, V.; Paulsen, I. T.; Palenik, B.; Xu, Y. *Nucleic Acids Res.* **2006**, *34*, 1050.
- (212) Silberbach, M.; Burkovski, A. J. *Biotechnol.* **2006**, *126*, 101.
- (213) Su, Z.; Olman, V.; Mao, F.; Xu, Y. *Nucleic Acids Res.* **2005**, *33*, 5156.
- (214) Dixon, R.; Kahn, D. *Nat. Rev. Microbiol.* **2004**, *2*, 621.
- (215) Lie, T. J.; Wood, G. E.; Leigh, J. A. *J. Biol. Chem.* **2005**, *280*, 5236.
- (216) Rodionov, D. A.; Dubchak, I. L.; Arkin, A. P.; Alm, E. J.; Gelfand, M. S. *PLoS Comput. Biol.* **2005**, *1*, e55.
- (217) Grose, J. H.; Bergthorsson, U.; Roth, J. R. *J. Bacteriol.* **2005**, *187*, 2774.
- (218) Kurnasov, O. V.; Polanuyer, B. M.; Ananta, S.; Sloutsky, R.; Tam, A.; Gerdes, S. Y.; Osterman, A. L. *J. Bacteriol.* **2002**, *184*, 6906.
- (219) Rossolillo, P.; Marinoni, I.; Galli, E.; Colosimo, A.; Albertini, A. M. *J. Bacteriol.* **2005**, *187*, 7155.
- (220) Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G. D.; Maltsev, N. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2896.
- (221) Galperin, M. Y.; Koonin, E. V. *Nat. Biotechnol.* **2000**, *18*, 609.
- (222) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. *Science* **1999**, *285*, 751.
- (223) Enright, A. J.; Ouzounis, C. A. *Genome Biol.* **2001**, *2*, Research0034.
- (224) Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D.; Yeates, T. O. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4285.
- (225) Overbeek, R.; Begley, T.; Butler, R. M.; Choudhuri, J. V.; Chuang, H. Y.; Cohoon, M.; de Crecy-Lagard, V.; Diaz, N.; Disz, T.; Edwards, R.; Fonstein, M.; Frank, E. D.; Gerdes, S.; Glass, E. M.; Goessmann, A.; Hanson, A.; Iwata-Reuyl, D.; Jensen, R.; Jamshidi, N.; Krause, L.; Kubal, M.; Larsen, N.; Linke, B.; McHardy, A. C.; Meyer, F.; Neuweger, H.; Olsen, G.; Olson, R.; Osterman, A.; Portnoy, V.; Pusch, G. D.; Rodionov, D. A.; Ruckert, C.; Steiner, J.; Stevens, R.; Thiele, I.; Vassieva, O.; Ye, Y.; Zagnitko, O.; Vonstein, V. *Nucleic Acids Res.* **2005**, *33*, 5691.
- (226) Gerdes, S. Y.; Kurnasov, O. V.; Shatalin, K.; Polanuyer, B.; Sloutsky, R.; Vonstein, V.; Overbeek, R.; Osterman, A. L. *J. Bacteriol.* **2006**, *188*, 3012.
- (227) Via, P.; Badia, J.; Baldoma, L.; Obradors, N.; Aguilar, J. *Microbiology* **1996**, *142*, 1833.
- (228) Egan, S. M.; Schleif, R. F. *J. Mol. Biol.* **1994**, *243*, 821.
- (229) Hugouvieux-Cotte-Pattat, N. *Mol. Microbiol.* **2004**, *51*, 1361.
- (230) Sadvokskaya, N. S.; Laikova, O. N.; Mironov, A. A.; Gelfand, M. S. *Mol. Biol.* **2001**, *35*, 862.
- (231) Campbell, J. W.; Cronan, J. E., Jr. *J. Bacteriol.* **2002**, *184*, 3759.
- (232) Warren, M. J.; Raux, E.; Schubert, H. L.; Escalante-Semerena, J. C. *Nat. Prod. Rep.* **2002**, *19*, 390.
- (233) Rodionov, D. A.; Vitreschak, A. G.; Mironov, A. A.; Gelfand, M. S. *J. Biol. Chem.* **2003**, *278*, 41148.
