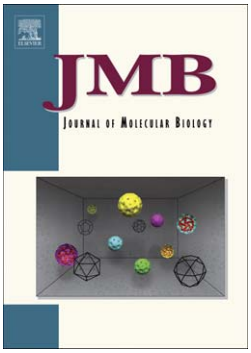

The role of DNA-binding specificity in the evolution of bacterial regulatory networks

Irma Lozada-Chávez, Vladimir Espinosa Angarica, Julio Collado-Vides, Bruno Contreras-Moreira

PII: S0022-2836(08)00427-0
DOI: doi: 10.1016/j.jmb.2008.04.008
Reference: YJMBI 60357

To appear in: *Journal of Molecular Biology*

Received date: 14 February 2008
Accepted date: 2 April 2008

Please cite this article as: Lozada-Chávez, I., Angarica, V.E., Collado-Vides, J. & Contreras-Moreira, B., The role of DNA-binding specificity in the evolution of bacterial regulatory networks, *Journal of Molecular Biology* (2008), doi: 10.1016/j.jmb.2008.04.008

The role of DNA-binding specificity in the evolution of bacterial regulatory networks

Running title: Transcription factor specificity in regulatory networks

Irma Lozada-Chávez[1], Vladimir Espinosa Angarica[1,2,3], Julio Collado-Vides[1] and Bruno Contreras-Moreira[1,4]

[1]Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Av. Universidad s/n, Cuernavaca, 62210 Morelos, México.

[2]Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza. Pedro Cerbuna, 12. 50009 Zaragoza, España.

[3]Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza. Corona de Aragón, 42. Edificio Cervantes, 50009 Zaragoza, España.

[4]Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas, Av. Montañana 1.005. 50059 Zaragoza, España

Online Supplementary Material: http://www.eead.csic.es/compbio/suppl/prok_specificity/

Corresponding authors: bcontreras@eead.csic.es, ilozada@ccg.unam.mx, telephone: +52 777 3291693, +34 976716089

Abbreviations footnote: TF = Transcription Factor, TG = Target Gene, TRN = Transcriptional Regulatory Network, bsDNA = DNA binding site, IC = information content, PWM = position-weight matrix, BDBH = bi-directional best hit, HS = high specificity, LS = low specificity.

ABSTRACT

Understanding the mechanisms by which transcriptional regulatory networks (TRNs) change through evolution is a fundamental problem. Here we analyze this question using data from *Escherichia coli* and *Bacillus subtilis*, finding that paralogy relationships are insufficient to explain the global or local role observed for transcription factors (TFs) within regulatory networks. Our results provide a picture in which DNA-binding specificity, a molecular property that can be measured in different ways, is a predictor of the role of transcription factors. In particular, we observe that global regulators consistently display low binding specificities, while displaying comparatively higher expression values in microarray experiments. In addition, in this work we find a strong negative correlation between binding specificity and the number of co-regulators which help coordinate genetic expression at a genomic scale. A close look at several orthologous TFs, including FNR, a regulator found to be global in *E. coli* and local in *B. subtilis*, confirms the diagnostic value of specificity in order to understand their regulatory function, and also highlights the importance of evaluating the metabolic and ecological relevance of effectors as another variable in the evolutionary equation of regulatory networks. Finally, a general model is presented that integrates some evolutionary forces and molecular properties, aiming to explain how regulons grow and shrink, as bacteria tune their regulation to increase adaptation.

Keywords: transcription, regulatory network, binding specificity, global regulator, paralogy

INTRODUCTION

The expression of genes can be controlled by transcriptional regulatory mechanisms in response to cellular stimuli. Transcriptional regulation in prokaryotes depends generally upon the recognition of specific DNA operator sites (bsDNA) by transcription factors (TFs). These protein-DNA interactions affect the synthesis of messenger RNA molecules of target genes (TG), which can be activated or repressed. Overall, the set of transcriptional regulatory interactions in a given organism is often called Transcriptional Regulatory Network (TRN). Genomic and statistical analysis of TRNs has shown that transcriptional proteins have a differential connectivity, in which a small set of TFs regulates a much larger set of TGs [1; 2; 3]. Even though different criteria have been proposed to define the property of connectivity [4], it is possible to assign TFs one of two functional roles, being either local or global regulators. On the basis of the number of TGs that a TF might regulate and additional features such as the different sigma-classes of promoters, the number of co-regulators and the number of conditions, highly connected TFs are called global regulators. In contrast, a large proportion of TFs in a network affects the expression of only one or few genes. These are called local regulators [5; 6].

It is thought that genetic duplication might be the main evolutionary mechanism rewiring transcriptional networks[7], and could also explain the origin of global and local regulators. In particular, Teichmann and Babu [8] have proposed that TRNs evolve by duplication of TFs and TGs which might conserve their regulation or rather gain new regulatory interactions. Genetic duplication indeed accounts for 52% of the TRN in *E. coli* [8]. However, Cosentino and coworkers [9] have concluded that the contribution of this mechanism to the network architecture is maximum within local regulators and TGs and otherwise minimal when global TFs are considered. Besides, although duplication events have been recognized in many different species, TRNs are poorly conserved across bacterial species [10; 11] not only because global regulators do not necessarily share similar evolutionary histories, but also because they do not necessarily regulate similar metabolic responses in different organisms [3; 12; 13; 14; 15; 16; 17]. Therefore, we find that there are still important questions to be answered regarding the evolution of regulatory networks. Here we take the two best annotated prokaryotic transcriptional networks, the gram-negative *Escherichia coli* K12 [18] and the gram-positive *Bacillus subtilis* [19], with remarkably different niches [20] and evolutionary histories [21; 22], in order to address this

subject. This work re-evaluates the contribution of genetic duplication, by asking how it is that paralogous TFs acquire different roles in regulatory networks. More explicitly, we aim at identifying distinctive properties required for TFs to evolve as global or local regulators.

Firstly, we take the collection of TFs from *E. coli* and *B. subtilis* in order to estimate their specificity, defined as the ability to discriminate binding sites along DNA molecules. The results obtained demonstrate that binding specificity is strongly correlated with the hierarchical role of TFs within regulatory networks, with global regulators consistently displaying low specificity (LS), while local regulators show high specificity (HS), as already anticipated by different groups. This observation suggests that the ability of TFs to conserve or gain new TGs might depend on this biochemical property. In addition, this work finds that regulatory proteins with low specificity show higher expression values in microarray experiments, perhaps as expected, since they bind to more DNA sites. Furthermore, we find that the degree of co-regulation by more than one TF in *E. coli* is negatively correlated with the specificity of DNA-binding, and we discuss several biological processes that might explain this observation. To examine our findings, we compare orthologous TFs for which sets of experimentally verified bsDNAs are available in both bacteria, with detailed insight into the FNR (fumarate and nitrate reduction) regulatory protein, confirming that the calculated specificity values are in agreement with their global or local roles. Finally, a general model is presented that summarizes some mechanisms that affect how regulons grow and shrink; in other words, how TFs might gain or lose regulatory interactions as bacteria tune their regulatory networks in order to better respond to their environmental and metabolic requirements. While this paper presents evidence about the importance of binding specificity and co-regulation, the model also includes two variables that must be involved in this evolutionary process: the rate of genomic mutations and the effectors sensed by bacterial TFs.

RESULTS AND DISCUSSION

Contribution of genetic duplication to the evolution of transcriptional networks

There is compelling evidence suggesting that gene duplication is a major force explaining the growth of TRNs[8] and it is also expected that this process will affect the connectivity distribution of these networks [23; 24], as has been seen in other biological networks. Here, we evaluate this hypothesis using data from *E. coli* and *B. subtilis* by asking whether there is any coupling between the occurrence of TF duplication events and the role of transcription factors within regulatory networks. To accomplish this goal it was firstly necessary to classify TFs in terms of paralogy. As explained in Materials and Methods, in *E. coli* we predicted 24 groups of complete paralogs from a set of 85 TFs for which, experimentally characterized bsDNAs are available. In *B. subtilis* we found 25 paralogous groups out of 91 TFs. In both cases there were a few TFs labeled as singletons, since no duplication evidence was found for them (15 in *E. coli* and 26 in *B. subtilis*).

Figure 1 tells that duplication events have occurred at all levels of TRNs, although they seem to be more frequent towards the low connectivity end of the regulatory hierarchy. This means that most TF duplication events have resulted in adding nodes to the base of the network, in agreement with recent observations [9]. Furthermore, this figure shows that most global regulators belong to different paralogous groups in the two species subject of this study. With the exception of CRP and FNR in *E. coli*, most global regulators have paralogs in the network, which in contrast have local regulatory roles. For instance, ArcA has eight related known TFs in the *E. coli* network, all of them thought to be local regulators. In *B. subtilis*, CcpA is another remarkable example, with five other known regulatory proteins supposed to be paralogously related. It is important to note that this methodology relies entirely on finding paralogous TFs and cannot separate duplication events from possible horizontal transfer events.

From these results it can be stated that identifying paralogy relationships neither helps understanding the role of TFs nor does it explain how network nodes become regulatory targets of previously existing TFs. In other words, we still need to know which distinctive properties of TFs make them more or less likely to gain or lose regulatory interactions, which is something

known to be happening in evolution [25; 26]. For this reason we focused on TF binding specificity, defined here as the ability of DNA-binding proteins to discriminate a small subset of DNA sequences from the vast repertoire of sequences found in a genome. There are different ways of approximating the specificity of DNA-binding proteins (see for instance [27; 28; 29]). As explained in next section, we tried different measures and obtained compatible results with all of them.

Specificity estimated through the observed diversity of DNA binding sites

A natural way of estimating the specificity of TFs is shown in Figure 2, provided that collections of binding sites are available. The actual property measured is the unadjusted information content (UIC) of sequence motifs, which is known to be a valid estimate of the relative specificity of DNA-binding proteins[30], commonly calculated for sequence logo representations of binding motifs. Both scatter plots show that the information content of sequence motifs is strongly correlated with the number of sites recognized by each TF. In other words, translating information content to specificity, proteins able to recognize many DNA sites show lower specificity than local regulators, which present high specificity. This result agrees with previous observations made by Sengupta and collaborators in *E. coli*[29]. Since some TFs bind only to one or two sites, and others to more than a hundred different genomic positions, this variable was log-transformed for convenience. In addition, as sequence motifs have different widths, the information content in these figures was normalized by dividing the raw IC by the motif width, as explained in Material and Methods. The correlation coefficient obtained for the *E. coli* data was -0.81 (pairs=67, R^2=0.66, p<10E-^{16}); for *B. subtilis* we also found a significant correlation coefficient of -0.81 (pairs=70, R^2=0.66, p<10E-^{17}). The results obtained with these two species, only remotely related with each other, suggest that this functional correlation between binding specificity and regulon size might be found in other bacterial species. However, other variables might be affecting the interpretation of these results as discussed in the following paragraph.

