Molecular BioSystems

This article was published as part of the

Computational and Systems Biology
themed issue

Please take a look at the full [table of contents](#) to access the
other papers in this issue.

Scaling relationship in the gene content of transcriptional machinery in bacteria†‡

Ernesto Pérez-Rueda,^a Sarath Chandra Janga^{*b} and Agustino Martínez-Antonio^{*c}

Received 14th April 2009, Accepted 9th June 2009

First published as an Advance Article on the web 17th July 2009

DOI: 10.1039/b907384a

The metabolic, defensive, communicative and pathogenic capabilities of eubacteria depend on their repertoire of genes and ability to regulate the expression of them. Sigma and transcription factors have fundamental roles in controlling these processes. Here, we show that sigma, transcription factors (TFs) and the number of protein coding genes occur in different magnitudes across 291 non-redundant eubacterial genomes. We suggest that these differences can be explained based on the fact that the universe of TFs, in contrast to sigma factors, exhibits a greater flexibility for transcriptional regulation, due to their ability to sense diverse stimuli through a variety of ligand-binding domains by discriminating over longer regions on DNA, through their diverse DNA-binding domains, and by their combinatorial role with other sigmas and TFs. We also note that the diversity of extra-cytoplasmic sigma factors and TF families is constrained in larger genomes. Our results indicate that most widely distributed families across eubacteria are small in size, while large families are relatively limited in their distribution across genomes. Clustering of the distribution of transcription and sigma families across genomes suggests that functional constraints could force their co-evolution, as was observed in sigma54, IHF and EBP families. Our results also indicate that large families might be a consequence of lifestyle, as pathogens and free-living organisms were found to exhibit a major proportion of these expanded families. Our results suggest that understanding proteomes from an integrated perspective, as presented in this study, can be a general framework for uncovering the relationships between different classes of proteins.

^a Departamento de Ingeniería Celular y Biocatálisis, IBT-UNAM. AP. 565-A, Cuernavaca, Morelos, 62210, México

^b MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 0QH. E-mail: sarath@mrc-lmb.cam.ac.uk

^c Departamento de Ingeniería Genética. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Irapuato, 36500, México. E-mail: amartinez@ira.cinvestav.mx

† This article is part of a *Molecular BioSystems* themed issue on Computational and Systems Biology.

‡ Electronic supplementary information (ESI) available: An extensive set of TFs from all 675 bacterial genomes (including redundant ones) and further supplementary material associated with this study. See DOI: <http://www.ibt.unam.mx/~erueda/ScalingTranscription.htm>

Introduction

Bacteria respond and adapt to diverse environmental conditions as a consequence of their gene repertoire and regulatory mechanisms, among other elements.^{1–3} The availability of their genome sequences has enabled the investigation of their differences at genetic, molecular and biochemical levels. Recent studies have shown that the evolutionary events associated with regulatory gene families, such as their expansion and contraction, contribute significantly to shaping the gene repertoire and genome size of different lineages of prokaryotes.^{4–7}

Ernesto Pérez-Rueda

B. subtilis, to understand the evolution of TFs and predict their functional roles. He has published several international publications on these topics.

Ernesto Perez-Rueda has been a professor at Universidad Nacional Autonoma de Mexico (UNAM) since 2004. He obtained his PhD at the Center for Genomic Sciences in UNAM and worked on the identification of functional residues in homeoproteins in his post-doctoral research at the Free University of Brussels. His research focuses on the analysis of DNA-binding transcription factors in diverse bacteria, such as *E. coli* and

Sarath Chandra Janga

bacteria at UNAM in Mexico. He has published more than 25 research manuscripts on various aspects of prokaryotic and eukaryotic biology in the fields of computational molecular and systems biology. His current research interests include understanding the design principles and constraints imposed on post-transcriptional and post-translational gene control in prokaryotic and eukaryotic organisms.

