4658–4670 *Nucleic Acids Research, 1999, Vol. 27, No. 23*

DNA-binding proteins and evolution of transcription regulation in the archaea

L. Aravind[1,2] and Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA and [2]Department of Biology, Texas A&M University, College Station, TX 77843, USA

Received May 3, 1999; Revised and Accepted September 20, 1999

ABSTRACT

Likely DNA-binding domains in archaeal proteins were analyzed using sequence profile methods and available structural information. It is shown that all archaea encode a large number of proteins containing the helix–turn–helix (HTH) DNA-binding domains whose sequences are much more similar to bacterial HTH domains than to eukaryotic ones, such as the PAIRED, POU and homeodomains. The predominant class of HTH domains in archaea is the winged-HTH domain. The number and diversity of HTH domains in archaea is comparable to that seen in bacteria. The HTH domain in archaea combines with a variety of other domains that include replication system components, such as MCM proteins, translation system components, such as the α-subunit of phenyl-alanyl-tRNA synthetase, and several metabolic enzymes. The majority of the archaeal HTH-containing proteins are predicted to be gene/operon-specific transcriptional regulators. This apparent bacterial-type mode of transcription regulation is in sharp contrast to the eukaryote-like layout of the core transcription machinery in the archaea. In addition to the predicted bacterial-type transcriptional regulators, the HTH domain is conserved in archaeal and eukaryotic core transcription factors, such as TFIIB, TFIIE-α and MBF1. MBF1 is the only highly conserved, classical HTH domain that is vertically inherited in all archaea and eukaryotes. In contrast, while eukaryotic TFIIB and TFIIE-α possess forms of the HTH domain that are divergent in sequence, their archaeal counterparts contain typical HTH domains. It is shown that, besides the HTH domain, archaea encode unexpectedly large numbers of two other predicted DNA-binding domains, namely the Arc/MetJ domain and the Zn-ribbon. The core transcription regulators in archaea and eukaryotes (TFIIB/TFB, TFIIE-α and MBF1) and in bacteria (the σ factors) share no similarity beyond the presence of distinct HTH domains. Thus HTH domains might have been independently recruited for a role in transcription regulation in the bacterial and archaeal/eukaryotic lineages. During subsequent evolution, the similarity between archaeal and bacterial gene/operon transcriptional regulators might have been established and maintained through multiple horizontal gene transfer events.

INTRODUCTION

The study of bacterial and phage operons gave rise to some of the basic tenets of our understanding of genetic control which include the use of DNA-binding proteins to positively or negatively regulate transcription initiation by the RNA polymerase. Since the determination of the crystal structures of the c1 and Cro repressor proteins of bacteriophage λ (1,2), the helix–turn–helix (HTH) domain has become one of the paradigms of DNA–protein interactions. Analysis of the structures of several HTH proteins showed that they contact DNA by means of insertion of a third, distal helix of a right-handed three-helix bundle into the major groove of double-stranded DNA (3,4). Subsequently, a combination of X-ray crystallography and protein sequence analysis of bacterial DNA-binding regulatory proteins has revealed that the HTH domain is a common theme present in most of them (5–8). Studies on developmental and differentiation regulators in eukaryotes have led to the identification of a number of transcription factors that possess conserved domains, such as the homeodomain, the Paired domain, the POU domain and the Myb/SANT domains. Sequence–structure studies indicate that although these proteins show very little or no detectable sequence similarity to bacterial transcription regulators, they contain the HTH structure in the core of their DNA-binding domains (9–13). This realization has led to the idea that HTH is an ancient DNA-binding module that has been utilized for a variety of transcription regulation processes in the course of evolution.

Subsequent studies on several DNA-binding proteins, such as the bacterial biotin operon regulator BirA (14) and eukaryotic forkhead and H1/H5 (15), have led to the identification of a modified derivative of the HTH domain known as the winged-helix domain (wHTH) (16). These proteins, in addition to the core three-helix bundle, contain a C-terminal β-hairpin called the wing. In some versions of this domain, the loop between helix 1 (H-1) and helix 2 (H-2) is extended to give rise to β-strand elements that interact with the two strands of the wing. Further sequence–structure analyses have shown that a wide range of eukaryotic DNA-binding proteins, such as HNF3, H1

*To whom correspondence should be addressed. Tel: +1 301 435 5913; Fax: +1 301 480 9241; Email: koonin@golem.nlm.nih.gov

Table 1. Families of archaeal HTH proteins

HTH family	Phyletic distribution[a]	Archaeal genes	Comments
HTH proteins conserved in all 4 archaeal species			
MBF-1	**Mj,Mta,Af,Ph** + Eukaryotes	MJ0586, MTH729, AF1977, PH0783	In addition to the cHTH domain, these archaeal proteins contain an N-terminal Zn-ribbon. Possible basal transcription factor.
TFIIB	**Mj,Mta,Af,Ph**(2) + Eukaryotes	MJ0782, MTH885, AF1299, PH1482, PH0864	A basal transcription factor containing 2 HTH domains which form the core of the cyclin fold; an N-terminal Zn-ribbon
TFIIE	**Mj,Mta,Af,Ph** + Eukaryotes	MJ0777, MTH1669, AF0751, PH0619	A basal transcripition factor with a C-terminal Zinc ribbon domain. The HTH is degenerate in eukaryotes
DtxR-like	**Mj,Mta(2),Af(3),Ph** + Bacteria	MJ0568, MTH214, MTH936, MTH1362,AF1785, AF1984, AF2395, PH1163	Related to the bacterial transcription regulators of the DtxR family; an iron-binding module C-terminal of the HTH
MJ1164	**Mj,Mta,Af,Ph**	MJ1164, MTH967, AF1787, PH1808	Archaea-specific family; only distant relationships with bacterial HTH domains
MJ1243	**Mj,Mta,Af,Ph**	MJ1243, MTH1178, AF0666, PH0803	Archaea-specific family; only distant relationship with bacterial HTH domains. Contain an N-terminal, AT-hook-like basic patch; the Mt protein contains a C-terminally fused Zn-ribbon.
MCM + HTH	**Mj(4),Mta,Af,Ph** (Eukaryotes)	MJ0363, MJ0961, MJ1489, MJECL13, MTH1770, AF0517, PH0606	The archaeal but not eukaryotic MCM proteins contain a HTH domain C-terminal of the AAA+ ATPase domain
Phe-RS (α-subunit)	**Mj,Mta,Af,Ph** + Eukaryotes	MJ0487, MTH742, AF1955, PH0658	A HTH with a possible RNA binding function fused to the N-terminus of phenylalanyl tRNA synthetase α-subunit. MJ0487 contains a highly-derived version of the HTH domain. Mj, Mt and Af encoded a second version of the α-subunit which lacks the HTH domain.
Conserved in 3 of the 4 archaeal species			
Tfx	**Mj,Mta,Ph**	MJ0173, MTH916, PH0763	Regulator of the formylmethylfuran dehydrogenase operon.
MJ0272	**Mj,Mta,Af**(2),Hsp + bacteria, bacteriophages	MJ0272,MTH1328,AF1793, AF1627	A typical cHTH with strong similarity to bacterial and phage repressors
MJ1545	**Mj,Ph,Af**	MJ1545, AF2414, PH1471	A cHTH distantly related both to bacteriophage repressors and to MBF1

and HSF, as well as bacterial transcription regulators, such as ompR, BirA and CAP, assume the wHTH structure (13,17,18).

In addition to numerous DNA-binding proteins, HTH or wHTH structures have been detected in RNA-binding proteins, such as ribosomal protein L11, RNA editing enzymes and some aminoacyl-tRNA synthetases, which is consistent with an ancient, general nucleic acid-binding function of these domains (19–21).

While the functions of the HTH domain in the crown-group eukaryotes and in bacteria are well understood, there is very little information about gene-specific transcriptional regulators in the third domain of life, the archaea. Detailed analyses of the archaeal genomes have revealed two categories of archaeal genes, namely those that are most similar to eukaryotic homologs and those that most closely resemble bacterial homologs (22,23). Some uncertainty due to possible large variations in evolutionary rates notwithstanding, this split most likely reflects a mixed evolutionary heritage of the archaea resulting from multiple gene exchanges with the bacteria followed by gene displacements. The 'eukaryotic' genes generally encode the stable core of proteins involved in translation, replication, repair and transcription, whereas most of the 'bacterial' genes encode the evolutionarily mobile 'shell' components, such as metabolic enzymes, structural proteins and signal transduction molecules (22,24,25).

