DNA Research **12**, 269–280 (2005) doi:10.1093/dnares/dsi016

Genome-Wide Survey of Transcription Factors in Prokaryotes Reveals Many Bacteria-Specific Families Not Found in Archaea

Yoshiaki Minezaki,[1] Keiichi Homma,[1,2] and Ken Nishikawa[1,*]

Laboratory of Gene-Product Informatics, Center for Information Biology–DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka, 411-8540 Japan[1] and Japan Science and Technology Corporation, 1–8, Honcho 4-chome, Kawaguchi City, Saitama, 332-0012, Japan[2]

(Received 2 July 2005; revised 28 September 2005; published online 10 January 2006)

Abstract

Assignment of all transcription factors (TFs) from genome sequence data is not a straightforward task due to the wide variation in TFs among different species. A DNA binding domain (DBD) and a contiguous non-DBD with a characteristic SCOP or Pfam domain combination are observed in most members of TF families. We found that most of the experimentally verified TFs in prokaryotes are detectable by a combination of SCOP or Pfam domains assigned to DBDs and non-DBDs. Based on this finding, we set up rules to detect TFs and classify them into 52 TF families. Application of the rules to 154 entirely sequenced prokaryotic genomes detected >18 000 TFs classified into families, which have been made publicly available from the 'GTOP_TF' database. Despite the rough proportionality of the number of TFs per genome with genome size, species with reduced genomes, i.e. obligatory parasites and symbionts, have only a few if any TFs, reflecting a nearly complete loss. Also the number of TFs is significantly lower in archaea than in bacteria. In addition, all but 1 of the 19 TF families present in archaea is present in bacteria, whereas 33 TF families are found exclusively in bacteria. This observation indicates that a number of new TF families have evolved in bacteria, making the transcription regulatory system more divergent in bacteria than in archaea.

Key words: transcription factor; domain organization; DNA binding; prokaryote; comparative genomics

1. Introduction

The genetic information contained in DNA is transcribed to RNA by a transcription complex including DNA-directed RNA polymerase (RNAP). A bacterial transcription initiation complex comprised of the core RNAP enzyme and a σ factor binds to a promoter and, upon initiation of RNA synthesis, releases the σ factor.[1] The archaeal transcription initiation machinery has a combination of different core RNAP proteins and basal transcription factors (TFs) such as TATA-box binding protein (TBP) and transcription factor B, each of which is homologous to the eukaryotic counterpart.[2] In either case, the transcriptional complex can be easily detected by homology because the essentiality of the transcriptional complex entails high conservation of amino acid sequence in every component. On the other hand, TFs such as repressors, activators and enhancer binding proteins, all of which bind to double-stranded DNA at specific sites to interfere or modulate RNAP function, display an enormous variation.[3,4] Different types of TFs have dissimilar 3D structures, with the only shared characteristic being the ability to bind double-stranded DNA. The aim of this study is 2-fold: one is to develop a method to systematically detect all kinds of TFs encoded by a genome with the highest possible accuracy, and the other is to compare results among all prokaryotic species, particularly focusing on any distinctions between bacteria and archaea. As we are conducting a separate investigation on eukaryotes, utilizing extensively compiled data of eukaryotic TFs in the TRANSFAC database,[5] we limit our attention to prokaryotes in this paper.

Although there have been hitherto no genome-wide surveys for TFs across all prokaryotes, several

Communicated by Osamu Ohara

* To whom correspondence should be addressed. Tel. +81-55-981-6859, Fax. +81-55-981-6889, E-mail: knishika@genes.nig.ac.jp

investigations have been carried out for particular species, taxons or TF families: Perez-Rueda and Collado-Vides[6] conducted keyword and PROSITE searches to assign 314 regulatory DNA binding proteins in *Escherichia coli*, and stored them in the RegulonDB database. Aravind and Koonin[7] examined all the archaeal genomes then available (*Methanocaldococcus jannaschii, Methanothermobacter thermoautotrophicum, Archaeoglobus fulgidus* and *Pyrococcus horikoshii*) and made a long list of HTH (helix–turn–helix) proteins, which include some non-TFs. Kyrpides and Ouzounis[8] investigated the same four archaeal species, found 280 transcription-associated proteins including not only TFs but also basal TFs and RNAPs, and classified them into 58 families. A larger scaled survey by Cases et al.[9] investigated 60 bacterial species and categorized proteins into several functional groups, including a group of transcription-related proteins. Ranea et al.[10] performed proteomics analyses on 56 prokaryotic species and classified proteins into CATH[11] superfamilies and functional categories containing one with transcriptional regulators. Recently, Martinez-Bueno et al.[12] exhaustively identified TFs of the AraC-XylS and TetR families from 123 genomes of archaea and bacteria, and deposited them in the BacTregulators database. All the above-mentioned identifications of TFs are in principle based on sequence alignment of the entire sequence. However, as the DBD of a TF constitutes only a small portion such approach often fails to properly identify a DBD and consequently leads to overidentification. On the other hand, Babu and Teichmann[13] used SCOP[14] domains aligned to DBDs as identifiers, listed 271 TFs of *E. coli* and classified them into 11 TF families including 1 with an RNA-binding domain. As the same SCOP domains are sometimes found in both DBDs and non-DBDs (see Results) this methodology also produces some overidentified cases.

In order to reliably detect TFs we developed a novel method employing a combination of a DBD and a contiguous non-DBD with specific SCOP or Pfam domains as the main identifiers of TFs. Our method detects all experimentally verified prokaryotic TFs and is considered to miss or erroneously assign TFs infrequently. Application of the rules for each kind of TFs (i.e. TF family) to entirely sequenced prokaryotes revealed that bacteria have many kingdom-specific families of TFs, whereas archaea share almost all of their TF families with bacteria.

2. Materials and Methods

2.1. Genome data

The main body of the dataset used in the present study comes from the GTOP database[15] (http://spock.genes.nig.ac.jp/~genome/gtop.html), which has been constructed and maintained in this laboratory. GTOP contains all the open reading frames (ORFs) assigned to each organism by the genome sequencing team, and provides structural and functional information on ORFs analyzed by homology search at the protein level. In GTOP, a PSI-BLAST[16] search was conducted for each query ORF against the public databases of PDB[17] (released on 14 November 2003) and SCOP[14] (version 1.65), together with a BLAST search against Swiss-Prot[18] (version 42.5). The *E*-value threshold was set at 0.001 in both search methods. The hidden Markov model program (HMMER)[19] was also utilized for the search against Pfam[20] (version 11). We used a version of GTOP (released on 19 May 2004) containing the genomic data of 18 archaeal and 136 bacterial species together with a number of eukaryotes and bacteriophages.