For instance, the catalogue of TF binding sites is probably incomplete for most TFs and biased towards regulatory proteins that play a role in physiological conditions that are more easily reproduced in experimental labs. How would this affect the analysis? We approached this question by randomly sampling the collection of available sites in both model organisms. The

idea was to repeat the analysis in Figure 2 after 100 rounds of resampling using only 30% of the reported sites for each TF. Of course this could only be done for TFs with at least 7 sites, but the resulting correlation coefficients are very similar in both species: -0.86 in *E. coli* (pairs=40, R^2=0.74, p<10E-^{12}) and -0.89 in *B. subtilis* (pairs=34, R^2=0.79, p<10E-^{11}). While this experiment shows that the number of available bsDNAs does not change the previously observed correlation between regulon size and TF specificity, it also proves that the actual IC measurements (i.e. specificities) may change depending on the collection of sites we have at hand. As an illustration, inspecting the data in Figure 2 we may conclude that DnaA has an IC of 0.71 in *E. coli*. However, if we take the mean IC after 100 random samples (Table 1) we might say that the specificity of DnaA is actually 1.12. If we must take these IC measurements as absolute values, then probably it is wiser to take the values compiled after sampling. Table 1 shows the specificity estimates in Figure 2 next to the mean IC after sampling.

The next variable considered was the geometry of the binding sites. Since TFs can bind to DNA in different ways –i.e as monomers or dimers, with or without spacers-, only the 10 most informative columns in each motif were taken in order to calculate the IC, ensuring a fair comparison of motifs. This approach would also compensate for potential errors in the annotation of motif widths. The analysis on the *E. coli* dataset yields a correlation coefficient of -0.82 (pairs=63, R^2=0.67, p<10E-^{15}). The picture is similar when using *B. subtilis* data, with a correlation coefficient of -0.79 (pairs=27, R^2=0.63, p<10E-^{6}). Again, a very significant correlation was found, reinforcing the initial observations.

Finally, we tried to estimate binding specificity using exactly two sites for each TF: the best and the worst sites when aligned to the corresponding sequence motif, in the form of a position-weight matrix. Here, the idea was to approximate the variability of sites recognized by any TF, expecting that highly specific proteins would bind to sites with similar scores, while LS regulators would recognize a broad range of sites. Thus, we calculated the PWM score variability for every TF finding once again significant correlations in both bacterial species with respect to the number of binding sites. In *B. subtilis* we find a correlation coefficient of 0.74 (pairs=46, R^2=0.54, p<10E^{-8}), compared to a coefficient of 0.91 (pairs=55, R^2=0.83, p=0) in *E. coli*. It is important to note that the same picture holds when coefficients of variation, less

sensitive to outliers, are calculated for each TF.

Diversity of DNA binding structural potentials as a measure of binding specificity

A rather different method for estimating binding specificity is shown in Figure 3, where the crystallographic structures of 11 *E. coli* protein-DNA complexes were used to thread the collection of RegulonDB binding sites for each of them. This collection includes TrpR, Rob, PurR, PhoB, NarL, MetJ, MarA, FadR, DnaA, CRP and LacR. As explained in Materials and Methods, each sequence was scored in terms of an estimate of the structural binding potential, and the observed score diversity plotted against the number of recognized binding sites. Despite the small number of complexes available, we observe a correlation coefficient of 0.92 (pairs=11, R^2=0.85, p=0.0004) between connectivity and the observed energy variability, supporting the hypothesis that global regulators are able to bind a larger collection of sites, at the cost of being less specific. These results provide new insights into the molecular recognition of DNA binding sites, suggesting that the array of interface contacts between protein and DNA counterparts, as captured in crystallographic complexes, can be utilized in order to estimate the specificity of TFs. Unfortunately, we cannot perform this analysis on *B. subtilis* due to the lack of structural data.

Contact-based estimations of binding specificity

Inspired by a previous work by Luscombe [31], we attempted to classify TFs according to their ratio of specific to non-specific protein-DNA contacts. A key difference in this approach is that no binding site knowledge is used. Instead, a large collection of protein-DNA complexes is required in order to build comparative models of TFs, which are then used to identify amino acid residues that are likely to contact nitrogen bases at the interface (specific contacts), as opposed to non-specific contacts, that usually include phosphate and sugar atoms. Despite the fact that this approach ignores indirect DNA readout mechanisms, it was used to estimate the specificity of 82 transcription factors (49 from *E. coli* and 33 from *B. subtilis*), yielding no correlation between contact-based specificity and connectivity, presumably as a result of using approximate theoretical models, instead of crystallographic structures. However, global TFs display low specificities and therefore these somewhat low-resolution results give further support to our previous observations and are important as they show that similar conclusions

might be reached using different data sources.

Adding co-regulation to binding specificity

So far these results suggest that highly connected TFs, those expected to have a larger impact on regulation, display relatively low binding specificities. However, by analyzing the curated data in RegulonDB [18] a more complex picture emerges, since a large fraction of *E. coli* promoters are subject to regulation by several TFs. Therefore, we should be studying binding specificity in the context of combinatorial regulation [32] (no such data is currently available for *B. subtilis*). Figure 4 shows a scatter plot of the number of co-regulators of TFs and the number of target genes in *E. coli*, revealing a correlation coefficient of 0.94 (pairs=153, R^2=0.90, p=0). This clearly means that highly connected TFs, those that seem to be less able to discriminate DNA sequences, co-regulate more often than other TFs.

However, can this distribution of co-regulating TFs be explained in terms of random combinations? Well, we find that 839/2861 (29%) of *E. coli* genes are subject to regulation by only one transcription factor. Conversely, 71% of the total number of genes is found to be regulated by two or more TFs. We can take these proportions in order to calculate the expected number of co-regulated TGs for any one TF. Consider the transcription factor NarL, known to be affecting the expression of 98 target genes. We should expect that around 70 of those genes are co-regulated by other TFs. However, RegulonDB tells that 96 of those TGs are actually co-regulated. What does this difference mean? If this calculation is done with all TFs in *E. coli* we fill a table and can then calculate the statistical significance of the differences between the expected and the observed co-regulation frequencies by means of a χ^2 test. Using this test we find a very small probability ($p<10E-^7$) that the observed differences happen by chance (if we take all TFs with 5 or more expected co-regulated TGs the probability is still $p<10E-^7$). Please note that most global regulatory proteins (with the exception of FIS) actually co-regulate more genes than could be expected by chance.

Since we have shown that highly connected TFs are less specific, these results can be interpreted as a sort of compensation mechanism: low specificity regulators have regulatory partners and even if can potentially bind to many DNA sequences, they will still need nearby co-regulating proteins in order to have an influence over transcription at several levels of the regulatory network. However, there are alternate ways of reading these results. Let us consider catabolite

repression, which involves the preferential use of certain carbon sources over others when a mixture of them is available to the microorganism for growth, by means of co-regulation mechanisms[33]. In *E. coli*, the transcriptional regulation of catabolite repression is carried out by CRP, a global regulator showing a low specificity (sampled normUIC values of 0.39); however, 83% of its TGs are co-regulated by other TFs. This high rate of co-regulation may be understood by at least three mechanisms. Firstly, when complexed with its effector cAMP, CRP binds to binding sites in the promoter of some TGs, interacting directly with RNA polymerase to initiate transcription[34]. Secondly, suboptimal cAMP-CRP binding sites may also be targeted by CRP homologues responding to other signals, for example the redox-sensor FNR, and *vice versa,* thus permitting a degree of cross-talk between bsDNAs belonging to promoters controlled by proteins of the same family [35]. Thirdly, the cAMP-CRP complex may also interact with promoter-specific TFs, such as the nucleoside-regulator CytR, increasing the DNA-binding specificity of its co-regulator i) by providing additional contacts through its surface, ii) by creating a DNA conformation that is better recognized by the co-regulator, or iii) by inducing a conformational change in the co-regulator that promotes its interaction with the bsDNA[36; 37]. To summarize, the complexity of co-regulation in prokaryotes prevents the formulation of a more general hypothesis that would explain the observed correlation with binding specificity, particularly when bacterial regulators usually include, apart from the DNA-binding domain, an effector-sensing domain that responds to particular ecological cues.

Low specificity transcription factors show high expression levels

Different sources of evidence presented here suggest that binding specificity is an important property of transcription factors that might help explain their biology. One arising prediction is that LS regulatory proteins are more likely to bind to genomic DNA sites, since their repertoire of recognized sequences is comparatively larger. However, the concentration of these proteins must also be considered, as this will ultimately limit the number of genomic sites bound [38]. The set of microarray experiments collected by Faith [39] allows us to check this prediction in *E. coli*, as they provide data for 60 non-redundant conditions. Indeed these data seem to support this hypothesis, as shown in Figure 5, in which mean normalized expression values for *E. coli* transcription factors are plotted against their number of reported binding sites, with a significant correlation coefficient of 0.66 (pairs=65, R^2=0.43, p<10E-^{8}). This scatter plot shows that

regulators such as CRP, with 207 binding sites reported in the genome, are expressed at higher levels than AraC, with only 13 sites reported. This coupling between mRNA expression levels and regulon size is a novel observation in bacteria, and was also predicted, although with little support from the data, in recent experiments in yeast [40]. However, this can only be indirect evidence, since we can merely infer transcription levels, not protein concentrations. Additional data, such as the rate of occupation of operator sites in the genome, would be required to further test the hypothesis.

DNA-binding specificity of orthologous transcription factors in *E. coli* and *B. subtilis*

The use of two bacterial models with remarkably different life styles [20] and long phylogenetic distance [21; 22] gives us the opportunity to explore our findings by comparing orthologous TFs. As listed in Table 2, we found eight pairs of orthologous TFs with two or more experimentally verified DNA binding sites. Here we examine these orthologous pairs in order to test whether global and local TFs really exhibit different specificities that can be compared across species. If we skip Lrp, a global regulatory protein in *E. coli* for which only one binding site is available in *B. subtilis* (AzlB), it is found that in 5 out of 7 cases the specificity estimates are congruent, as lower values correspond to more binding sites. The values for DnaA are not congruent, but in both genomes it is clearly a very high specific transcription factor, with values greater than 1.1. However, CytR and CcpA have very similar specificity values in both species while the regulon sizes are 10 and 48, respectively. We now look at these examples with more detail.

