Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria

Nancy Rivera-Gómez, Lorenzo Segovia and Ernesto Pérez-Rueda

Correspondence

Ernesto Pérez-Rueda
erueda@ibt.unam.mx

Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

The ability of bacteria to deal with diverse environmental changes depends on their repertoire of genes and their ability to regulate their expression. In this process, DNA-binding transcription factors (TFs) have a fundamental role because they affect gene expression positively and/or negatively depending on operator context and ligand-binding status. Here, we show an exhaustive analysis of winged helix–turn–helix domains (wHTHs), a class of DNA-binding TFs. These proteins were identified in high proportions and widely distributed in bacteria, representing around half of the total TFs identified so far. In addition, we evaluated the repertoire of wHTHs in terms of their partner domains (PaDos), identifying a similar trend, as with TFs, i.e. they are abundant and widely distributed in bacteria. Based on the PaDos, we defined three main groups of families: (i) monolithic, those families with little PaDo diversity, such as LysR; (ii) promiscuous, those families with a high PaDo diversity; and (iii) monodomain, with families of small sizes, such as MarR. These findings suggest that PaDos have a very important role in the diversification of regulatory responses in bacteria, probably contributing to their regulatory complexity. Thus, the TFs discriminate over longer regions on the DNA through their diverse DNA-binding domains. On the other hand, the PaDos would allow a great flexibility for transcriptional regulation due to their ability to sense diverse stimuli through a variety of ligand-binding compounds.

Received 19 April 2011

Revised 26 May 2011

Accepted 27 May 2011

INTRODUCTION

Bacteria are unicellular organisms that have a ubiquitous distribution. In recent years, more than 1000 organisms from diverse phylogenetic divisions have been completely sequenced, showing that the genomic organization resulting in the contemporary systems is a product of diverse evolutionary events, such as gene expansion, gene loss and lateral gene transfer. In this regard, two important factors have been identified as being responsible for the plasticity of the genomes, i.e. the gene repertoire and the regulatory mechanisms (Bengtsson, 2004; Lynch & Conery, 2003; Lynch, 2006). In general, gene regulation at the transcription initiation level in bacteria is primarily mediated by sigma (σ) factors and by DNA-binding transcription factors (TFs) (Browning & Busby, 2004). σ Factors provide the specificity for promoter recognition and DNA melting needed for transcription initiation (Gruber & Gross, 2003; Ishihama, 2000; Wösten, 1998), although they perform these functions only when bound to the RNA polymerase. In contrast, TFs affect gene expression positively and/or negatively depending on operator

context and ligand-binding status (Martínez-Antonio *et al.*, 2006; Miroslavova & Busby, 2006; Wall *et al.*, 2004).

Comparative bacterial genome analyses have shown that TFs may vary considerably in abundance and distribution (Aravind *et al.*, 2005; Levine & Tjian, 2003; Madan Babu *et al.*, 2006). In this regard, diverse studies have suggested that the abundance of TFs increases with increasing organismal complexity (Brown *et al.*, 2002; Changizi, 2001; Levine & Tjian, 2003; van Nimwegen, 2003; West & Brown, 2005), and the different proportions of these regulatory elements suggest that interplay among them defines the complexity of a regulatory network.

Although TFs are related to a wide diversity of functions, including differentiation, DNA restoration and cellular maintenance, and information on a large number of functional descriptions has accumulated, many questions remain unanswered, mainly associated with bacterial regulatory network organization and the repertoire of TFs. In this regard, various authors have described that bacteria share common principles of gene regulation across large phylogenetic distances, such as in *Escherichia coli* and *Bacillus subtilis* (Janga & Pérez-Rueda, 2009; Moreno-Campuzano *et al.*, 2006).

In this work, we identified the DNA-binding TFs that belong to the winged helix–turn–helix (wHTH) domain in

Abbreviations: DBD, DNA-binding domain; PaDo, partner domain; TF, transcription factor; wHTH, winged helix–turn–helix domain.

Three supplementary tables and three supplementary figures are available with the online version of this paper.

668 sequenced bacterial genomes representing a large diversity of divisions, lifestyles and genome sizes. wHTH domains have been classified as extensions of HTH domains, which are characterized by the presence of a third α -helix and an adjacent β -sheet and are central components in DNA binding. The recognition helix binds as in the regular HTH motifs, and the extra secondary structural elements provide additional contacts with the DNA backbone (Brennan & Matthews, 1989a, b). This structure has been identified in almost all micro-organisms, from Bacteria to Archaea, and is composed of diverse families, such as the catabolite gene activator (CAP) family, the heat shock and E2F/DP TFs, and the Ets domain family, among others (Brennan, 1993). In addition to the DNA-binding domain (DBD), TFs usually contain additional domains, called partner domains (PaDos), that are associated with diverse functions, such as protein–protein interactions, ligand-binding and/or catalytic activities (Madan Babu & Teichmann, 2003). This kind of structural organization is in agreement with previous reports which suggested that about two-thirds of proteins in prokaryotes are multidomain proteins (Tordai *et al.*, 2005).