Analysis of the readily identifiable components of the archaeal transcription machinery revealed a clear eukaryotic affinity. The components shared with eukaryotes, to the exclusion of bacteria, include the unique RNA polymerase subunits, the TATA-binding protein, TFIIS, TFIIB (designated TFB in archaea; 26) and TFIIE-α (Table 1; 27). Furthermore, phylogenetic analyses of the large RNA polymerase subunits unequivocally demonstrated the affiliation between archaea and eukaryotes (28,29). Functional analysis of these subunits and basal transcription factors has suggested that they assemble and function in a very similar way in archaea and eukarya (30). These results

Table 1. *Continued*

MJ1053	**Mj,Af,Ph,Aae**	MJ1053, PH0283, AF0796	Outside archaea, highly similar homologs detected only in *Aquifex*; likely horizontal gene transfer from archaea to *Aquifex* (Aravind et al., 1998??)
MJ0558	**Mj,Mta,Af**	MJ0558, MTH163, AF0998	wHTH with a conserved, uncharacterized C-terminal domain
MJ0723	**Mj, Af, Ph + bacteria**	MJ0723, AF1723, PH1592	wHTH of the AsnC class, a number of highly conserved bacterial homologs
AF1148	**Af, Mta, Ph,** Sso + bacteria	AF1148, MTH1193, PH1692	wHTH of the AsnC class, an ortholog in *Sulfolobus* and a number of highly conserved bacterial homologs
wHTH+CBS	**Mj,Mta,Af,**Sso	MJ1232,MTH126,AF0111	The fusion of a HTH domain with 2 CBS domains is a unique archaeal feature also conserved in the crenachaeon Sulfolobus
HTH + Orotate phosphoribosyl-transferase	**Mj,Mta,Af**	MJ1646,AF0386, MTH876	N-terminal fusion of a HTH domain to a phosphoribosyltransferase, a unique archaeal domain architecture
HTH + Rio1-like protein kinase	**Mj,Af,Ph** + eukaryotes	MJ1073,AF2426, PH0512	N-terminal fusion of an HTH domain to a S/T protein kinase of the RIO1 family
PH-type ATPase+HTH	**Mj,Mta,Ph**(7),Sso	MJECL04, MTH196, PH0436, PH0976, PH0977, PH0846, PH0539, PH1067,PH0437	HTH fused to ATPases of a distinct archaeal family also seen in Sulfolobus
AF1298	**Mta,Ph**(2),**Af**	MTH1288,AF1298,PH1744, PH0180	Related to the ArsR family of bacterial transcriptional regulators
MTH1569	**Mj,Mta,Af**	MTH1569, MJ0159, AF2227	HTH fused to an uncharacterized duplicated domain at the C-terminus
MJ0723	**Mj,Ph,Af +** bacteria	MJ0723, PH1592, AF1723	A distinct family within the Lrp/AsnC class of HTH proteins that are common in both archaea and bacteria
MJ1475	**Mj,Ph,Mta**	MJ1475, MTH589,PH1560	Contain a conserved, AT-hook-like basic patch C-terminal of the HTH domain
DNA methylase + HTH	**Ph,Af,Mj**(2),Aae,Taq	MJ1273, MJ0675, AF1611, PH0728	C terminal fusion of a HTH domain with an unusual predicted DNA methylase. Specific to thermophilic archaea and bacteria, likely horizontal dissemination
Conserved in 2 of the 4 archaeal species			
cHTH+CBS	**Ph,Af**	AF2118,PH1748	Highly significant similarity to bacterial cHTH proteins (e.g. HI0659 from H. influenzae) but unique domain arrangement, with a cHTH domain fused to 2 C-terminal CBS domains

have led to the general concept of a eukaryote-type transcription mechanism in the archaea.

Detailed comparisons of the first completely sequenced archaeal genome, that of *Methanococcus jannaschii*, to bacterial genomes and a subsequent comparative analysis of the four sequenced archaeal genomes have revealed the presence of large numbers of HTH proteins that were more similar to bacterial HTH/wHTH proteins than to any of their known eukaryotic counterparts (22,24,31). Furthermore, preliminary experimental characterization of some of these archaeal DNA-binding regulators has suggested that they regulate transcription similarly to their bacterial counterparts (32,33). These observations stood out against the similarity of the core transcription machinery of the archaea to that of the eukaryotes and prompted us to investigate the archaeal transcription machinery, and particularly the transcriptional regulators, in detail. Here, we use profile search methods to identify the HTH proteins from the four completely sequenced euryarchaeal genomes and classify them using clustering by sequence similarity, phylogenetic tree topology and domain architecture. We also investigate the presence of other domains co-occurring with HTH domains in archaeal proteins and predict the possible functions of the respective proteins. Additionally, we describe other potential transcriptional regulators encoded in the archaeal genomes, such as the Met/Arc transcription factors (34,35) and Zn-ribbons (36). We discuss the implications of these observations for the evolution of transcription regulation in general.

Table 1. *Continued*

MJ1553	**Mj**,**Af**(2)	MJ1553,AF1697, AF2136	wHTH of the ArsR class
MJ1325	**Mj**,**Mta**,Aae + Bacteria	MJ1325, MTH899	Cadmium resistance operon regulator
MJ0905??	**Mj**,**Ph**	MJ0905,PH0544	
MJ0432	**Mj**,**Ph**,Sso	MJ0432, PH1061	cHTH conserved in Sulfolobus ; related to the bacterial MarR family
AF1264	**Mta**,**Af**	AF1264, MTH864	
AF2203	**Mta**,**Af**+bacteria	AF2203,MTH281	cHTH with highly conserved homologs in several bacteria
MJ1120	**Mj**,**Mta**,Aae + Bacteria	MJ1120,MTH1545	Belongs to the LysR class of HTH proteins
MJ0300	**Mj**,**Af** + Bacteria	MJ0300,AF2127	Belongs to the LysR class of HTH proteins
PhoU + HTH	**Mj**,**Mta**,	MJ1641, MTH1724	HTH domain fused to a C terminal PhoU domain (phosphate uptake regulator); unique archaeal domain architecture
AAA+ ATPase + HTH	**Mj**,**Ph**	MJ0774, PH0212	AAA+ ATPase domain fused to a C-terminal wHTH domain of the ArsR class
AF2008	**Af**,**Ph**	AF2008, PH0046	
MJ1398	**Af**,**Mj**	MJ1398, AF1564	Duplicated wHTH domain
AF0396	**Af**(3),**Mta**(3)	MTH1579,MTH1844,AF0396, MTH1843, AF2271, AF0102	wHTH domain fused to a conserved C terminal domain with a possible ezymatic function
MJ1503	**Mj**(2),**Af**(2), Hsp	MJ1503, MJ1082, AF1263,AF2143	wHTH protein conserved also in Halobacterium
AF1580	**Af**,**Ph** + bacteria	AF1580, PH1891	wHTH highly conserved in several bacterial species
MJ0287	**Mj**(2),**Af**(3)	MJ0287, MJ0290, AF2083,AF1663, AF1459	
AF2106	**Af**,**Ph**	AF2106, PH0660	HTH domain fused to a conserved C- terminal domain that is also present as a stand-alone protein in Mt
AF1043	**Af**,**Ph**	AF1043, PH0494	
MJ0621	**Mj**,**Af**	MJ0621, AF0643	
MJ1165	**Mj**,**Mta**	MJ1165, MTH1553	cHTH fused to a C-terminal ferredoxin domain

[a]Unless specifically indicated otherwise, all families of HTH proteins should be considered archaea-specific, with only a generic (and not necessarily statistically significant in direct pairwise comparisons) similarity to bacterial HTH domains. The numbers in parentheses indicate the number of paralogs (whenever more than one) in the given archaeal species. Species name abbreviations: Af, *Archaeoglobus fulgidus*; Mj, *Methanococcus jannaschii*; Mta, *Methanobacterium thermoautotrophicum*; Ph, *Pyrococcus horikoshii*; Aae, *Aquifex aeolicus*; Hsp, *Halobacterium* sp.; Sco, *Sulfolobus solfataricus.*

MATERIALS AND METHODS

Sequences, databases and sequence analysis

The predicted protein sequences (proteomes) derived from the four complete euryarchaeal genomes, namely *M.jannaschii* (Mj) (37), *Methanobacterium thermoautotrophicum* (Mta) (38), *Archaeoglobus fulgidus* (Af) (39) and *Pyrococcus horikoshii* (Ph) (40), were obtained from the genomes division of Entrez and corrected for frameshifts or incorrect starts wherever necessary. Sequence analysis was performed using a combination of local alignment searches using the BLAST family of programs and profile searches using the PSI-BLAST (41) and MoST (42) programs. Briefly, the procedure was as follows. Using a variety of known HTH domains as queries, the non-redundant (NR) protein sequence database was searched using the PSI-BLAST program for six iterations, with the cut-off for inclusion of sequences into the profile set at the expectation (*e*) value of 0.01. The outputs were examined for the presence of false positives, i.e. proteins whose pattern of amino acid residue conservation and predicted secondary structure were incompatible with those seen in HTH domains. The sequences that gave maximum depth of detection and no false positives in six iterations were chosen for saving the profile using the –C option of PSI-BLAST. These profiles were then individually run against the four archaeal proteomes using the –R option and preliminary lists of predicted HTH proteins were extracted. These proteins were run against the NR database using the gapped BLAST program in order to detect other domains present in them and to confirm the presence of the HTH. As an independent mode of detection, an alignment of the HTH domains was extracted from a library of known bacterial HTH transcription factors using the Gibbs sampling approach as implemented in the PROBE program (43). This alignment was used to derive a position-specific weight matrix that was run against the four individual archaeal proteomes using the MoST program. This approach was also used as the principal method for detecting the MetJ/Arc family members from archaeal and bacterial genomes.