The first cases are LexA and DnaA, two regulators that respond to DNA cleavage in both bacteria and bind DNA with high specificity, suggesting that indeed are local TFs with similar roles in different genomes. The second case is Fur, a local regulator in *E. coli* and *B. subtilils* that coordinates the expression of iron uptake and homeostasis pathways in response to available iron [41; 42; 43]. Fur shows high specificity values in both organisms, as expected for such a specialized regulatory role.

The next cases are two orthologous TFs that are part of two-component regulatory systems. The first system, CpxR (CpxA) in *E. coli*, responds to several conditions associated with envelope stress, such as alkaline pH and overproduction of secreted proteins, and also to attachment of

cells to surfaces or the assembly of structures on the cell surface, folding or degradation of misfolded proteins in the periplasm and pili subunits as well as monitoring of porin status [44]. This system also responds to exposure to copper[45] and EDTA[46] in *E. coli*, while its *B. subtilis* counterpart YycF (YycG) is involved in the control of genes for cell wall metabolic processes, cell membrane composition and cell division[47]. The second, PhoB (PhoR), regulates the phosphate regulon in *E. coli*[48], while its counterpart in *B. subtilis*, ResD (ResE), is involved in nitrate respiration in response to oxygen limitation or nitric oxide [49]. Both orthologous TFs have high specificity values, as expected for local regulators, even when they can respond to different effectors.

The remaining orthologous TFs have different positional roles in both organisms. Let us first see CcpA, which is a global regulator in *B. subtilis*, controlling carbon catabolite repression (as CRP in *E. coli*) [50] with a specificity estimate of 0.88, while the orthologous CytR, a local regulator in *E. coli* [37], has a similar specificity value of 0.85. As mentioned earlier, these appear to be incongruent specificity estimates, as CcpA is known to bind to 48 sites, while CytR binds to 10. However, it should be mentioned that CytR, in co-regulation with CRP, has been described as the most promiscuous DNA-binder of the LacI familiy[37].

Finally, we analyze the transcription factor FNR (fumarate and nitrate reduction), a global TF in *E. coli* (FNR_{eco}) which is local in *B. subtilis* (FNR_{bsu}). FNR_{eco} has been extensively annotated in RegulonDB, while Reents and coworkers have been exhaustively studied the FNR_{bsu} regulon via transcriptomic analysis in combination with bioinformatics-based binding site prediction [16]. From 35 TGs identified as part of the FNR regulon during the transition of *B. subtilis* to anaerobic growth conditions, only eight genes are seen to be directly regulated via a *cis*-acting FNR_{bsu} box in the corresponding promoter regions as demonstrated previously by Cruz-Ramos and coworkers via construction of fusions and mutant strains [51; 52]. Indeed, the red dots in the Figures 2 show that FNR is relatively low specific in *E. coli* (sampled normUIC values of 0.63 for FNR_{eco} and 1.38 for FNR_{bsu}), in agreement with the fact that FNR regulates a much larger set of genes in *E. coli* than in *B. subtilis*. The amino acid residues presumed to be recognizing specific FNR sites change from *E. coli* to *B. subtilis*, and as a consequence the sequence logos are partially different. However, we still ignore why this protein, that senses O_2 via a Cysteine-

$[4Fe\text{-}4S]^{2+}$ cluster located in the amino terminus in FNR_{eco} [53] and the carboxyl terminus in FNR_{bsu} [16], has a major regulatory role in *E. coli* and only a minor effect in the TRN of *B. subtilis* (see Table 2). We believe that the answer to this question lies on the ecological niches of both bacteria. *E. coli* has adapted to live inside the host's gut and must be able to grow rapidly in the ileum under aerobic conditions but also in competition for limited nutrients under anaerobic conditions in the colon [54]. Therefore, it seems that shifting between these two environments is part of the species lifestyle, and FNR regulates this by affecting the expression of 135 genes in *E. coli* [18]. In contrast, *B. subtilis* usually dwells in the soil, where fluctuations in the availability of oxygen are not that frequent or periodic, depending mostly on the soil's water content [20]. Presumably this is why in this species FNR regulates the transcription of only 8 genes required for adaptation to low oxygen tension [16; 19].

To summarize, although orthologous proteins are generally thought to have the same function in different species, it has been previously reported that TFs are not conserved between phylogenetically distant species, specially the global regulators, that are gained or lost rapidly through evolution [10; 11; 55]. Even in small phylogenetic distances, such as Proteobacteria for *E. coli* or Firmicutes for *B. subtilis*, it has been found that global regulators do not necessarily share similar evolutionary histories nor they regulate similar metabolic responses[3; 12; 13; 14; 15; 16; 17]. In this section we have presented a DNA-binding specificity assessment of the set of orthologous TFs present in *E. coli* and *B. subtilis*, suggesting that the correlations described throughout the paper can be of practical use for the task of characterizing the role of regulatory proteins in prokaryotes. Our data allows us to claim that it is possible to infer the function of a TF as global or local if we can confidently measure its binding specificity. However, the DNA-binding domain can only tell us about one half of the evolutionary and functional history of a bacterial TF. The sensing/allosteric domain is most likely the result of several evolutionary processes, perhaps dominated by the environmental relevance of the corresponding effector, as illustrated by the FNR analysis. In some cases, the evolutionary history of allosteric domains might be a much better guide in order to define the functional role of a TF, as perhaps the cases of CytR and CpxR suggest.

A conceptual model for the evolution of transcriptional regulatory networks

The presented results provide a picture of bacterial regulatory networks in which binding specificity is a predictor of the hierarchy of any TF. Our data suggest that the ability of TFs to conserve or gain new TGs is not inherited from their paralogous counterparts, but it is at least correlated to their power to discriminate DNA sequences. Here we approximated the specificity of transcription factors using three different approaches, observing that global regulators (including nucleoid-associated proteins[56]) from two bacterial models with remarkably different life styles and long phylogenetic distance consistently display low binding specificities, and that specificity values of most orthologous TFs between *E. coli* and *B. subtilis* are congruent with their global or local role. We have also found that low specificity regulators are transcribed at relative high levels in *E. coli*, perhaps as a consequence of these proteins not being co-localized with their TGs in the genome, suggesting that an efficient occupancy of binding sites may be achieved by high copy number instead [38; 40; 57]. In addition, it is clear from Figure 4 that less specific TFs have more co-regulators, other TFs that help translate their global control to more specialized subsets of target genes, adding one more variable to this evolutionary scenario. However, it seems obvious that other variables will be conditioning the evolution of regulatory networks. Of special interest are variables that might be restricting or enhancing the ability of TFs to gain, conserve or even lose regulatory interactions.

For instance, the mechanisms that generate or delete genomic binding sites should also be considered to fully understand this question, as already envisaged by Sengupta and collaborators[29]. In this respect, Figure 6 shows a scatter plot of the theoretically estimated probability of site generation and the number of cognate binding sites of transcription factors in both *E. coli* and *B. subtilis*, predicting that LS regulators are more likely to bind to DNA sites appearing as a result of point mutations. A protein such as CRP, able to recognize 90 different oligonucleotides, will bind a randomly generated sequence with a probability roughly two orders of magnitude larger than CaiF, able to discriminate only 2 sequences. A different view to the same numbers could be that poor DNA sequence discriminators, with large sets of targets genes, are less vulnerable to random genomic mutations, since more mutations are needed to disable a binding site. Moreover, it should be noted that bacterial genomes are plastic and experience genomic rearrangements that modify the composition and orientation of operons, providing means for creating or destroying binding sites beyond point mutations[27; 58]. Our specificity estimations might be indicating that local regulators, in evolutionary time scales, are

more likely to gain binding sites as a result of such genomic rearrangement events. However, this hypothesis would require further testing and we have no direct evidence supporting it.

In addition, as bacterial regulators usually include a signal-sensing allosteric domain, it is likely that the metabolic and ecological relevance of these effectors will largely affect the evolution of TFs and their regulons. In other words, as introduced in the previous section, the evolutionary fate of transcription factors will depend on both the DNA-binding and the allosteric domains. We anticipate two ways in which sensing domains might have and impact over the network evolution. Firsty, they might induce conformational changes on the attached DNA-binding domains upon binding of effector molecules. For instance, it has been demonstrated that CRP increases its specificity after binding to cyclic AMP molecules[34]. Similar evidence has been found for LacI[59] or Cbl[60]. In this sense, it seems that allosteric domains might be regulating specificity, somewhat compensating the intrinsic promiscuity of some DNA-binding domains. Secondly, not all signals sensed by regulatory proteins are equally relevant for the species adaptation, nor they evenly describe the species's ecological niche. This conceptual model predicts that TFs are more likely to conserve or gain new target genes if they increase adaptation by logically linking allosteric effectors to the expression of new regulatory targets or operons..
In summary, the model in Figure 7 attempts to summarize the evolutionary variables that make regulons grow and shrink between species, such as FNR in *E. coli* and *B. subtilis*, as bacteria tune their regulatory networks in order to better respond to their environment and their metabolic requirements.

MATERIALS AND METHODS

Regulatory network collection

We downloaded the transcriptional regulatory interactions of *E. coli* K12 from RegulonDB release 5.5 [18]. We also obtained the regulatory interactions of *B. subtilis* from the Database of transcriptional regulation in *B. subtilis* (DBTBS) release 4.1 [19]. Both databases compile

experimental information curated from the literature. We considered only regulatory interactions where the DNA binding sites have been experimentally characterized. For *E. coli* we collected a total of 85 transcription factors regulating 1593 target genes through 1314 DNA binding sites, while we collected a total of 91 TFs regulating 732 TGs through 944 bsDNA in *B. subtilis* (see Table S1 from Supplementary Material).

Detection of paralogy and orthology of transcription factors

Search of paralogues

In order to detect possible TF duplication events in the genomes of *E. coli* and *B. subtilis*, we used both sequence and three-dimensional structural domain assignments of the proteins in the network as a measure of paralogy. Therefore, if two proteins had exactly the same domain composition and the same number of domains, we assumed that they were derived from genetic duplication of a common ancestor. As bacterial regulators usually have at least two protein domains, conservation of the DNA-binding domain was not considered sufficient to detect paralogy. We defined domains according to the structural annotation system of the SUPERFAMILY database [61], based on the domain classification scheme of SCOP [62], and according to the sequence annotations of the PFAM database [63]. Both assignment schemes rely on the use of libraries of hidden Markov models (HMM) to represent domains.

