3434–3445 *Nucleic Acids Research, 2006, Vol. 34, No. 12*
doi:10.1093/nar/gkl423

Bacterial regulatory networks are extremely flexible in evolution

Irma Lozada-Chávez, Sarath Chandra Janga* and Julio Collado-Vides

Programa de Genomica Computacional, Centro de Ciencias Genomicas, Universidad Nacional
Autonoma de Mexico, Apdo. Postal 565-A, Avenue Universidad, Cuernavaca, Morelos,
62100 Mexico, Mexico

Received February 16, 2006; Revised April 7, 2006; Accepted May 29, 2006

ABSTRACT

Over millions of years the structure and complexity of the transcriptional regulatory network (TRN) in bacteria has changed, reorganized and enabled them to adapt to almost every environmental niche on earth. In order to understand the plasticity of TRNs in bacteria, we studied the conservation of currently known TRNs of the two model organisms *Escherichia coli K12* and *Bacillus subtilis* across complete genomes including Bacteria, Archaea and Eukarya at three different levels: individual components of the TRN, pairs of interactions and regulons. We found that transcription factors (TFs) evolve much faster than the target genes (TGs) across phyla. We show that global regulators are poorly conserved across the phylogenetic spectrum and hence TFs could be the major players responsible for the plasticity and evolvability of the TRNs. We also found that there is only a small fraction of significantly conserved transcriptional regulatory interactions among different phyla of bacteria and that there is no constraint on the elements of the interaction to co-evolve. Finally our results suggest that majority of the regulons in bacteria are rapidly lost implying a high-order flexibility in the TRNs. We hypothesize that during the divergence of bacteria certain essential cellular processes like the synthesis of arginine, biotine and ribose, transport of amino acids and iron, availability of phosphate, replication process and the SOS response are well conserved in evolution. From our comparative analysis, it is possible to infer that transcriptional regulation is more flexible than the genetic component of the organisms and its complexity and structure plays an important role in the phenotypic adaptation.

INTRODUCTION

Evolution is the result of variation and selection of the components and structure of organisms through time. Transcriptional regulation plays a prominent role in the expression of genetic information. Its primary role in microbial organisms is controlling the response to environmental changes, such as nutritional status and several stresses. An important idea emerging in post-genomic biology is that transcriptional regulation can be viewed as a complex network of interactions among diverse types of molecules like proteins, DNA and metabolites (1–4). In this work we try to assess the evolution of the structure and plasticity of the transcriptional regulatory network (TRN) across species at three distinct levels: individual components of the TRN, pairs of regulatory interactions and regulons [A regulon is defined as the group of all genes regulated by a transcription factor (TF).], through a comparative analysis of their conservation.

The basic unit of gene regulatory interaction consists of three components: a TF, its DNA-binding site (operator) and the target gene (TG). Topologically, the TRN is complex because genes may be regulated by more than one TF and some TFs may control more than one gene through DNA-binding site(s) (5–7). The TRN comprises a significant proportion of the genome in each organism and it constitutes a major component of the genetic basis for the evolution of diverse aspects of bacterial phenotypes. It is important to learn how the TRN evolves as it would enable us to study the molecular evolutionary ecology of regulatory diversification by examining both the extent and pattern of regulatory gene diversity, the phenotypic effects of molecular variation and their ecological consequences.

It is also important to recognize that, although abundant sequence data and complete genomes are available, the experimental determination of TRNs has been limited to a few organisms even in prokaryotes. Besides, there is no clear relationship between the presence of a TF, its TG and DNA-binding site(s), and their structural and biochemical characteristics that could have been transferred between genomes. It is also difficult to evaluate a specific measure between

*To whom correspondence should be addressed. Tel: +52 777 3132063; Fax: +52 777 3291694; Email: sarath@ccg.unam.mx sequence homology, function and interaction transfer for any two proteins involved in a regulatory interaction (6,8,9). However, several groups have recently examined the transfer of regulatory interaction annotations from one organism to another using comparative genomic approaches (9,10). The transfer of such interactions involves assigning functional roles to TFs and TGs, based on protein sequence similarity and on the conservation of topological patterns of the TRN, such as motifs and modules (8,11).

The *Regulog* approach uses cross-species data to predict DNA–protein interactions across genomes. A TF and TG interaction in one species is predicted to occur in another species if their best sequence matches have been determined in the target group of genomes. The presence of just one of the components of the regulatory interaction is not enough to transfer the interaction annotation, it is necessary that both TF and its TG(s) are detected in another organism. Using this approach, Yu *et al.* (9) have shown that orthologous TFs and TGs of *Saccharomyces cerevisiae* and *Drosophila melanogaster* tend to share the same regulatory interaction if the eukaryotic TFs have minimal sequence identities of 30–60% depending on the protein family. More recently, Sharan *et al.* (12) associated functions to proteins using network-level conservation of protein–protein interactions in eukaryotic genomes. This implies that high sequence similarity does not necessarily mean that the function is conserved; but conservation at the level of network modules allows more confident function determination from the context. Therefore, the best matches are not always present within conserved protein clusters enforcing the notion that it is advantageous to increase the detection of conserved functions by including paralogous family expansion and contraction, and even gene loss. The high specificity of the predictions attained by Sharan *et al.* (12) can be maintained because conservation is evaluated in the context of a protein interaction subnetwork and not independently for each interaction. However, it has been shown that the patterns of conservation between protein–protein interactions versus protein–DNA interactions is different (9), and that the transcriptional regulatory logic differs radically between Eukaryotes and Prokaryotes (13). As a consequence, the performance of transcriptional interaction mapping methods cannot be currently assessed at a large scale (7,9).

Given the increasing number of sequenced genomes, it is possible and quite important to have a broader perspective of the evolution of TRNs by mapping the changes in the components of the regulatory interactions, which might differ from the common reconstruction of the metabolic, structural and some transcriptional histories of the organisms. Understanding the evolution of TRNs will not only improve our insight over the biological constraints different organisms have acquired over time but also enable us to decipher the basic design principles underlying them. Besides, one can reconstruct a regulatory history from the core of the transcriptional regulatory interactions that have been shared in the cellular processes of bacteria.

We used the TRNs of two different model Bacteria. One of these is the TRN of the Gram-negative bacterium *Escherichia coli* K12 contained in RegulonDB, which is probably the best known in bacteria (14). This database contains experimental information corresponding to nearly 20% of the TRN of *E.coli* (5). The second best studied Prokaryote in terms of transcriptional regulation is the Gram-positive *Bacillus subtilis*. We obtained the complete set of regulatory interactions in this bacterium documented in DataBase of Transcriptional regulation in *Bacillus subtilis* (DBTBS) (15). It is interesting to note that even though both are free-living bacteria and require similar concentrations of oxygen and temperature levels, *E.coli* has adapted to thrive inside its host while *B.subtilis* has adapted to soil environments. In this work we used a modified version of the Regulog approach described above to identify the interaction pairs and regulons of these networks through a comparison against complete genomes of Bacteria, Archaea and Eukarya.

MATERIALS AND METHODS

Protein sequence collection

A total of 204 completely sequenced genomes, including *E.coli* K12 and *B.subtilis*, were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG, ftp://ftp.genome.ad.jp/pub/kegg/genomes/) (16). For details about the 204 completely sequenced genomes used in this study see Table 1 in the Supplementary Data.