Thus, TFs are two-headed proteins, with a DBD and a PaDo. DBDs have been widely used to classify TFs into families (Kummerfeld & Teichmann, 2006; Pérez-Rueda *et al.*, 2004; Pérez-Rueda & Janga, 2010). In contrast, few analyses are available on the additional domains, or PaDos, despite their importance in the regulatory response. PaDos have been associated with diverse functions, such as allosteric regulation of TFs across binding to a wide variety of functional compounds, in protein–protein interactions, or with enzymic properties (Madan Babu & Teichmann, 2003), and they are a fundamental link to environmental conditions and the functional conformational changes in the regulators (Taraban *et al.*, 2008).

In order to evaluate the abundance and distribution across all bacterial taxonomic divisions, the repertoire of wHTH TFs was analysed in terms of their domain organization. We evaluated the PaDos that are involved in DNA-binding activity, because relatively few of them have been explored in regulatory families, such as in the GntR family, for which four subfamilies have been identified that correlate with the functions of the regulated genes (Rigali *et al.*, 2002, 2004). From a global perspective, diverse types of PaDos, such as those associated with small-molecule binding, protein–protein binding and enzymic domains, and those with unknown function, have been identified in the regulatory network of *E. coli* K-12 (Madan Babu & Teichmann, 2003). The results obtained here provide insights into the functional and evolutionary constraints imposed on the expansion patterns of the TFs with a wHTH in bacteria. We believe that an improved understanding of the evolution of the transcriptional regulatory machinery across bacterial genomes will improve our knowledge about the evolutionary constraints that play a role in the formation of regulatory networks.

METHODS

Genome sequences. The complete list of genomes evaluated was obtained from the website <ftp://ftpncbi.nlm.nih.gov/genomes/bacteria>. We considered annotated genes as those with open reading frames that encode predicted protein sequences (the proteome) in all bacteria.

Identification of TFs. In order to identify the repertoire of TFs in the sequenced bacterial genomes, we used a combination of information sources and bioinformatics tools. As a first step, we identified and evaluated 295 wHTH TFs in three bacterial models, *E. coli* K-12, *B. subtilis* and *Corynebacterium glutamicum*, from three different databases, RegulonDB v 6.0 (Gama-Castro *et al.*, 2008), DBTBS v 5.0 (Sierro *et al.*, 2008) and CoryneRegnet v 4.0 (Baumbach, 2007), and their domain assignments were obtained from the Superfamily database (version 25 April 2010). From these, 107 TFs belong to *E. coli*, 118 to *B. subtilis* and 70 to *C. glutamicum*. These TFs clustered into families according to their PFAM assignments (Finn *et al.*, 2008), and their domain organization was defined, leaving 176 PaDos, clustered in 27 superfamilies, with 15 in *E. coli*, 23 in *B. subtilis* and 11 in *C. glutamicum*.

In the second stage, TFs associated with the wHTH domain were identified in 668 complete bacterial genomes, based on specific Hidden Markov model searches and from the regulators deposited in the DBD and Superfamily databases (Kummerfeld & Teichmann, 2006). For the purposes of our study, 428 nonredundant organisms were considered. The organisms classified were from 19 phyla: *Acidobacteria*, *Actinobacteria*, *Aquificales*, *Bacterioidetes*, *Chlamydia*, *Chlorobi*, *Chloroflexi*, *Cyanobacteria*, *Deinococcus*, *Thermus*, *Dictyoglomi*, *Elusimicrobia*, *Fusobacteria*, *Planctomycetes*, *Spirochaetes*, *Tenericutes*, *Thermotogae*, *Verrucomicrobia* and *Firmicutes*, and also *Proteobacteria*. Because of the large diversity and numerous organisms that have been completely sequenced and are associated with *Firmicutes* and *Proteobacteria*, their classes were also considered: *Bacillus* and *Clostridium* for *Firmicutes*, and *Alpha*-, *Beta*-, *Delta*-, *Epsilon*- and *Gamma*proteobacteria (Supplementary Table S1, available with the online version of this paper). In this work, we refer to nonredundant genomes as representative bacterial species, as previously defined by Janga & Moreno-Hagelsieb (2004). In brief, if there are diverse strains of the same species, a representative genome is considered; however, the order of elimination follows the importance of certain species as model organisms (such as *E. coli* K-12 and/or *B. subtilis*), and then the order of importance follows the highest number of genes having orthologues across phyla. For instance, *Mycobacterium avium* strain 104 is representative of diverse *Mycobacterium* strains (*M. avium* paratuberculosis, *M. bovis*, *M. bovis* BCG Pasteur 1173P2, *M. leprae* and *M. smegmatis* MC2 155) (Supplementary Table S1).

Clustering of families of regulatory factors and PaDos. In order to evaluate the distribution and abundance of TF families and their corresponding PaDos across 428 nonredundant bacterial genomes, a hierarchical complete linkage clustering algorithm was applied with correlation uncentred as the similarity measure. Analyses were performed using the program Mev (multiexperiment viewer; <http://www.tm4.org/mev>). In order to determine the relative abundance of the families of TFs and their associated PaDos by phylum, we calculated the fraction of genomes in the group that had at least one member versus the number of representative organisms. Thus, the following formula was considered: relative abundance by phylum = (total no. of TFs or PaDos identified)/(total no. of representative organisms by phylum). Thus, a value of 1 corresponds to presence and 0 represents absence. Because our aim was to evaluate the taxonomical distribution of TFs and PaDos, 24 taxonomical divisions were considered.