The HTH proteins were classified using two methods. (i) Single linkage clustering as implemented in the Grouper program of the SEALS package (44). This program uses the bit score of the gapped BLAST alignment as the criterion for clustering the

proteins. Serial bit score cut-offs in the range 40–75 were used to divide the proteins into clusters of increasing stringency. (ii) In order to delineate orthologous families, the proteins were compared to the NR database and the symmetry of the best hits between different genomes was examined. Specifically, if protein X1 from proteome X gave best hits to proteins Y1, Z1 and Q1 in proteomes Y, Z and Q in a search against the NR database, it was determined whether proteins Y1, Z1 and Q1 in similar searches against the NR gave the other three proteins as the best hits from their respective proteomes. Such a distribution of best hits is termed 'symmetric' and is consistent with the hypothesis that these proteins form a family of orthologs.

Multiple alignments were constructed by combining PSI-BLAST-generated pairwise alignments and adjusted using the information on the secondary elements in HTH and Arc/MetJ domains with well-resolved structure, such as LexA, BirA, Cro, c1, CAP and Arc. These alignments were used to build phylogenetic trees using the neighbor-joining method as implemented in CLUSTALW/X (45) and the PHYLIP package (46). As the information in the alignment comprising only the HTH domain was not sufficient to resolve higher order branching, only the terminal branches with bootstrap support >70% in 500 trials were considered.

Secondary structure predictions and structural database threading were performed using the PHD program (47,48). Protein homology modeling was performed using the PROMOD-II program (49). The SWISS-PDB viewer was used to manipulate the pdb files and provide the alignments for model building. Protein structures were visualized using the MOLSCRIPT program (50).

RESULTS AND DISCUSSION

Sequence–structure-based detection and classification of HTH domains

In order to investigate the archaeal HTH domain-containing proteins in detail, we used structural criteria (see the SCOP database; 51) to divide them into three sets: (i) the winged-HTH (wHTH) domains that contain the C-terminal β-hairpin and, in some cases, the additional elements of the β-sheet between H-1 and H-2; (ii) Cro-like HTH (cHTH) domains that have no additional elements at the C-terminus and contain a tight loop between H-1 and H-2; (iii) miscellaneous HTH (mHTH), a heterogeneous collection of domains that differ from the typical HTH in terms of certain features, such as shortened helices or incorporation into other super-structures, e.g. the cyclin fold in TFIIB (52; see Supplementary Material, Fig. S1). For each of these classes, we investigated the neighbors in protein sequence space using iterative PSI-BLAST searches with the structurally characterized representatives as starting points (six iterations, *e* value cut-off of 0.01 for including sequences in the profile).

Using structurally characterized proteins that represent distinct families within the wHTH class as queries for iterative searches, we found that all these families recognized each other; a similar result was seen with the cHTH. In the case of the heterogeneous mHTH category, we used separate profiles for each family that included the TFIIB/TFB, MerR and MCM-type families. Some of the sequences that were detected in searches with profiles for the cHTH and wHTH were shown, by means of subsequent multiple alignments and secondary structure prediction, to represent distinct families of HTH domains and, accordingly, were included in the heterogeneous mHTH category. Most of these relationships were independently recognized using multiple alignments constructed with the Gibbs sampling method implemented in the PROBE program as the input for iterative database searching using MoST. The consistent detection of sequence similarity with different searching methods, together with the direct evidence for structural similarity of diverse proteins, strongly suggest that all these HTH domains are homologous. In addition to known HTH domains and their close homologs, this analysis detected a number of uncharacterized proteins that can be predicted to adopt the HTH structure, particularly that of the wHTH class (see below).

These searches detected primarily bacterial and archaeal (predicted) HTH domains and only a few eukaryotic proteins. The main classes of detected eukaryotic HTH proteins were the members of the Paired and CENP-B families (13), both of which are distinct and apparently relatively recent derivatives of the transposon integrase HTH domain family (53,54). The POU domain and the homeodomain were detectable with marginal statistical significance (*e* value ~0.01) in some of the searches that were initiated with the transposase-type HTH domains (53,55). Thus, with a few exceptions, bacterial and archaeal HTH domains appear to be much more closely related to each other than either of them are to the eukaryotic structural counterparts. In searches initiated with eukaryotic queries, such as the homeodomain, myb, H1, forkhead and ETS, the prokaryotic proteins were largely undetectable, with the exception of two Myb-like proteins from *Bacillus subtilis* (YwfN, 732380; YlbO, 2340012). Having established an apparent affinity of the predicted archaeal HTH proteins with the bacterial ones, we investigated them in greater detail.

As a result of the searches described above, the predicted HTH domains were extracted from the four complete sets of archaeal proteins. The individual sequences of these domains were run against the NR database using the single-pass gapped BLAST program to identify the closest homologs. To verify the prediction of the HTH structure, multiple alignments were made for several of such families and used as queries for secondary structure prediction using the PHD program. In conjunction with the available 3-dimensional structures of HTH proteins, the resulting predictions were used as a guide for constructing multiple alignments of all predicted archaeal wHTH, cHTH and mHTH domains (Fig. S2a–c).

When the alignments of all the different types of the HTH domains are superimposed, the sequence signatures underlying the structural conservation become apparent (Fig. S2). Specifically, H-1 contains a triad of medium sized to bulky residues that typically consists of a polar residue followed by two hydrophobic ones (positions 10–12 in Fig. S2). In the cHTH domains, the distal position of this triad is in some cases occupied by a basic residue. The most conserved signature of the HTH proteins, the dyad of an acidic and a hydrophobic residue (most frequently, glutamic acid–isoleucine), is located in the middle of H-2 (Fig. S2, positions 31–32). This signature is followed by the characteristic turn which is signaled by a position at the end of H-2 that is most often occupied by residues with high turn-forming propensity, such as glycine, asparagine, lysine or aspartate (Fig. S2, position 38). H-3 is the recognition helix

that makes the major groove contact with DNA and contains the residues that, in part, are responsible for the DNA-binding specificity. This helix contains a highly conserved hydrophobic position in its N-terminal part (Fig. S2, position 44) and a bulky position in the C-terminal part. In the wHTH proteins, the latter position is occupied by a hydrophobic residue, whereas in the cHTH domains, it is a charged residue (Fig. S2, position 51).

In addition to the above signatures that are seen in all HTH domains, there are other features that are restricted to specific classes. As already mentioned, the wHTH domains show considerable variability in the region between H-1 and H-2, and in some of them, one or more β-strands are found in this region. The most characteristic feature of the wHTH domains is the C-terminal β-sheet, the wing. The junction between H-3 and the first strand of this sheet is marked by a highly conserved glycine (position 55 in Fig. S2a), which is a signature of the wHTH class of HTH domains. While strand 1 is easily discernible and moderately conserved in the wHTH sequences, there is considerable variability in the loop between the two strands (Fig. S2a). In the cHTH domains, the region between H-1 and H-2 does not show much variability but is characterized by a conserved glycine that maintains the structure in a closed conformation (Fig. S2b, position 26). Thus the classical cHTH domain is smaller and more compact than the wHTH domain. These features, in addition to the specificity of the profiles for each of these types of HTH, helped in distinguishing these two classes of HTH proteins encoded in the archaeal genomes. In a few of the wHTH domains, the C-terminal sheet appeared to show some signs of degeneration, but their evolutionary affinities nevertheless could be established on the basis of their specific sequence similarity to other wHTH proteins.