Interaction data

We obtained the transcriptional regulatory interactions of *E.coli* K12 from RegulonDB version 4.0 (14), which compiles experimental information extracted from the literature. We also obtained the regulatory interactions of *B.subtilis* from DBTBS (15). We only considered regulatory interactions where regulators and TGs encode a polypeptide. Hence, interactions involving tRNA and other non-polypeptide coding TGs were ignored. Similarly, there are some TFs that activate and repress the same TG due to the presence of more than one DNA-binding site (e.g. Crp regulating *galE* as activator or repressor depending on two different DNA-binding sites); we considered them as redundant interactions and only one interaction was used to represent them in the final dataset. Therefore, a total of 1678 non-redundant regulatory interactions that represent 119 TFs acting on 850 TGs were included in this work for *E.coli*. While a total of 785 non-redundant interactions representing 99 TFs affecting 666 TGs were included from the *B.subtilis* TRN.

Detection of potential TFs and TGs across species

It has been extensively reported that (i) duplication of sequences, (ii) divergence and (iii) recombination are major sources of functional variation in protein evolution (17,18). However, it is important to note that the definition of 'function' has often been vague and different approaches have been considered in comparative genomics (19–21). In this work, we assigned functional roles to TFs and TGs in other genomes by using an intersection of three criteria for the detection of orthologous proteins: (i) bi-directional best hits (BDBHs), (ii) coverage in the BLASTP (22) alignment and (iii) detection of PFAM (23) conserved domains.

Orthologs are defined as proteins in different species that evolved from a common ancestor by speciation (24) and usually have the same function. Proteins that evolved recently from a common ancestor by duplication previous to any speciation event are called 'outparalogs' and hence are less likely to maintain the same function (25). In contrast, 'inparalogs' are defined as those which have evolved by gene duplications that happened after the speciation event and are more likely to conserve their function. Operationally, both inparalogs and orthologous sequences are usually defined as best-matching homologs or BDBHs in another organism (26–28). Sequences in the same genome with >95% identity estimated with the CD-HIT program (29) were considered in this work as 'inparalogs' and grouped into clusters. To identify orthologs we use the BDBHs definition through depurated genomes at 95% identity, with a significant BLASTP *E*-value ($\leqslant 10^{-3}$) using the WU-BLAST program (22). However, functional assignment is not yet complete with this approach since identifying orthologs for TFs is not always straightforward.

It is well known that conserved domains inside a protein determine their specific function and that these can represent evolutionary units especially for proteins with more than one domain where the pattern of functional conservation is more complex. Therefore, proteins are more likely to share functions if they contain the same domains in a similar arrangement (20,30,31). However, it is very important to consider that an increase in the number of domains can change the original function of a protein (32). We defined the conserved domains for all sequences analyzed in this work by hidden Markov models (HMM) taken from PFAM version 10 (23), using the HMMER 2.3.1 program (33) with an *E*-value $\leqslant 10^{-3}$. In addition, we required that at least 70% of the PFAM model is covered by the sequence.

Operationally we identified orthologs as those proteins that satisfy the following four conditions:

(i) Sequences of the target genome that have a BDBH in the query genome with a significant BLASTP *E*-value ($\leqslant 10^{-3}$).
(ii) At least 70% of the query sequence is included in the BLASTP alignment.
(iii) Target sequences share the PFAM domains of their query counterparts. Target sequences having one or more domains which match the orientation and arrangement to that of the query sequence and do not increment the total size of the protein in >100 residues were also considered in the analysis.
(iv) All the sequences included previously in the inparalog cluster were considered candidates that maintain the function only if the conditions 1, 2 and 3 are true for the representative sequence of the cluster.

We predicted the orthologs and PFAM domains of 119 TFs and 850 TGs of the TRN of *E.coli*, 99 TFs and 666 TGs of the TRN of *B.subtilis* as well as for the rest of the proteins from the complete genomes of *E.coli* and *B.subtilis* across the complete genomes of 175 bacteria, 19 archaea and 10 eukaryotes.

Data management

To facilitate the display of results, we only show 110 complete genomes in all the figures, obtained by filtering out strains and species of the same bacterial genus keeping the strain or species with the maximum number of genes among a given genera of organisms. The evolutionary distance from *E.coli* and *B.subtilis* to all organisms was obtained according to the evolutionary branching process reported previously by Brown *et al.* (34). The evolutionary distance between any two organisms is related to the sum of the distances between each organism and its closest common ancestor.

Conservation of orthologs

To normalize the extent of conservation of the components (TFs and TGs) of the regulatory network in comparison with the total genome, we devised a simple metric called Conservation Index (CI) defined as follows:

$$\mathrm{CI} = \frac{x/\mathrm{TC}}{y/\mathrm{GC}},$$

where *x* is the number of orthologs present in the target genome from the total number of components (TC = TFs or TGs) of the TRN under consideration, and *y* is the total number of orthologs detected in the target genome from all protein coding genes (GC) in the genome under consideration, which in the case of *E.coli* would stand at 4248 and 4079 for *B.subtilis*. Therefore, CI is a measure of conservation of the components of the TRN of a genome pondered respect to the conservation of its genes. A CI near to 0 would indicate that the regulatory network components are poorly conserved in comparison with the genomic conservation, while a CI close to 1 would suggest that both the TF and TG are conserved to the same extent.

Prevalence of TF–TG orthologous pairs across genomes

The huge differences in genome size and gene content across Bacteria, Archaea and Eukarya, or between parasitic, symbiotic and free-living organisms, can introduce bias when calculating the frequency distribution of the shared regulatory interactions across organisms. To correct for this problem, a factor of distance (*D*) was considered for weighting the presence or absence of transcriptional regulators and their TGs across genomes:

$$D_x = \mathrm{A}'/(\mathrm{A}' + \mathrm{A} \cap \mathrm{B}) \text{ and } D_\mathrm{y} = \mathrm{B}'/(\mathrm{B}' + \mathrm{A} \cap \mathrm{B}),$$
$$\text{where } \mathrm{A}' = \mathrm{A} - (\mathrm{A} \cap \mathrm{B}) \text{ and } \mathrm{B}' = \mathrm{B} - (\mathrm{A} \cap \mathrm{B}).$$

For the TF TF-X which regulates a TG TG-Y, A denotes the set of all organisms from 110 non-redundant genomes in which an ortholog is found for TF-X and B denotes the set of all organisms in which an ortholog is seen for TG-Y. A′ represents the subset of organisms which has an ortholog for X but not for Y and B′ represents the subset of organisms for which an ortholog of Y is found but not X. AnB represents the set of organisms in which both orthologs are found. As an example, consider the case of an interacting pair, TF-X and TG-Y, where the TF distance (D_x) is higher than TG distance (D_y) because TF contains a higher number of orthologs than the TG. Clearly, D_x should contain most of the orthologs corresponding to that of the D_y, and the unique number in the D_y (B′) ought to be very small. In the limit, if the D_y has no unique orthologs relative to the D_x, the distance D_y would reach zero. A similar procedure for weighting has been used by others in the past, but focusing on domain contents (35).

The distances for each pair of TF and TG for the complete TRNs generated by the above approach were classified into three classes: (i) TF and TG co-occur and hence the TF is likely to regulate the TG ($D_{TF} = D_{TG}$), (ii) TF is more conserved than TG ($D_{TF} > D_{TG}$) and (iii) TG is more conserved than TF ($D_{TF} < D_{TG}$) based on pre-defined thresholds (see below and Supplementary Data, Method 1). To evaluate the statistical significance of the conservation of the regulatory interactions in these three different classes, we compared against 1000 randomly constructed regulatory networks for *E.coli* and *B.subtilis* each composed of the same number of interactions as the original TRNs but by switching the edges while maintaining the degree of each node the same as in the known TRN. It should be noted that this method of randomization preserves the in and out degree of the node and hence topologically resembles known TRNs. In the entire analysis we excluded the interactions where TFs are auto-regulated as they would generate a bias when calculating the co-occurrence effect of TF–TG pairs. So the final set of interactions analyzed in this approach included 1620 TF–TG pairs in *E.coli* and 738 TF–TG pairs in *B.subtilis*.