Sarath Chandra Janga is a PhD student at the MRC Laboratory of Molecular Biology and University of Cambridge. Sarath obtained his Bachelors and Masters in biochemical engineering and biotechnology at the Indian Institute of Technology, Delhi in 2003. Prior to starting his PhD, Sarath worked extensively and co-ordinated a number of research projects on transcriptional regulation, genome organization and comparative genomics in

Based on comparative genomics, it has been shown that genes associated with transcriptional regulation increase in a quadratic proportion with respect to the genome size.^{8–10} These observations become pertinent given that the regulation of transcription initiation in bacteria is primarily mediated by sigma factors (σ), which provide most of the specificity for promoter recognition and DNA melting needed for transcription initiation.^{11–13} In fact, sigma factors perform these functions only when bound to the RNA polymerase (RNAP). On the other hand, DNA-binding transcription factors (TFs)¹⁴ affect gene expression by blocking or allowing the access of the RNAP to the promoter, depending on the operator context and ligand-binding status.^{15–18} Usually, most gene transcription in exponentially growing bacteria is initiated by RNAP carrying a housekeeping σ , similar to *E. coli* σ^{70} or *B. subtilis* σ^A . Alternative σ s typically redirect the RNAP towards a subset of genes required during specific conditions, such as stress response or growth transitions, among others.^{11–13} TFs represent a class of proteins devoted to sense and bind signals to regulate genes, in response to specific compounds.^{17,19} Although there is extensive evidence for the existence of alternative regulatory mechanisms in diverse bacterial systems from post-transcriptional regulation,^{20–22} they are not considered in this study, as we focus on the specific role of TFs in mediating regulatory mechanisms in a wide range of completely sequenced bacterial genomes.

It has been previously suggested that the abundance of TFs increases with an increase in an organism's complexity^{8,23–26} as a consequence of different evolutionary events, such as gene expansion, gene loss and lateral gene transfer.^{24,27,28} On the other hand, the repertoire of TFs, depending on their hierarchical position in the network of transcriptional interactions, have also been shown to play an important role in shaping the organization of genes on bacterial chromosomes.^{29–32} In this study, we analyze the repertoires of σ s and TFs in 291 eubacterial genomes and compare their distribution in relation to the genome size to understand their contribution to gene regulation in different lineages and lifestyles. The results obtained here provide insights into the functional and evolutionary constraints imposed on different classes of regulatory factors in bacterial organisms.

Agustino Martínez-Antonio

Agustino Martínez-Antonio obtained his doctoral degree in biochemical sciences from UNAM. After postdoctoral research at UNAM and INSERM, he is currently a professor at the Research and Advanced Studies Centre of the National Polytechnic Institute (CINVESTAV-IPN). His interests include understanding the design principles governing the structure and function of regulatory networks in prokaryotes.

Results

The abundance of sigma factors and TFs correlates with genome size in bacteria

To study the abundance and diversity of regulatory proteins controlling transcription initiation, the repertoires of σ s and TFs were obtained in 291 non-redundant (NR) bacterial genomes (see Materials and methods section for details). A comparison of regulatory elements across genomes suggested that they increase almost quadratically with genome size (Fig. 1). In particular, we found that the repertoire of TFs is roughly 10 times higher than σ s (hundreds vs. tens) when we considered the general profiles in all the genomes analyzed, suggesting a proportion in the order of 1 σ : 10 TFs : 100 annotated ORFs per genome, although some genomes deviate from this trend. This observation suggests that possible functional relationships between TFs and σ s, on one hand, and bacterial lifestyles, on the other, could both be influencing the observed trend. We discuss the impact of both of these scenarios in the following sections.

The variation in the extent of conservation of σ s compared to TFs might be explained based on their regulatory roles at transcription initiation

Firstly, the differences in the abundance of repertoires of σ s and TFs in bacteria might be attributed to the different regulatory roles associated with them. Transcription starts when a σ interacts with RNAP to recognize its specific sequence promoter (Fig. 1). This promoter recognition stage imposes the existence of at least one σ per organism, which

Fig. 1 The distribution of the number of TFs and σ s in bacterial genomes as a function of genome size. Genomes are sorted on the x-axis by the number of ORFs. The abundance of TFs and σ s in each genome is shown on the y-axis (each dot corresponds to one genome). σ s are shown in pink and transcription factors in blue.

typically belongs to the σ^{70} family.¹³ As a result, bacterial systems might be able to switch between different transcriptional programs based exclusively on their repertoire of σ factors. Nonetheless, the transcriptional programs mediated uniquely *via* σ s would be restricted, as a result of their limited repertoire and the small collection of ligands they can recognize, such as guanosine tetraphosphate (ppGpp).³³ As a consequence, σ s exhibit a limited ability to directly couple the environmental conditions with gene transcription. In addition, σ s have a constrained DNA-binding region in terms of length and the diversity of sequences they recognize, as they need to be structurally-coupled to the RNAP in the promoter zone. These restricted zones of action divide the universe of σ s into promoters recognized by σ^{70} and those recognized by σ^{54} (the binding zones correspond to about -10 to -35 bp for σ^{70} and -12 to -24 for σ^{54} , relative to the transcription start site).^{34,35}