2.2. TFs for analyses

We deal with all kinds of TFs bound to the double-stranded DNA at specific sites to regulate RNAP function but not those bound to DNA in a non-specific manner or involved in the initiation complex of RNAP itself. Accordingly, bacterial σ factors or archaeal general TFs (TBP) were excluded from our analyses. Also omitted are DNA binding proteins that function non-specifically, such as bacterial HU and archaeal histones, as they affect the transcription process in general.[21] Moreover factors controlling transcription termination like Rho,[22] which bind to RNA rather than DNA, were kept out of this study. Ambiguous cases were decided individually by consulting the literature; we excluded those that were originally regarded as TFs, but were recently revealed not to be TFs, such as cold-shock proteins[23] and TenA.[24] Following the current classification of TFs, each TF family was defined based on the kind of DBD it contains and was named after a representative gene (protein) belonging to the family (see Table 1).

2.3. Selection of TFs with experimental evidence

We first surveyed all the prokaryotic entries of Swiss-Prot to select TFs annotated with direct experimental evidence, excluding those annotated 'by similarity'. The selected entries were manually checked as to whether the literature cited therein provided genuine experimental evidence, and those failing such confirmation were removed. The final count of the Swiss-Prot entries that satisfy our aforementioned criteria of TFs is 382, of which 135 were from *E. coli* and 48 came from *Bacillus subtilis*. The remaining 199 TFs were taken from various species, including 13 TFs from six archaeal species: *Sulfolobus solfataricus*, *Methanosarcina acetivorans*, *Halobacterium sp. NRC-1*, *M. jannaschii*, *Methanococcus maripaludis* and *Pyrococcus furiosus*. To find more TFs with experimental evidence, we conducted an additional survey of TF candidates in Pfam and SCOP database with experimental evidence, and added their homologs to form a set of TF candidates. Only those candidate TFs that meet the criteria of the previous

Table 1. List of high confidence TF families characterized by Pfam and SCOP domains.

TF family	Total count	Pfam		SCOP		Representative TF
		DBD	Non-DBD	DBD	Non-DBD	
LysR	2530	**PF00126**	PF03466	**a.4.5.8**	c.94.1.1	LysR_ecol
TetR/AcrR	1681	**PF00440**	—	**a.4.1.9**	—	AcrR_ecol
GntR	1394	**PF00392**	PF00155; PF00532; PF07702	**a.4.5.6**	c.67.1.1; c.93.1.1	GntR_tten
AraC	1375	**PF00165**	PF01965; PF02311; PF02805	**a.4.1.8**	b.82.1.9; c.23.16.2; c.55.7.1; g.48.1.1	AraC_ecol
CRO/CI/Xre	1259	**PF01381**	—	**a.35.1.2; a.35.1.3**	—	DicA_ecol
OmpR	1241	**PF00486**	PF00072	**a.4.6.1**	c.23.1.1; c.23.1.3	OmpR_ecol
LuxR/NarL	1071	**PF00196**	PF00072; PF03472	**a.4.6.2**	c.23.1.1; d.110.5.1	NarL_ecol
MarR	948	**PF01047**	—	**a.4.5.28**	—	MarR_ecol
LacI	750	**PF00356**	PF00532	**a.35.1.5**	c.93.1.1	LacI_ecol
ArsR	622	**PF01022**	—	**a.4.5.5; a.4.5.36**	—	ArsR_ecol
Fis	570	**PF02954**	PF00072; PF00158; PF01590; PF06506	**a.105.1.1**	c.23.1.1; c.37.1.20; d.110.2.1	Fis_ecol
MerR	549	**PF00376**	—	**a.6.1.3**	—	MerR_styp
AsnC/Lrp	439	—	—	**a.4.5.28; a.4.5.32**	d.58.4.2	AsnC_ecol
DeoR	355	—	—	**a.4.5.1; a.4.5.24**	c.35.1.2; c.63.1.3	DeoR_ecol
Crp/Fnr	353	**PF00325**	PF00027	**a.4.5.4**	b.82.3.1; b.82.3.2	Crp_ecol
Fur	265	**PF01475**	—	—	—	Fur_ecol
PadR	253	**PF03551**	—	—	—	PadR_bsub
RpiR	227	**PF01418**	PF01380		c.80.1.1; c.80.1.3	RpiR_ecol
Rrf2	218	**PF02082**	—	—	—	IscR_ecol
DnaA	139	—	—	**a.4.12.2**	c.37.1.20	DnaA_ecol
BolA/YrbA	121	**PF01722**	—	**d.52.6.1**	—	BolA_ecol
ROK/NagC/XylR	118	—	PF00480	**a.4.5.1; a.4.5.24; a.4.5.28; a.4.5.32; a.4.5.36**	—	NagC_ecol
LytTR	115	**PF04397**	PF00072		c.23.1.1	LytT_bsub
SorC	113	—	PF04198	**a.4.5.4; a.4.13.2**	—	SorC_sfle
ArgR	98	**PF01316**	PF02863	**a.4.5.3**	d.74.2.1	ArgR_ecol
DtxR	92	**PF01325**	PF02742; PF04023	**a.4.5.24**	a.76.1.1; b.34.1.2	DtxR_cdip
LexA	86	**PF01726**	PF00717	**a.4.5.2**	b.87.1.1	LexA_ecol
TrmB	68	**PF01978**	—	—	—	AF1009_aful
BirA	67	—	PF02237; PF03099	**a.4.5.1**	b.34.1.1; d.104.1.2	BirA_ecol
PenR/BlaI/MecI	59	**PF03965**	—	**a.4.5.28**	—	BlaI_bant
SfsA	57	**PF03749**	—	—	—	SfsA_ecol
Nlp	42	—	—	**a.35.1.2**	—	SfsB_ecol
Archaeal HTH-10	40	**PF04967**	—	—	—	AF0805_aful
CopG/RepA	38	**PF01402**	—	—	—	NikR_ecol
PutA	38	—	PF00171; PF01619	**a.176.1.1**	c.1.23.2; c.82.1.1	PutA_ecol
ModE	31	**PF02573**	PF03459	**a.4.5.8**	b.40.6.1; b.40.6.2	ModE_ecol
PaiB	30	**PF04299**	—	—	—	PaiB_bsub
CtsR	28	**PF05848**	—	—	—	CtsR_bsub
AfsR/DnrI/RedD	27	**PF00486**	PF03704	—	—	EmbR_mtub
CodY	27	**PF06018**	—	—	—	CodY_bsub
TrpR	25	**PF01371**	—	**a.4.12.1**	—	TrpR_ecol
MtlR	24	**PF05068**	—	—	—	MtlR_ecol
ROS/MUCR	22	**PF05443**	—	—	—	Ros_atum
MetJ	21	**PF01340**	—	**a.43.1.2**	—	MetJ_ecol
GutM	17	**PF06923**	—	—	—	GutM_ecol
Crl	16	**PF07417**	—	—	—	Crl_ecol
ComK	14	**PF06338**	—	—	—	ComK_bsub
FlhD	14	**PF05247**	—	**a.145.1.1**	—	FlhD_ecol
RtcR	11	**PF06956**	PF00158	—	—	RtcR_ecol
Spo0A	9	—	PF00072	**a.4.6.3**	c.23.1.1	Spo0A_bsub
DctR	6	—	—	**a.4.5.1**	c.23.1.1	DctR_bsub
NifT/FixU	6	**PF06988**	—	—	—	NifT_anab

DBDs characterized by Pfam/SCOP ID codes are indicated in boldface. Only those ID codes of Pfam and SCOP that were used to identify proteins as TFs are indicated in the table. A representative TF for each TF family is shown as a protein name plus the shortened name of the species. The full species names are given in GTOP_TF.

section and that are not annotated as TFs in Swiss-Prot were retained. We then conducted a literature search for every remaining candidate to identify TFs with direct experimental evidence. With 12 TFs found in this process, we obtained a total of 394 TFs with experimental evidence and designated it as the set of experimentally verified TFs.