Clustering the conserved interactions

For each TF in *E.coli* and *B.subtilis*, we calculated the percentage of total interactions conserved in its regulon across genomes. To represent this distribution we clustered by the extent of TRN and regulon conservation using Centroid Linkage Clustering method with an Uncentered Correlation as distance metric from the Cluster program (36). Other distance metrics were also evaluated but were not found to be significantly different in their ability to group lineages and regulons. Clustering data represent 118 regulons in *E.coli* and 93 regulons in *B.subtilis* conserved across genomes.

RESULTS

Conservation of TFs and TGs across species

Based on experimental information from 119 TFs and 850 TGs in *E.coli* K12 and 99 TFs and 666 TGs in *B.subtilis*, forming the components of their respective TRNs, we predicted their counterparts in 204 complete genomes, including 175 bacteria, 19 archaea and 10 eukaryotes (for details see Supplementary Table 1 and Figure 1). Figure 1 shows the distribution of orthologous conservation of the components (TFs and TGs) of the TRN from *E.coli* and *B.subtilis* across 110 non-redundant genomes representing 23 different phyla of the three cellular domains based on the phylogenetic reconstruction from Brown *et al.* (34).

From the perspective of *E.coli* (Figure 1a), the closest phylum includes 76 different Proteobacteria grouped from five subdivisions (15α, 10β, 42γ, 4δ and 5ε). The extent of conservation in these groups is the highest of all analyzed phyla, where just over 30% of both TFs and TGs were conserved with the exception of parasitic and endosymbiotic organisms, which share only 10% of TFs and 20% TGs of the TRN from *E.coli*. Firmicutes from four different classes (10 Mollicutes, 22 Bacillales, 15 Lactobacillales and 4 Clostridia) were included too, which were found to have 20–30% of conserved TGs and 10–20% of conserved TFs, with the exception of parasitic and endosymbiotic organisms from Mollicutes, Mycobacterium and Tropheryma, which present <5% conservation of both TFs and TGs. Similar fractions of orthologs were detected in Actinobacteria as in Firmicutes. Other phyla like Bacteroidetes, Fusobacteria, Plactomyces, Cyanobacteria, Deinococci, Aquificae and Thermotogae share 10–25% of TGs and 5–15% of TFs. Parasitic phyla that include Chlamydiae and Spirochaetes have <15 and 5% of conserved TGs and TFs, respectively. Among the 19 archaeal genomes which comprise 4 Crenarchaeota and 14 Euryarchaeota, we found 7–15% TGs and <3% of the TFs. The only known archaeal parasite, *Nanoarchaeum equitans* shares <1% TGs and TFs. Finally, 11 eukaryotic genomes which included 2 Protists, 4 Fungi, 2 Plants, 1 Insect and 1 Nematode share between 8 and 18% of TGs with the exception of the obligate intracellular parasite *Encephalitozoon cuniculi* that shows only 2% of TGs. Only *S.cerevisiae, Cyanidioschyzon merolae*, *Arabidopsis thaliana* and *Caenorhabditis elegans* contain <1% of TF orthologs to those in *E.coli*.

From the perspective of *B.subtilis* (Figure 1b), although there seem to be fluctuations in the distribution of orthologs across genomes, >25% of TFs and TGs were found conserved in the Bacillus and Lactobacillus lineages. Parasitic and endosymbionts organisms in Mollicutes, Chlamydia, Spirochete and αγ Proteobacteria share <10% of the TGs and 5% of TFs. Conservation of the TFs and TGs across proteobacteria seems to be roughly constant, despite variations in phylogenetic distances with respect to *B.subtilis* until *Bdellovibrio bacteriovorus*. Beyond bacterial lineages we found that the conservation of the TFs drops off rapidly with no TFs conserved in Eukarya.

Irrespective of the variations in the conservation of the TFs and TGs across various phyla from the perspective of both the genomes, we observe that TGs tend to be more conserved than TFs as the phylogenetic distance increases while in closely related lineages TFs seem to be more conserved than TGs. This suggests that the majority of the transcriptional regulatory machinery in Bacteria could be lineage specific strengthening a previous observation made at the level of taxa (37). The TGs of the experimentally characterized TRN of *E.coli* correspond to 20% of its complete genome, while the characterized TFs correspond to 3%. In general, the measure of conservation (CI) (Materials and Methods and Figure 1) shows that there is a steady increase in the conservation of the proportion of regulated component (TGs) of the cell in comparison with the regulatory component (TFs). Another interesting observation is that there is a certain fraction of the regulated component which is conserved in all lineages irrespective of the extent of genomic conservation of genes. However, it should be noted that the conserved fraction need not necessarily correspond to the same set of genes. From the view point of *B.subtilis,* although the decrease in TF conservation with phylogenetic distance is not as clear until far off lineages, the distribution of TG conservation seem to be more like that of *E.coli*.

Evolution of global regulators across bacterial species

Here we consider the definition of global regulators (GRs) for *E.coli* from Martínez-Antonio and Collado-Vides (5), based