On the other hand, TFs define a different regulatory level compared to σ s. These proteins exhibit diverse structural and functional domains, where one of them specifically binds to DNA and the other can sense and bind one or more ligand compounds from endogenous and/or exogenous sources,¹⁷ such as the TyrR of *E. coli*, which bind to three aromatic amino acids and ATP.³⁶ In addition, TFs associate combinatorially, not only with σ s, but also with a number of other TFs and DNA-binding sites,^{37,38} thus allowing the rewiring of a transcriptional network depending on the environmental conditions; for instance, *sodA*, a gene encoding for superoxide dismutase in *E. coli*, is regulated by up to eight different TFs responsible for various cellular responses, including Fur (ferric uptake regulation protein), Arc (aerobic respiratory control) and Fnr (fumarate nitrate reduction/regulator of anaerobic respiration).^{39,40} Finally, the diversity of sequences that TFs can recognize is enormous and can occur anywhere from a few bases downstream of the promoter zone to up to hundreds of bases upstream of the transcription start site (Fig. 1).^{41,42} For instance, the global regulator CRP (catabolic repressor protein) in *E. coli* can regulate promoters associated with four out of the seven possible σ

and co-regulate with more than 50 different TFs.^{43,44} In summary, TFs constitute a class of proteins whose space of action is more flexible than that of σ s, not only in sensing diverse environmental and endogenous stimuli, but also in recognizing a wide range of binding site sequences over a larger zone on the DNA around the transcription start site.

Lifestyles explain the abundance of σ s and TFs in bigger genomes

The results of the previous sections suggest that regulatory complexity should increase in larger genomes and might be associated with bacterial lifestyles, as the environment should influence the bacterial genome structure and function. Thus, we analyzed the genomes in relation to the four global classes of lifestyles.⁴⁵ These included extremophiles (21 genomes), intracellular bacteria (28 genomes), pathogens (109 genomes) and free-living bacteria (133 genomes). To understand how the complexity of gene regulation depends on the number of σ s and/or TFs, as a function of increasing genome size and how they are associated to lifestyle, we calculated the ratio of TFs/number of genes (T/G) and σ s/number of genes (S/G), (Fig. 2). From this analysis, we found that the increase in regulatory complexity in intracellular (I) and extremophilic (E) bacteria depends almost exclusively on the TF repertoire (no correlation was observed for an increase in σ with genome size for these lifestyles). On the other hand, in pathogenic (P) bacteria, the regulatory repertoire is contributed to by TFs and to some extent by σ s. In contrast, σ s and TFs contributed almost equally to the regulatory repertoire in free-living (F) bacteria. Thus, TFs contribute significantly to the regulatory complexity of bacteria belonging to different lifestyles, whereas σ s contribute more significantly to the transcriptional machinery of regulation in pathogens and free-living bacteria. These results agree with previous observations, which suggest that few regulatory elements identified in small genomes would compensate the regulation of the entire genome with an increase in the number of DNA-binding sites per element, in contrast to the large number of elements identified in large genomes that control a lesser proportion of DNA-binding sites

Fig. 2 The ratio of regulatory factors to the total number of ORFs per genome. The number of genes encoding for TFs and σ s were normalized with respect to the total number of ORFs per genome (T/G and S/G , respectively), and these ratios are shown for bacteria belonging to four different lifestyles: free-living (F) ($\blacktriangle$), extremophiles (E) ($\blacksquare$), pathogens (P) ($\blacklozenge$) and intracellular (I) ($\bullet$).

on average.¹⁰ In addition, genes in small genomes are organized into large operons, simplifying the transcriptional machinery necessary for gene expression. This is in contrast to large genomes, which have a reduced number of genes in operons, influencing the proportion of σ s and TFs in those organisms,⁴⁶ suggesting that complex lifestyles would require a higher proportion of TFs and transcription units to better orchestrate a response to changing conditions.