3. Results

3.1. Detection of TFs based on selection rules

Examining many TFs, we noticed that most of them can be detected by a combination of a DBD and a contiguous non-DBD with specific SCOP or Pfam domains as the identifier of TFs (Fig. 1a). We note that, besides the directionality of DBDs and non-DBDs in the figure, a DBD may be present at the C-terminal end of a non-DBD.[13] Homology search for the entire sequence of gene/protein may detect three proteins (MglB, LacI and AraR) as homologs having similar functions, because a large part of the sequence can be aligned to each other. In reality, however, MglB is a periplasmic ligand binding protein, whereas both LacI and AraR function as TFs (repressors). We note that LacI and AraR are classified into different TF families (the LacI and GntR families, respectively) based on the difference in the N-terminal DBDs, although they possess a common structural domain (SCOP ID: c.93.1.1) at the C-termini. This example illustrates that despite the small size of a DBD, its presence is crucial for a protein to function as a TF. It should be emphasized that the existence of a SCOP or Pfam domain characteristic of a DBD is necessary, but it is not a sufficient condition for a TF: for example, despite having the same SCOP domain (a.4.6.3) as assigned to the DBD of a TF, SCOP domain (c.55.7.1) is annotated as methylated-DNA-protein-cysteine methyltransferase in Swiss-Prot and is not a TF (Fig. 1b). The combination approach thus reduces the number of non-TFs erroneously detected as TFs in approaches involving the DBD only.

We individually determined the domain organization pattern of each TF family from the set of experimentally verified TFs. We classified TFs into families according to the specific combinations of Pfam and/or SCOP domains corresponding to the DBDs and non-DBDs they contain. Each TF family consists of members of the set of experimentally verified TFs that have both a DBD and a contiguous non-DBD with specific Pfam or SCOP domains. For example, the rule for the Spo0A family (Fig. 1b and Table 1) stipulates that proteins contain at least one of the combinations of a DBD and a non-DBD, a.4.6.3-PF00072 and a.4.6.3-c.23.1.1, and thereby excludes Ogt_bsub, because this protein has neither a PF00072 nor a c.23.1.1 domain juxtaposed to the DBD.

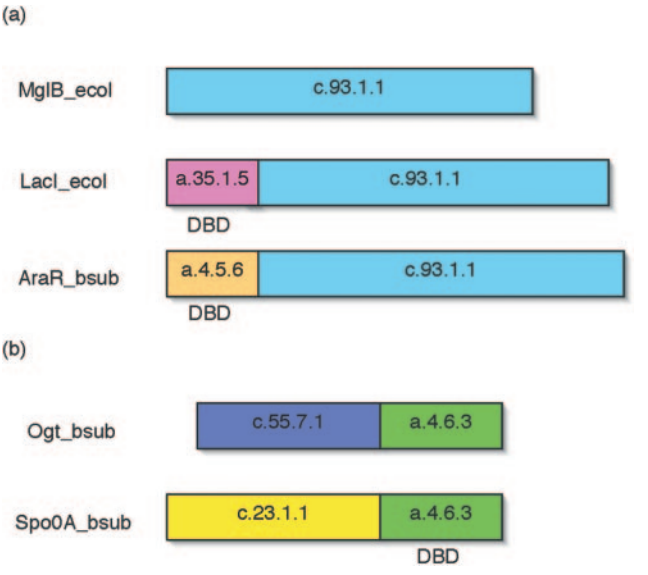

Figure 1. Domain organization of TFs. (**a**) LacI and AraR are TFs with DBDs, while MglB is not a TF because it has no DBD. (**b**) Despite the presence of the same SCOP domain (a.4.6.3) in Ogt and spo0A, the former is not a TF, while the latter is.

In this way, we set up selection rules for individual TF families through analyses of the domain organization of experimentally verified TFs. We determined a rule for each TF family by trial and error; when evident false positives were found, then the selection rule was revised and the process was repeated until there remained no false positives. By Pfam or SCOP combinations, we can detect 86.1% of TFs in the experimentally verified set. In addition, the following cases were included in the selection rules. Many TFs with only DBD assignment (e.g. those in the TetR/AcrR family) represent those whose non-DBDs have yet to be assigned. A small number of TFs, e.g. those in the PaiB family, are short and clearly possess DBDs only, because non-DBDs cannot possibly be present in the remainder of the small protein. Both kinds will be called 'solitary' TF families below, and together constitute 9.8% of the TFs. The rest in the set of experimentally verified TFs can be detected using alignment to existing members of TF families with the requirement that the length of the hit sequences differ from that of the query by <20% (expressed as 'aligned to a TF of comparable length' in the following). The length requirement is intended to ensure alignment for nearly the entire length including the DBD, which tends to be short, and thereby to minimize over-assignment. COG number(s) (from the COG database,[25] 20 June 2004 release) specifically assigned to most of the TF families serve to verify sequence alignments. Table 1 presents a list of 52 TF families, each with Pfam and SCOP identifiers and a representative TF.

Based on the analysis of the experimentally verified TFs, we set up the following criteria for TF selection.

A protein with a DBD and a contiguous non-DBD with a combination of Pfam or SCOP domains specific to each TF family (Table 1) is said to be a high confidence TF. A protein with a DBD of a Pfam or SCOP domain specific for a solitary TF family is considered as a high confidence TF, too, if it is aligned to an experimentally verified TF of comparable length belonging to the same family. Furthermore a protein with a Pfam or SCOP domain characteristic of a TF family but without non-DBD assignment is also classified as a high confidence TF, if it is aligned to a TF family member of comparable length. On the other hand, a protein lacking an identifiable DBD is judged to be a highly probable TF if it is aligned to an experimentally verified TF of comparable length. Finally a protein is regarded as a probable TF if it is detected with a DBD of a Pfam or SCOP domain characteristic of a TF family, but with no homologs in the set of experimentally verified TFs. The high quality set is defined to include only high confidence and highly probable TFs. Our analyses of TFs are based exclusively on high quality TFs (see below).

The aforementioned TF assignment procedure was applied to 154 wholly sequenced prokaryotic genomes in GTOP. The total number of TFs detected in the high quality set was 18 577, 95% of which were detected as high confidence TFs, while the remainder (5%) were classified as highly probable TFs. Table 1 lists the total count of high confidence TFs in each TF family. The largest TF family is LysR, followed by TetR/ArcR and GntR. Though we did not take the order of DBDs and non-DBDs into account, each TF family turned out to have a specific order with absolutely no exceptions. This finding is in agreement with a previous investigation[26] reporting that the order of domains is nearly fixed in prokaryotic proteins. The number distribution of TF families shows that a small number of major TF families and a large number of minor ones exist. This tendency is in good agreement with the reported distribution of *E. coli* TFs[13] that obeys the power law.[27] Detailed data of high confidence and highly probable TFs as well as the numbers of TFs and TF families assigned in each genome are available from the GTOP_TF database (http://spock.genes.nig.ac.jp/~gtop_tf/index2.html) and the Supplementary Table (http://spock.genes.nig.ac.jp/~gtop_tf/tableS1.pdf).