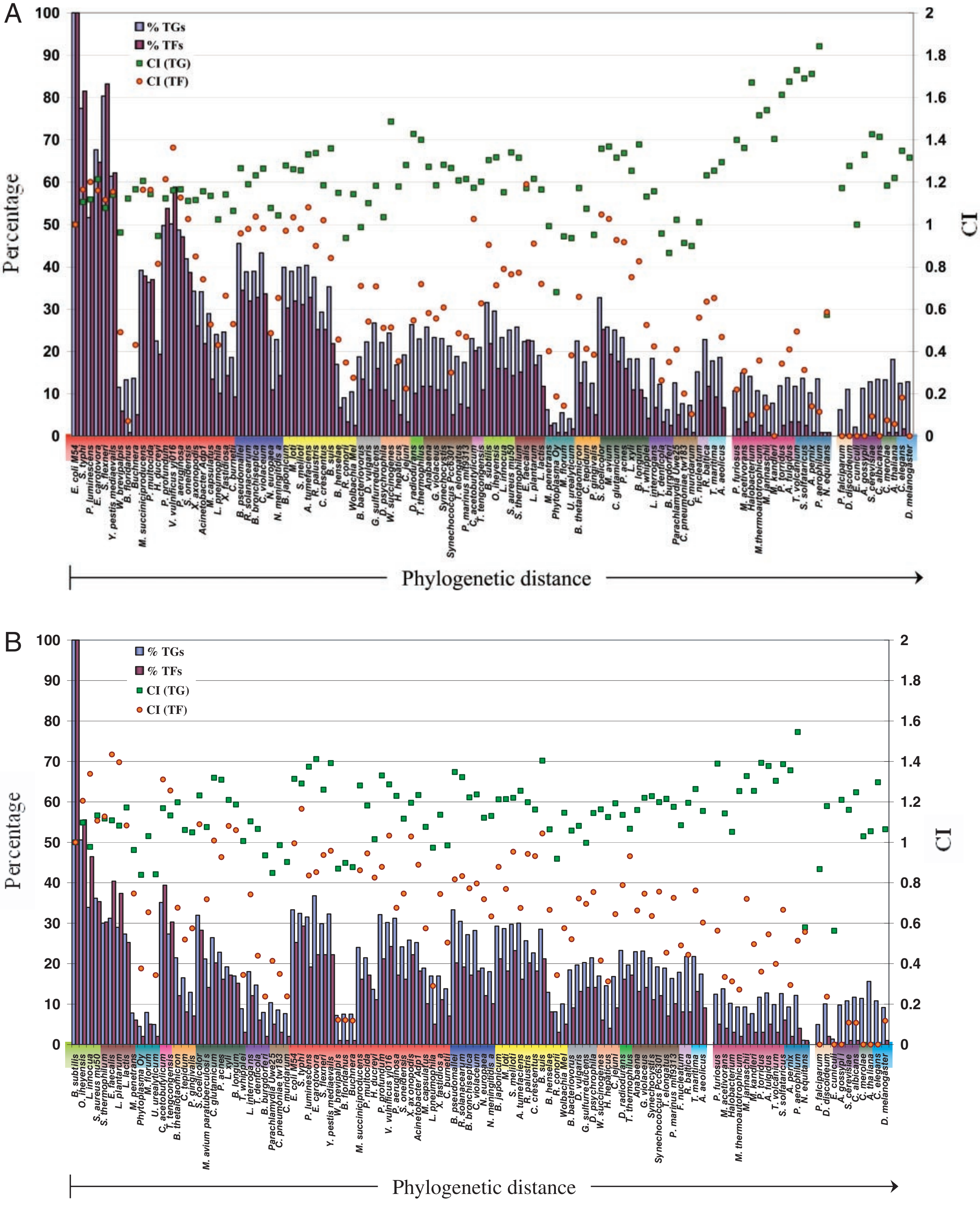


Figure 1. Conservation of the components of the TRN (TFs and TGs) across the three domains of life for (**a**) *E.coli* K12 and (**b**) *B.subtilis*. In *X*-axis are 110 non-redundant genomes ordered by phylogenetic distance (Materials and Methods). In *Y*-axis (to the left) is the percentage of conservation of the elements (TFs and TGs) of the TRNs. CI values (shown to the right on the *Y*-axis) represent a measure of conservation of the components of the TRN of a genome with respect to the conservation of its genes. Color codes on *X*-axis represent different phylogenetic clades as described in Supplementary Data.

on the number of genes they regulate and additional factors, such as the number of co-regulators and the number of conditions. Owing to the absence of sufficient information for *B.subtilis* to classify TFs on the same basis, we considered those TFs as GRs which regulate the highest number of genes in the known TRN (>20 regulatory interactions). GRs regulate the activity of 51% of the known TRN in *E.coli* (5), so we aimed at understanding the conservation spectrum of these genes. There are seven GRs in *E.coli*, Crp, Fnr, Ihf, Fis, ArcA, Hns and Lrp, and eight in *B.subtilis*, CcpA, AbrB, ComK, Fur, PhoP, TnrA, CodY and PurR. The predicted orthologs of GRs vary in their extent of conservation across the phylogenetic spectrum, although none of them seem to be conserved in eukaryotes (Figure 2a and b). However, their TGs are conserved in the three cellular domains suggesting that these TGs could be regulated in those organisms by analogous or paralogous TFs. It is interesting to note that none of the global TFs are homologously related at the sequence level between *E.coli* and *B.subtilis*, indicating that global TFs need not be conserved among phylogenetically distant genomes. This observation could imply that global TFs evolve in different lineages independently, according to the requirements in different conditions in which the organisms dwell. Of all the GRs only Lrp and the Ihf subunits (HimD or HimA) were found to occur in Archaea suggesting that most of these GRs originated in bacterial lineages. Curiously, orthologs of Crp and Fnr, which are paralogs in *E.coli,* seem to have an alternating distribution beyond Proteobacteria, in Firmicutes and Cyanobacteria, possibly indicating a substitution of their roles in these lineages or horizontal gene transfer of one of the members of the Crp family from one to another. In *B.subtilis* only the GRs, CcpA, Fur and PhoP seem to show their presence in phylogenetically distant genomes, suggesting an ancient origin compared to its other GRs. Finally, Fis, ArcA and Hns in *E.coli* have a limited distribution in other bacterial species, specifically restricted to Proteobacteria, while AbrB, ComK, CodY and PurR in *B.subtilis* are restricted to Bacillus and Lactobacillus lineages. A recent work in this direction shows the poor conservation of the hubs in regulatory networks of prokaryotic genomes (38).

The case of the phylogenetic distribution of Lrp extending to Archaea, needs further discussion as it is the only monomeric GR that is well conserved across lineages. Previously, homologs of Lrp-like transcriptional regulators were identified although their presence was detected only in Prokaryotes (39). The wide phyletic distribution of Lrp homologs among Archaea and Bacteria suggests that an Lrp-type regulator was present in the last common ancestor of Bacteria and Archaea. Nevertheless the distribution of Lrp-type regulators seems to vary across organisms (e.g. 20 copies in *Mesorhizobium loti*, 3 in *E.coli* and none in the strains of *Buchnera, Mycoplasma* and *Chlamydia*). The latter ones are bacterial endosymbionts and completely depend on their host for the supply of amino acids and other key metabolites. They have reduced genomes which could explain the absence of Lrp members. In spite of their conservation in various phyla, even in closely related species its global regulatory mechanism does not seem to be conserved as has been demonstrated by the analysis of the Lrp ortholog of *Haemophilus influenzae* (40). These observations point to the conclusion that even in organisms of the same phyla there is no real constraint for the conservation of the functions of GRs, although the GRs themselves might be well conserved at the level of sequence. Hence providing insight that at the level of transcriptional regulatory machinery orthologous genes in different genomes could play different functional roles.

Evolution of TF–TG pairs in TRNs

We implemented a distance (D) as described in Materials and Methods for a comparative analysis of the orthology distribution of a TF and its TG. As mentioned earlier, we studied the conservation of the transcriptional regulatory interactions across bacterial species by assigning each TF–TG pair to one of the three different categories: (i) when a TF and its TG are both present or absent together, (ii) when a TF is conserved in more species than its TG, and (iii) when a TG is conserved in more species than its TF (see examples in Figure 3). Ideally if the regulatory interaction is co-occurring across species, one would expect that D_{TF} and D_{TG} should both be equal to zero, but for several reasons like horizontal transfer events, loss or duplication of genes and errors involved in the detection of orthologs, one could obtain biases in the distribution of co-evolving TF–TG pairs. In order to take into account these factors and to determine a threshold for identifying co-evolving TF–TG pairs we used pairs of genes in metabolic pathways from KEGG (16) as a control (for details about generation of thresholds see Supplementary Data). It is known that genes in the same metabolic pathway often co-evolve and are well conserved (41,42). Based upon the thresholds determined for each genome we identified the co-evolving TF–TG pairs and then included the rest of the interactions into one of the two classes based on whether D_{TF} is higher or lower than D_{TG}.