The contribution of sigma factors to the transcriptional machinery trend

In order to assess the contribution of σ s to the trends described in Fig. 1 and Fig. 2, they were divided into three main groups based on their sequence and function. As described in the previous section, we then computed the ratio of the number of σ s/number of genes (S/G) in all the genomes for each group of σ s, namely σ^{54} , σ^{70} and extra cytoplasmic function (ECF) sigma factors.¹³ From this analysis, we found that the abundance of σ s is primarily determined by the number of ECFs and σ^{70} s, as the number of σ^{54} members was found to be roughly constant and often occurred in no more than a single copy in most genomes (Fig. 3(a)). ECFs were highly abundant in free-living and pathogenic bacteria, with genomes containing more than 2000 genes, and might be the result of massive gene duplications.^{47,48} The extent of conservation of different types of σ s across bacteria suggests a functional role for each, depending on their distribution. For instance, σ^{70} is indispensable to the adequate maintenance of a

cell and is the only sigma identified in small genomes with less than 800 genes, whereas ECFs are factors associated with the regulation of functional processes beyond the basal ones. In obligate intracellular pathogens, such as *Mycoplasma sp.*, *Streptococcus* mutants or *Lactobacillus plantarum*, there is only one housekeeping σ^{70} and no alternative σ s. σ^{54} factors were found to exhibit an almost constant distribution of one copy per genome, except in some pathogens and free-living eubacteria, where they were identified in two-copies (see the ESI†). σ^{54} factors require the assistance of specialized activators of the EBP (enhancer binding protein) family of TFs, and this might have constrained the number of genes regulated by σ^{54} , i.e. promoters associated with σ^{54} frequently require the bending of long intergenic DNA stretches via IHF, resulting in a specific physical proximity between the RNAP and TFs.^{49,50} Thus, evolutive mechanisms working for chromosome compactness might be working against the increased use of σ^{54} promoters in bacteria.

To analyze the specific contribution of the different families of σ s to gene transcription, we computed the ratio of the number of σ s/genomes (S/G) in all the genomes. Fig. 3(b) shows, as expected, that σ^{70} s have a higher proportion of genes to transcribe in small genomes, but that as genome size increases, this proportion diminishes; ECF is the only family whose proportion of regulated genes increases in larger genomes. Most of the diversification of ECFs corresponds to free-living and pathogenic genomes with ~5000 ORFs.

Fig. 3 The distribution of families of σ s in bacterial genomes. (a) Genome size is shown on a log scale on the x-axis. The y-axis shows the number of σ factors in each family per genome. (b) The ratio of the number of sigma factors from each family to the total number of ORFs per genome; the three outliers, with a high number of ECFs, correspond (from left to right) to β -proteobacteria (*N. europaea*) and two bacteriodes (*B. fragilis* NCTC9434 and *B. thetaiotaomicron* VPI-5482).

The abundance of TFs does not correlate with the diversity of families, and large families are not the most widely distributed

An appealing hypothesis is that a high diversity of TF families would contribute more significantly to regulatory plasticity than σ s. In line with this hypothesis, an analysis of 93 TF families, comprising of a total of 46 255 TFs across all the genomes analyzed in this study, showed a reduced diversity of families in small genomes, with an increasing proportion in larger ones, especially in pathogens (P) and free-living organisms (F) (Fig. 4(a)). The diversity of families reaches a maximum in genomes with around 5000 ORFs. The higher number of TFs in larger genomes does not necessarily imply the diversity of families beyond this plateau, but instead an increase in the size of some families of TFs. Congruent with this observation, Fig. 4(b) shows that the average number of TFs per family increases linearly, with a few families of TFs expanding disproportionately. These families comprise of LysR and TetR, which represent about 24% of the total set of TFs identified (11 078 of 46 255 proteins). Members of these two families increase abruptly in larger genomes, as shown in Fig. 4(c), which also shows three other most-populated

families of TFs in eubacteria for the sake of comparison. The increase in the size of these two families in larger genomes coincides with the plateauing of the diversity of families in these bacterial genomes (marked by arrows in Fig. 4(a), (b), and (c)). Another feature associated with large families is that they are not widely distributed among bacteria, despite their role in controlling important processes, such as cell–cell communication (LuxR), the response to external conditions by two-component systems (OmpR), the sensing, uptake and metabolism of external food sources (GntR and LysR), or resistance to antibiotics (TetR). On the other hand, some families with an average size of a few copies per genome, such as DnaA, LexA and IHF from *E. coli*, proposed to be essential in standard growth conditions in this bacterium and in keeping its DNA and nucleoid integrity,^{51,52} can be considered to be conserved across bacteria. This is because they were identified in at least 86% of the genomes, suggesting probable gene loss events in bacteria where they are absent (Fig. 5).