RESULTS

The proportion of wHTHs contributes significantly to the total repertoire of TFs in bacteria

In order to gain insights into the commonalities and differences in gene regulation between bacterial species from the perspective of TFs, we compared the repertoires of TFs identified in 428 nonredundant bacterial sequenced genomes by using diverse bioinformatics tools. From this analysis, we found that the wHTH comprises 48 % of the total TFs identified in bacteria, being the most abundant superfamily of DNA-binding structures described so far in this cellular domain (Table 1 and Supplementary Table S1). Indeed, alternative DNA-binding structures have been identified in minor proportions, such as the lambda repressor, homeodomain-like domain and C-terminal effector domain. This result not only correlates with the fact that the DBDs are associated with TFs, and in particular wHTHs are among the most ancient domains, probably derived from a relatively small set of folds (Aravind & Koonin, 1999; Madan Babu & Teichmann, 2003; Pérez-Rueda & Collado-Vides, 2001), but also shows that the wHTHs have been highly successful domains in nature.

Based on this apparent overrepresentation of wHTH proteins in the repertoire of TFs, we evaluated their abundance in the context of genome size. In this regard, it was previously described that TFs follow a quadratic distribution, i.e. large genomes contain a high proportion of TFs and, vice versa, small genomes have a small repertoire of TFs (Pérez-Rueda *et al.*, 2004, 2009). Thus, we asked

whether wHTH abundance correlates with the number of open reading frames. From this analysis, we found that wHTHs follow a similar distribution to the total repertoire of TFs, suggesting that this superfamily contributes significantly to this distribution (see Fig. 1). Indeed, the wHTH corresponds to around 50 % of the total of TFs per genome and in some organisms more than 70 % of the total repertoire of TFs corresponds to wHTH proteins, such as in *Mycoplasma genitalium*, *Prochlorococcus marinus* NATL1A, *Bordetella bronchiseptica* RB50 and *Thermosynechococcus elongatus* BP-1. Based on these results, we consider that the abundance of TFs and wHTHs may be associated with organismal diversity, i.e. organisms with free-living lifestyles and a large genome size contain a major proportion of transcriptional regulators, compared with parasitic or symbiont organisms with their small genome sizes, as described previously (Pérez-Rueda *et al.*, 2009). For instance, diverse free-living bacteria, such as *Burkholderia* sp. 383, with its large genome size, contain a high proportion of wHTH TFs. Recent studies suggest that organisms with free lifestyles require a large proportion of regulatory proteins as a consequence of an increase in the genome size, which also increases the number of putative regulatory interactions (Croft *et al.*, 2003). Alternatively, organisms with smaller genome sizes, such as *Mycoplasma* species, contain a small proportion of wHTH TFs, between one and six wHTH TFs. This proportion of TFs correlates with the fact that symbionts and/or parasitic obligate bacteria have substantially reduced genome sizes. Indeed, in some organisms with a substantial reduction in the gene repertoire, we were not able to identify TFs or wHTH TFs, such as in the *Buchnera*, *Rickettsia* and *Wolbachia* genera,

Table 1. Superfamilies associated with DNA-binding TFs in bacteria

Superfamily	Total TFs (%)	No. of families	Total TFs with PaDos	No. of different PaDos
wHTH DBD	30642 (46.33)	39	17304	79
Homeodomain-like	17110 (25.87)	10	8452	75
Lambda repressor-like DBD	6437 (9.73)	6	2875	49
C-terminal effector domain of the bipartite response regulators	5057 (7.64)	3	3747	39
Putative DBD	2085 (3.15)	2	299	23
IHF-like	1782 (2.69)	1	0	0
Nucleic acid-binding proteins	1294 (1.95)	1	21	3
Glucocorticoid receptor-like (DBD)	473 (0.71)	2	473	2
Ribbon-helix-helix	434 (0.65)	4	64	6
AbrB/MazE/MraZ-like	248 (0.37)	3	38	5
KorB DBD-like	230 (0.34)	1	176	5
TrpR-like	107 (0.16)	1	–	–
ACT-like	68 (0.10)	1	63	1
Flagellar transcriptional activator FlhD	64 (0.096)	1	–	–
Haemolysin expression modulating protein H	35 (0.052)	1	–	–
A DBD in eukaryotic transcription factors	25 (0.037)	1	8	2
DBD	24 (0.036)	1	18	1
p53-like transcription factors	6 (0.009)	1	5	4
DBD of <i>Mtuf</i> -box binding protein MBP1	5 (0.008)	1	–	–

Fig. 1. Distribution of wHTH TFs and total TFs. *Bordetella pertussis* (Bpe), *Burkholderia* sp. 383 (Bur) and ‘*Candidatus* Protochlamydia amoebophila’ (Cpr) are included in this illustration as reference points. On the x axis (log scale), genomes are sorted by size from smallest to largest. On the y axis (log scale) are the corresponding numbers of TFs. A linear regression was calculated using the Pearson correlation (r^2) between the number of genes and the total number of TFs. Each dot represents a bacterial genome; wHTHs (blue) and total TFs (red) are indicated.

suggesting that they exhibit alternative regulatory mechanisms beyond TFs. Although the wHTH contributes significantly to the total TFs, probably following a similar path of duplication events to the rest of the genes in Bacteria (as illustrated in Fig. 1), we identified organisms such as *Bordetella pertussis* and ‘*Candidatus* Protochlamydia amoebophila’ in which the wHTH does not represent the most abundant superfamily. In fact, in these organisms the homeodomain-like superfamily is the most abundant DBD associated with TFs, suggesting alternative means for regulation of gene expression.