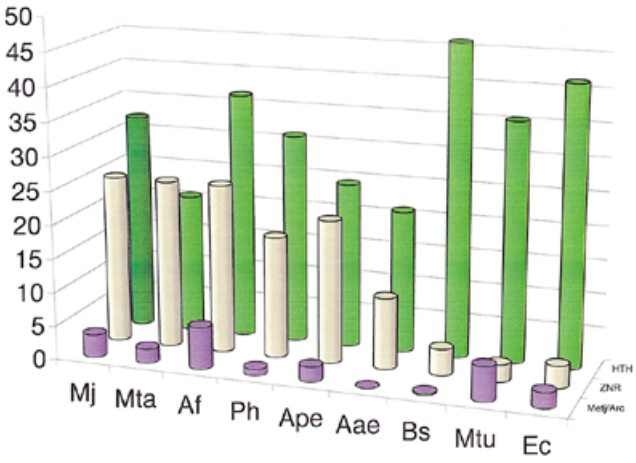

Figure 1. Relative abundance of HTH, Arc/MetJ and Zn-ribbon domains in archaea and bacteria. Vertical axis: number of detected domains per 1000 proteins. Bacterial species abbreviations: Aae, *Aquifex aeolicus*; Bs, *Bacillus subtilis*; Mtu, *Mycobacterium tuberculosis*; Ec, *Escherichia coli*. Note the striking differences in the distributions of the Arc/MetJ and Zn-ribbon proteins in contrast to that of the HTH proteins. Also note the low number of HTH proteins in *Methanobacterium* relative to the other species. The preliminary data for the crenarchaeon *Aeropyrum pernix* (Ape) were included (see Addendum). This genome has been reported to encode 2694 proteins. However, since this results in a gene density that is much lower than in other archaea and bacteria and also because an unusually large fraction of these predicted proteins have no detectable homologs in other species (L.Aravind and E.V.Koonin, unpublished observations), we consider this to be a gross overestimate. The data shown in the figure are based on an estimated 1800 proteins encoded in the *A.pernix* genome.

The diversity of HTH proteins in archaea and trends in their evolution

Our analysis revealed a much greater diversity of HTH domains among the archaea than previously suspected, in terms of both numbers and relationships. The counts of detected HTH domains range from 39 in *M.thermoautotrophicum* to 88 in *A.fulgidus* (Fig. 1). Even when the counts were normalized by the total number of genes in the genome of each archaeon, there was a clear difference in the abundance of the HTH domains (Fig. 1). *Archaeoglobus fulgidus*, which has the largest genome, encodes more than twice as many HTH domains per 1000 proteins as *M.thermoautotrophicum*. The numbers of HTH domains in *P.horikoshii* and *M.jannaschii* are closer to that in *Archaeoglobus*, suggesting that *Methanobacterium* has a deficit of HTH proteins. In each case, the normalized count of HTH proteins in archaea is within the range seen in free-living bacteria with proteome sizes comparable to those of the archaea and only slightly lower than in bacteria with large genomes and versatile metabolism, such as *B.subtilis* and *Escherichia coli* (Fig. 1).

The wHTH domains are by far the most abundant class of HTH domains in the archaea (Fig. S2). There are five or more times as many wHTH domains than there are Cro-like HTH domains in each of the archaeal species; the count of the Cro-like domains is about the same in each of the archaea, namely 7–11. The remaining types of HTH domains (the miscellaneous class) account for only a small fraction of the total number; some of them are specialized, conserved versions of the HTH domain, such as those seen in the MCM proteins and in TFIIB/TFB (Fig. S2c).

To establish orthologous relationships among the archaeal HTH-containing proteins, we investigated the phyletic distribution of the taxon-specific best hits with an *e* value threshold of 10^{-2} for all archaeal HTH domains. Among the 214 archaeal HTH domains that gave database hits above the cut-off, <10% were reliably (*e* value difference of two orders of magnitude or greater from the next best hit) most similar to bacterial homologs, whereas for the rest, the best hit was another archaeal protein. Thus, most of the archaeal HTH domains appear to have undergone at least some evolution within the archaeal superkingdom. The evolutionary events involved might have included both vertical inheritance and horizontal transfer amidst the archaea (24). Those archaeal HTH-containing proteins that did give best hits to bacterial proteins might be cases of horizontal transfer from bacteria into archaea (see below).

Those archaeal proteins that form symmetrical pairs of intergenomic best hits can be considered likely orthologs. This notion is reinforced in cases where a consistent clique of symmetrical best hits connects three or all four of the archaeal genomes (24,56). The validity of these putative orthologous clusters was additionally evaluated by using single linkage clustering of archaeal HTH domains and neighbor-joining tree analysis. Using these procedures, we identified eight clusters of likely orthologs that are present in all four archaea, 15 that

Table 2. Likely origin of archaeal HTH proteins by horizontal gene transfer from bacteria

Archaeal protein	Phyletic distribution of homologs	Best database hits[a]		Comment
		In bacteria (eukaryotes)	In archaea	
AF0074	**Af** + bacteria	BirA_Bs (1146239); 2e-33	MTH1916; 1e-28	HTH domain fusion with biotin synthetase (bifunctional protein BirA); domain organization conserved in Af and several bacterial species but not in other archaea that have a single-domain biotin synthetase
MTH1285	**Mta** + bacteria	YkvN_Bs (2633747); 5e-26	NONE	A family of small cHTH proteins highly conserved in a number of bacterial species
PH0952	**Ph** + eukaryotes, actinomycetes, Bs	I2C-2_Le (2258317); 2e-05[b]	NONE	Eukaryotic/actinomycete-type HTH-AP-ATPase also containing a TPR-repeat domain
MTH659, MTH700	**Mta**(2) + bacteria	SC6A5.19_ Scoe (4539174); 5e-08[c]	NONE	cHTH domain fused to a C- terminal AraC-like double-stranded β-helix domain; likely duplication in Mt subsequent to horizontal gene transfer from bacteria
PH1691	**Ph** + bacteria	PurR_Bs (2127038); 1e-15	Apt_Mj (2127744); 7e-06	cHTH fused to a C- terminal phosphoribosyltransferase domain; domain arrangement conserved in Ph and Gram-positive bacteria
AF0673	**Af** + bacteria	HmrR_Rl (4633809); 2e-12	NONE	Belongs to the MerR/NolA family of transcription factors
AF1022	**Af**, Rc, Av, Ec	MopB_Rc (585505); 4e-09	NONE	The Af protein is highly similar to the HTH of the bacterial homologs but lacks the molybdenum-binding domain present in the latter

[a]Database hits with an associated *e* value <0.01 in a single pass, gapped BLAST search are included.

[b]The C-terminal portion of PH0952 that consists of TPR repeats and shows similarity to a variety of proteins was removed prior to the database search. Database hits are indicated by the gene name followed by an abbreviated species name and the gene identification number (in parentheses); the *e* value is also indicated ($e - n = 10^{-n}$). Species name abbreviations not used in Table 1: Av, *Azotobacter vinelandii*; Bs, *Bacillus subtilis*; Ec, *Escherichia coli*; Le, *Lycopersicon esculentum* (tomato); Rc, *Rhodobacter capsulatus*; Rl, *Rhizobium leguminosporum.*

[c]Data for MTH659.

are represented in three species and 22 that are found in two species (Table 1).

Notably, among the eight orthologous clusters of HTH-containing proteins that are conserved in all four archaea, four have obvious orthologs in eukaryotes. These clusters include the orthologs of the basal transcription factors TFIIB (mHTH), TFIIE-α (wHTH) and MBF1 (cHTH) and phenylalanyl-tRNA synthetase α-subunit (wHTH). Of the remaining four all-archaeal clusters, one includes the DtxR-type HTH domains that have closely related bacterial orthologs involved in iron metabolism regulation (57), and the other three appear to be archaea-specific.

Beyond the small set of HTH domains that are components of the core transcription machinery, the rest are not universally conserved (Table 1). The orthologous sets account for only ~25% of the total number of HTH-containing proteins in the four archaea; this suggests prominent roles for lineage-specific duplication and horizontal transfer in the evolution of the observed diversity. The results of single linkage clustering, tree analysis for different HTH classes and the TAX_COLLECTOR analysis indicate that lineage-specific gene duplication has been a major force in the evolution of archaeal HTH-containing proteins. These specific expansions are particularly prominent in *Archaeoglobus* and *Pyrococcus* (see below).

Horizontal transfers between archaeal species are hard to detect due to the general background conservation of protein sequences. In contrast, horizontal transfers from bacteria are apparent (Table 2). Clear-cut examples include the BirA-like protein in *Archaeoglobus*, the HTH-containing, apoptotic-type (AP) ATPase from *Pyrococcus* and proteins containing the double-stranded β-helix domain in *Methanobacterium*. A conservative estimate of at least seven cases of horizontal transfer of bacterial genes encoding HTH proteins into one of the four archaeal lineages can be made on the basis of the BLAST search results, clustering and domain organization of the HTH proteins. Additional, ancient horizontal transfers of bacterial genes might have been obscured by subsequent inter-archaeal horizontal transfer. Candidates for such ancient transfers are the LysR family (9) and the CadR family that are prevalent in bacterial genomes but are represented by only one or two members in two of the archaeal genomes (Tables 1 and 2).