We searched for protein domains in the complete genomes of *E. coli* and *B. subtilis* using HMMs taken from PFAM version 20.0 and SUPERFAMILY version 1.69, using the HMMER 2.3.1 program [64] with an expectation value $\leq 10^{-3}$. This cut-off value has been used previously to define TFs families in bacteria [3; 65; 66], although it is less stringent than the E-value $\leq 10^{-4}$ used to reduce the total number of superfamilies assigned to major clades (Archaea, Bacteria, and Eukarya) by Yang and co-workers [21]. E-values here also serve as a confidence level for every candidate identified as a paralogue within an organism.

Thus, we predict groups of paralogues that include the set of 85 know TFs and 1593 TGs of *E. coli* from RegulonDB release 5.5 and the set of 91 know TFs and 732 TGs of *B. subtilis* from BDTBS release 4.1. In order to group putative paralogous regulatory proteins, we required that each group included the same resulting members after both PFAM and SUPERFAMILY domain assignments, except in the cases of seven *E. coli* and one *B. subtilis* TFs that have no SUPERFAMILY assignments with our cut-off value. In those cases only PFAM assignments

were considered in order to group them.

Search of orthologues

The search for orthologues was carried out as reported previously [10], assigning functional roles to TFs in other genomes by first filtering intraspecific paralogues and then using an intersection of three criteria for the detection of orthology: (i) bi-directional best hits (BDBHs), (ii) coverage of BLASTP [67] pairwise alignments and (iii) conservation of PFAM domains. Accordingly, we identified orthologues as pairs of *B. subtilis* and *E. coli* proteins that satisfy the following conditions:

(i) Sequences of the target genome that have a BDBH in the query genome with a significant BLASTP E-value ($<10^{-3}$).

(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 more than 100 residues were also considered in the analysis.

Estimation of transcription factor specificity based on the information content of DNA sequence motifs

Here we describe a way to estimate the observed DNA binding specificity of transcription factors for which we have at least two experimentally characterized binding sites. The process is essentially the same for our two bacterial datasets, with minor differences justified by the different annotation detail of *E. coli* and *B. subtilis* sites.

For *E. coli* we had a collection of 67 TFs with at least 2 reported sites, with 25 having more than 10 annotated sites. We used the computer program CONSENSUS [68] to build optimized sequence motifs with equiprobable prior nucleotide frequencies. We used the motif widths defined in RegulonDB 5.5 for each TF. CONSENSUS returns the unadjusted information content for each motif (UIC), that can be width-normalized so that different motifs can be

directly compared, using the expression IC_{norm} = IC / width. This is necessary as the motifs used in this work have widths that range from 7 (for instance NarL) to more than 20, and this variable ultimately limits the information content of motifs.

For *B. subtilis* we had a collection of 70 TFs with a minimum of 2 known sites, of which 23 have more than 10 associated sites, all extracted from DBTBS 4.1. Since sites for the same TF can have different widths in this data source, we used the program WCONSENSUS [68] to build sequence motifs with a prior %GC content of 43. This program attempts to find the optimal motif width in terms of information content.

In order to estimate the variability of scores for sites recognized by every TF we took the position weight matrices (PWM) generated by CONSENSUS (*E. coli*) and WCONSENSUS (*B. subtilis*) and aligned all available sites for each TF against them, by running the program PATSER [68] and recording the scores. The highest and lowest scores were kept, as well as the standard deviation, and the variability calculated with Equation 1:

variability(scores) = max(scores) – min(scores) / standard_dev(scores) (Equation 1)

Note that these variability measurements are normalized by the standard deviation of scores for a given TF, so they are comparable for different TFs.

Estimation of transcription factor specificity by estimating DNA binding potential

A modified version of the DNASITE program [69], that uses full atom detail and identifies hydrogen bonds and hydrophobic interactions, was used to estimate DNA-binding potentials (manuscript under review). Briefly, the program threads experimentally characterized DNA binding sites from RegulonDB 5.5 into crystallographic protein-DNA complexes for 11 transcription factors in *E. coli* and scores each site using H-bond and Van der Waals weight matrices. These matrices give log-likelihood scores to pairs of interacting atoms in the protein-DNA interface and were compiled on a set of non-redundant protein-DNA complexes. The sum of weights over a protein-DNA interface, linearly combined with indirect readout DNA deformation, is regarded as the potential of binding of a given site. As before, we calculate score

variability for a TF using Equation 1. These are the eleven TFs used here, with the number of binding sites for each indicated in parenthesis: TrpR (10), Rob (6), PurR (15), PhoB (16), NarL (73), MetJ (23), MarA (13), FadR (10), DnaA (8), CRP (182) and LacR (3). The list of corresponding Protein Data Bank complexes is: 1TRO [70], 1D5Y [71], 2PUA [72], 1GXP [73], 1JE8 [74], 1CMA [75], 1XS9 [76], 1H9T [77], 1J1V [78], 1CGP [79] and 1EFA [80].

Estimation of mean expression values from microarray experiments

A set of 60 published non-redundant expression profiles for *E. coli* was provided by the authors [39], already normalized using the robust multi-array analysis (RMA) procedure, that allows direct comparisons between them. Most of these conditions are independent single-gene over-expression experiments. The mean expression value across 60 conditions was then calculated for all those *E. coli* transcription factors for which an information content estimate of specificity was available, to produce the scatter plot shown in Figure 7.

Calculation of correlation coefficients

All correlation coefficients mentioned in this paper correspond to Pearson coefficients calculated using the function *cor.test* in the R package for statistical computing (http://www.r-project.org/).

Calculation of probabilities of site generation

The collection of binding sites for every TF was aligned using CONSENSUS with a fixed motif width of 10 columns, to make them all directly comparable. Alignments are then parsed in order to count the number of different sites of length 10 found, a number called diffN, that is an approximation of the sequence space recognized by any TF. The probability of generating sites for any one TF is then calculated by dividing diffN by 4^{10}, the total number of possible oligonucleotides of that length.

Online Supplementary Material: http://www.eead.csic.es/compbio/suppl/prok_specificity/

ACKNOWLEDGEMENTS

We thank Heladia Salgado and Sarath Chandra Janga for their help in obtaining RegulonDB and microarray expression data. We are also grateful to the Computational Genomics Group and an anonymous referee for comments and suggestions to improve this work. The Computational Genomics group is supported by NIH grant RO1-GM071962. B.C.M. was funded by a postdoctoral fellowship from Universidad Nacional Autónoma de México and by Fundación Aragón I+D. V.E.A. was supported by Red Iberoamericana de Bioinformática and CYTED and is now recipient of a doctoral fellowship awarded by Banco Santander Central Hispano, Fundación Carolina and Universidad de Zaragoza.

AUTHOR CONTRIBUTIONS

I.L.C. and B.C.M. designed and performed research; V.E.A. provided matrices of atomic protein-DNA contacts and I.L.C.,V.E.A., J.C.V. and B.C.M. wrote the paper.

FIGURE LEGENDS

Figure 1. Paralogous groups of transcriptions factors in the TRNs from *E. coli* (a) and *B. subtilis* (b). Global regulators are on the top row while local regulators are in the bottom. Black lines indicate directed transcriptional regulation between TFs, only for cases with evidence reported in RegulonDB and DBTBS. 24 paralogous families for *E. coli* and 25 for *B. subtilis* are circumscribed in shady rectangles in the figure. Paralogous families involving global regulators are shown as yellow ovals. Paralogous groups in which only one member of the family has experimental evidence are shown as green ovals. Finally, 15 TFs in *E. coli* and 26 TFs in *B. subtilis* predicted to be singletons, with no paralogous copies in the genome, are shown as blue ovals. This figure highlights the importance of duplication/horizontal transfer events across regulatory networks, since there are many paralogous groups. Note that several global regulators in both species either are part of groups in which other TFs are not global or are singletons (i.e. CodY and ComK in *B. subtilis*). This is important as it shows that recognizing paralogy gives little information about the evolutionary fate of TFs.

Figure 2. Scatter plot of normalized information content *versus* number of binding sites in *E. coli* (a) and *B. subtilis* (b). A linear fit is also plotted to illustrate the observed correlation coefficients of -0.81. The red dot highlights FNR, a protein that regulates respiration in both species, being labeled as global in *E. coli* (a) and as local in *B. subtilis* (b).

Figure 3. Scatter plot of binding energy variability versus log (number of binding sites), obtained from 11 *E. coli* TF-DNA complexes. The most variable transcription factor is CRP, whilst the most specific regulators are LacR and Rob.

Figure 4. *Scatter plot of co-regulators versus the number of regulated target genes in E. coli for each transcription factor.* Data was taken from RegulonDB 5.5, removing regulatory interactions without experimentally-determined binding sites associated to them.

Figure 5. Mean expression value of *E. coli* transcription factors (across 60 non-redundant microarray experiments) plotted versus the number of reported binding sites within the genome. As expected, in general local regulators are relatively less expressed when compared to global regulators.

Figure 6. Theoretical estimates of the probability of random generation of genomic binding sites in *E. coli* (A) and *B. subtilis* (B). Note that probabilities vary up to two orders of magnitude between specific and low specific DNA binders.

Figure 7. Evolutionary model for regulatory networks.
This plot shows variables that affect the evolution of transcriptions factors and their regulons. Two main variables are considered here, binding specificity and frequency of co-regulation, normalized in a [0-1] scale. Note that a scatter plot of these two variables clearly separates global transcription factors (plotted in red) from the other regulatory proteins, highlighting their potential diagnostic value. The subplot B summarizes the main observations of this paper, together with a theoretical variable that is not easily measured, effector relevance, that we anticipate can play an important role here. The model proposes to use the degree of co-regulation as an indirect measure of effector relevance, similarly to mutation resistance, which is represented as being inversely proportional to binding specificity. This evolutionary model lets us more realistically define the functional (global or local) role for any TF as a function of different evolutionary forces, rather than isolated properties that can misestimate the importance of TFs.

The subplot in the left bottom corner summarizes the main observations of this paper, together with a theoretical variable that is not easily measured, effector relevance, that we anticipate can play an important role here. Similarly to mutation resistance, which is represented as being inversely proportional to binding specificity, the model proposes to use the degree of co-regulation as an indirect measure of effector relevance.
potential diagnostic value. The subplot in the left bottom corner summarizes the main observations of this paper, together with a theoretical variable that is not easily measured,

effector relevance, that we anticipate can play an important role here. The model proposes to use the degree of co-regulation as an indirect measure of effector relevance. Similarly to mutation resistance which is represented as being inversely proportional to binding specificity. This evolutionary model can let us define the functional (global or local) role for any TF from more realistic situations by considering the role TF as a function of different evolutionary forces, rather than isolated properties that can overestimate/underestimate the function of a TF in the TRNs.