Table 1 shows the Z-scores of conservation for each category of TF–TG pairs in both the genomes computed upon comparing with the randomly generated TRNs as described earlier (Materials and Methods). It can be seen that there is a relatively small fraction of the TRN in both genomes which is conserved and co-evolving across genomes. However the significance of co-evolution from the perspective of *B.subtilis* seems to be smaller than in *E.coli* as seen from the P-value (Table 1), which might be due to the under representation of the number of genomes in Firmicutes compared to Proteobacteria or due to the difference in the size of the TRNs being used in the two genomes. A roughly equal proportion of TF–TG pairs occur in the categories of TF > TG and TF < TG in both genomes (Supplementary Figure 2). The Z-scores in the respective categories suggest that there is a no clear tendency for either TF > TG or TF < TG in both the genomes, as the Z-scores in each case correspond to no >3–4 SDs except that of TF < TG in *E.coli*. This indicates that there is no constraint on the co-evolution of a TF and its TG in an interacting pair for majority of the interactions. Note that this is different from the quantitative analysis of the conservation of individual elements (TFs and TGs) conserved from the TRN as here we are interested in the co-evolution of the pairs of interactions.

The above analysis suggests that there is a small but well-conserved fraction of the TRN which is present in

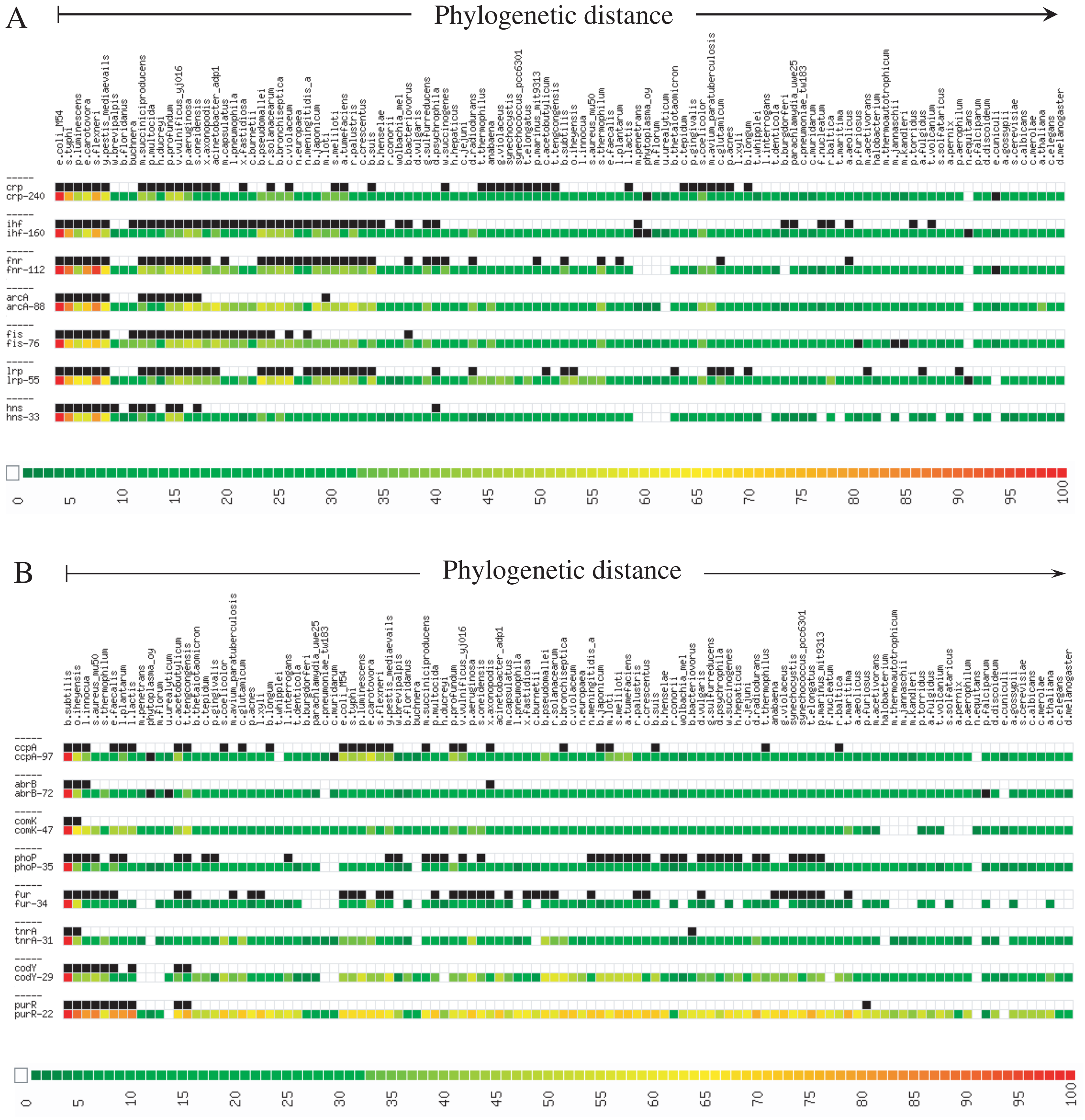


Figure 2. Conservation of GRs and their regulons across genomes for (**a**) *E.coli* K12 and (**b**) *B.subtilis*. Note that for GRs only presence (in black) or absence (in white) is shown (upper section) while for regulons percentage of interactions conserved is shown (lower section for each GR). Genomes are arranged in increasing order of phylogenetic distance with respect to the organism of reference.

diverse phyla and that the majority of the TF–TG pairs evolve independently.

Conservation of regulons across bacterial species

In Figure 4, we show the conservation of regulatory interactions at the level of regulons for both *E.coli* and *B.subtilis* across 110 non-redundant complete genomes representing various phyla, clustered horizontally by the extent of TRN conservation across genomes and vertically by the extent of regulon conservation (Materials and Methods). In general, it can be seen that the TRNs share few regulons across the phylogenetic spectrum, although the conservation is more in closely related lineages.

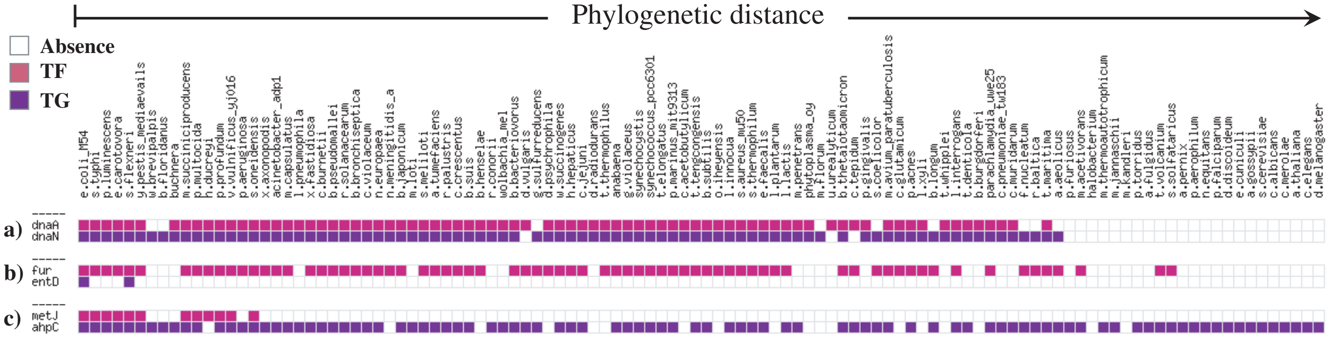


Figure 3. Classification of TF–TG pairs into three different categories. Examples of TF–TG pairs distributed in to different classes based on their co-evolution pattern: (**a**) TFs and TGs co-evolve [*dnaA* and *dnaN* in *E.coli*, where $D_{\text{dnaA}} = 3/(3 + 75) = 0.038$ and $D_{\text{dnaN}} = 9/(9 + 75) = 0.107$]; (**b**) TF is evolutionarily more conserved than TG [*fur* and *entD* in *E.coli*, where $D_{\text{fur}} = 70/(70 + 2) = 0.972$ and $D_{\text{entD}} = 0/(0 + 2) = 0$] and (**c**) TF is less conserved than TG [*metJ* and *ahpC* in *E.coli*, where $D_{\text{metJ}} = 1/(1 + 11) = 0.083$ and $D_{\text{ahpC}} = 81/(81 + 11) = 0.88$].