The analysis of orthologous clusters suggests that the common ancestor of the four euryarchaeal species most likely encoded 20–30 HTH proteins. This number is consistent with a free-living prokaryote that, in addition to the core transcription machinery, encoded transcription regulators for some basic metabolic processes that might have included, among others, iron metabolism (57). The subsequent radiation of the archaeal lineages seems to have involved considerable loss of genes

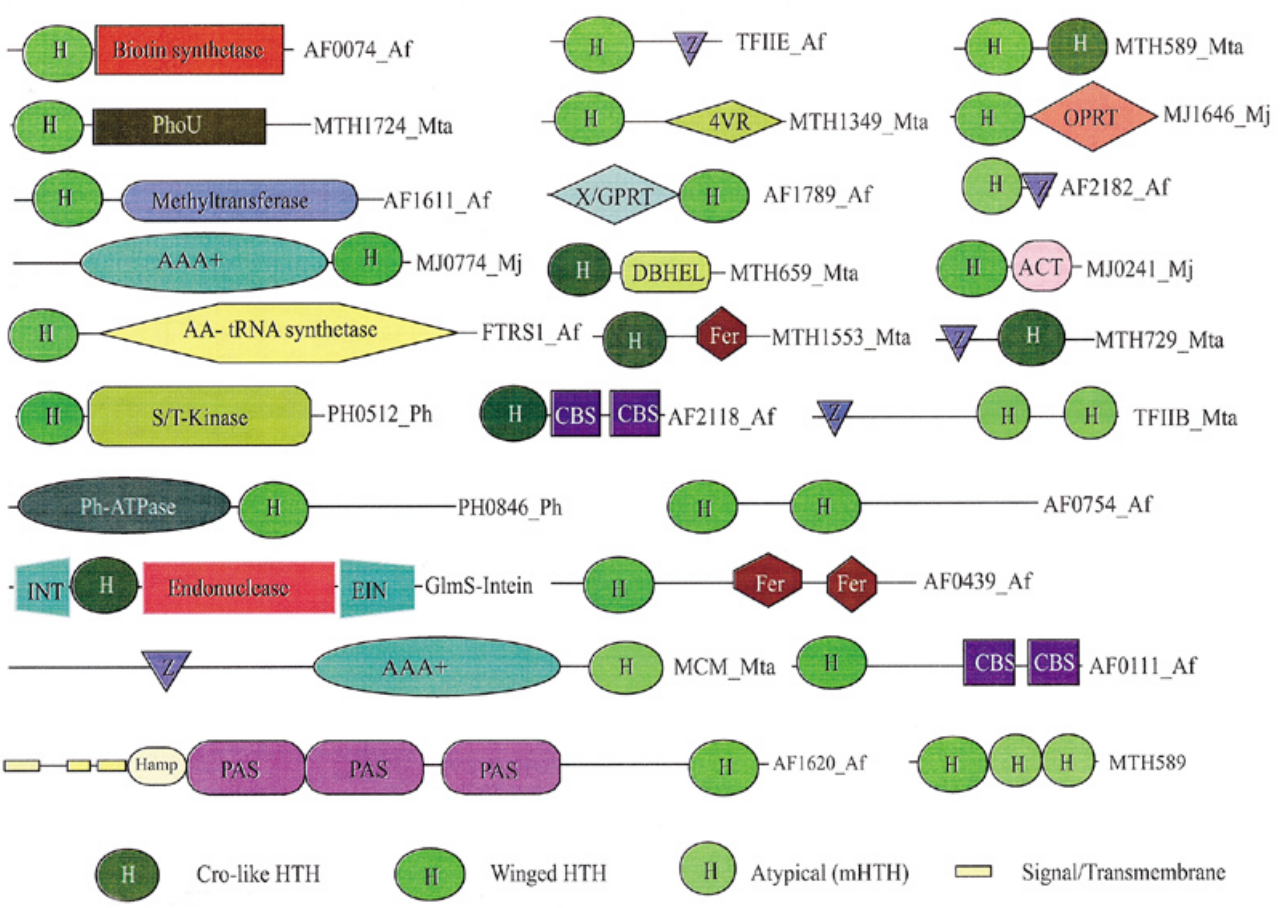

Figure 2. Distinct domain architectures of archaeal proteins containing predicted HTH domains. The diverse domain architectures of the archaeal HTH proteins are shown with the different domains drawn to scale. The wHTH, cHTH and the heterogeneous mHTH categories are distinguished from each other. The abbreviations for some of the domains shown in the figure are: 4VR, 4-vinyl reductase; Fer, ferredoxin; HAMP, histidine kinase-adenylyl cyclase-methyl accepting protein-phosphatase domain; OPRT, orotate phosphoribosyl transferase; X/GPRT, xanthine-guanine phosphoribosyltransferase; DBHEL, AraC-like double-stranded β-helix; Z, Zn-ribbon. The rest of the domains are described in the text; the HINT module is shown as split INT+EIN to indicate the insertion of other domains into it.

coding for HTH-containing proteins (note the 13 orthologous clusters that include three archaeal species in Table 1). This was compensated for by independent, lineage-specific duplications that resulted in the accumulation of certain families in particular species, along with some horizontal transfer from bacterial and other archaeal genomes.

Domain architectures and predicted functions of the archaeal HTH proteins

Although it is formally possible that some of the archaeal HTH domains are involved in protein–protein interactions rather than in nucleic acid binding, the latter is the predominant, if not the only role of HTH proteins in bacteria. Given the clear relationships between bacterial and archaeal HTH domains, in terms of sequences, and in many cases, domain organization (see below), the prediction that most of the archaeal HTH proteins are DNA-binding transcription regulators seems to be justified.

About one-third of the archaeal HTH proteins possess distinct additional domains that are fused to the HTH domain. The rest are small proteins that have no distinct globular domains other than HTH or contain predicted non-globular domains; a few also contain predicted globular domains whose identity could not be discerned by sequence analysis. The additional globular domains that are fused to the HTH domains can be classified into enzymatic and interaction-mediating ones. Figure 2 shows a representative set of domain architectures of these proteins from different archaea.

In several archaeal proteins, a wHTH domain is fused to apparently active (as judged by the conservation of all essential catalytic residues; data not shown) metabolic enzymes, such as biotin synthase, orotate phosphoribosyl transferase, xanthine-guanine phosphoribosyltransferase and threonine synthase. These proteins might combine the catalytic function with that of transcription regulation of the biosynthetic genes in response to the respective metabolite in the environment. Several similar cases are seen in bacteria (13,57,58) but some of the combinations of HTH domains with nucleotide biosynthesis enzymes thus far appear to be unique to the archaea.

A conserved family of archaeal proteins (e.g. AF1611 shown in Fig. 2), which is also represented in thermophilic bacteria, contains a wHTH domain fused to the N-terminus of a predicted DNA methyltransferase (58). This might be an adaptation of the wHTH for targeting the methyltransferase to DNA for thermophile-specific DNA modification.

The other prominent group of enzymatic domains that are associated with the archaeal HTH domains are ATPases. One such fusion is seen in the archaeal orthologs of MCM proteins. MCM proteins play critical roles in the initiation of DNA replication in eukaryotes (59,60). The archaeal MCM orthologs contain a distinct HTH domain at the C-terminus, which is not detectable in the eukaryotic MCM proteins, and either represents an archaea-specific fusion or might have been eroded in eukaryotes. Perhaps the most notable of the ATPase–HTH combinations is the C-terminal fusion with a novel class of archaeal ATPases (Fig. 2). This domain architecture is represented in the genomes of *Pyrococcus*, *Methanococcus* and *Methanobacterium* and also in the incomplete genome of the crenarchaeon *Sulfolobus sulfotaricus*, suggesting a widespread distribution in the archaea. These ATPases are particularly abundant in *Pyrococcus*, which encodes at least eight members of this family (24). The combination of the ATPase domain with the HTH domain suggests that these proteins might possess a helicase-like activity and play a regulatory role in DNA unwinding or might be involved in reorganizing DNA-associated protein complexes.

Another remarkable domain architecture is the fusion of an HTH domain to the N-terminus of a predicted S/T-type protein kinase domain of the Rio1 family (61) that is conserved in *A.fulgidus*, *P.horikoshii* and *M.jannaschii*. An orthologous protein with the same domain architecture was also detected in eukaryotes (yeast YNL207w). This unusual protein kinase might be an ancient regulatory protein that could function through phosphorylation of DNA-associated protein complexes.

The fusion of the wHTH domain to the N-terminus of the phenylalanyl-tRNA synthetase α-subunit is a relatively rare case when an HTH domain seems to be used for RNA-binding rather than transcriptional regulation. The wHTH domain in these proteins is similar to those seen in other RNA-binding proteins, such as the RNA-editing deaminase and the eukaryotic ribosomal protein S10 (21).