In summary, our results suggest that a family's abundance and distribution is associated with evolutionary events in bacteria. For instance, small families widely distributed among

Fig. 4 Characteristics of TF families in bacterial genomes. (a) The number of TF families as a function of the number of ORFs in each genome, grouped according to the lifestyle of the organism: E (extremophiles), I (intracellular), P (pathogens) and F (free-living bacteria). (b) The average number of TFs per family as a function of the number of ORFs in each genome, grouped according to the lifestyle of the organism, as in (a). (c) The ratio of the number of TFs to ORFs per genome for the five most abundant families of TFs in bacterial genomes.

Fig. 5 The diversity and conservation of TF families in bacteria. The occurrence of a TF family across genomes as a function of the total number of TFs identified. Some families of TFs conserved in a few copies per genome are circled in pink. Note that these are also the most conserved families of TFs in the analyzed genomes. In contrast, some families (circled in blue) are the most populated, though are less conserved, in comparison to those circled in pink across genomes.

Functional relationships might impose evolutionary constraints

Since some proteins tend to work together in a functional context, we analyzed the distributions of different families, as this would give us an indication about the co-evolution of regulatory factors. Hence, we clustered the co-occurrence of the regulatory protein families (TFs and σ s) in all 291 bacterial

genomes, as shown in Fig. 6. From this analysis, we found that the distribution of σ^{54} , IHF and EBP families is correlated, supporting the functional interdependence discussed above (and inset in Fig. 6) and probable co-evolution, where members and mechanisms have been preserved along the course of evolution. A second cluster including σ^{70} , the ECF family of sigma factors and other highly abundant families (more than 15 members per genome) responsible for regulating diverse mechanisms of stress responses (MarR), antibiotic resistance (TetR), osmotic response (OmpR) and quorum sensing response (LuxR), among other processes, were also found to be clustered as a result of this analysis. This suggests a strong functional relationship among these σ and TF families. These clusters, in addition, give insights into the functional interdependence between regulatory proteins from different families, which could help in the characterization of regulators in poorly studied genomes.

Materials and methods

Genome sequences

Predicted proteomes for 291 eubacteria were obtained from the entrez genome database of the NCBI (<ftp://ncbi.nlm.nih.gov/genomes/bacteria>).⁵⁴

A complete list of non-redundant genomes can be obtained at http://popoluh.wlu.ca/Phyl_Profiles/NR_genomes/REDUNDANCY.html. In brief, two genomes are considered redundant if they share a genomic similarity score (GSS) higher than 0.95, where GSS is defined as the ratio of the sum of all the BLAST bit-scores for protein coding genes that have orthologs between two genomes being compared and reaches a maximum of one if all the proteins of one organism are identical to their corresponding orthologs of another organism. This would be the case when the proteomes are identical.^{55,56} A complete list of genomes analyzed and their repertoire of TFs is provided as ESI.†

Fig. 6 The clustering of transcription and sigma factor families across bacterial genomes based on their co-occurrence profiles. A clear co-occurrence distribution is observed for IHF, EBP and σ^{54} families, suggesting a functional interdependence between them. The co-regulatory mode of action for these regulatory proteins is shown in the inset.

The identification of families of DNA-binding transcription factors (TFs)

To identify and analyze the repertoire of TFs in bacterial genomes, a combination of information from different sources and bioinformatics tools were used. Firstly, 45 088 putative TFs were collected from the transcription factor DB,⁵⁷ a database devoted to the identification and classification of DNA-binding TFs by means of the SUPERFAMILY library and PFAM hidden Markov models (HMMs). In a second phase, 90 family-specific HMMs previously reported from *E. coli* K12 and 57 family-specific HMMs from *B. subtilis*^{5,58} were used to scan the complete genome sequences (E-value threshold = 10^{-3}) with the *hmmsearch* module of the HMMer suite program (<http://hmmer.janelia.org>). TF families were identified based on their DNA-binding domains: in a first step, if a protein shared more than 25% of the identity in its DNA-binding region with any member of the well-characterized TFs of *E. coli* and/or *B. subtilis*, it was included in this particular family. In order to include distant homologs and to decrease the bias associated with the over-representation of TFs from specific organisms, these families were expanded by Blast searches⁵⁹ against the SwissProt database⁶⁰ using an E-value threshold of 10^{-6} . Proteins retrieved were filtered at 100% to exclude redundancy using the program CD-hit⁶¹ and aligned with ClustalW.⁶² Proteins with less than 50% similarity against their corresponding HMM were excluded. This step is important to explore potential TFs not identified through the first approach and *vice versa*, *i.e.* the coverage of the DBD database corresponds to approximately 70% of the universe of TFs and can be complemented with family-specific HMMs.⁶³ Previous studies using this approach for predicting new TFs suggest that these models are successful in identifying a significant fraction of experimentally confirmed TFs in different lineages,^{40,64} confirming the value of these predictions for studying genome-scale patterns. An extensive set of TFs from all 675 bacterial genomes (including redundant ones) and supplementary material associated with this study is available.[‡]