3.2. Comparison with previous studies

In order to check the precision of the TF detection method, it is important to compare the results with those reported in the literature. Such comparisons are, however, not straightforward to make because the definitions of TFs and TF families often differ from one study to another. For instance, Babu and Teichmann[13] employed the SCOP domain as the identifier of TFs in their thorough TF analysis for *E. coli*. However, as their definition of the TF families is based on DBDs specified by the SCOP superfamilies and not the SCOP and Pfam families as in our case, it is difficult to compare the two results. It is nevertheless possible to compare the two results at the individual protein level using one-to-one correspondence. Babu and Teichmann[13] detected 271 TFs from *E. coli*, 233 of which were identical to our high quality set (Fig. 2). Out of the 38 proteins found exclusively in their set, 13 proteins [e.g. CspA (cold-shock protein), Hns (histon-like protein) and RpoE (sigma-E factor)] were attributable to the different definition they use, while the remaining 25 were categorized as 'probable' TFs and therefore not included in the high quality set. Almost all of these probable TFs, e.g. YafY, YaiV and YbaQ, are products of hypothetical genes and have been poorly characterized. In contrast 19 TFs with experimental support, including 11 annotated in Swiss-Prot (e.g. BolA, CaiF, and CdaR), were missed by Babu and Teichmann presumably because the unavailability of the structure of the DBDs in these proteins made it impossible to identify them as TFs by SCOP search alone.

Genome sequencing teams provide detailed lists of individual genes assigned to the genomes and classify them into functional groups. We found 11 such species, including 2 archaea (*M. acetivorans* and *A. fulgidus*), and compared the assignment of TFs with ours at the TF family level (Table 2). The results of genome-wide manual annotations and our TF assignments generally agree.

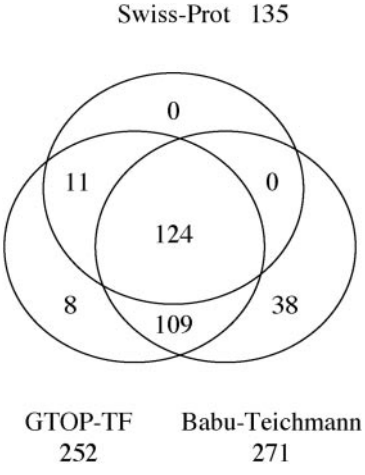

Figure 2. Relationship of *E. coli* TFs annotated in Swiss-Prot, the high quality TFs (present study) and TFs annotated by Babu and Teichmann.[13] The 11 proteins shared by Swiss-Prot and GTOP_TF are BolA, CaiF, CdaR Crl, DcuR, DpiA, GutM, MalY, MtlR, NikR and PutA. The 8 proteins assigned only in the present study are SfsA, YehT, YggD, YjgJ, YjhU, YpdB, YqjI and YrbA. The 38 proteins exclusively assigned by Babu and Teichmann[13] are CspA, FecI, FrvR, Hns, IhfA, IhfB, PinQ, RacR, RpoE, StpA, YafY, YagA, YaiV, YbaQ, YdaW, YdcQ, YddM, YeiI, YfeC, YfeD, YffH, YffS, YfgA, YfjR, YgeH, YheO, YhgG, YhiE, YiiF, YjjM, YmfN, YqeH, YqeI, B0373, B0502, B0540, B1027 and B1146.

Table 2. Comparison of our TF assignments at the TF family level with those in primary annotations.

Species	TF family	Primary annotation	This study	Species	TF family	Primary annotation	This study
M. acetivorans[36]				*Lactobacillus johnsonii*[43]			
	TetR	15	16		GntR	9	9
	MarR	11	13		LacI	7	7
	Lrp	1	3		RpiR	5	5
A. fulgidus[37]					ArsR	3	3
	TetR	1	1		LysR	4	4
	MarR	1	2		AraC	4	4
	Lrp	14	8	*Lactococcus lactis*[44]			
S. coelicolor[38]					LacI	7	5
	LacI	33	35		LysR	9	6
	WhiB	8	13		AraC	3	3
Bifidobacterium longum[39]					GntR	4	5
	LacI	22	21		DeoR	4	3
	LysR	5	5		MarR	11	12
	AraC	1	1	*Fusobacterium nucleatum*[45]			
	WhiB	2	3		TetR	6	6
	MerR	3	4		GntR	6	5
	Fur	1	1		DeoR	5	2
B. subtilis[40]					LuxR/LysR	2	2
	GntR	20	20		MarR	2	2
	LysR	19	19		Crp	2	2
	LacI	12	11		MerR	2	1
	AraC	11	12	*Rhodopseudomonas palustris*[46]			
	Lrp	7	6		AraC	23	23
	DeoR	6	4		DeoR	1	1
Clostridium acetobutylicum[41]					LuxR	11	11
	AcrR/TetR	28	27		LysR	27	27
	MarR/RmrR	22	18		MarR	17	17
	LysR	14	14		ArsR	9	9
	LacI	9	7		AsnC	5	5
Photorhabdus luminescens[42]					Crp	15	15
	LuxR	32	40		GntR	13	13
	LysR	20	29		IclR	7	6
					MerR	3	3
					TetR	39	38
					Fis	13	13
					CopG	2	1

The ROK, TetR and KorSA/GntR families in *S. coelicolor*, the LuxS family in *B. longum* and the Xre family in *C. acetobutylicum* are omitted because the definitions of the TF families in the studies differ from those in our investigation.

Moreover, the agreement is as good in species of archaea as those of bacteria. We therefore claim that the method developed in this research excludes doubtful cases and detects more TFs than those reported in the literature. The results also indicate that our procedure works as well on archaeal species as on bacteria despite the presence of only a small number of archaeal proteins in the set of experimentally verified TFs.