Table 1. Statistical significance of conservation for the different categories of TF-TG pairs in the TRNs of *E.coli* K12 and *B.subtilis*

E.coli K12			*B.subtilis*		
Category	Interactions	*Z*-score (*P*-value)	Category	Interactions	*Z*-score (*P*-value)
TF = TG	15	5.44 (<0.0001)	TF = TG	15	2.99 (0.0028)
TF > TG	813	−5.06 (<0.0001)	TF > TG	363	3.24 (0.0012)
TF < TG	759	3.68 (0.00023)	TF < TG	349	−3.60 (0.00032)

We compared the distribution of the genomes in the horizontal axis which was based on the extent of conservation of the TRNs, with that of the phylogenetic distribution generated by the method of Brown *et al.* (34) and found that several lineages were appropriately grouped, suggesting that TRN conservation can aid in segregating the major bacterial kingdoms and that phylogenetic distance provides a measure of the extent of TRN conservation. It is interesting to note that the clades closest to *E.coli* (Figure 4a), which includes several Proteobacteria, share ∼40% of the transcriptional regulatory interactions from *E.coli* except for several parasitic and endosymbiotic organisms, which were grouped together and show poor conservation of the TRN. The Proteobacteria *Blochmannia floridanus*, mollicutes *Mesoplasma florum* and *Ureaplasma urealyticum*, the Archaea *Methanopyrus kandleri*, *Methanococcus jannaschi*, *Methanobacterium thermoautotropicum*, *Halobacterium* sp, *Pyrococcus furiosus*, *Aeropyrum pernix*, *N.equitans* and the 10 analyzed eukaryotic organisms do not seem to share any regulatory interaction with *E.coli*.

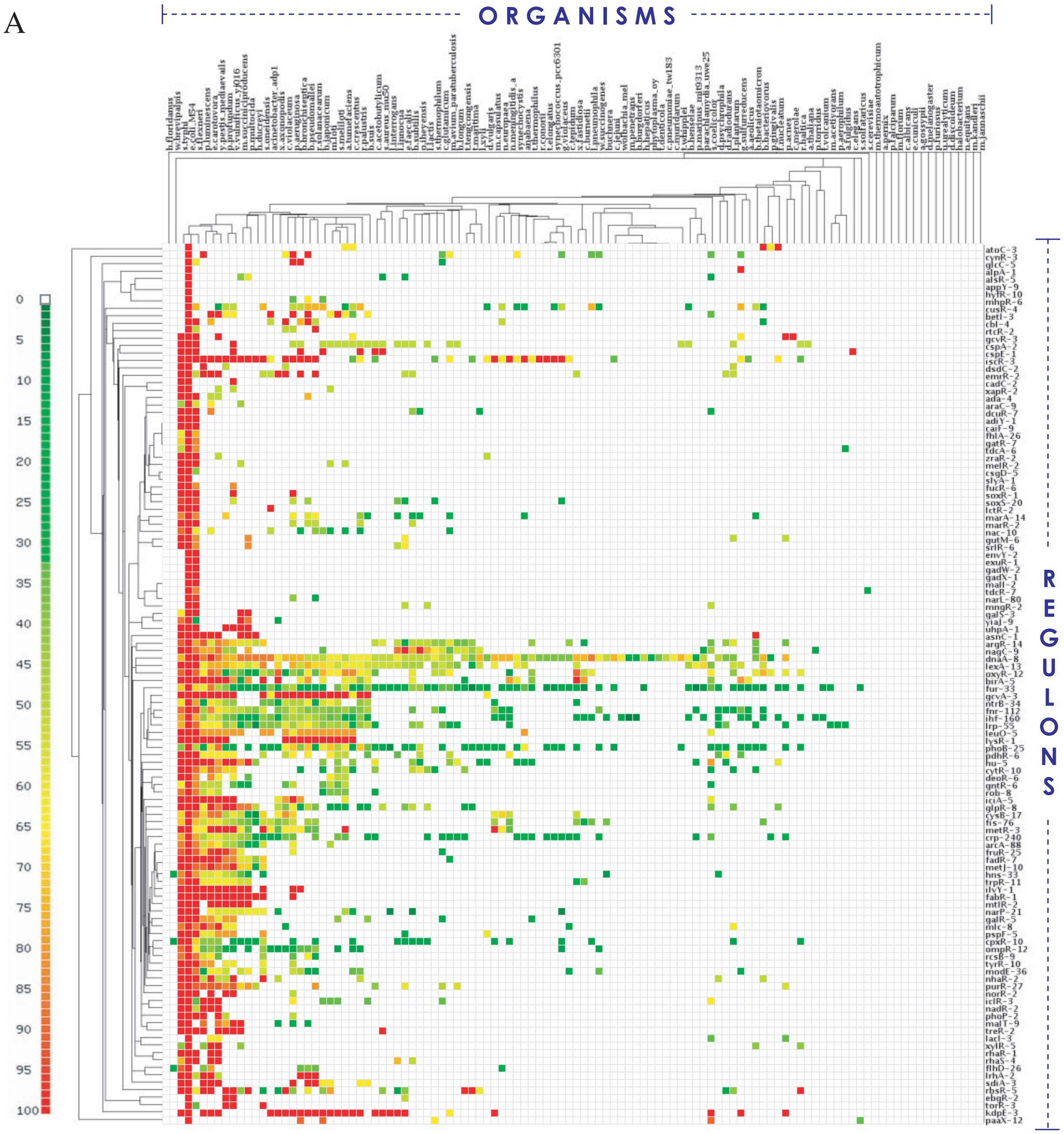
From the perspective of *B.subtilis* (Figure 4b) the closest clades share ∼30% of the transcriptional regulatory interactions except for some parasitic and endosymbiotic organisms from the Bacillus and Lactobacillus lineages. The Mollicute *U.urealyticum*, the Archaea *Picrophilus torridus*, *Pyrobaculum aerophilum, N.equitans* and the 10 analyzed eukaryotic organisms do not seem to share any regulatory interaction with the TRN of *B.subtilis*.

The horizontal distribution in Figure 4 which shows the conservation of regulons in the respective TRNs points out that certain regulons are widely conserved across species, although this fraction seems to be higher from the perspective of *E.coli*. Highly conserved regulons from *E.coli*'s TRN include metabolic and structural components like IscR, ArgR, AsnC, BirA, Crp, DnaA, Fnr, Fur, GlpR, Ihf, LexA, Lrp, NagC, OxyR, OmpR, PhoB, KdpE and RbsR. Within these conserved regulons there is only a small set of ancient conserved interactions among different bacterial phyla, which represent ∼6% of the TRN of *E.coli*. These interactions regulate important cellular processes in *E.coli* such as synthesis of arginine, asparagine, biotin and ribose, transport of amino acids and iron, availability of phosphate, replication process and the SOS response system (for additional information about these anciently conserved interactions from both the genomes see Supplementary Table 2). Highly conserved regulons from *B.subtilis* include DnaA, LexA, HrcA, PerR, BirA, AzlB, YwfK, AhrC (ArgR), CcpA, Fur, ResD, YycF, PhoP, DegU and MntR. These regulons are involved in the regulation of the synthesis of arginine and biotin; transport of manganese, availability of phosphate, heat shock response genes, global regulatory functions, replication process and the SOS response system. Some of the conserved regulons found here such as ArgR, BirA and LexA have been reported previously to be found in various phyla (6,43,44). However, this repertoire of conserved regulons should enhance our understanding of the conservation at the level of regulons and guide further experimental studies to characterize them.