The identification of σ factors

Three HMMs were used to identify σ^{70} , σ^{54} and ECF-like sigma factors across genomes. σ^{70} and σ^{54} models were retrieved from the PFAM database.⁶⁵ ECFs have been considered as a separate group of σ^{70} proteins because of their significant sequence divergence from the σ^{70} family. Thus, we constructed a specific ECF HMM based on the well-known repertoire of ECF proteins in *B. subtilis*. These proteins were used to run the motif discovery and search system, MEME/MAST (using default parameters), to identify specific regions associated with this group. We selected two motifs to construct HMMs and to scan the whole repertoire of bacterial genomes. The motifs and HMMs are available in the ESI.[‡]

Clustering of families of regulatory factors

To analyze the distribution of σ s and TF families across the 291 bacterial genomes, they were first saved as a matrix. This matrix was then loaded into the cluster 3.0 program⁶⁶ to identify groups of families that correlate in terms of their

occurrence profile across all the bacterial genomes. A hierarchical complete linkage clustering algorithm was run with an uncentered correlation as a similarity measure. The clustering results were then visualized using the Treeview program.⁶⁶

Conclusions

To understand the relationship between the expansion patterns of different regulatory factors involved in gene regulation at transcription initiation, 291 completely sequenced bacterial genomes, which represent adaptive designs for different lifestyles, were analyzed. We showed that the distribution of σ s and TFs follows a trend, with a ratio of 1 σ per 10 TFs and 100 ORFs in all the genomes analyzed, coinciding with our present knowledge that σ s direct RNAP to a small repertoire of binding sites in sequence and location, compared to the diversity provided by the collection of TFs at the promoters in a genome. For instance, in *E. coli*, around 95% of its genes are transcribed by σ^{70} , with the fine tuning of their expression mediated by TFs.⁴⁴ In addition, we found that, in large genomes, there is a decrease in the number of different families of TFs, *i.e.*, in the diversity of families, than would otherwise be expected. In this context, abundant families are not widely distributed across all bacteria. In contrast, some small families are the most widely distributed. This difference might be associated with different phenomena, such as evolutionary constraints by regulatory mechanisms, as discussed in the case of DnaA or LexA and EBP families. Our results also suggest that in larger genomes, regulatory complexity may possibly increase as a result of the increasing number of members from the ECF family and some TF families. However, it is unclear if this increase would correspond to an increase in complexity by means of multiple parallel switches and feed-forward loops in regulatory networks (as shown for carbon sources in *E. coli*⁶⁷), as long regulatory cascades, or as a combination of both. Overall, the analyses presented here will not only contribute to improving our understanding of the influence of design on the regulation of gene expression, but also support the basis for a comprehensive modelling of transcriptional regulatory networks in bacteria. The observations discussed in this study should be valid for a wide-range of bacteria in most genomic studies; the analysis of over 100 genomes is reported to be sufficient and robust enough to be generalized.⁶⁸

Abbreviations

σ s	Sigma factors
TFs	Transcription factors
EBP	Enhancer binding proteins
ECF	Extra-cytoplasmic sigma factors

Acknowledgements

EP-R was financed by a grant (IN-217508) from DGAPA-UNAM and by grants given to Lorenzo Segovia. S. C. J. acknowledges financial support from the Cambridge Commonwealth Trust and the MRC-Laboratory of Molecular

Biology. A. M.-A. acknowledges financial support from CONCYTEG (young researcher) and CINVESTAV (multi-disciplinary). S. C. J. thanks colleagues at MRC-LMB for critically reading the manuscript and providing helpful comments. We thank Jose Antonio Ibarra and Gabriel Moreno-Hagelsieb for their helpful comments in the preparation of the manuscript, and Derek Wilson for providing us with the collection of TFs available through the DBD database.