REFERENCES

1. **Thieffry, D., Huerta, A. M., Perez-Rueda, E. & Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays* 20, 433-40.**
2. **Guelzim, N., Bottani, S., Bourgine, P. & Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 31, 60-3.**
3. **Moreno-Campuzano, S., Janga, S. C. & Perez-Rueda, E. (2006). Identification and analysis of DNA-binding transcription factors in Bacillus subtilis and other Firmicutes--a genomic approach. *BMC Genomics* 7, 147.**
4. **Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science* 286, 509-12.**
5. **Gottesman, S. (1984). Bacterial regulation: global regulatory networks. *Annu Rev Genet* 18, 415-41.**

6. **Martinez-Antonio, A. & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* 6, 482-9.**
7. **Foster, D. V., Kauffman, S. A. & Socolar, J. E. (2006). Network growth models and genetic regulatory networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 73, 031912.**
8. **Teichmann, S. A. & Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet* 36, 492-6.**
9. **Cosentino Lagomarsino, M., Jona, P., Bassetti, B. & Isambert, H. (2007). Hierarchy and feedback in the evolution of the Escherichia coli transcription network. *Proc Natl Acad Sci U S A* 104, 5516-20.**
10. **Lozada-Chavez, I., Janga, S. C. & Collado-Vides, J. (2006). Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res* 34, 3434-45.**
11. **Madan Babu, M., Teichmann, S. A. & Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 358, 614-33.**
12. **Tobisch, S., Zuhlke, D., Bernhardt, J., Stulke, J. & Hecker, M. (1999). Role of CcpA in regulation of the central pathways of carbon catabolism in Bacillus subtilis. *J Bacteriol* 181, 6996-7004.**
13. **Morales, G., Linares, J. F., Beloso, A., Albar, J. P., Martinez, J. L. & Rojo, F. (2004). The Pseudomonas putida Crc global regulator controls the expression of genes from several chromosomal catabolic pathways for aromatic compounds. *J Bacteriol* 186, 1337-44.**
14. **Friedberg, D., Midkiff, M. & Calvo, J. M. (2001). Global versus local regulatory roles for Lrp-related proteins: Haemophilus influenzae as a case study. *J Bacteriol* 183, 4004-11.**
15. **Suh, S. J., Runyen-Janecky, L. J., Maleniak, T. C., Hager, P., MacGregor, C. H., Zielinski-Mozny, N. A., Phibbs, P. V., Jr. & West, S. E. (2002). Effect of vfr mutation on global gene expression and catabolite repression control of Pseudomonas aeruginosa. *Microbiology* 148, 1561-9.**
16. **Reents, H., Munch, R., Dammeyer, T., Jahn, D. & Hartig, E. (2006). The Fnr regulon of Bacillus subtilis. *J Bacteriol* 188, 1103-12.**
17. **Derouaux, A., Dehareng, D., Lecocq, E., Halici, S., Nothaft, H., Giannotta, F., Moutzourelis, G., Dusart, J., Devreese, B., Titgemeyer, F., Van Beeumen, J. & Rigali, S. (2004). Crp of Streptomyces coelicolor is the third transcription factor of the large CRP-FNR superfamily able to bind cAMP. *Biochem Biophys Res Commun* 325, 983-90.**
18. **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. (2006). RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34, D394-7.**
19. **Makita, Y., Nakao, M., Ogasawara, N. & Nakai, K. (2004). DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res* 32, D75-7.**
20. **Nakano, M. M. & Zuber, P. (1998). Anaerobic growth of a "strict aerobe" (Bacillus subtilis). *Annu Rev Microbiol* 52, 165-90.**
21. **Yang, S., Doolittle, R. F. & Bourne, P. E. (2005). Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A* 102, 373-8.**

22. Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* 311, 1283-7.
23. Amoutzias, G. D., Weiner, J. & Bornberg-Bauer, E. (2005). Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. *Gene* 347, 247-53.
24. Madan Babu, M. & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* 31, 1234-44.
25. Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X. Y., Biggin, M. D. & Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* 2, e130.
26. Doniger, S. W. & Fay, J. C. (2007). Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3, e99.
27. Espinosa, V., Gonzalez, A. D., Vasconcelos, A. T., Huerta, A. M. & Collado-Vides, J. (2005). Comparative studies of transcriptional regulation mechanisms in a group of eight gamma-proteobacterial genomes. *J Mol Biol* 354, 184-99.
28. Rajewsky, N., Socci, N. D., Zapotocky, M. & Siggia, E. D. (2002). The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Res* 12, 298-308.
29. Sengupta, A. M., Djordjevic, M. & Shraiman, B. I. (2002). Specificity and robustness in transcription control networks. *Proc Natl Acad Sci U S A* 99, 2072-7.
30. Stormo, G. D. & Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* 23, 109-13.
31. Luscombe, N. M. & Thornton, J. M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* 320, 991-1009.
32. Bilu, Y. & Barkai, N. (2005). The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol* 6, R103.
33. Cases, I. & de Lorenzo, V. (1998). Expression systems and physiological control of promoter activity in bacteria. *Curr Opin Microbiol* 1, 303-10.
34. Garges, S. & Adhya, S. (1988). Cyclic AMP-induced conformational change of cyclic AMP receptor protein (CRP): intragenic suppressors of cyclic AMP-independent CRP mutations. *J Bacteriol* 170, 1417-22.
35. Sawers, G., Kaiser, M., Sirko, A. & Freundlich, M. (1997). Transcriptional activation by FNR and CRP: reciprocity of binding-site recognition. *Mol Microbiol* 23, 835-45.
36. Kallipolitis, B. H., Norregaard-Madsen, M. & Valentin-Hansen, P. (1997). Protein-protein communication: structural model of the repression complex formed by CytR and the global regulator CRP. *Cell* 89, 1101-9.
37. Pedersen, H. & Valentin-Hansen, P. (1997). Protein-induced fit: the CRP activator protein changes sequence-specific DNA recognition by the CytR repressor, a highly flexible LacI member. *Embo J* 16, 2108-18.
38. Evangelisti, A. M. & Wagner, A. (2004). Molecular evolution in the yeast transcriptional regulation network. *J Exp Zoolog B Mol Dev Evol* 302, 392-411.
39. Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J. & Gardner, T. S. (2007). Large-scale mapping and validation

of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5, e8.

40. Aurell, E., d'Herouel, A. F., Malmnas, C. & Vergassola, M. (2007). Transcription factor concentrations versus binding site affinities in the yeast S. cerevisiae. *Phys Biol* 4, 134-43.
41. Bagg, A. & Neilands, J. B. (1987). Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in Escherichia coli. *Biochemistry* 26, 5471-7.
42. Baichoo, N., Wang, T., Ye, R. & Helmann, J. D. (2002). Global analysis of the Bacillus subtilis Fur regulon and the iron starvation stimulon. *Mol Microbiol* 45, 1613-29.
43. Ollinger, J., Song, K. B., Antelmann, H., Hecker, M. & Helmann, J. D. (2006). Role of the Fur regulon in iron transport in Bacillus subtilis. *J Bacteriol* 188, 3664-73.
44. Batchelor, E., Walthers, D., Kenney, L. J. & Goulian, M. (2005). The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC. *J Bacteriol* 187, 5723-31.
45. Yamamoto, K. & Ishihama, A. (2006). Characterization of copper-inducible promoters regulated by CpxA/CpxR in Escherichia coli. *Biosci Biotechnol Biochem* 70, 1688-95.
46. DiGiuseppe, P. A. & Silhavy, T. J. (2003). Signal detection and target gene induction by the CpxRA two-component system. *J Bacteriol* 185, 2432-40.
47. Howell, A., Dubrac, S., Noone, D., Varughese, K. I. & Devine, K. (2006). Interactions between the YycFG and PhoPR two-component systems in Bacillus subtilis: the PhoR kinase phosphorylates the non-cognate YycF response regulator upon phosphate limitation. *Mol Microbiol* 59, 1199-215.
48. Makino, K., Shinagawa, H., Amemura, M., Kawamoto, T., Yamada, M. & Nakata, A. (1989). Signal transduction in the phosphate regulon of Escherichia coli involves phosphotransfer between PhoR and PhoB proteins. *J Mol Biol* 210, 551-9.
49. Baruah, A., Lindsey, B., Zhu, Y. & Nakano, M. M. (2004). Mutational analysis of the signal-sensing domain of ResE histidine kinase from Bacillus subtilis. *J Bacteriol* 186, 1694-704.
50. Lulko, A. T., Buist, G., Kok, J. & Kuipers, O. P. (2007). Transcriptome analysis of temporal regulation of carbon metabolism by CcpA in Bacillus subtilis reveals additional target genes. *J Mol Microbiol Biotechnol* 12, 82-95.
51. Cruz Ramos, H., Hoffmann, T., Marino, M., Nedjari, H., Presecan-Siedel, E., Dreesen, O., Glaser, P. & Jahn, D. (2000). Fermentative metabolism of Bacillus subtilis: physiology and regulation of gene expression. *J Bacteriol* 182, 3072-80.
52. Cruz Ramos, H., Boursier, L., Moszer, I., Kunst, F., Danchin, A. & Glaser, P. (1995). Anaerobic transcription activation in Bacillus subtilis: identification of distinct FNR-dependent and -independent regulatory mechanisms. *Embo J* 14, 5984-94.
53. Khoroshilova, N., Popescu, C., Munck, E., Beinert, H. & Kiley, P. J. (1997). Iron-sulfur cluster disassembly in the FNR protein of Escherichia coli by O2: [4Fe-4S] to [2Fe-2S] conversion with loss of biological activity. *Proc Natl Acad Sci U S A* 94, 6087-92.

54. Schaechter, M. (2000). Escherichia coli, General Biology. In *Encyclopedia of Microbiology* Second Edition edit. (Lederberg, J., ed.), Vol. 1 A-C, pp. 260-269. Academic Press, New York.

55. Price, M. N., Dehal, P. S. & Arkin, A. P. (2007). Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput Biol* 3, 1739-50.

56. Dame, R. T. (2005). The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol Microbiol* 56, 858-70.

57. Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand, M. S. & Mirny, L. A. (2007). How gene order is influenced by the biophysics of transcription regulation. *Proc Natl Acad Sci U S A* 104, 13948-53.