3.3. Diversity of TFs in prokaryotes

For comparison of different species, Fig. 3 plots the number of TFs per genome (the figures are given in GTOP_TF) against the total number of ORFs, which serves as an indicator of genome size in prokaryotes. Excluding archaea and a few other species, the graph shows an initial lag up to ∼1500 ORFs per genome and

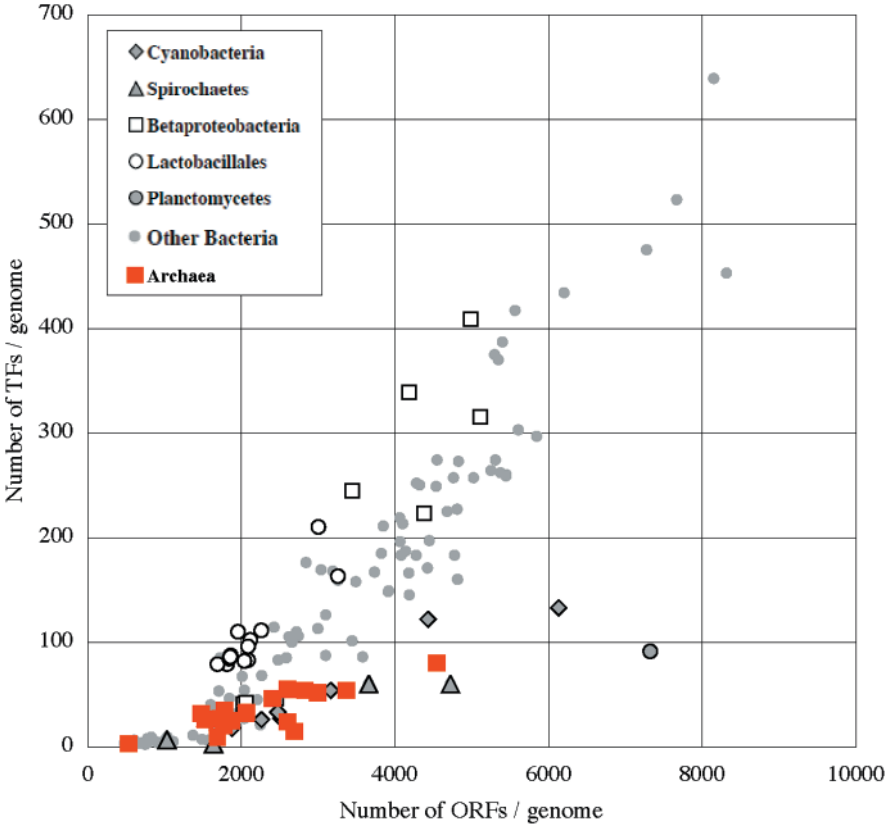

Figure 3. Relationship between the number of TFs and genome size. The total number of high quality TFs per genome is plotted against genome size (an indicator of the total number of ORFs).

then a nearly linear increase. This observation is consistent with the quadratic relationship previously reported.[10] The initial lag section contains those species that have only a few TFs with a small genome size, i.e. 1500 ORFs or less. These species are all parasitic and symbiotic organisms: seven Chlamydiae; seven Mollicutes; two *Rickettsias* species in Alphaproteobacteria; three *Buchneras* species, *Wigglesworthia brevipalpis* and *Candidatus Blochmannia floridanus* in Gammaproteobacteria; two *Tropherymas* species in Actinobacteria and *Nanoarchaeum equitans* in Nanoarchaea. The number of TFs in these organisms ranges from 2 to 11 (see GTOP_TF). In sharp contrast to its complete absence in archaea, DnaA is found in all bacterial species presumably because this protein plays the dual role of an essential factor involved in DNA replication and of a transcription regulator,[28] although the latter function is unlikely to be essential. Besides DnaA, the HrcA repressor (a negative regulator of heat shock genes) is universally present in Chlamydiae and Mollicutes species, but is completely absent in *Buchnera* species. We think it likely that these parasitic and symbiotic organisms have shed most TFs as they retained only the minimum sets of genes for their dependent life style.

The nearly linear section (open symbols and small gray dots in Fig. 3) consists of the majority of bacteria. The positive intersection with the x-axis corresponding to the lag shows that TFs are needed in proportion to the genome size above a certain number of ORFs (∼1500). The TFs may be considered as factors that regulate complex cell functions beyond the minimum level. Though Betaproteobacteria (8 species) and Lactobacillales (13 species) belong to this group, they are shifted upward (open symbols) from the rest of the group. Four species with large genomes including two Actinobacteria (*Streptomyces avermitilis* and *Streptomyces coelicolor*) and two Alphaproteobacteria (*Bradyrhizobium japonicum* and *Mesorhizobium loti*) have 453–639 TFs (see also GTOP_TF), although the number of TF families they have is limited to 33–35, which is less than those of many Gammaproteobacteria including *E. coli*. Figure 4 presents this point more clearly; the number of TF families levels off with large numbers of TFs. In other words, in species with >300 TFs the number of TF families remains nearly constant (35–40), suggesting the divergence of similar TFs by gene duplication. It is also noticed that Gammaproteobacteria and Bacillales (open symbols in Fig. 4) have relatively more TF families than others.

On the other hand, the group lying close to the abscissa (filled symbols in Fig. 3) has no clear lag phase, but instead shows a roughly linear dependence over the entire range with a slope less steep than that in the nearly linear section described above. This group consists of Cyanobacteria (8 species), Spirochaetes (4),

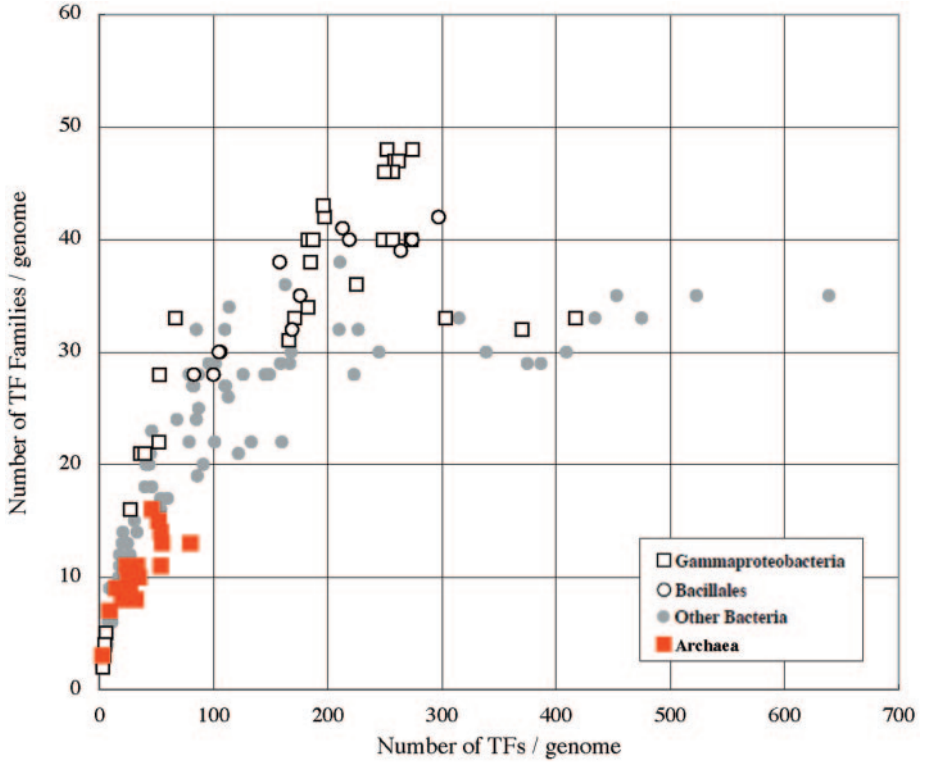

Figure 4. Relationship between the number of TF families and the number of TFs. The number of TF families of each species is plotted against the total number of high quality TFs in the genome.