Another important question that emerges from the distribution of wHTHs is whether the abundance of these kinds of proteins in all the genomes also reflects a large proportion of families. In this sense, we found that the abundance of wHTH proteins is derived preferentially by duplication events more than by the existence of different families. Therefore, we classified the repertoire of wHTH TFs into 39 different families, and their distribution among all bacteria was evaluated (Supplementary Table S2, available with the online version of this paper). From this analysis, we describe in the next two sections the most evident results.

(i) We found that the abundance of TFs in larger genomes does not necessarily involve diversity in the repertoire of families, but it does suggest an increase in the size of the family, i.e. whereas there is a large proportion of TFs as a consequence of genome size, the number of different families per genome is almost constant, from 1 to 26, with an average of 14 families in bacteria. However, in some cases we identified a large number of families, such as for the bacterium *Saccharopolyspora erythraea*, a filamentous soil microbe used for industrial-scale production of the antibiotic erythromycin (Oliynyk *et al.*, 2007), for which 26 different wHTH families were identified, the maximum number of families per organism. A probable identification of abundant families might suggest diverse duplication events in Bacteria, whereas small families would suggest gene loss, lateral gene transfer or invention *de novo* (Supplementary Table S1).

(ii) Seven families include 80% of the total TFs with wHTHs: ArsR, MarR, LysR, biotin, CAP, GntR and Lrp (Fig. 2). From these families, LysR represents an interesting evolutionary group, because it contains the most abundant family of TFs identified so far as well as a large proportion of proteins with dual activity (repressor and activator proteins). This family is mainly responsible for the regulation of basic and ancient physiological processes, such as amino acid biosynthesis, associated with the last common ancestor of Bacteria, Archaea and Eukarya (Hernández-Montes *et al.*, 2008). In contrast, small families were also identified, such as iron, ArgR and LexA, proposed to be essential under standard growth conditions and in maintaining DNA integrity in *E. coli*. Based on these data, a family’s abundance not only suggests ancient evolutionary events in Bacteria but also reveals a limit in the number of wHTH families in all bacteria. Thus, it seems that there is a limit of expansion for all families in bacteria, independent of the genome size and an increase in the number of duplication events associated with each family, suggesting that the TF repertoire in bacteria is associated mainly with events of duplication, recombination and lateral gene transfer. An interesting observation is that in cell division in *Chlamydiae*, during which massive gene loss events have been identified, the HrcA family of TFs was exclusively identified. HrcA is a small family that contains a large proportion of negative regulators of class I heat-shock genes (*grpE-dnaK-dnaJ* and *groELS* operons) that prevent heat-shock induction of these operons (Fischer *et al.*, 2002). We suggest that proteins of this family may be associated with alternative functions beyond heat-shock induction, playing an important role in the adaptation of parasitic bacteria to their hosts.

A universal pattern of distribution is observed in bacterial wHTH families

In order to evaluate the distribution of the 39 TF families in bacteria, we analysed their distribution across 24 taxonomical divisions (Fig. 2; see also Supplementary Tables S1

Fig. 2. Clustering of the co-occurrence of wHTH TFs in all the bacterial genomes. (a) A hierarchical centroid linkage clustering algorithm was applied with Manhattan metric distance as the similarity measure and complete linkage metric distance (Eisen *et al.*, 1998). In the upper section, the names of the 24 bacterial divisions are indicated. The names of the 39 families are also shown. The clusters are indicated with arrows. The families and their relative abundance levels are indicated.

and S2). The relative abundance of families was calculated by phylum, with a value of 0 representing absence and 1 representing presence. Based on this analysis we identified a cluster of eight families widely distributed in all the bacterial divisions: the LexA, LysR, FUR, ArsR, MarR, CAP, HrcA and Rrf2 families. An interesting finding was that the seven most abundant families described previously were included in this cluster, together with families that are less abundant. All these families regulate a plethora of functions, such as amino acid metabolism (LysR), carbon source assimilation (CAP, Körner *et al.*, 2003; Maddocks & Oyston, 2008), resistance to diverse stresses (MarR, Ellison & Miller, 2006; LexA,

Shimoni *et al.*, 2009; HrcA, Ellison & Miller, 2006), cysteine biosynthesis and benzoate degradation (Rrf2, Even *et al.*, 2006; Peres & Harwood, 2006) and metal assimilation (FUR, Pennella & Giedroc, 2005) and resistance (ArsR, Busenlehner *et al.*, 2003). It is probable that all these families could have been present in the last common ancestor of Bacteria and as a consequence the bacterial cenancestor would have a high capability to contend with diverse, challenging environments and also be a self-supporting system.

In a second cluster, families with a large distribution pattern, except in small organisms, were identified: GntR,

Lrp, biotin and IclR families. These families regulate biotin synthesis, carbon source assimilation and amino acid biosynthesis, among other processes.

A third cluster with a low distribution pattern was identified and included the ArgR, Iron and Rex families. These families regulate genes associated with carbon source assimilation, arginine biosynthesis, iron uptake and responses to changes in the cellular NADH/NAD⁺ redox state, respectively, and include few members per organism.