For example, among these conserved regulons there are at least two hypothetical TFs: YwfK and YycF whose function is yet unknown. These conserved patterns at the level of regulons could be used to understand and characterize these TFs through a combination of experimental and computational methods thereby aiding in the determination of their function. Computationally, one can identify the function of the TF from the functional context of its regulated genes or their conserved orthologs in a way similar to what has been demonstrated earlier for several TFs from genomic context (45). Regulatory binding sites can be identified through a phylogenetic foot printing analysis of the upstream regions of putative TGs in closely related genomes. Experimental evidence added to computational predictions can elevate the quality of the predictions as has been shown in the case of LexA regulon (6).

DISCUSSION

The complexity of the TRNs in bacterial organisms is largely affected by their adaptation to the dynamically changing environmental stresses that are characteristic of an organism's niche. For example, enteric bacteria, soil bacteria and other free-living bacteria live in complex environments and have correspondingly complex sensor-response-control subsystems (46). In contrast, the narrow ecological ranges and frequent population bottlenecks of obligate pathogens and symbionts have resulted in increased rates of genetic drift and reduced selective constraint on gene function and number (47–49). Our results indicate that obligate symbiotic



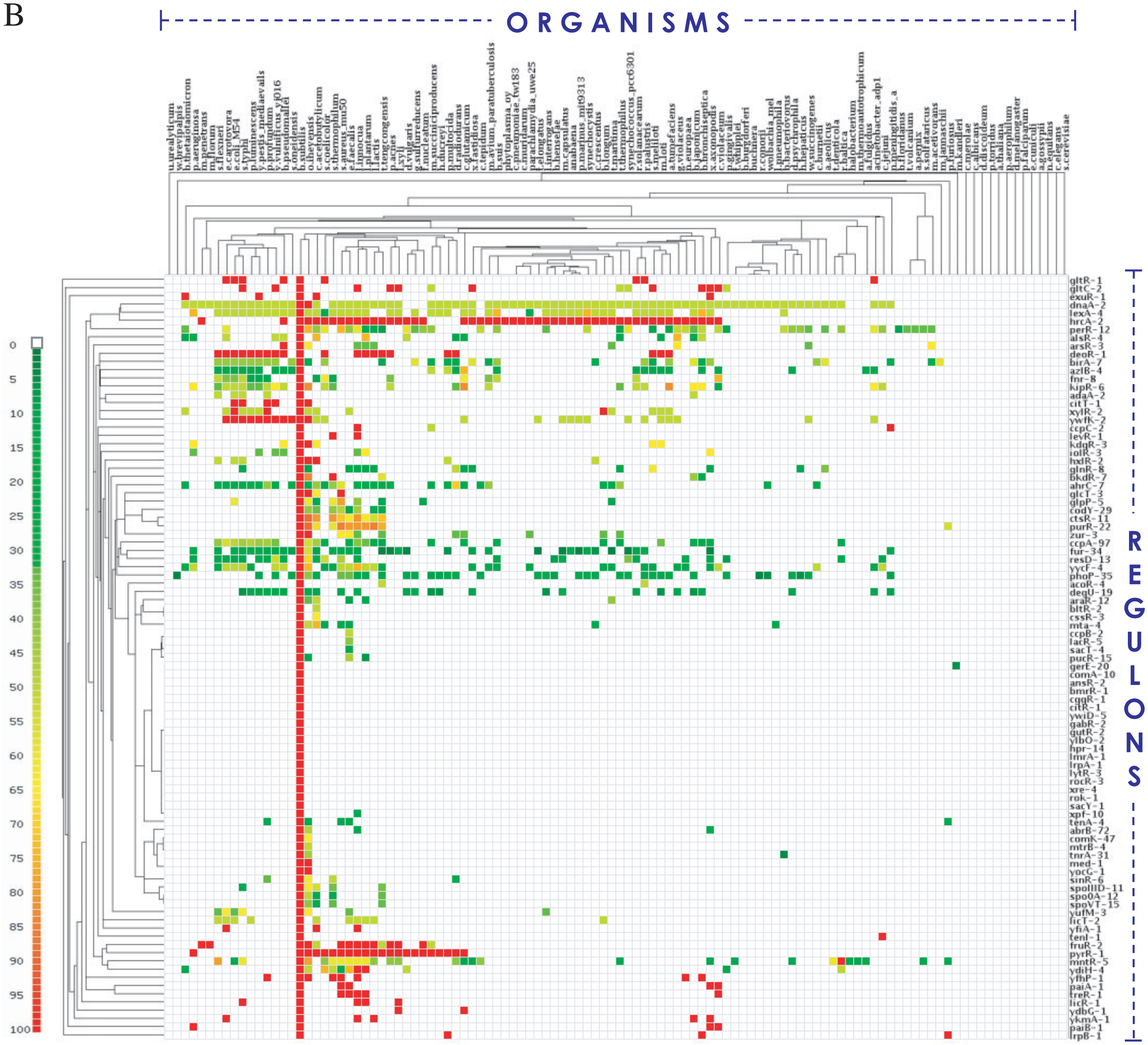


Figure 4. Conservation of regulons across genomes clustered by the extent of TRN and regulon conservation for (**a**) *E.coli* K12 and (**b**) *B.subtilis*. The intensity of the color for each regulon in each genome indicates the percentage of total interactions in the regulon shared.

as well as parasitic life styles share only ∼10% of the orthologous components of the TRNs of *E.coli* and *B.subtilis*. The loss of regulatory elements may reflect a relative constancy in the host environment, allowing these organisms to have a simplified regulatory structure (46,50). According to our results, the loss of TFs more than TGs could be the main cause of these dramatic changes in the TRN. This can also be seen from the specific scenario of the conservation of GRs of *E.coli* which have a limited biological distribution although they directly modulate the expression of ∼51% of its genes (5).

As reported previously, the conservation of genes and regulatory interactions is related to the phylogenetic distance and to the life style of the organisms (10,38). Based on our results, we can see that quantitatively the TFs are less conserved than the TGs as phylogenetic distance increases, which could suggest that transcriptional regulation of genes changes faster through evolution than the genes themselves. Related to this, Maslov *et al.* (51) found that the rate of evolutionary differentiation of transcriptional regulatory interactions proceeds faster than that of TGs and their protein interactions. However, an analysis of the conservation of pairs of regulatory interactions across genomes indicated different tendencies in the conservation of TF–TG pairs, suggesting that TF–TG pairs often do not co-evolve in the evolution of TRNs. Nevertheless it should be clear that in the first case, when a TF is conserved in different species without its corresponding TG, it would imply that the TF is indeed involved in the regulation of a different set of TGs than those in the genome under consideration and in the second case, when a TG is conserved and its TF is lost, it could imply that the TG is regulated by an analogous or paralogous factor. Both cases suggest a level of plasticity that TRNs can impose on the evolution of genomes to different environments. The evolutionary reasons for the observed tendencies in the conservation of TF–TG pairs need to be analyzed more specifically.