Planctomycetes (1) and all 18 archaeal species. Species of this group have characteristically small numbers of TFs. The most remarkable among them is the *Pirellula sp.*, the only entirely sequenced species in Planctomycetes,[29] as it has only 91 TFs despite the large genome size (7325 ORFs). The number should be compared with 500 TFs or more expected in bacteria of similar genome sizes. A key to account for this discrepancy lies in the fact that *Pirellula sp.* has as many as 50 σ factors besides TFs.[29] Scarcity of TFs is also notable in Cyanobacteria, especially in the *Anabaena sp.* (*Nostoc sp.*), which has only 133 TFs in 6132 ORFs. The scantiness of TFs in Cyanobacteria can be explained by the fact that Cyanobacteria have a highly developed two-component signal transduction system.[30] Throughout all the archaeal species examined, the numbers of TFs and TF families are small (Figs 3 and 4). We consider this to be the most significant finding in the present study and discuss it in the following section.

4. Discussion

In contrast to previously publicized methods, the newly developed method uses a combination of Pfam or SCOP domains assigned to the DBDs and the non-DBDs of proteins to select TFs. Compared with the single use of the DBDs, this combination approach drastically reduces over-assigned cases. To exclude over-assigned cases in solitary TF families, we introduced an additional requirement, namely alignment to an experimentally verified TF of comparable length. We anticipate a very low number of mistakenly detected cases in the set of TFs selected by our method.

The sensitivity and therefore the extent of under-assignment of the TF detection method developed in the present study depends mainly on the sensitivity of the homology-search tools utilized, i.e. BLAST, PSI-BLAST and HMMER. In the GTOP database, the average residue-wise fraction of proteins aligned to known a 3D structure (PDB) by PSI-BLAST in the genomes of prokaryotic species is 45.5%, but this fraction is increasing with time as the number of PDB entries increases. The corresponding residue-wise fraction aligned to Pfam is even higher (52.0% if HMMER is used for alignment) presumably reflecting the fact that the Pfam database is constructed independently of the known protein structures. Although the fraction of known domains is considerable and increasing, a significant number of protein domains in each genome remain unknown. As we cover all the experimentally verified TFs in prokaryotes by our detection rules, cases missed by our method arise only from the failure to detect TFs caused by faulty performance of alignment programs or the lack of known homologs. The failure of all the alignment programs to detect homology with known proteins is considered

to be infrequent, thanks to the high sensitivity of PSI-BLAST and HMMER. We expect the alignment programs to work almost equally reliably on proteins in archaea as those in bacteria despite the phylogenic remoteness and generally sparser protein annotation in archaea than in bacteria, particularly in *E. coli* and *B. subtilis*. In fact, the average residue-wise coverage of proteins by PDB with the use of PSI-BLAST is 42.3% in archaea and is comparable to the corresponding figure, 46.1%, in bacteria. Consequently the only possible major source of under-assignment is the failure to detect DBDs because they belong to unknown families. However, as bacterial TFs are well studied it is unlikely to find many new DBD families in bacteria. Furthermore, considering the fact that more than half of proteins in archaea have SCOP and Pfam domain assignments and that almost no archaea-specific TF families have been found in structurally aligned proteins, we think it improbable to discover many archaea-specific TF families in the future with more PDB data. Comparison of TFs detected by our method with previously reported TFs (Fig. 2 and Table 2) supports the conclusion of theoretical considerations.

As shown in Figs 3 and 4, the numbers of TFs and TF families are lower in archaea than in the majority of bacteria. For example, *M. acetivorans*, whose genome is larger than that of *E. coli*, has only 80 high quality TFs classified into 13 TF families (see GTOP_TF). To compare bacteria and archaea more systematically, the phylogenetic pattern for each TF family across major taxa of prokaryotes is presented in Fig. 5, where only those taxa having four or more species of known genome are included. The following observations remain valid even if taxa with less than four species are included (Supplemental Table S1 and GTOP_TF). Surprisingly there was only one minor TF family specific to archaea found in this study, Archaeal HTH-10 (the third row, Fig. 5). We note that there are no TF families commonly shared by archaea and eukaryotes besides those shown in Fig. 5; although all protein sequences of archaea were searched in the GTOP database against all kinds of Pfam domains including eukaryotic DBDs (e.g. zinc-finger, homeobox, and leucine-zipper), no eukaryotic DBDs were detected in archaea. As depicted in Fig. 5, archaea and bacteria share 18 TF families including 10 out of a total of 20 major ones (rows 4–21, in which the major TF families are yellow-tinted), including the nearly ubiquitous LysR, TetR/AcrR and GntR families. Since the same domain organization is kept throughout the prokaryotes, as shown with the AsnC/Lrp family,[31,32] these TF families must have descended from the common ancestor of bacteria and archaea.[8] Notably there are no essential differences between Crenarchaeota and Euryarchaeota (columns 2 and 3), although some TF families (e.g. LysR, MerR and BirA) are absent in Crenarchaeota. The remaining 10 major as well as 23 minor families in the list are unique to bacteria and are also absent in eukaryotes according to our preliminary analyses. Thus, they must have evolved in bacteria after branching off from archaea. Interestingly, well-known TFs such as OmpR, AraC, LacI and Fis fall in this group. At the same time, we notice several taxon-specific TF families: ROS/MUCR (specific to Alphaproteobacteria), AfsR/DnrI/RedD (Actinobacteria) and ComK (Bacillales) as well as MtlR, MetJ and Crl (Gammaproteobacteria), as previously reported.[33] It should be noted, however, that these taxon/species-specific TF families are all minor families, and are more or less biased to the present knowledge of TFs with experimental evidence. It is relevant to note that Gammaproteobacteria tend to have more TF families detected than other bacteria with comparable numbers of TFs per genome (Fig. 4), which generally reflect the genome size (Fig. 3). We consider it probable that this difference reflects the better-studied nature of Gammaproteobacteria, especially *E. coli*, than other bacteria. Minor TF families unique to individual taxa or species will increase in the future as experimental evidence accumulates, especially in bacteria other than Gammaproteobacteria. However, the distribution of major TF families is unlikely to be altered, since only minor TF families in bacteria and archaea are possibly not covered by the present selection rules as reasoned in the preceding paragraph. Therefore we consider it probable that approximately half of the major TF families exist exclusively in bacteria, while the rest are shared by the two kingdoms.