- (234) Rodionov, D. A.; Hebbeln, P.; Gelfand, M. S.; Eitinger, T. *J. Bacteriol.* **2006**, *188*, 317.
- (235) Hebbeln, P.; Eitinger, T. *FEMS Microbiol. Lett.* **2004**, *230*, 129.
- (236) Zayas, C. L.; Woodson, J. D.; Escalante-Semerena, J. C. *J. Bacteriol.* **2006**, *188*, 2740.
- (237) Sudarsan, N.; Hammond, M. C.; Block, K. F.; Welz, R.; Barrick, J. E.; Roth, A.; Breaker, R. R. *Science* **2006**, *314*, 300.
- (238) Borovok, I.; Gorovitz, B.; Schreiber, R.; Aharonowitz, Y.; Cohen, G. J. *Bacteriol.* **2006**, *188*, 2512.
- (239) Masse, E.; Gottesman, S. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4620.
- (240) Rodionov, D. A.; Gelfand, M. S.; Todd, J. D.; Curson, A. R.; Johnston, A. W. *PLoS Comput. Biol.* **2006**, *2*, e163.
- (241) Rodionov, D. A.; Dubchak, I.; Arkin, A.; Alm, E.; Gelfand, M. S. *Genome Biol.* **2004**, *5*, R90.

- (242) Akanuma, G.; Nanamiya, H.; Natori, Y.; Nomura, N.; Kawamura, F. *J. Bacteriol.* **2006**, *188*, 2715.
- (243) Nanamiya, H.; Akanuma, G.; Natori, Y.; Murayama, R.; Kosono, S.; Kudo, T.; Kobayashi, K.; Ogasawara, N.; Park, S. M.; Ochi, K.; Kawamura, F. *Mol. Microbiol.* **2004**, *52*, 273.
- (244) Owen, G. A.; Pascoe, B.; Kallifidas, D.; Paget, M. S. *J. Bacteriol.* **2007**, *189*, 4078.
- (245) Shin, J. H.; Oh, S. Y.; Kim, S. J.; Roe, J. H. *J. Bacteriol.* **2007**, *189*, 4070.
- (246) Paulsen, I. T.; Sliwinski, M. K.; Saier, M. H., Jr. *J. Mol. Biol.* **1998**, *277*, 573.
- (247) Caspi, R.; Foerster, H.; Fulcher, C. A.; Hopkinson, R.; Ingraham, J.; Kaipa, P.; Krummenacker, M.; Paley, S.; Pick, J.; Rhee, S. Y.; Tissier, C.; Zhang, P.; Karp, P. D. *Nucleic Acids Res.* **2006**, *34*, D511.
- (248) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354.
- (249) Ren, Q.; Chen, K.; Paulsen, I. T. *Nucleic Acids Res.* **2007**, *35*, D274.
- (250) Mauchline, T. H.; Fowler, J. E.; East, A. K.; Sartor, A. L.; Zaheer, R.; Hosie, A. H.; Poole, P. S.; Finan, T. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17933.
- (251) Rodionov, D. A.; Vitreschak, A. G.; Mironov, A. A.; Gelfand, M. S. *Nucleic Acids Res.* **2004**, *32*, 3340.
- (252) Guillen-Navarro, K.; Araiza, G.; Garcia-de Los Santos, A.; Mora, Y.; Dunn, M. F. *FEMS Microbiol. Lett.* **2005**, *250*, 209.
- (253) Hebbeln, P.; Rodionov, D. A.; Alfandega, A.; Eitinger, T. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2909.
- (254) Kreneva, R. A.; Gelfand, M. S.; Mironov, A. A.; Iomantas, Iu. A.; Kozlov, Iu. I.; Mironov, A. S.; Perumov, D. A. *Genetika* **2000**, *36*, 1166.
- (255) Burgess, C. M.; Slotboom, D. J.; Geertsma, E. R.; Duurkens, R. H.; Poolman, B.; van Sinderen, D. *J. Bacteriol.* **2006**, *188*, 2752.
- (256) Snel, B.; van Noort, V.; Huynen, M. A. *Nucleic Acids Res.* **2004**, *32*, 4725.
- (257) Lozada-Chavez, I.; Janga, S. C.; Collado-Vides, J. *Nucleic Acids Res.* **2006**, *34*, 3434.
- (258) Hershberg, R.; Margalit, H. *Genome Biol.* **2006**, *7*, R62.