HTH domains in the archaea are also fused to a variety of non-enzymatic domains that are predicted to mediate interactions with other proteins, nucleic acids or small molecule ligands (Fig. 2). In bacteria, HTH domains typically are fused to domains recognizing small molecules that regulate the expression of various operons. Homologs of some of these proteins with similar domain combinations were detected in the archaea. These include the α-helical iron-binding domain and the C-terminal β-barrel domain present in the DtxR-like proteins (57) (e.g. MJ0568 and MTH936) and the β-jelly roll domain (62) that is fused to LysR-like HTH domains (AF2127, MTH1545 and MJ1120). *Methanobacterium thermoautotrophicum* contains two proteins (MTH659 and MTH700) with Cro-like HTH domains that are fused to the N-terminus of a double-stranded β-helix domain similar to those found in vicilin and phaseolin and probably involved in sugar recognition (63). A similar domain is fused to the N-terminus of the wHTH in AraC-type transcription factors, suggesting that at least twice in evolution the double-stranded β-helix domain had been independently fused to a DNA-binding HTH domain.

In addition to the proteins that have clear bacterial counterparts, we detected several archaea-specific fusions of HTH domains to small ligand-binding domains (Fig. 2). In the AF1620 protein, for example, a wHTH domain is combined with three PAS domains that are likely to respond to light or redox potential by binding a ligand (64). Several archaeal proteins (e.g. MTH1553 and AF0439) contain fusion of an iron-binding ferredoxin domain to either wHTH or cHTH domains (Fig. 2). Given the importance of iron-redox metabolism in the archaea, it seems possible that the proteins that contain the HTH–ferredoxin combination function as regulators of the redox response. Several independent fusions of cHTH and wHTH domains with cystathionine β synthase (CBS) domains, which are predicted to function in signal transduction via ligand binding (65), were detected in the archaea. This domain combination might play a conserved sensor role in the archaea since the cHTH–CBS architecture is conserved not only in three of the four completely sequenced genomes but also in *S.sulfotaricus*. Similarly, the aspartokinase-chorismate mutase-TyrR (ACT) domain (66), a predicted amino acid-binding domain involved in allosteric regulation of a number of enzymes, is fused to a HTH in the MJ0241 protein. The tendency of HTH domains to combine with various small ligand-binding domains as well as metabolic enzymes suggests that in archaea, transcription responds to the environment in a manner akin to that in bacteria.

An interesting observation is the detection of cHTH domains in a number of archaeal inteins (e.g. those in the replication factor C and initiation factor 2 from *M.jannaschii*). These HTH domains are sandwiched between the N-terminal part of the Hedgehog–intein (HINT) module and the inserted homing endonuclease domain (67,68; Fig. 2). This association with the endonuclease domain suggests that the HTH might play a role in the recognition of target sequences by the endonuclease in the process of homing.

Some archaeal HTH proteins also contain another predicted DNA-binding domain, namely the Zn-ribbon. The core transcription factors TFIIE-α and TFB are combinations of different forms of the HTH domain with the Zn-ribbon. The archaeal orthologs of the eukaryotic MBF1 protein (69,70) also combine a cHTH with a Zn-ribbon which is absent in their eukaryotic counterparts.

A unique family of archaea-specific HTH proteins is defined by MJ1243 and its orthologs and is represented by one member in all four archaeal genomes. These HTH proteins contain a basic patch that is located N-terminally of the HTH and is predicted to adopt an AT-hook-like minor groove-binding surface (71). This combination of major and minor groove-binding motifs echoes the synergism between the AT-hook and other DNA-binding domains in eukaryotes (71).

Other nucleic acid-binding proteins with a possible role in archaeal transcription

The Arc/MetJ family of transcription factors. While HTH is the predominant DNA-binding domain in prokaryotes, several other unrelated transcription factor families have been found in bacteria. One of these includes the MetJ/Arc proteins that act by inserting an N-terminal β-strand into the major groove of DNA (34,72). In addition to this strand, they possess two C-terminal

helices whose orientation is reminiscent of the two C-terminal helices of HTH proteins (73). However, rather than participating in DNA binding as they do in the classical HTH domains, in the MetJ/Arc family proteins, these helices are involved in dimerization of the domain (Fig. S3a). Most of these proteins are very small and divergent in sequence, although they seem to adopt a conserved tertiary structure; this makes it difficult to identify them by sequence comparison alone.

To investigate this family as completely as possible, we performed iterative PSI-BLAST searches using all known Arc/MetJ proteins as queries and retrieved all homologs detectable by this approach. An alignment of the DNA-binding domains from all these proteins was constructed using the PROBE program. Even more divergent members were then detected using the MoST program to iteratively search the database with a weight matrix derived from a conserved ungapped block extracted from the alignment. The alignment was subsequently refined using secondary structure prediction. PDB database threading using the PHD program with multiple alignments as input provided strong support for these proteins adopting the Arc/MetJ fold. Specifically, the Arc/MetJ structures consistently figured as the best hits with *Z* scores in the range 2.1–2.8.

Examination of the multiple alignment and comparison with the 3-dimensional structure of the Arc protein from phage P22 shows that the conservation pattern of this family correlates with both DNA interactions and structure-stabilizing properties of the residues (Fig. S3b). In both the first strand and the two following helices, a significant part of the conservation rests on hydrophobic residues whose side chains point to the interior of the structure and stabilize it via hydrophobic interactions. The conserved hydrophobic residues in H-2 are likely to stabilize inter-monomer contacts. The conservation of all these residues in the Arc/MetJ superfamily is consistent with these proteins adopting a similar fold and dimeric organization. Another set of conserved residues have small side chains that favor the turn-like conformation between strand-1 and H-1 and the two C-terminal helices. Most of the remaining conserved polar residues are in positions that are compatible with DNA interaction. These positions include the residue in strand-1 that is inserted into the major groove of DNA and forms hydrogen bond contacts with the bases. This polar position is likely to play a role in DNA-binding specificity of these proteins (Fig. S3a and b), along with the non-conserved part of strand-1. The other polar residues, such as the one at the beginning of H-1 and the one after the turn between the two helices (Fig. S3) are likely to influence the interactions with the backbone of the DNA.

Our analysis showed that Arc/MetJ proteins are far more common in both archaea and bacteria than previously suspected (Fig. 1). These proteins were detected in all archaea, including *Sulfolobus*, with the greatest number (14) found in *A.fulgidus*. This suggests that a distinct family of Arc/MetJ-like DNA-binding proteins was already present in the common ancestor of the archaea. We found that this family is nearly ubiquitous in bacteria also, with the maximum number (20 proteins) present in *Mycobacterium*. Plasmid-encoded Arc-MetJ family proteins appear to regulate the transcription of plasmid genes by interacting with plasmid-encoded proteins of another family (74) that we have shown to contain the PIN (PilT N-terminal) domain (24). The PIN domain family is greatly expanded in *Archaeoglobus* and *Mycobacterium*, which also encode the maximum number of Arc/MetJ-like proteins. Thus, the cooperation between the Arc/MetJ and PIN domains in transcription regulation might be a general phenomenon.

The rubredoxin-like Zn-ribbons. The rubredoxin-like Zn-ribbon is a small metal-binding module that is widespread in proteins from all three domains of life (51,75). Zn-ribbons participate in many different functions that include: (i) DNA binding, e.g. the C-terminal domains of topoisomerase I, TFIIS and RNA polymerase subunits (36); (ii) RNA binding, e.g. ribosomal proteins S27, L37, L40 and aminoacyl-tRNA synthetases (74); (iii) redox reactions, e.g. rubredoxins, rubrerythrins and nigrerythrins (75); (iv) Zn-ribbons found as inserts within other domains where they probably play a strucutural role or aid in catalysis, as in the case of adenylate kinase (78) and pyruvate formate lyase. The structure of this domain is typified by rubredoxin from *Desulfovibrio gigas* (79) and has as its core a Zn^{2+} ion coordinated by two pairs of cysteines, each associated with a more or less symmetric pair of short β-strands (Fig. S4). Beyond the four strands that surround the bound Zn^{2+} ion, different members of this family show considerable structural elaboration in the form of insertions between the two cysteine-containing pairs of strands. This makes the spacing between the cysteine pairs extremely variable.

Due to this variability, comprehensive detection of Zn-ribbon domains based on sequence alone is particularly challenging. We attempted to overcome this difficulty, at least in part, by using transitive searches with PSI-BLAST followed by evaluation of the conserved residue patterns with respect to their compatibility with the Zn-ribbon structure. Additionally, the conservation of the cysteines makes it possible to apply pattern searches for the detection of Zn-ribbons. Given the multiple functions of these modules, the presence of a Zn-ribbon, as such, is not indicative of DNA binding or a function in transcription regulation for the respective protein. However, given that they are common in transcription factors and RNA polymerase subunits from archaea, eukaryotes and bacteria, we systematically surveyed the archaeal genomes for Zn-ribbons. This analysis revealed remarkable abundance of Zn-ribbon proteins in archaeal proteomes which, when normalized for the gene number, exceeded the number of such proteins in bacterial genomes by at least a factor of 2 (Fig. 1). Based on sequence conservation among Zn-ribbons themselves and the presence of additional domains, we classified the Zn-ribbon-containing proteins into groups that are likely to represent their functions in transcription, DNA replication and repair, translation or RNA binding, redox processes and other miscellaneous processes (Fig. S4). This grouping shows that a significant number of archaeal proteins that contain Zn-ribbons are likely to be involved in nucleic acid binding and, more specifically, in transcription (Table S1).