58. Watanabe, H., Mori, H., Itoh, T. & Gojobori, T. (1997). Genome plasticity as a paradigm of eubacteria evolution. *J Mol Evol* 44 Suppl 1, S57-64.

59. Daber, R., Stayrook, S., Rosenberg, A. & Lewis, M. (2007). Structural analysis of lac repressor bound to allosteric effectors. *J Mol Biol* 370, 609-19.

60. Stec, E., Witkowska-Zimny, M., Hryniewicz, M. M., Neumann, P., Wilkinson, A. J., Brzozowski, A. M., Verma, C. S., Zaim, J., Wysocki, S. & Bujacz, G. D. (2006). Structural basis of the sulphate starvation response in E. coli: crystal structure and mutational analysis of the cofactor-binding domain of the Cbl transcriptional regulator. *J Mol Biol* 364, 309-22.

61. Gough, J. & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res* 30, 268-72.

62. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-40.

63. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* 30, 276-80.

64. Eddy, S. R. (1996). Hidden Markov models. *Curr Opin Struct Biol* 6, 361-5.

65. Perez-Rueda, E. & Collado-Vides, J. (2000). The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res* 28, 1838-47.

66. Perez-Rueda, E., Collado-Vides, J. & Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem* 28, 341-50.

67. Lopez, R., Silventoinen, V., Robinson, S., Kibria, A. & Gish, W. (2003). WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res* 31, 3795-8.

68. Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563-77.

69. Contreras-Moreira, B. & Collado-Vides, J. (2006). Comparative footprinting of DNA-binding proteins. *Bioinformatics* 22, e74-80.

70. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* 335, 321-9.

71. Kwon, H. J., Bennik, M. H., Demple, B. & Ellenberger, T. (2000). Crystal structure of the Escherichia coli Rob transcription factor in complex with DNA. *Nat Struct Biol* 7, 424-30.

72. **Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices.** ***Science*** **266, 763-70.**
73. **Blanco, A. G., Sola, M., Gomis-Ruth, F. X. & Coll, M. (2002). Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator.** ***Structure*** **10, 701-13.**
74. **Maris, A. E., Sawaya, M. R., Kaczor-Grzeskowiak, M., Jarvis, M. R., Bearson, S. M., Kopka, M. L., Schroder, I., Gunsalus, R. P. & Dickerson, R. E. (2002). Dimerization allows DNA target site recognition by the NarL response regulator.** ***Nat Struct Biol*** **9, 771-8.**
75. **Somers, W. S. & Phillips, S. E. (1992). Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands.** ***Nature*** **359, 387-93.**
76. **Dangi, B., Gronenborn, A. M., Rosner, J. L. & Martin, R. G. (2004). Versatility of the carboxy-terminal domain of the alpha subunit of RNA polymerase in transcriptional activation: use of the DNA contact site as a protein contact site for MarA.** ***Mol Microbiol*** **54, 45-59.**
77. **van Aalten, D. M., DiRusso, C. C. & Knudsen, J. (2001). The structural basis of acyl coenzyme A-dependent regulation of the transcription factor FadR.** ***Embo J*** **20, 2041-50.**
78. **Fujikawa, N., Kurumizaka, H., Nureki, O., Terada, T., Shirouzu, M., Katayama, T. & Yokoyama, S. (2003). Structural basis of replication origin recognition by the DnaA protein.** ***Nucleic Acids Res*** **31, 2077-86.**
79. **Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees.** ***Science*** **253, 1001-7.**
80. **Bell, C. E. & Lewis, M. (2000). A closer view of the conformation of the Lac repressor bound to operator.** ***Nat Struct Biol*** **7, 209-14.**

TABLES 1-2

Bacillus subtilis			
	#sites	normUIC	sampled_normUIC
ComK	120	0.72	0.75
PhoP	65	0.62	0.68
CcpA	48	0.84	0.88
MtrB	41	1.1	1.2
SpoIIID	39	0.69	0.86
Spo0A	38	0.64	0.8
TnrA	37	0.98	1.03
AbrB	34	0.52	0.78
CodY	34	0.66	0.9
DegU	37	0.94	0.91
GerE	31	0.71	0.88
Fur	26	0.85	0.95
PurR	24	0.78	1
SpoVT	16	1.4	1.4
DnaA	16	1.06	1.28
Hpr	14	0.75	1.21
PerR	13	0.81	1.25
ResD	12	0.85	1.25
GlnR	12	1.07	1.35
SnR	12	0.99	1.35
CssR	12	0.95	1.34
YlbO	12	0.89	1.39
AraR	11	0.91	1.28
RocR	11	1.22	1.35
CtsR	10	1.06	1.28
ComA	10	0.97	1.35
GlpP	10	1.2	1.35
LexA	10	0.94	1.38
Rok	10	1.34	1.34
CcpC	9	1.26	1.27
Fnr	8	1.26	1.38
YycF	8	1.2	1.32
CitT	8	0.91	1.27
PucR	7	0.83	1.28

Escherichia coli			
	#sites	normUIC	sampled_normUIC
CRP	207	0.36	0.39
Fis	121	0.29	0.35
IHF	78	0.5	0.57
ArcA	77	0.48	0.57
NarL	76	0.71	0.78
FNR	75	0.56	0.63
Lrp	54	0.28	0.42
Fur	47	0.8	0.9
H-NS	34	0.45	0.64
CpxR	33	0.55	0.71
LexA	23	0.69	0.84
MetJ	23	1.04	1.16
OmpR	20	0.55	0.78
ArgR	18	0.66	0.88
SoxS	18	0.57	0.79
GlpR	18	0.55	0.8
PhoP	18	0.58	0.82
PhoB	17	0.54	0.81
TyrR	17	0.6	0.82
PurR	16	0.83	1.06
MarA	16	0.44	0.81
MalT	15	0.55	0.83
NarP	14	0.93	1.12
AraC	13	0.59	0.98
FruR	12	0.78	1.04
TrpR	10	1.03	1.14
Nac	10	0.66	1
CytR	10	0.51	0.85
NagC	10	0.68	0.97
GntR	10	0.71	0.99
FadR	10	0.7	0.98
IclR	10	1.02	1.25
OxyR	9	0.5	1.03
CysB	8	0.68	1.14
DnaA	8	0.71	1.12
IscR	8	0.62	1.09
Mlc	7	0.9	1.16
DeoR	7	0.73	1.15
GalR	7	0.83	1.15
GalS	7	0.93	1.21

Table 1. Normalized information content (specificity) of transcription factors in *B. subtilis* and *E. coli* with 7+ reported binding sites. The left column in each species shows mean IC values computed after sampling 100 times taking only 30% of the available sites.

Species	Name(s)	ID(s)	bsDNAs (#)	Role	Effectors / others	Domains	TF-DNA contacts	Specificity
E. coli	**FNR**	b1334 16129295	75	GLOBAL [1]	O_2 [2; 3]	**PF00325.11** PF00027.18	197-G, 207-V, **208-E**	**0.63**
B. subtilis	**FNR**	BG11343 16080784	8	LOCAL [4]	O_2 [4]	**SF46785** SF51206	178-Q, 188-R, **189-E**	**1.38**
E. coli	**Lrp** (AlsB, LstR)	b0889 16128856	54	GLOBAL [5]	Leucine, alanine [6]	PF01037.10	-	**0.42**
B. subtilis	**AzlB** (YrdG)	BG11914 16079725	1	LOCAL [7]	Unknown [7]	**SF46785** SF54909	24-L, 34-P, 35-S	-
E. coli	**CytR**	b3934 16131772	10	LOCAL [8]	Cytidine [9]	**PF00356.11** PF00532.11	**23-A**, 33-D, 61-V, 62-K	**0.85**
B. subtilis	**CcpA** (GraR, AlsA)	BG10376 16080026	48	GLOBAL [10; 11]	[HPr (Ser-P)] and [Crh (Ser-P)] [12], Frc 1,6-P_2 and Glc-6P [13]	**SF47413** SF53822	15-S, 16-G, **17-A**, 21-R, 55-L, 56-A	**0.88**
E. coli	**Fur**	b0683 16128659	47	LOCAL	Fe^{2+} [14]	**PF01475.9**	-	**0.9**
B. subtilis	**Fur** (YqkL)	BG11766 16079409	26	LOCAL [15; 16]	Fe^{2+} [17]	**SF46785**	-	**0.95**

E. coli	**LexA** (ExrA,LexA, Spr,TsI,UmuA)	b4043 16131869	23	LOCAL	Self-cleavage [18] (*in vivo* requires recA)	**PF01726.7** PF00717.13	-	**0.84**
B. subtilis	**LexA** (DinR)	16078848 BG10678	10	LOCAL	Self-Cleavage [19] (*in vivo* requires recA)	**SF46785** SF51306	-	**1.38**
E. coli	**DnaA**	b3702 16131570	8	LOCAL	ATP and ADP [20;21]	**PF08299.1** PF00308.8	**398-R**, **422-P**, **432-D**, **433-H**, **434-T**, **435-T**, 437-L, **438-H**	**1.12**
B. subtilis	**DnaA** (DnaH, DnaJ, DnaK)	BG10065 16077069	16	LOCAL	ATP and ADP [22]	**SF48295** SF52540	**378-R**, **402-P**, **412-D**, **413-H**, **414-T**, **415-T**, 417-I, **418-H**	**1.28**
E. coli	**CpxR** (YiiA)	b3912 16131752	33	LOCAL	Phosphorylated by CpxA (Cu ions – e.i $CuSO_4$) [23]; pH and EDTA (Digiuseppe and Silhavy, 2003) * [two-components]	**PF00072.13** PF00486.18 **SF52172**	**194-R**, 195-A, 198-M, 202-N, **221-R**	**0.71**
B. subtilis	**YycF**	BG10001 16081093	8	LOCAL	Phosphorylated by YycG (Not yet determinated [24]) [two-components]		**195-R**, 196-T, 199-V, 203-R, **222-R**	**1.32**
E. coli	**PhoB** (PhoT)	b0399 16128384	17	LOCAL	Phosphorylated by *1)* PhoR (ATP) [25] [two-components]	**PF00072.13** PF00486.18	**192-R**, **193-T**, 196-V, **200-R**, 218-R	**0.81**
B. subtilis	**ResD** (YpxD)	BG10534 16079369	12	LOCAL	Phosphorylated by ResE (O_2? and NO) [26] [two-components]	SF52172	**200-R**, **201-T**, 204-T, **208-R**, 228-W	**1.25**

Table 2. Orthologous TFs shared between *Escherichia coli* and *Bacillus subtilis*.