Finally, a high diversity of families with erratic distribution patterns were included in diverse clusters, such as Rio2, associated with *Bacillales*, Ets, associated with *Actinobacteria*, and MetH, exclusively associated with *Actinobacteria*. In addition, we identified probable families with lateral gene transfer events, including Vacu, which is associated with *Bacillales* and *Actinobacteria*. This finding suggests that diverse lineage-specific TFs are involved in specific and important processes, such as sporulation in bacilli, or in some specific amino acid biosynthesis routes. It is interesting that the absence of TFs for several important amino acid biosynthesis routes in *B. subtilis* and other *Firmicutes* may have been complemented by the invention of novel regulatory mechanisms, such as transcription attenuation (Gollnick *et al.*, 2005; Merino & Yanofsky, 2005; Rodionov *et al.*, 2004). Indeed, a large diversity of regulatory mechanisms beyond TFs was recently described, including antisigma factors, RNAs and protein–protein interactions, among others (Martínez-Núñez *et al.*, 2010).

Multidomain proteins are highly abundant in bacterial TFs

At the present time, the considerable diversity of sequenced bacterial genomes available to the scientific community offers an invaluable source of information for evaluating the abundance and diversity of the repertoire of regulatory proteins controlling gene expression at the level of transcription initiation. These proteins can be analysed to evaluate their influence on bacterial adaptation and responses to environmental stimuli. Therefore, in order to evaluate the contributions of these domains in the total set of proteins identified as TFs, the domain repertoire beyond the DBD was analysed. From this analysis we identified different groups based on domain architectures, i.e. 57 % of the TFs exhibited more than one structural domain (the DBD and PaDos), whereas 43 % of the total repertoire was associated only with the DBD (Table 2 and Supplementary Table S3, available with the online version of this paper). The monodomain proteins can be further subdivided into two categories: the first one includes proteins for which more than 94 % of the protein is occupied by the DBD and the second category includes proteins for which the DBD covers only 50 % of the sequence. These latter proteins may exhibit additional domains not identified using structural data. Therefore, in Fig. 3 we present the distribution of multidomain TFs in all the bacterial genomes. From this illustration, it is evident

that the abundance of multidomain TFs follows a similar distribution to total TFs, i.e. their abundance correlates with the genome size. For instance, organisms with small genomes may contain a lower proportion of multidomain proteins than do larger genomes. This result reinforces the notion that small genomes are associated with stable environments, where a limited number of TFs are necessary to regulate gene expression. In contrast, larger genomes contain a great proportion of multidomain TFs, suggesting that these domains contribute to functional adaptations to environmental changes. Based on this analysis, the diversity and abundance of TF families and their PaDos would contribute significantly to the regulatory diversity.

Therefore, 79 PaDos were identified in the whole collection of wHTH TFs associated with bacteria, and in order to evaluate the distribution of these PaDos in bacteria, 22 divisions were analysed. From this analysis we identified diverse groups. The most representative clusters are shown in Fig. 4 (see also Supplementary Figs S1 and S2, available with the online version of this paper). The first group contains four different PaDos, such as periplasmic-binding protein-like II (PBP II), cAMP-binding domain-like, GAF domain-like and LexA/signal peptidase domains. PBP II and cAMP domains are associated with the LysR and CAP families, whereas GAF is associated with diverse families, such as as IclR, HrcA, Plan, FUR and others. The LexA/signal peptidase domain is associated with the LexA family. It is important to mention that the first three PaDos are highly abundant in all bacteria, representing 70 % of the total PaDos, and they are also related to large TF families. These findings suggest that these PaDos are very successful in all the bacteria and are intimately related to large families of TFs.

The second cluster includes eight different PaDos: dimeric α - and β -barrel (dimeric), PLP-dependent transferases (PLP), NagB/RpiA/CoA transferase-like (NagB), C-terminal domain of transcriptional repressors, C-terminal domain of arginine repressor, class II aaRS and biotin synthetases, iron, and NAD(P)-binding Rossmann fold domains. From these domains, dimeric, PLP and NagB have been identified as highly abundant, representing 15 % of total PaDos. An interesting observation for these domains is that they exhibit a similar distribution pattern to the PaDos included in the first cluster, except they are absent in parasites, symbionts and, in general, in small genomes, suggesting probable gene loss events. The third cluster is integrated by MOP-like, S-adenosyl-L-methionine-dependent methyltransferases and acyl-CoA N-acyltransferases (Nat), which have been mainly identified in the divisions *Proteobacteria* and *Acidobacteria*. In the subsequent clusters, we identified diverse PaDos, including the PRTase-like, ribokinase-like and CBS domains and the putative transcriptional regulator TM1602, C-terminal domain, constrained to *Firmicutes* and *Fusobacteria*. Finally, we found the rhodanese/cell cycle control phosphatase, fatty acid-responsive transcription factor FadR C-terminal, SIS, phosphorytosine, Bet v1-like, and nucleotidyltransferase domains clustered together.