The conserved archaeal Zn-ribbon proteins that have well-understood functions in transcription are components of the core machinery that include TFIIB/TFB, TFIIE-α, TFIIS/RPOM (two repeats), MBF1 and the RPOE″ and RPB′ subunits of RNA polymerase. In addition to these core transcription components, the Zn-ribbon is fused to a number of HTH domains, and the respective proteins are predicted to function as transcription regulators (Fig. 2). Some conserved Zn-ribbon proteins that are similar to the small RNA polymerase subunits and are shared by archaea and eukaryotes, such as the MJ1474 and MJ0890 families (Fig. S4 and Table S1), may represent yet uncharacterized core transcription factors. Examination of the

phylogenetic distribution of the Zn-ribbon proteins shows that at least seven of them must have been functioning in transcription in the common ancestor of archaea and eukaryotes (Table S1). In addition, there have been at least four independent fusions of Zn-ribbons to HTH domains that are unique to archaea, which emphasizes the importance of Zn-ribbons in archaeal transcription (Figs 2 and S4).

In addition to these Zn-ribbons that seem to have a recognizable role in transcription, the domain organization of several proteins is compatible with such a function although there is no direct evidence to support this hypothesis. These include the protein family that is conserved in all four archaea and contains Zn-ribbons fused to two N-terminal CBS domains in a manner that is reminiscent of a similar fusion of the HTH and CBS domains (Figs 2 and S4). These proteins might be novel transcription factors that sense a ligand using the CBS domains and bind DNA via the Zn-ribbons. Zn-ribbons that are likely to bind DNA or RNA were also detected in several archaeal proteins that are involved in DNA replication/repair and translation.

Apart from these Zn-ribbons that can be predicted, with considerable confidence, to bind DNA or RNA, there are other such modules in archaeal proteins that may or may not have a role in transcription. Many of these are associated with enzymatic domains, such as the ABC ATPase, PP-ATPase and threonine synthase, while others occur as the only detectable feature in uncharacterized small proteins. Many of these 'stand-alone' proteins show specific similarity to rubredoxins and are likely to participate in the autotrophic iron-dependent respiration that is typical of many archaea (80).

Evolutionary implications of the diversity of archaeal DNA-binding domains

The results of the present analysis of predicted archaeal DNA-binding proteins, in particular those that contain the HTH domain, may have significant implications for the evolution of transcription in general. The most important observation seems to be that there is an extensive diversity of HTH domains in the archaea, which is of the same order of magnitude as the diversity seen in bacteria. Most of the predicted archaeal HTH-containing proteins form archaea-specific families, but collectively, they show significantly greater similarity to bacterial than to eukaryotic HTH-containing transcription factors. This is in striking contrast to the core transcription machinery, which shows clear eukaryotic affinities. Our estimation of the number of HTH domains that might have been encoded by the common ancestor of the euryarchaea suggests that a significant archaea-specific set of HTH domains emerged early in the evolution of this domain of life.

Another notable observation is the detection of the HTH domain in several conserved transcription factors, such as TFIIE-α, TFIIB/TFB and MBF1. TFIIE-α (81) and TFIIB (82) are well-characterized factors that are required for transcription in eukaryotes but the exact role of MBF1 has not been studied extensively. Experiments in *Bombyx mori* and yeast have suggested that the function of MBF1 is to link transcription activators to TBP (69,70). Its extreme conservation in the eukaryotes and the presence of an ortholog in all four archaea suggests that MBF1 is another basal transcription factor associated with the core transcription initiation complex. Even in these core transcription factors, there are clearly discernible differences between the eukaryotic and archaeal versions. In the eukaryotic TFIIB, the HTH domain has greatly diverged in sequence, and in several families of paralogs, such as the cyclins and the retinoblastoma-like proteins (83), is hardly recognizable at the sequence level. In contrast, the archaeal orthologs of TFIIB contain a HTH domain that is readily recognizable by sequence searches (Fig. S2c). Similarly, the HTH domain has degenerated in eukaryotic TFIIE-α, although the Zn-ribbon that is shared with the archaeal ortholog is relatively more conserved. In the case of MBF1, the converse is observed, i.e. the Zn-ribbon is missing in eukaryotes, whereas the HTH domain is highly conserved throughout. Taken together, these observations seem to indicate that the archaeal core transcription machinery is closer to that of the common ancestor of archaea and eukaryotes.

Other than the HTH domain, the only other prominent DNA-binding domain that is seen both in archaea and in eukaryotes is the Zn-ribbon. As shown above, this domain is present in several core components of the transcription machinery, namely RNA polymerase subunits, such as RPOM and RPOB′, essential transcription factors TFIIB, TFIIS and MBF1, and the conserved transcriptional regulator Sir2. In all these proteins, with the sole exception of MBF1, the Zn-ribbon is conserved in archaea and eukaryotes. In contrast, there are no orthologs of any of these transcription factors in bacteria. In terms of both absolute and particularly normalized numbers, archaea have more Zn-ribbon proteins than bacteria and eukaryotes. This correlates with their iron-dependent metabolism (80) in which the Zn-ribbon proteins could bind iron ions that undergo redox transformations. It seems plausible that in the common ancestor of archaea and eukaryotes, as well as during the subsequent evolution of the archaea, a number of Zn-ribbons have been adapted for DNA-binding and transcription regulation from ancestral iron/zinc-binding proteins that were originally involved in redox processes. This is reminiscent of the presence of intact ferredoxin domains in the archaeal RNA polymerase 30/40K subunits (84); the ferredoxin domain has degenerated in the eukaryotic and bacterial orthologs of this RNA polymerase subunit (the α-subunit in bacteria) (L.Aravind and E.V.Koonin, unpublished observations).

Beyond the core components, the similarity between the archaeal and bacterial transcription systems in terms of gene/operon-specific, HTH-containing transcription factors might have been maintained by horizontal gene exchange (22,24). Much of this postulated gene transfer probably occurred before the divergence of the currently known euryarchaeota and was followed by considerable lineage-specific evolution.

While the HTH domain is the most important DNA-binding domain in archaea and bacteria, in eukaryotes (at least those that belong to the crown group), it represents only one class of DNA-binding proteins, along with several other, eukaryote-specific, non-HTH transcriptional regulators. Even in the eukaryotic lineages in which derived HTH domains are numerically expanded, as in the case of Homeodomain, HNF3, Paired and SANT/MYB, this appears to have happened relatively late in evolution, i.e. after the radiation of the crown group lineages (85). It seems that at some point in the evolution of the eukaryotes, the old transcription regulators inherited from the common ancestor of the archaea and eukaryotes have been largely lost. New regulators might have evolved by rapid duplication and divergence of a few survivors and from ubiquitious HTH domains of transposases.

SUPPLEMENTARY MATERIAL

See Supplementary Material available at NAR Online.

ADDENDUM

While this manuscript was under revision, the first complete sequence of a crenarchaeote genome, that of *Aeropyrum pernix*, became available (86). We performed a preliminary tally of predicted DNA-binding domains encoded in this genome. The resulting values are within the range typical of euryarchaeota and most bacteria (Fig. 1) which reinforces our conclusion that a considerable repertoire of DNA-binding domains should have been present in the common ancestor of the archaea. Furthermore, the distinctive layout of the archaeal transcription system is supported by the finding, in *A.pernix*, of orthologs of most of the predicted transcriptional regulators that are conserved in four or three euryarchaeal species. The lists of the Gene Identification numbers of *A.pernix* proteins containing predicted DNA-binding domains are available upon request.

An independent evolutionary study of the archaeal transcription system has been published during the revision of this manuscript (87). Some of the predicted archaeal HTH described here were detected in this work but, in general, the indicated numbers of HTH domains appear to be underestimates.

ACKNOWLEDGEMENTS

We thank Drs K. S. Makarova and R. L. Tatusov for valuable contributions at the initial stages of this analysis.