Figure 3 indicates that Cyanobacteria and Spirochaetes belong to the same group as archaea. Figure 5 shows that these bacteria (Cy and Sp) indeed have a limited number of TF families, but have a different repertoire (e.g. OmpR) from that of archaea. The life of obligatory or semi-obligatory parasites, such as Chlamydiae, Mycoplasma and *Buchnera*, depends heavily on host cells for nutrition, resulting in deletion of many genes from the genomes.[34] Hence, it is natural to think that many kinds of TFs have also been deleted in these species. A recent study[35] revealed that *Buchnera* has lost all genes but one (metR) that regulate transcriptions of genes involved in syntheses of various amino acids. The idea is corroborated by the TF family distribution of Mollicutes, which include Mycoplasma and Chlamydiae (Mo* and Ch* in Fig. 5). The fact that they have only a few TFs implies that expression of all of their genes is mostly under no specific regulation just as in the case of housekeeping genes. On the other hand, autotrophic species in Gammaproteobacteria, Bacillales and other phyla tend to have many kinds of TFs to control cellular processes in response to environmental changes. Assuming that the variety in the TF family a given species possesses is indicative of the complexity of the species, Gammaproteobacteria are the most diversified form of prokaryotes

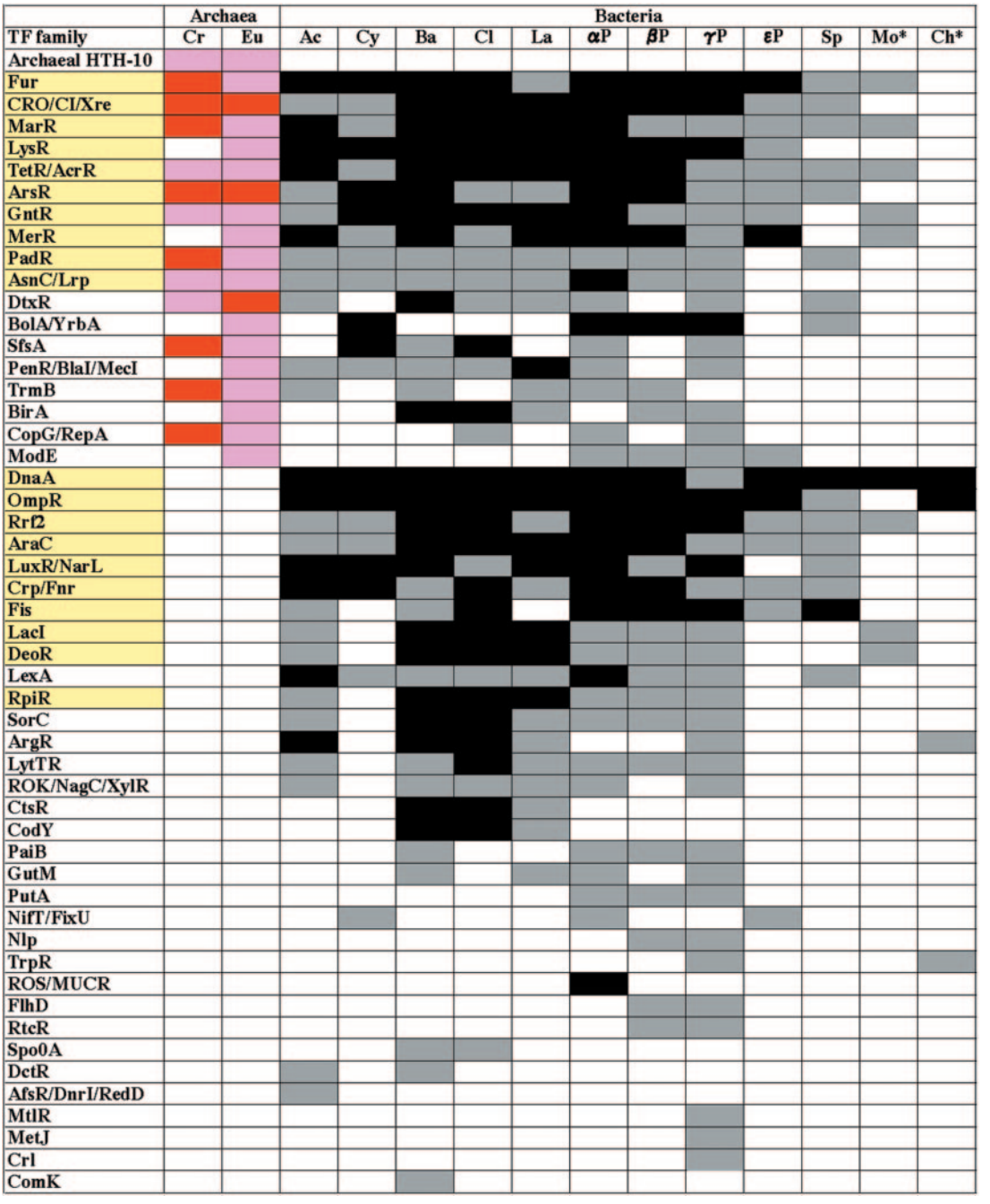

Figure 5. Phylogenetic pattern of high confidence TF families. Presence or absence of each TF family in major phyla is shown. The abbreviation of each phylum with the number of species after removing those of obligatory and semi-obligatory parasites in brackets is Crenarchaeota, Cr (4); Euryarchaeota, Eu (13); Actinobacteria, Ac (10); Cyanobacteria, Cy (8); Bacillales, Ba (12); Clostridia, Cl (4); Lactobacillales, La (13); Alphaproteobacteria, αP (9); Betaproteobacteria, βP (8); Gammaproteobacteria, γP (29); Epsilonproteobacteria, εP (5): Spirochaetes, Sp (4); Mollicutes, Mo* (7); and Chlamydiae, Ch* (7). Two major phyla of archaea, 12 major phyla of bacteria as well as two bacterial phyla consisting of obligatory and semi-obligatory parasites are placed from left to right. The TF family unique to archaea is placed at the top, followed by those shared by archaea and bacteria, and lastly by those unique to bacteria. The cells of major TF families, i.e. the top 20 families in Table 1, are colored yellow. A box colored black or red shows that all the species of the phylum excluding those of obligatory and semi-obligatory parasites have TFs of the corresponding family. A box is tinted grey or pink if at least one species, but not as many species as to be painted black or red, belonging to the phylum has TFs of the family. In archaea, red and pink boxes are used, while in bacteria, black and grey boxes are utilized. A blank box signifies the complete absence of the TF family in the phylum.

(Figs 4 and 5), although some of the minor TF families must be biased to good annotation of *E. coli* as described above. Thus, by the same measure, we can say that archaea are less diverse than bacteria[8] (Fig. 5), since it is unlikely that our method failed to detect any archaea-specific major TFs families, as argued above.

Finally, we should point out another fact revealed in the present study. That is, all but one DBDs of known structure fall in the all-α type or class a in the SCOP classification (Table 1). The only exception to this rule is the DBD of the BolA/YrbA family, which belongs to the $\alpha + \beta$ type (class d) in SCOP. Remarkably TFs having DBDs of the all-α type, particularly HTH proteins,[7] are predominant throughout prokaryotes. This implies two things: one is the phylogenetic continuity of bacteria and archaea, and the other is a presumable dichotomy between prokaryotes and eukaryotes because TFs in eukaryotes are known to have various kinds of DBDs including those of the all-β type (class b) as well as an abundance of zinc-finger types,[4,33] in addition to the all-α type such as the homeobox domain. Comparison of TFs between prokaryotes and eukaryotes will be the subject of our next report.

Acknowledgements: The authors thank Dr Nobuyuki Fujita for the valuable advice. This work was supported by a grant-in-aid for the National Project of Protein 3000 from the MEXT, Japan.