Eight orthologous transcription factors with experimentally verified DNA binding sites available were found in both bacteria. Name(s), identification number(s) and the number of bsDNA were compiled from RegulonDB and DBTBS. Information concerning the global and local roles of TFs was taken from Martinez-Antonio and Collado-Vides [27] and Moreno-Campuzano *et al.* [28]. TF-DNA contacts were predicted using the TFmodeller software [29], marking conserved interface residues in bold. PFAM (PF) and SUPERFAMILY (SF) DNA-binding domains are marked in bold type. Information about effectors was compiled from literature. Symbols → **O_2**: oxygen; **[HPr (Ser-P)]**: Histidine-containing Protein (Ser46-phosphorylated); **[Crh (Ser-P)]**: Catabolite repression HPr (Ser46-phosphorylated); **Frc-1,6-P_2**: Fructose-1,6-bisphosphate; **Glc-6P**: Glucose 6-phosphate; **Fe^{2+}**: ionic iron (II); **ATP**: Adenosine TriPhosphate; **ADP**: Adenosine DiPhosphate; **Cu**: copper ions (i.e. copper sulfate -**$CuSO_4$**-); **NO**: Nitric oxide. *Other inducers for the CpxR-A two-component system have been identified, such as the accumulation of misfolded pilus subunits PapG and PapE and of lipid II ECA intermediate, as well as decrease levels of phosphatidylethanolamine; however, it is not known if these inducers generates a unique signal that is sensed by Cpx system[30]. Specificity data were taken from Table 1.