- (259) Panina, E. M.; Vitreschak, A. G.; Mironov, A. A.; Gelfand, M. S. *FEMS Microbiol. Lett.* **2003**, *222*, 211.
- (260) Permina, E. A.; Kazakov, A. E.; Kalinina, O. V.; Gelfand, M. S. *BMC Microbiol.* **2006**, *6*, 49.
- (261) Ravcheev, D. A.; Gerasimova, A. V.; Mironov, A. A.; Gelfand, M. S. *BMC Genomics* **2007**, *8*, 54.
- (262) Marincs, F.; Manfield, I. W.; Stead, J. A.; McDowall, K. J.; Stockley, P. G. *Biochem. J.* **2006**, *396*, 227.
- (263) LaMonte, B. L.; Hughes, J. A. *Microbiology* **2006**, *152*, 1451.
- (264) Kovaleva, G. Yu.; Gelfand, M. S. *Mol. Biol.* **2007**, *41*, 126.
- (265) Shelper, D.; Rajagopal, L.; Harris, T. O.; Rubens, C. E. *J. Bacteriol.* **2003**, *185*, 6592.
- (266) Sperandio, B.; Polard, P.; Ehrlich, D. S.; Renault, P.; Guedon, E. *J. Bacteriol.* **2005**, *187*, 3762.
- (267) Vitreschak, A. G.; Lyubetskaya, E. V.; Shirshin, M. A.; Gelfand, M. S.; Lyubetsky, V. A. *FEMS Microbiol. Lett.* **2004**, *234*, 357.
- (268) Herrera, M. C.; Ramos, J. L. *J. Mol. Biol.* **2007**, *366*, 1374.
- (269) Akers, J. C.; Tan, M. *J. Bacteriol.* **2006**, *188*, 4236.
- (270) Gollnick, P.; Babitzke, P.; Antson, A.; Yanofsky, C. *Annu. Rev. Genet.* **2005**, *39*, 47.
- (271) Ramseier, T. M. *Res. Microbiol.* **1996**, *147*, 489.
- (272) Saier, M. H., Jr. *FEMS Microbiol. Lett.* **1996**, *138*, 97.
- (273) Hantke, K. *Curr. Opin. Microbiol.* **2001**, *4*, 172.
- (274) Rudolph, G.; Hennecke, H.; Fischer, H. M. *FEMS Microbiol. Rev.* **2006**, *30*, 631.
- (275) Johnston, A. W.; Todd, J. D.; Curson, A. R.; Lei, S.; Nikolaidou-Katsaridou, N.; Gelfand, M. S.; Rodionov, D. A. *Biomaterials* **2007**, *20*, 501.
- (276) Mustonen, V.; Lassig, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15936.
- (277) Danilova, L. V.; Lyubetsky, V. A.; Gelfand, M. S.; Laikova, O. N. *Mol. Biol.* **2003**, *37*, 716.
- (278) Yellaboina, S.; Ranjan, S.; Vindal, V.; Ranjan, A. *FEBS Lett.* **2006**, *580*, 2567.
- (279) Studholme, D. J.; Pau, R. N. *BMC Microbiol.* **2003**, *3*, 24.
- (280) Su, Z.; Olman, V.; Mao, F.; Xu, Y. *Nucleic Acids Res.* **2005**, *33*, 5156.
- (281) Erill, I.; Jara, M.; Salvador, N.; Escibano, M.; Campoy, S.; Barbe, J. *Nucleic Acids Res.* **2004**, *32*, 6617.
- (282) Permina, E. A.; Mironov, A. A.; Gelfand, M. S. *Gene* **2002**, *293*, 133.
- (283) Ravcheev, D. A.; Gelfand, M. S.; Mironov, A. A.; Rakhmaninova, A. B. *Genetika* **2002**, *38*, 1203.
- (284) Yuan, Z. C.; Zaheer, R.; Morton, R.; Finan, T. M. *Nucleic Acids Res.* **2006**, *34*, 2686.
- (285) Liu, J.; Tan, K.; Stormo, G. D. *Nucleic Acids Res.* **2003**, *31*, 6891.
- (286) Hallin, P. F.; Ussery, D. *Bioinformatics* **2004**, *20*, 3682.

CR068309+