References

- 1 M. Lynch and J. S. Conery, *Science*, 2003, **302**, 1401–1404.
- 2 M. Lynch, *Annu. Rev. Microbiol.*, 2006, **60**, 327–349.
- 3 B. O. Bengtsson, *J. Theor. Biol.*, 2004, **231**, 271–278.
- 4 Y. Minezaki, K. Homma and K. Nishikawa, *DNA Res.*, 2005, **12**, 269–280.
- 5 E. Perez-Rueda, J. Collado-Vides and L. Segovia, *Comput. Biol. Chem.*, 2004, **28**, 341–350.
- 6 D. A. Rodionov, *Chem. Rev.*, 2007, **107**, 3467–3497.
- 7 J. A. Oguiza, K. Kiil and D. W. Ussery, *Trends Microbiol.*, 2005, **13**, 565–568.
- 8 E. van Nimwegen, *Trends Genet.*, 2003, **19**, 479–484.
- 9 O. X. Cordero and P. Hogeweg, *Trends Genet.*, 2007, **23**, 488–493.
- 10 N. Molina and E. van Nimwegen, *Genome Res.*, 2008, **18**, 148–160.
- 11 M. M. Wosten, *FEMS Microbiol. Rev.*, 1998, **22**, 127–150.
- 12 A. Ishihama, *Annu. Rev. Microbiol.*, 2000, **54**, 499–518.
- 13 T. M. Gruber and C. A. Gross, *Annu. Rev. Microbiol.*, 2003, **57**, 441–466.
- 14 D. F. Browning and S. J. Busby, *Nat. Rev. Microbiol.*, 2004, **2**, 57–65.
- 15 N. S. Miroslavova and S. J. Busby, *Biochem. Soc. Symp.*, 2006, 1–10.
- 16 M. E. Wall, W. S. Hlavacek and M. A. Savageau, *Nat. Rev. Genet.*, 2004, **5**, 34–42.
- 17 A. Martinez-Antonio, S. C. Janga, H. Salgado and J. Collado-Vides, *Trends Microbiol.*, 2006, **14**, 22–27.
- 18 S. C. Janga and J. Collado-Vides, *Res. Microbiol.*, 2007.
- 19 A. Goelzer, F. Bekkal Briki, I. Martin-Verstraete, P. Noirot, P. Bessieres, S. Aymerich and V. Fromion, *BMC Syst. Biol.*, 2008, **2**, 20.
- 20 A. Gutierrez-Preciado, T. M. Henkin, F. J. Grundy, C. Yanofsky and E. Merino, *Microbiol. Mol. Biol. Rev.*, 2009, **73**, 36–61.
- 21 R. R. Breaker, *Science*, 2008, **319**, 1795–1797.
- 22 C. A. Wakeman, W. C. Winkler and C. E. Dann, 3rd, *Trends Biochem. Sci.*, 2007, **32**, 415–424.
- 23 J. H. Brown, V. K. Gupta, B. L. Li, B. T. Milne, C. Restrepo and G. B. West, *Philos. Trans. R. Soc. London, Ser. B*, 2002, **357**, 619–626.
- 24 M. Levine and R. Tjian, *Nature*, 2003, **424**, 147–151.
- 25 M. A. Changizi, *J. Theor. Biol.*, 2001, **211**, 277–295.
- 26 G. B. West and J. H. Brown, *J. Exp. Biol.*, 2005, **208**, 1575–1592.
- 27 L. Aravind, V. Anantharaman, S. Balaji, M. M. Babu and L. M. Iyer, *FEMS Microbiol. Rev.*, 2005, **29**, 231–262.
- 28 M. Madan Babu, S. A. Teichmann and L. Aravind, *J. Mol. Biol.*, 2006.
- 29 S. C. Janga, H. Salgado, J. Collado-Vides and A. Martinez-Antonio, *J. Mol. Biol.*, 2007, **368**, 263–272.
- 30 S. C. Janga, H. Salgado and A. Martinez-Antonio, *Nucleic Acids Res.*, 2009, **37**, 3680–3688.
- 31 G. Kolesov, Z. Wunderlich, O. N. Laikova, M. S. Gelfand and L. A. Mirny, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 13948–13953.
- 32 C. Marr, M. Geertz, M. T. Hutt and G. Muskhelishvili, *BMC Syst. Biol.*, 2008, **2**, 18.
- 33 L. Jores and R. Wagner, *J. Biol. Chem.*, 2003, **278**, 16834–16843.
- 34 J. D. Gralla, *Curr. Opin. Genet. Dev.*, 1996, **6**, 526–530.
- 35 G. Lloyd, P. Landini and S. Busby, *Essays Biochem.*, 2001, **37**, 17–31.