Table 2. Monodomain and multidomain families identified in this work

Family name	Multidomain TFs (%)*	No. of different PaDos*	Most abundant PaDo	Proportion of PaDos (%)*
DsvD	0	0	—	0
Ets	0	0	—	0
H1/H5	0	0	—	0
Meth	0	0	—	0
RTP	0	0	—	0
Vacu	0	0	—	0
Biot	51	24	Class_II	21
ArsR	14	22	PBP II	27
Mj22	6	13	GAF	29
E-II	47	15	cAMP	36
MarR	5	22	Acyl-CoA	37
MotA	42	5	Actin-like_ATPase	40
Iron	83	16	Iron	41
SelB	39	8	GAF	41
Heli	61	9	GAF	43
SCF	30	2	NAD(P)	50
FUR	1	5	PLP	50
Arch	56	4	GAF	56
Lrp	67	28	Dimeric	58
Plan	49	6	GAF	59
P4	100	4	P-loop	62
Rrf2	2	2	PLP	63
Z-DNA	54	9	GAF	64
RPA32	50	3	PLP	67
F93	2	2	Actin-like_ATPase	67
AF2	25	7	NagB	70
ArgR	92	7	Arginine_repressor	75
ModE	70	3	MOP	79
GntR	23	22	PLP	79
HrcA	97	9	GAF	81
IclR	88	20	GAF	83
Peni	5	2	SIS domain	88
CAP	92	15	cAMP	89
LexA	73	9	LexA/Signal	92
PurR	99	4	PRT	94
Rex	99	2	NAD(P)	99
Rio2	100	1	NAD(P)	100
LysR	99	7	PBP II	99
FokI	63	1	NagB	100

*0 represents monodomain families.

These domains have been identified in only low proportions, mainly in *Proteobacteria*.

In summary, we identified six PaDos that represent 84 % of the total domains identified in bacterial TFs (PBP II, GAF, dimeric, PLP, NagB and cAMP domains), many of which are universally distributed in bacteria and intimately associated with specific families. It is also interesting that these PaDos are found in abundant families, reinforcing their probable roles in basic physiological processes, such as PBP II being exclusively associated with the largest family of TFs identified so far, LysR (see Supplementary Fig. S2). This finding suggests that PaDos and wHTH domains

probably coevolved, based on their pattern distributions. Alternatively, we identified some PaDos as probably lineage specific, such as the PTS-reg, TM1602, PTS-sys, PRTase-like, CBS, HPr kN-T and Thio/thiol domains. These latter domains were exclusively identified in *Firmicutes*. We suggest that these PaDos may be the consequence of invention *de novo*, because of their specific distribution in *Firmicutes*. In addition, to reinforce these previous observations, we asked if the PaDos were specifically associated with wHTH TFs. To obtain further insights into the specific and general associations of these domains and wHTHs, we evaluated all DNA-binding structures reported so far and classified them as homeodomain-like, lambda

Fig. 3. Distribution of multidomain TFs in bacterial genomes. On the *x* axis (log scale), genomes are sorted by size from smallest to largest. On the *y* axis (log scale) are the corresponding numbers of TFs. A linear regression was calculated using the Pearson correlation (r^2) between the number of genes and the total number of TFs. Each dot represents a bacterial genome; wHTHs (blue) and multidomain proteins (red) are indicated.

repressor or nucleic acid-binding domains (Table 1) in order to identify the distributions of these domains. Based on this analysis, we determined that almost 40 % of the total PaDos associated with wHTHs are exclusive to this class of structural domains, suggesting that these domains have been preferentially recruited by wHTHs. A similar finding has been identified in other superfamilies, for instance, 30 % of the total PaDos are lambda specific, and 33 % of the total PaDos are homeodomain-like specific (see Supplementary Fig. S3, available with the online version of this paper).

The diversity of PaDos defines three main groups of families

In order to evaluate the diversity of TFs in terms of their structural domains, we defined three main groups of families: (i) monodomain families, those families that exhibit only the DBD; (ii) monolithic families, in which most of the protein members exhibit the DBD and a PaDo, usually in the same domain; and (iii) promiscuous families, those families with a large diversity of domains. We next describe each of these categories in more detail.

(i) Monodomain families. Twelve families were considered monodomain, including MarR, FurR and ArsR (Table 2). An interesting observation is that most of the families included in this category contain proteins of small sizes, around 150 amino acids in length, where the DBD covers most of the sequence.

(ii) Multidomain families. Multidomain families can be further subdivided into two groups, based on the presence of multiple domains: (ii) monolithic, where at least 80 % of their members exhibit a predominant PaDo associated with the DBD [such as the LysR family, in which the PBP II is present in 99 % of its members (Table 2 and Supplementary Figs S1 and S2)]; and (iii) promiscuous families, such as GntR or Lrp, for which diverse domains

are associated with the DBD. Therefore, the diversity in the repertoire of regulatory proteins with wHTHs is influenced by the organization and combination with the PaDos, and the families can be classified into three main groups. There are monodomain families, where the multimerization and ligand-binding sites are included in the DBD, and multidomain families, which can be divided into two groups. In monolithic families, the DBD has undergone a similar evolution process to the PaDos with few recombination events, such as the LysR family. Indeed, the domain organization is also conserved, where the DBD is located in the N terminus whereas the PLP II is located in the C terminus. Both monodomain and monolithic families include the most abundant families identified in the repertoire of TFs. Finally, we identified a large proportion of families that do not represent the most abundant families but that include a large diversity of PaDos.