References

1. Ishihama, A. 2000, Functional modulation of *Escherichia coli* RNA polymerase, *Annu. Rev. Microbiol.*, **54**, 499–518.
2. Ouhammouch, M. 2004, Transcriptional regulation in archaea, *Curr. Opin. Genet. Dev.*, **14**, 133–138.
3. Riechmann, J. L., Heard, J., Martin, G., et al. 2000, *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes, *Science*, **290**, 2105–2110.
4. Coulson, R. M. and Ouzounis, C. A. 2003, The phylogenetic diversity of eukaryotic transcription, *Nucleic Acids Res.*, **31**, 653–660.
5. Matys, V., Fricke, E., Geffers, R., et al. 2003, TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, **31**, 374–378.
6. Perez-Rueda, E. and Collado-Vides, J. 2000, The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12, *Nucleic Acids Res.*, **28**, 1838–1847.
7. Aravind, L. and Koonin, E. V. 1999, DNA-binding proteins and evolution of transcription regulation in the archaea, *Nucleic Acids Res.*, **27**, 4658–4670.
8. Kyrpides, N. C. and Ouzounis, C. A. 1999, Transcription in archaea, *Proc. Natl Acad. Sci. USA*, **96**, 8545–8550.
9. Cases, I., de Lorenzo, V., and Ouzounis, C. A. 2003, Transcription regulation and environmental adaptation in bacteria, *Trends Microbiol.*, **11**, 248–253.
10. Ranea, J. A., Buchan, D. W., Thornton, J. M., and Orengo, C. A. 2004, Evolution of protein superfamilies and bacterial genome size, *J. Mol. Biol.*, **336**, 871–887.
11. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. 1997, CATH—a hierarchic classification of protein domain structures, *Structure*, **5**, 1093–1108.
12. Martinez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L., and Tobes, R. 2004, BacTregulators: a database of transcriptional regulators in bacteria and archaea, *Bioinformatics*, **20**, 2787–2791.
13. Babu, M. M. and Teichmann, S. A. 2003, Evolution of transcription factors and the gene regulatory network in *Escherichia coli*, *Nucleic Acids Res.*, **31**, 1234–1244.
14. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. 1995, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, **247**, 536–540.
15. Kawabata, T., Fukuchi, S., Homma, K., et al. 2002, GTOP: a database of protein structures predicted from genome sequences, *Nucleic Acids Res.*, **30**, 294–298.
16. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
17. Berman, H. M., Westbrook, J., Feng, Z., et al. 2000, The Protein Data Bank, *Nucleic Acids Res.*, **28**, 235–242.
18. Bairoch, A., Apweiler, R., Wu, C. H., et al. 2005, The Universal Protein Resource (UniProt), *Nucleic Acids Res.*, **33**, D154–D159.
19. Eddy, S. R. 1998, Profile hidden Markov models, *Bioinformatics*, **14**, 755–763.
20. Bateman, A., Coin, L., Durbin, R., et al. 2004, The Pfam protein families database, *Nucleic Acids Res.*, **32**, D138–D141.
21. Reeve, J. N. 2003, Archaeal chromatin and transcription, *Mol. Microbiol.*, **48**, 587–598.
22. Kaplan, D. L. and O'Donnell, M. 2003, Rho factor: transcription termination in four steps, *Curr. Biol.*, **13**, R714–R716.
23. Phadtare, S. 2004, Recent developments in bacterial cold-shock response, *Curr. Issues Mol. Biol.*, **6**, 125–36.
24. Itou, H., Yao, M., Watanabe, N., and Tanaka, I. 2004, Structure analysis of PH1161 protein, a transcriptional activator TenA homologue from the hyperthermophilic archaeon *Pyrococcus horikoshii*, *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 1094–1100.
25. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.
26. Vogel, C., Berzuini, C., Bashton, M., et al. 2004, Supra-domains: evolutionary units larger than single protein domains, *J. Mol. Biol.*, **336**, 809–823.
27. Qian, J., Luscombe, N. M., and Gerstein, M. 2001, Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model, *J. Mol. Biol.*, **313**, 673–681.
28. Messer, W. and Weigel, C. 1997, DnaA initiator—also a transcription factor, *Mol. Microbiol.*, **24**, 1–6.
29. Glockner, F. O., Kube, M., Bauer, M., et al. 2003, Complete genome sequence of the marine planctomycete *Pirellula sp.* strain 1, *Proc. Natl Acad. Sci. USA*, **100**, 8298–8303.
30. Ohmori, M., Ikeuchi, M., Sato, N., et al. 2001, Characterization of genes encoding multi-domain proteins in the

genome of the filamentous nitrogen-fixing cyanobacterium *Anabaena sp.* strain PCC 7120, *DNA Res.*, **8**, 271–284.
31. Brinkman, A. B., Ettema, T. J., de Vos, W. M., and van der Oost, J. 2003, The Lrp family of transcriptional regulators, *Mol. Microbiol.*, **48**, 287–294.
32. Koike, H., Ishijima, S. A., Clowney, L., and Suzuki, M. 2004, The archaeal feast/famine regulatory protein: potential roles of its assembly forms for regulating transcription, *Proc. Natl Acad. Sci. USA*, **101**, 2840–2845.
33. Coulson, R. M. R., Enright, A. J., and Ouzounis, C. A. 2001, Transcription-associated protein families are primarily taxon-specific, *Bioinformatics*, **17**, 95–97.
34. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. 2000, Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp.* APS, *Nature*, **407**, 81–86.
35. Moran, N. A., Dunbar, H. E., and Wilcox, J. L. 2005, Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*, *J. Bacteriol.*, **187**, 4229–4237.
36. Galagan, J. E., Nusbaum, C., Roy, A., et al. 2002, The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity, *Genome Res.*, **12**, 532–542.
37. Klenk, H. P., Clayton, R. A., Tomb, J. F., et al. 1997, The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, *Nature*, **390**, 364–370.
38. Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., et al. 2002, Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2), *Nature*, **417**, 141–147.
39. Schell, M. A., Karmirantzou, M., Snel, B., et al. 2002, The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract, *Proc. Natl Acad. Sci. USA*, **99**, 14422–14427.
40. Kunst, F., Ogasawara, N., Moszer, I., et al. 1997, The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*, *Nature*, **390**, 249–256.
41. Nolling, J., Breton, G., Omelchenko, M. V., et al. 2001, Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*, *J. Bacteriol.*, **183**, 4823–4838.
42. Duchaud, E., Rusniok, C., Frangeul, L., et al. 2003, The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*, *Nat. Biotechnol.*, **21**, 1307–1313.
43. Pridmore, R. D., Berger, B., Desiere, F., et al. 2004, The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533, *Proc. Natl Acad. Sci. USA*, **101**, 2512–2517.
44. Bolotin, A., Wincker, P., Mauger, S., et al. 2001, The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403, *Genome Res.*, **11**, 731–753.
45. Kapatral, V., Anderson, I., Ivanova, N., et al. 2002, Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586, *J. Bacteriol.*, **184**, 2005–2018.
46. Larimer, F. W., Chain, P., Hauser, L., et al. 2004, Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*, *Nat. Biotechnol.*, **22**, 55–61.