- 36 J. Pittard, H. Camakaris and J. Yang, *Mol. Microbiol.*, 2005, **55**, 16–26.
- 37 S. Adhya, *Sci. STKE*, 2003, **2003**, pe22.
- 38 A. Barnard, A. Wolfe and S. Busby, *Curr. Opin. Microbiol.*, 2004, **7**, 102–108.
- 39 I. Compan and D. Touati, *J. Bacteriol.*, 1993, **175**, 1687–1696.
- 40 H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio and J. Collado-Vides, *Nucleic Acids Res.*, 2006, **34**, D394–397.
- 41 M. Madan Babu and S. A. Teichmann, *Trends Genet.*, 2003, **19**, 75–79.
- 42 J. Collado-Vides, B. Magasanik and J. D. Gralla, *Microbiol. Rev.*, 1991, **55**, 371–394.
- 43 S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penalzoa-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Treviño-Quintanilla and J. Collado-Vides, *Nucleic Acids Res.*, 2008, **36**, D120–124.
- 44 A. Martinez-Antonio and J. Collado-Vides, *Curr. Opin. Microbiol.*, 2003, **6**, 482–489.
- 45 I. Cases, V. de Lorenzo and C. A. Ouzounis, *Trends Microbiol.*, 2003, **11**, 248–253.
- 46 G. Moreno-Hagelsieb, *Curr. Genomics*, 2006, **7**, 163–170.
- 47 D. Missiakas and S. Raina, *Mol. Microbiol.*, 1998, **28**, 1059–1066.
- 48 J. D. Helmann, *Adv. Microb. Physiol.*, 2002, **46**, 47–110.
- 49 S. V. Kuznetsov, S. Sugimura, P. Vivas, D. M. Crothers and A. Ansari, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 18515–18520.
- 50 S. Sugimura and D. M. Crothers, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 18510–18514.
- 51 Y. Yamazaki, H. Niki and J. Kato, *Methods Mol. Biol.*, 2008, **416**, 385–389.
- 52 S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balazsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A.-L. Barabási, Z. N. Oltvai and A. L. Osterman, *J. Bacteriol.*, 2003, **185**, 5673–5684.
- 53 M. B. Miller and B. L. Bassler, *Annu. Rev. Microbiol.*, 2001, **55**, 165–199.
- 54 R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin, *Nucleic Acids Res.*, 2001, **29**, 22–28.
- 55 G. Moreno-Hagelsieb and S. C. Janga, *Proteins*, 2008, **70**, 344–352.
- 56 G. Moreno-Hagelsieb and J. Collado-Vides, *Bioinformatics*, 2002, **18**(suppl. 1), S329–336.
- 57 S. K. Kummerfeld and S. A. Teichmann, *Nucleic Acids Res.*, 2006, **34**, D74–81.
- 58 S. Moreno-Campuzano, S. C. Janga and E. Perez-Rueda, *BMC Genomics*, 2006, **7**, 147.
- 59 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
- 60 A. Bairoch and R. Apweiler, *Nucleic Acids Res.*, 2000, **28**, 45–48.
- 61 W. Li, L. Jaroszewski and A. Godzik, *Bioinformatics*, 2002, **17**, 77–82.
- 62 J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, 1994, **22**, 4673–4680.
- 63 E. Sonnhammer, S. Eddy and R. Durbin, *Proteins*, 1997, **28**, 405–420.
- 64 N. Sierro, Y. Makita, M. de Hoon and K. Nakai, *Nucleic Acids Res.*, 2008, **36**, D93–96.
- 65 J. Mistry and R. Finn, *Methods Mol. Biol.*, 2007, **396**, 43–58.
- 66 M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 14863–14868.
- 67 A. Martinez-Antonio, S. C. Janga and D. Thieffry, *J. Mol. Biol.*, 2008, **381**, 238–247.
- 68 D. E. Whitworth, *Trends Microbiol.*, 2008, **16**, 512–519.