Conclusions

What are the roles of the diversity and distribution of TFs in the regulatory plasticity of bacteria? The answer to this question is important in order to determine the contribution of these regulatory families in the evolution of these organisms and in the context of their responses to diverse environmental stimuli. Thus, based on the repertoire of DNA-binding TFs associated with the wHTH domains in 668 completely sequenced bacterial genomes, representing adaptative designs for different lifestyles, we have attempted to gain an understanding of the relationships between the DBD and PaDo distribution patterns involved in the modelling of transcriptional regulatory networks.

We have shown that TF families expand or contract in a lineage-specific manner to adapt to the varied environmental needs of the organisms. Similar trends have been observed in previous comparative studies on TF families in plants versus animals and at the level of taxa (Coulson

Fig. 4. Clustering of the co-occurrence of PaDos in bacterial genomes. (a) A hierarchical centroid linkage clustering algorithm was applied with Manhattan metric distance as the similarity measure and complete linkage (Eisen *et al.*, 1998). PaDo abundance levels are indicated. In the upper section, the names of the 24 bacterial divisions are indicated (as in Fig. 2). The names of the 79 PaDos are also shown. (b) The PaDos and their relative abundance levels are presented (see also Supplementary Table S3).

et al., 2001). A more general perspective regarding lineage-specific expansion of protein families and the implications for the diversification of organisms has also been described for eukaryotic species (Lespinet *et al.*, 2002). In this regard, the abundance levels of families and their domain organizations were evaluated in bacteria. From this analysis, we identified that TF families have preferentially suffered diverse expansion events more than invention *de*

novo, i.e. in all organisms evaluated in our study there was a similar distribution of families, but the members were different among the families. Alternatively, PaDo families are present at between 1 and 26 per organism, contributing significantly to the diversity of the regulatory machinery. Finally, three main categories of families were defined on the basis of their domain architecture: monodomain, monolithic and promiscuous. We suggest that the interplay

of all these elements, TF abundance, recombination of DBDs, and PaDos, and also duplication events, allows bacteria to adapt to changing environmental conditions and shows that they are models of regulatory networks.

ACKNOWLEDGEMENTS

This work was supported by the Mexican Science and Technology Research Council (CONACYT) with a doctoral scholarship (164293) awarded to N.R.-G. E.P.-R. was financed by grants from the DGAPA-UNAM (grant nos IN-217508 and IN-209511). We would like to thank Miguel Angel Ramirez, Tomas Villaseñor and Yalbi Martinez for critical reading of the manuscript.

REFERENCES

- Aravind, L. & Koonin, E. V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* 27, 4658–4670.
- Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* 29, 231–262.
- Baumbach, J. (2007). CoryneRegNet 4.0 – a reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics* 8, 429.
- Bengtsson, B. O. (2004). Modelling the evolution of genomes with integrated external and internal functions. *J Theor Biol* 231, 271–278.
- Brennan, R. G. (1993). The winged-helix DNA-binding motif: another helix–turn–helix takeoff. *Cell* 74, 773–776.
- Brennan, R. G. & Matthews, B. W. (1989a). The helix–turn–helix DNA binding motif. *J Biol Chem* 264, 1903–1906.
- Brennan, R. G. & Matthews, B. W. (1989b). Structural basis of DNA-protein recognition. *Trends Biochem Sci* 14, 286–290.
- Brown, J. H., Gupta, V. K., Li, B. L., Milne, B. T., Restrepo, C. & West, G. B. (2002). The fractal nature of nature: power laws, ecological complexity and biodiversity. *Philos Trans R Soc Lond B Biol Sci* 357, 619–626.
- Browning, D. F. & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* 2, 57–65.
- Busenlehner, L. S., Pennella, M. A. & Giedroc, D. P. (2003). The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. *FEMS Microbiol Rev* 27, 131–143.
- Changizi, M. A. (2001). Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. *J Theor Biol* 211, 277–295.
- Coulson, R. M., Enright, A. J. & Ouzounis, C. A. (2001). Transcription-associated protein families are primarily taxon-specific. *Bioinformatics* 17, 95–97.
- Croft, L. J., Lercher, M. J., Gagen, M. J. & Mattick, J. S. (2003). Is prokaryotic complexity limited by accelerated growth in regulatory overhead. *Genome Biol* 5, P2.
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863–14868.
- Ellison, D. W. & Miller, V. L. (2006). Regulation of virulence by members of the MarR/SlyA family. *Curr Opin Microbiol* 9, 153–159.
- Even, S., Burguière, P., Auger, S., Soutourina, O., Danchin, A. & Martin-Verstraete, I. (2006). Global control of cysteine metabolism by CymR in *Bacillus subtilis*. *J Bacteriol* 188, 2184–2197.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R. & other authors (2008). The Pfam protein families database. *Nucleic Acids Res* 36 (Database issue), D281–D288.
- Fischer, B., Rummel, G., Aldridge, P. & Jenal, U. (2002). The FtsH protease is involved in development, stress response and heat shock control in *Caulobacter crescentus*. *Mol Microbiol* 44, 461–478.
- Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-Rascado, L., Martínez-Flores, I. & other authors (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 36 (Database issue), D120–D124.
- Gollnick, P., Babitzke, P., Antson, A. & Yanofsky, C. (2005). Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu Rev Genet* 39, 47–68.
- Gruber, T. M. & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* 57, 441–466.
- Hernández-Montes, G., Díaz-Mejía, J. J., Pérez-Rueda, E. & Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome Biol* 9, R95.
- Ishihama, A. (2000). Functional modulation of *Escherichia coli* RNA polymerase. *Annu Rev Microbiol* 54, 499–518.
- Janga, S. C. & Moreno-Hagelsieb, G. (2004). Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res* 32, 5392–5397.
- Janga, S. C. & Pérez-Rueda, E. (2009). Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families. *Comput Biol Chem* 33, 261–268.
- Körner, H., Sofia, H. J. & Zumft, W. G. (2003). Phylogeny of the bacterial superfamily of Crp–Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev* 27, 559–592.
- Kummerfeld, S. K. & Teichmann, S. A. (2006). DBD: a transcription factor prediction database. *Nucleic Acids Res* 34 (Database issue), D74–D81.
- Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* 12, 1048–1059.
- Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* 424, 147–151.
- Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60, 327–349.
- Lynch, M. & Conery, J. S. (2003). The origins of genome complexity. *Science* 302, 1401–1404.
- Madan Babu, M. & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res* 31, 1234–1244.
- Madan Babu, M., Teichmann, S. A. & Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 358, 614–633.
- Maddocks, S. E. & Oyston, P. C. (2008). Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* 154, 3609–3623.
- Martínez-Antonio, A., Janga, S. C., Salgado, H. & Collado-Vides, J. (2006). Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol* 14, 22–27.