Despite poor conservation of the regulatory interactions across genomes, certain individual interactions have been well conserved across different eubacterial phyla, which could regulate essential transcriptional processes in Bacteria. Most of these processes are well characterized and are related directly or indirectly to the translational, structural and transcriptional machinery of the cell, suggesting a cause for their conserved nature across wide phylogenetic distances. Despite the type of regulation (repressor or activator) and that DNA-binding site(s) can change across genomes, it is reasonable to think that it is important to maintain the regulation of these core processes through the same elements, as in the case of BirA and DnaA regulators which seem to be a result of common ancestry in all bacteria.

The TRN appears to evolve in a step-wise manner, with loss and gain of individual interactions probably playing a greater role than loss and gain of whole motifs or modules of interactions. As Teichmann and Babu (52) reported previously, most network motifs have risen by convergent evolution and not by genetic duplication of ancestral circuits. Thus, with the exception of a small fraction of the TRN, it could be possible that large portions of the TRNs might have evolved through extensive changes and re-connections among the components of the network in the evolution of the species. Here we demonstrate that individual elements, interacting pairs and groups of interactions are not conserved, in fact even in closely related species. This reflects that in each speciation event to adapt to environmental changes, transcriptional regulation is more flexible than the genetic component of the organisms for phenotypic adaptation. This work should provide a perspective of the plasticity of the TRN in bacteria, which could contribute to understand the transcriptional basis of natural variation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online. Supplementary material including the complete set of regulons in *Escherichia coli K12* and *Bacillus subtilis* analyzed in this work and predicted regulons in complete genomes can be downloaded from: http://www.ccg.unam.mx/Computational_Genomics/TRNS/conservation/.

ACKNOWLEDGEMENTS

We would like to thank Heladia Salgado Osorio for supplying the interaction data from RegulonDB and to Kenta Nakai and Yuko Makita for providing us the data of transcriptional regulation from DBTBS. We would also like to thank J. Javier Díaz Mejía, Gabriel Moreno-Hagelsieb, Cei Abreu Goodger and Bruno Contreras Moreira for insightful discussions and comments on the manuscript. We would also like to acknowledge support from CONACyT program number 185993. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Fell,D.A. and Wagner,A. (2000) The small world of metabolism. *Nat. Biotechnol*., **18**, 1121–1122.
2. Thieffry,D., Huerta,A.M., Perez-Rueda,E. and Collado-Vides,J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays*., **20**, 433–440.
3. Ouzounis,C.A. and Karp,P.D. (2000) Global properties of the metabolic map of *Escherichia coli*. *Genome Res*., **10**, 568–576.
4. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet*., **31**, 64–68.
5. Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol*., **6**, 482–489.
6. Erill,I., Jara,M., Salvador,N., Escribano,M., Campoy,S. and Barbe,J. (2004) Differences in LexA regulon structure among Proteobacteria through *in vivo* assisted comparative genomics. *Nucleic Acids Res*., **32**, 6617–6626.
7. Herrgard,M.J., Covert,M.W. and Palsson,B.O. (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr. Opin. Biotechnol*., **15**, 70–77.
8. Mazurie,A., Bottani,S. and Vergassola,M. (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome. Biol*., **6**, R35.
9. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res*., **14**, 1107–1118.
10. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol*., **14**, 283–291.
11. Yeger-Lotem,E., Sattath,S., Kashtan,N., Itzkovitz,S., Milo,R., Pinter,R.Y., Alon,U. and Margalit,H. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. *Proc. Natl Acad. Sci. USA*, **101**, 5934–5939.
12. Sharan,R., Suthram,S., Kelley,R.M., Kuhn,T., McCuine,S., Uetz,P., Sittler,T., Karp,R.M. and Ideker,T. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.
13. Struhl,K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1–4.
14. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al*. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.
15. Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res*., **32**, D75–D77.
16. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
17. Teichmann,S.A., Murzin,A.G. and Chothia,C. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol*., **11**, 354–363.
18. Chothia,C., Gough,J., Vogel,C. and Teichmann,S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701–1703.
19. Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol*., **283**, 707–725.
20. Wilson,C.A., Kreychman,J. and Gerstein,M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol*., **297**, 233–249.
21. Lan,N., Montelione,G.T. and Gerstein,M. (2003) Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Curr. Opin. Chem. Biol*., **7**, 44–54.

22. Lopez,R., Silventoinen,V., Robinson,S., Kibria,A. and Gish,W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, **31**, 3795–3798.
23. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
24. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
25. Sonnhammer,E.L. and Koonin,E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
26. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
27. Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.
28. Janga,S.C. and Moreno-Hagelsieb,G. (2004) Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res.*, **32**, 5392–5397.
29. Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
30. Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.
31. Bornberg-Bauer,E., Beaussart,F., Kummerfeld,S.K., Teichmann,S.A. and Weiner,J.,III (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol. Life Sci.*, **62**, 435–445.
32. Wheelan,S.J., Marchler-Bauer,A. and Bryant,S.H. (2000) Domain size distributions can predict domain boundaries. *Bioinformatics*, **16**, 613–618.
33. Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.
34. Brown,J.R., Douady,C.J., Italia,M.J., Marshall,W.E. and Stanhope,M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nature Genet.*, **28**, 281–285.
35. Yang,S., Doolittle,R.F. and Bourne,P.E. (2005) Phylogeny determined by protein domain content. *Proc. Natl Acad. Sci. USA*, **102**, 373–378.
36. de Hoon,M.J., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
37. Coulson,R.M., Enright,A.J. and Ouzounis,C.A. (2001) Transcription-associated protein families are primarily taxon-specific. *Bioinformatics*, **17**, 95–97.
38. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, **358**, 614–633.
39. Brinkman,A.B., Ettema,T.J., de Vos,W.M. and van der Oost,J. (2003) The Lrp family of transcriptional regulators. *Mol. Microbiol.*, **48**, 287–294.
40. Friedberg,D., Midkiff,M. and Calvo,J.M. (2001) Global versus local regulatory roles for Lrp-related proteins: *Haemophilus influenzae* as a case study. *J. Bacteriol.*, **183**, 4004–4011.
41. Jeong,H., Tombor,B., Albert,R., Oltvai,Z.N. and Barabasi,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.
42. Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.*, **21**, 1055–1062.
43. Makarova,K.S., Mironov,A.A. and Gelfand,M.S. (2001) Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol.*, **2**, RESEARCH0013.
44. Rodionov,D.A., Mironov,A.A. and Gelfand,M.S. (2002) Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. *Genome Res.*, **12**, 1507–1516.
45. Doerks,T., Andrade,M.A., Lathe,W.,III, , von Mering,C. and Bork,P. (2004) Global analysis of bacterial transcription factors to predict cellular target processes. *Trends Genet.*, **20**, 126–131.
46. Cases,I., de Lorenzo,V. and Ouzounis,C.A. (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol.*, **11**, 248–253.
47. Moran,N.A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA*, **93**, 2873–2878.
48. Itoh,T., Martin,W. and Nei,M. (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc. Natl Acad. Sci. USA*, **99**, 12944–12948.
49. Andersson,S.G. and Kurland,C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol.*, **6**, 263–268.
50. Wilcox,J.L., Dunbar,H.E., Wolfinger,R.D. and Moran,N.A. (2003) Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Mol. Microbiol.*, **48**, 1491–1500.
51. Maslov,S., Sneppen,K., Eriksen,K.A. and Yan,K.K. (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol. Biol.*, **4**, 9.
52. Teichmann,S.A. and Babu,M.M. (2004) Gene regulatory network growth by duplication. *Nature Genet.*, **36**, 492–496.