Figure 1A

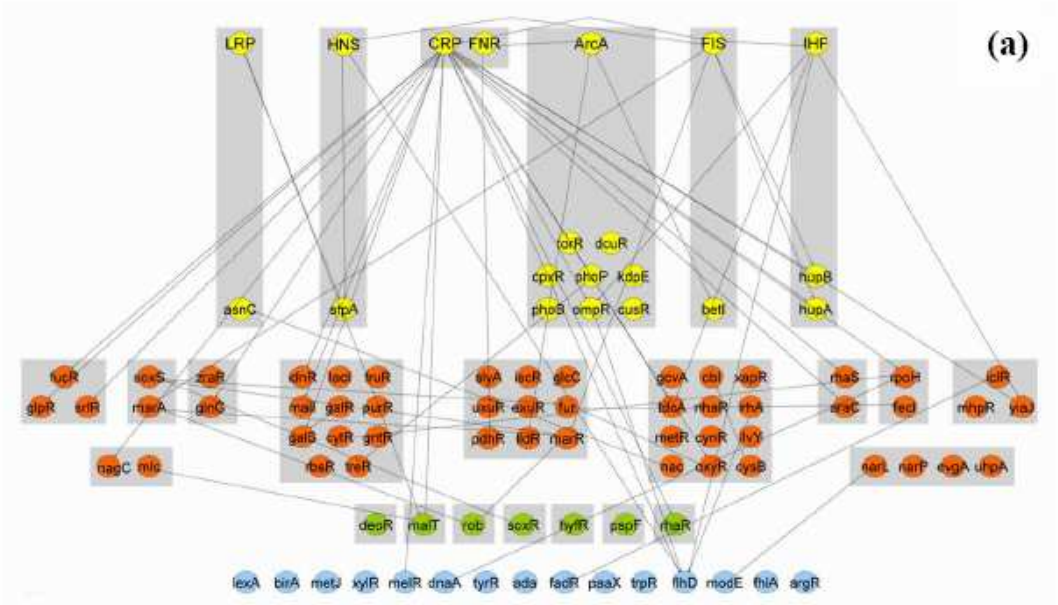

Figure 1B

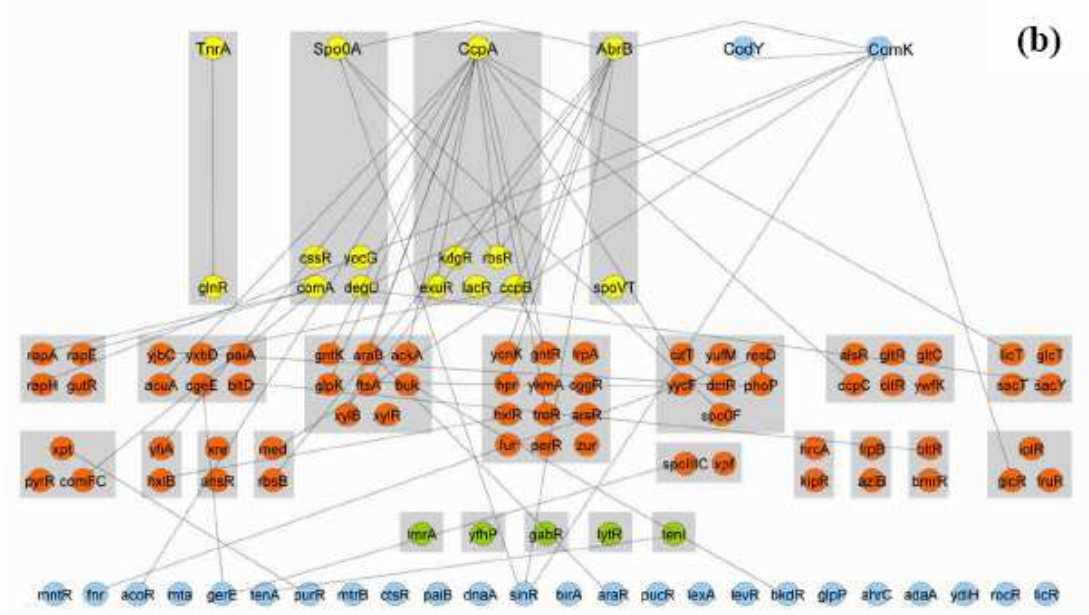

Figure 2A

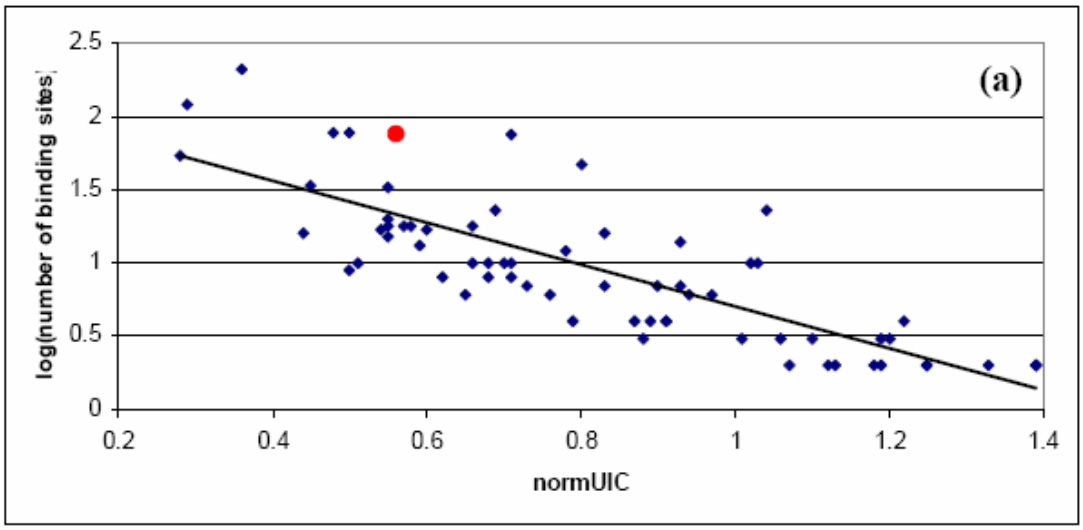

Figure 2B

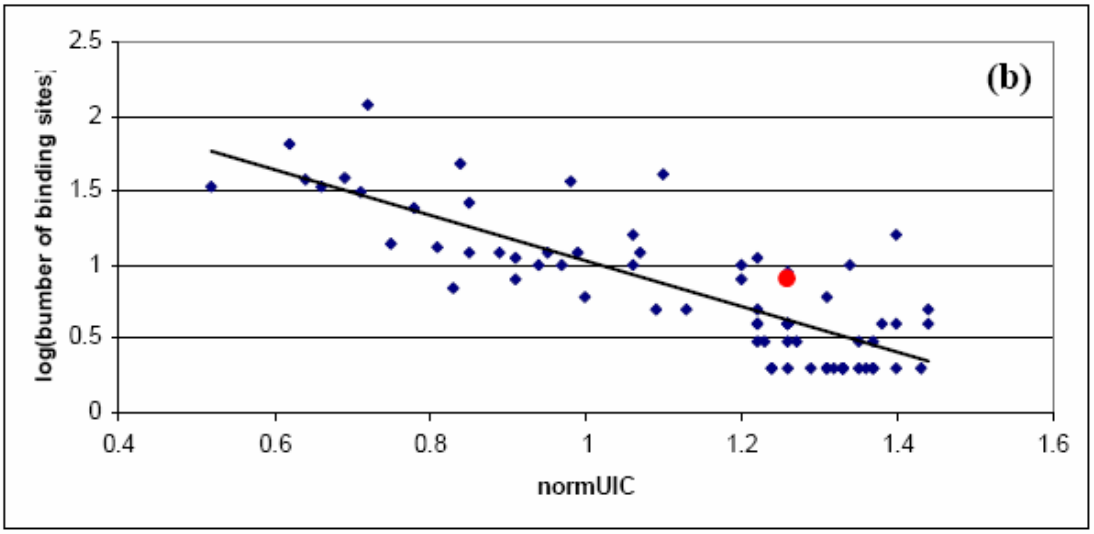

Figure 3

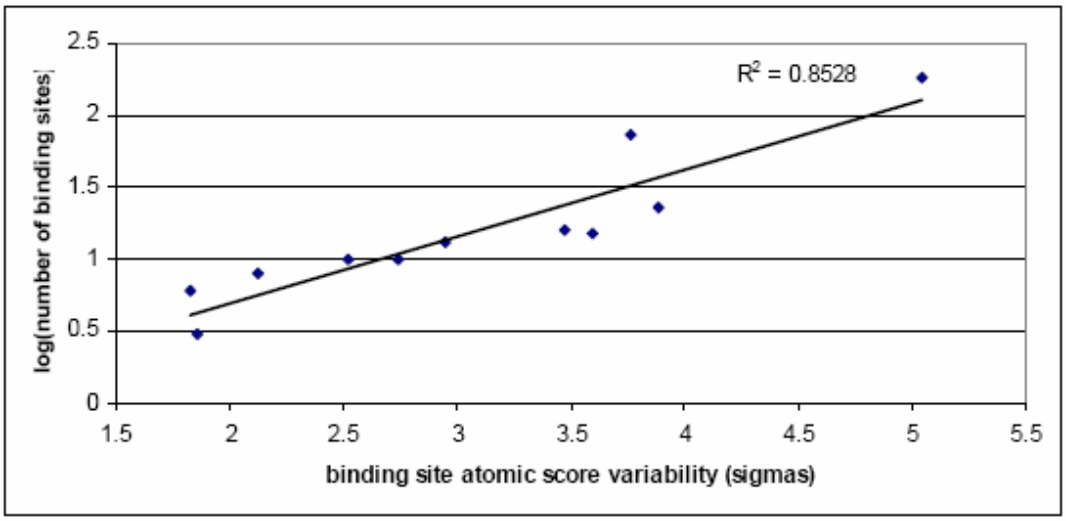

Figure 4

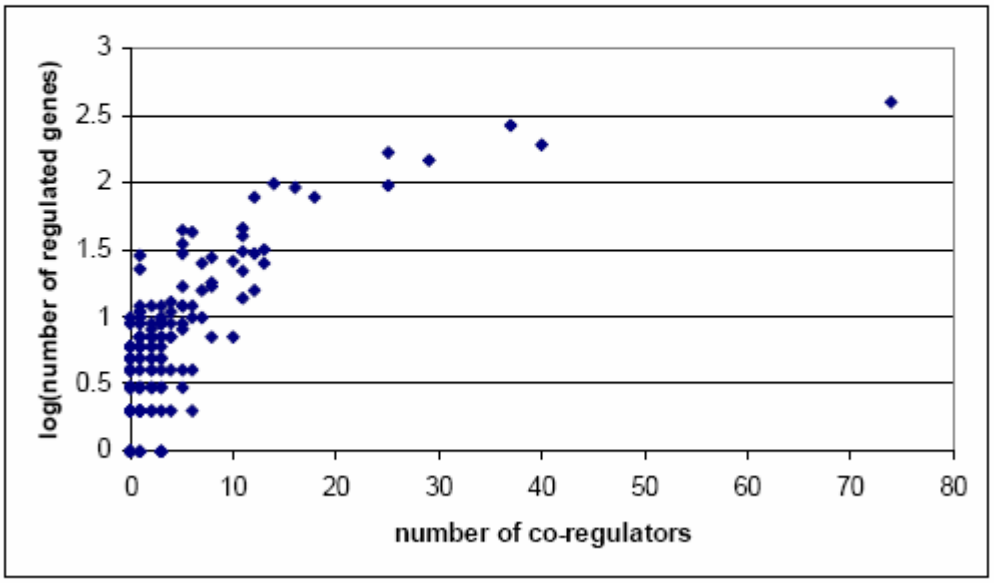

Figure 5

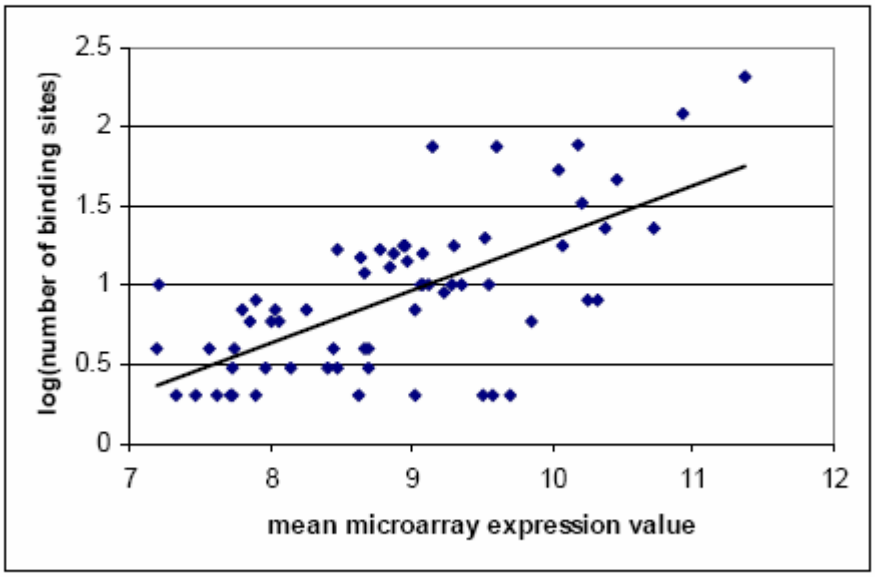

Figure 6A

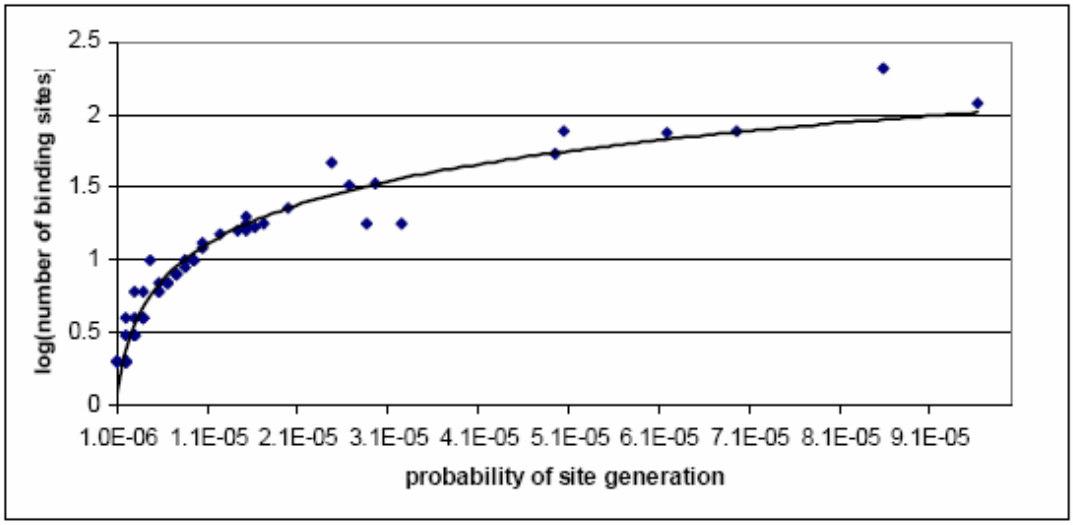

Figure 6B

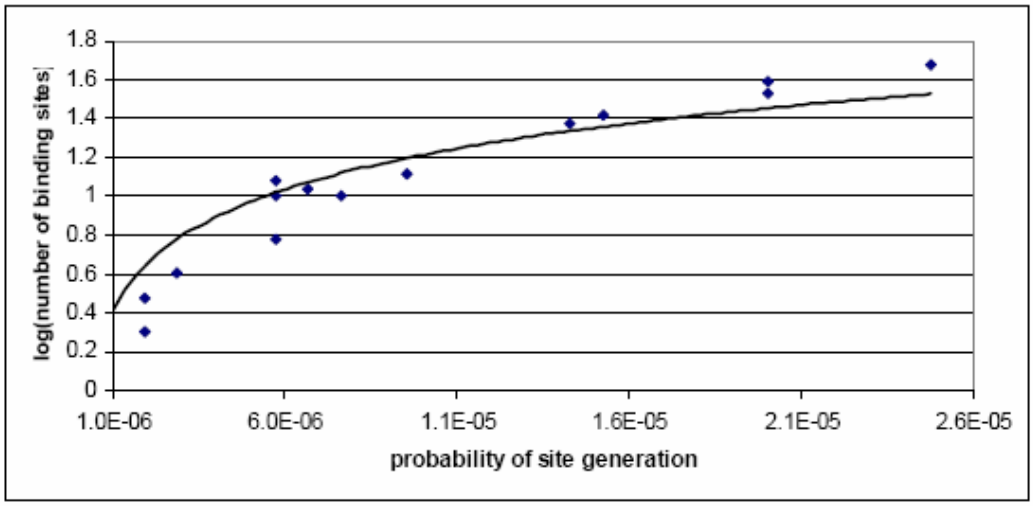

Figure 7A

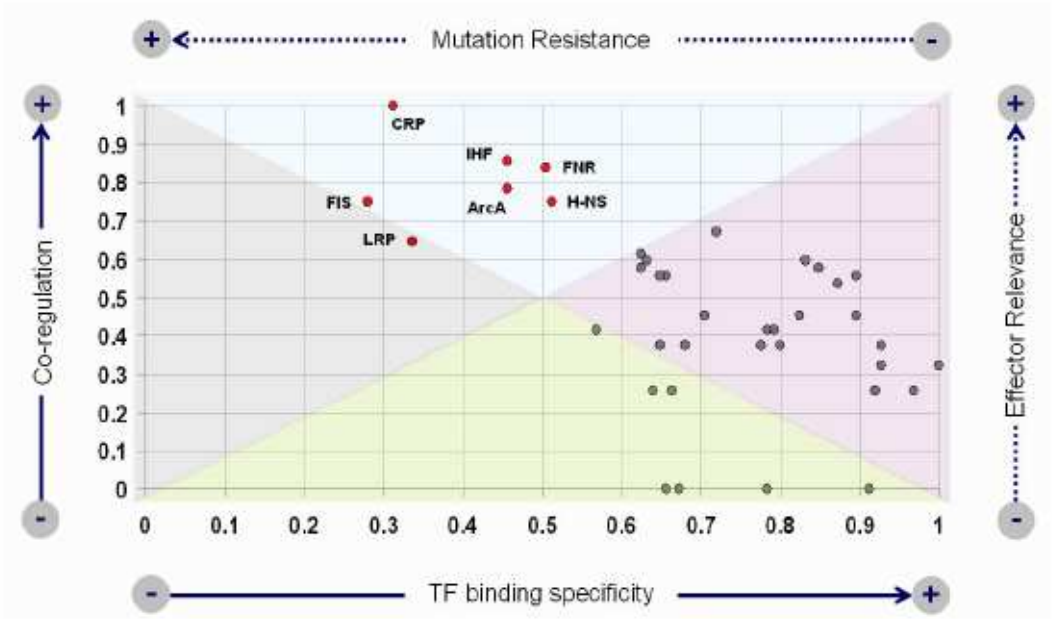

Figure 7B