- Martínez-Núñez, M. A., Pérez-Rueda, E., Gutiérrez-Ríos, R. M. & Merino, E. (2010).** New insights into the regulatory networks of paralogous genes in bacteria. *Microbiology* **156**, 14–22.
- Merino, E. & Yanofsky, C. (2005).** Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet* **21**, 260–264.
- Miroslavova, N. S. & Busby, S. J. (2006).** Investigations of the modular structure of bacterial promoters. *Biochem Soc Symp* (73), 1–10.
- Moreno-Campuzano, S., Janga, S. C. & Pérez-Rueda, E. (2006).** Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes – a genomic approach. *BMC Genomics* **7**, 147.
- Oliynyk, M., Samborsky, M., Lester, J. B., Mironenko, T., Scott, N., Dickens, S., Haydock, S. F. & Leadlay, P. F. (2007).** Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. *Nat Biotechnol* **25**, 447–453.
- Pennella, M. A. & Giedroc, D. P. (2005).** Structural determinants of metal selectivity in prokaryotic metal-responsive transcriptional regulators. *Biometals* **18**, 413–428.
- Peres, C. M. & Harwood, C. S. (2006).** BadM is a transcriptional repressor and one of three regulators that control benzoyl coenzyme A reductase gene expression in *Rhodospseudomonas palustris*. *J Bacteriol* **188**, 8662–8665.
- Pérez-Rueda, E. & Collado-Vides, J. (2001).** Common history at the origin of the position–function correlation in transcriptional regulators in archaea and bacteria. *J Mol Evol* **53**, 172–179.
- Pérez-Rueda, E. & Janga, S. C. (2010).** Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol Biol Evol* **27**, 1449–1459.
- Pérez-Rueda, E., Collado-Vides, J. & Segovia, L. (2004).** Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem* **28**, 341–350.
- Pérez-Rueda, E., Janga, S. C. & Martínez-Antonio, A. (2009).** Scaling relationship in the gene content of transcriptional machinery in bacteria. *Mol Biosyst* **5**, 1494–1501.
- Rigali, S., Derouaux, A., Giannotta, F. & Dusart, J. (2002).** Subdivision of the helix–turn–helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J Biol Chem* **277**, 12507–12515.
- Rigali, S., Schlicht, M., Hoskisson, P., Nothhaft, H., Merzbacher, M., Joris, B. & Titgemeyer, F. (2004).** Extending the classification of bacterial transcription factors beyond the helix–turn–helix motif as an alternative approach to discover new *cis/trans* relationships. *Nucleic Acids Res* **32**, 3418–3426.
- Rodionov, D. A., Dubchak, I., Arkin, A., Alm, E. & Gelfand, M. S. (2004).** Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. *Genome Biol* **5**, R90.
- Shimoni, Y., Altuvia, S., Margalit, H. & Biham, O. (2009).** Stochastic analysis of the SOS response in *Escherichia coli*. *PLoS ONE* **4**, e5363.
- Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. (2008).** DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* **36** (Database issue), D93–D96.
- Taraban, M., Zhan, H., Whitten, A. E., Langley, D. B., Matthews, K. S., Swint-Kruse, L. & Trewthella, J. (2008).** Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. *J Mol Biol* **376**, 466–481.
- Tordai, H., Nagy, A., Farkas, K., Bányai, L. & Patthy, L. (2005).** Modules, multidomain proteins and organismic complexity. *FEBS J* **272**, 5064–5078.
- van Nimwegen, E. (2003).** Scaling laws in the functional content of genomes. *Trends Genet* **19**, 479–484.
- Wall, M. E., Hlavacek, W. S. & Savageau, M. A. (2004).** Design of gene circuits: lessons from bacteria. *Nat Rev Genet* **5**, 34–42.
- West, G. B. & Brown, J. H. (2005).** The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *J Exp Biol* **208**, 1575–1592.
- Wösten, M. M. (1998).** Eubacterial sigma-factors. *FEMS Microbiol Rev* **22**, 127–150.

Edited by: D. W. Ussery