REFERENCES

1. Pabo,C.O. and Lewis,M. (1982) *Nature*, **298**, 443–447.
2. Anderson,W.F., Ohlendorf,D.H., Takeda,Y. and Matthews,B.W. (1981) *Nature*, **290**, 754–758.
3. Ohlendorf,D.H., Anderson,W.F., Fisher,R.G., Takeda,Y. and Matthews,B.W. (1982) *Nature*, **298**, 718–723.
4. Sauer,R.T., Yocum,R.R., Doolittle,R.F., Lewis,M. and Pabo,C.O. (1982) *Nature*, **298**, 447–451.
5. Dodd,I.B. and Egan,J.B. (1990) *Nucleic Acids Res.*, **18**, 5019–5026.
6. Gribskov,M. and Burgess,R.R. (1986) *Nucleic Acids Res.*, **14**, 6745–6763.
7. Dodd,I.B. and Egan,J.B. (1987) *J. Mol. Biol.*, **194**, 557–564.
8. Henikoff,S., Haughn,G.W., Calvo,J.M. and Wallace,J.C. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 6602–6606.
9. Klemm,J.D., Rould,M.A., Aurora,R., Herr,W. and Pabo,C.O. (1994) *Cell*, **77**, 21–32.
10. Xu,W., Rould,M.A., Jun,S., Desplan,C. and Pabo,C.O. (1995) *Cell*, **80**, 639–650.
11. Kissinger,C.R., Liu,B.S., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990) *Cell*, **63**, 579–590.
12. Ogata,K., Morikawa,S., Nakamura,H., Sekikawa,A., Inoue,T., Kanai,H., Sarai,A., Ishii,S. and Nishimura,Y. (1994) *Cell*, **79**, 639–648.
13. Suzuki,M. and Brenner,S.E. (1995) *FEBS Lett.*, **372**, 215–221.
14. Wilson,K.P., Shewchuk,L.M., Brennan,R.G., Otsuka,A.J. and Matthews,B.W. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 9257–9261.
15. Ramakrishnan,V., Finch,J.T., Graziano,V., Lee,P.L. and Sweet,R.M. (1993) *Nature*, **362**, 219–223.
16. Clark,K.L., Halay,E.D., Lai,E. and Burley,S.K. (1993) *Nature*, **364**, 412–420.
17. Holm,L., Sander,C., Ruterjans,H., Schnarr,M., Fogh,R., Boelens,R. and Kaptein,R. (1994) *Protein Eng.*, **7**, 1449–1453.
18. Martinez-Hackert,E. and Stock,A.M. (1997) *Structure*, **5**, 109–124.
19. Schade,M., Turner,C.J., Lowenhaupt,K., Rich,A. and Herbert,A. (1999) *EMBO J.*, **18**, 470–479.
20. Xing,Y., Guha Thakurta,D. and Draper,D.E. (1997) *Nature Struct. Biol.*, **4**, 24–27.
21. Wolf,Y.I., Aravind,L., Grishin,N.V. and Koonin,E.V. (1999) *Genome Res.*, **9**, 689–710.
22. Koonin,E.V., Mushegian,A.R., Galperin,M.Y. and Walker,D.R. (1997) *Mol. Microbiol.*, **25**, 619–637.
23. Feng,D.F., Cho,G. and Doolittle,R.F. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 13028–13033.
24. Makarova,K.S., Aravind,L., Galperin,M.Y., Grishin,N.V., Tatusov,R.L., Wolf,Y.I. and Koonin,E.V. (1999) *Genome Res.*, **9**, 608–628.
25. Doolittle,W.F. and Logsdon,J.M.,Jr (1998) *Curr. Biol.*, **8**, R209–R211.
26. Qureshi,S.A., Khoo,B., Baumann,P. and Jackson,S.P. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 6077–6081.
27. Thomm,M. (1996) *FEMS Microbiol. Rev.*, **18**, 159–171.
28. Puhler,G., Leffers,H., Gropp,F., Palm,P., Klenk,H.P., Lottspeich,F., Garrett,R.A. and Zillig,W. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 4569–4573.
29. Langer,D., Hain,J., Thuriaux,P. and Zillig,W. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 5768–5772.
30. Eloranta,J.J., Kato,A., Teng,M.S. and Weinzierl,R.O. (1998) *Nucleic Acids Res.*, **26**, 5562–5567.
31. Kyrpides,N.C. and Ouzounis,C.A. (1997) *J. Mol. Evol.*, **45**, 706–707.
32. Hochheimer,A., Hedderich,R. and Thauer,R.K. (1999) *Mol. Microbiol.*, **31**, 641–650.
33. Napoli,A., van der Oost,J., Sensen,C.W., Charlebois,R.L., Rossi,M. and Ciaramella,M. (1999) *J. Bacteriol.*, **181**, 1474–1480.
34. Raumann,B.E., Rould,M.A., Pabo,C.O. and Sauer,R.T. (1994) *Nature*, **367**, 754–757.
35. Suzuki,M. (1995) *Protein Eng.*, **8**, 1–4.
36. Wang,B., Jones,D.N., Kaine,B.P. and Weiss,M.A. (1998) Structure, **6**, 555–569.
37. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N.S.M. and Venter,J.C. (1996) *Science*, **273**, 1058–1073.
38. Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K., Harrison,D., Hoang,L., Keagle,P., Lumm,W., Pothier,B., Qiu,D., Spadafora,R., Vicaire,R., Wang,Y., Wierzbowski,J., Gibson,R., Jiwani,N., Caruso,A., Bush,D., Reeve,J.N. *et al.* (1997) *J. Bacteriol.*, **179**, 7135–7155.
39. Klenk,H.P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D., Richardson,D.L., Kerlavage,A.R., Graham,D.E., Kyrpides,N.C., Fleischmann,R.D., Quackenbush,J., Lee,N.H., Sutton,G.G., Gill,S., Kirkness,E.F., Dougherty,B.A., McKenney,K., Adams,M.D., Loftus,B., Venter,J.C. *et al.* (1997) *Nature*, **390**, 364–370.
40. Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A., Nagai,Y., Sakai,M., Ogura,K., Otsuka,R., Nakazawa,H., Takamiya,M., Ohfuku,Y., Funahashi,T., Tanaka,T., Kudoh,Y., Yamazaki,J., Kushida,N., Oguchi,A., Aoki,K. and Kikuchi,H. (1998) *DNA Res.*, **5**, 55–76.
41. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
42. Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
43. Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) *Nucleic Acids Res.*, **25**, 1665–1677.
44. Walker,D.R. and Koonin,E.V. (1997) *Intelligent Syst. Mol. Biol.*, **5**, 333–339.
45. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) *Nucleic Acids Res.*, **25**, 4876–4882.
46. Felsenstein,J. (1996) *Methods Enzymol.*, **266**, 418–427.
47. Rost,B. (1996) *Methods Enzymol.*, **266**, 525–539.
48. Rost,B., Schneider,R. and Sander,C. (1997) *J. Mol. Biol.*, **270**, 471–480.
49. Guex,N. and Peitsch,M.C. (1997) *Electrophoresis*, **18**, 2714–2723.
50. Kraulis,P.J. (1991) *J. Appl. Crystallogr.*, **24**, 946–950.
51. Hubbard,T.J., Ailey,B., Brenner,S.E., Murzin,A.G. and Chothia,C. (1999) *Nucleic Acids Res.*, **27**, 254–256.
52. Lagrange,T., Kapanidis,A.N., Tang,H., Reinberg,D. and Ebright,R.H. (1998) *Genes Dev.*, **12**, 34–44.
53. Smit,A.F. and Riggs,A.D. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 1443–1448.
54. Franz,G., Loukeris,T.G., Dialektaki,G., Thompson,C.R. and Savakis,C. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 4746–4750.
55. Pietrokovski,S. and Henikoff,S. (1997) *Mol. Gen. Genet.*, **254**, 689–695.

56. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
57. Qiu,X., Verlinde,C.L., Zhang,S., Schmitt,M.P., Holmes,R.K. and Hol,W.G. (1995) *Structure*, **3**, 87–100.
58. Aravind,L., Tatusov,R.L., Wolf,Y.I., Walker,D.R. and Koonin,E.V. (1998) *Trends Genet.*, **14**, 442–444.
59. Liang,C. and Stillman,B. (1997) *Genes Dev.*, **11**, 3375–3386.
60. Koonin,E.V. (1993) *Nucleic Acids Res.*, **21**, 2541–2547.
61. Leonard,C.J., Aravind,L. and Koonin,E.V. (1998) *Genome Res.*, **8**, 1038–1047.
62. Tyrrell,R., Verschueren,K.H., Dodson,E.J., Murshudov,G.N., Addy,C. and Wilkinson,A.J. (1997) *Structure*, **5**, 1017–1032.
63. Gane,P.J., Dunwell,J.M. and Warwicker,J. (1998) *J. Mol. Evol.*, **46**, 488–493.
64. Ponting,C.P. and Aravind,L. (1997) *Curr. Biol.*, **7**, R674–R677.
65. Shan,X. and Kruger,W.D. (1998) *Nature Genet.*, **19**, 91–93.
66. Aravind,L. and Koonin,E.V. (1999) *J. Mol. Biol.*, **287**, 1023–1040.
67. Hall,T.M., Porter,J.A., Young,K.E., Koonin,E.V., Beachy,P.A. and Leahy,D.J. (1997) *Cell*, **91**, 85–97.
68. Pietrokovski,S. (1998) *Protein Sci.*, **7**, 64–71.
69. Takemaru,K., Harashima,S., Ueda,H. and Hirose,S. (1998) *Mol. Cell. Biol.*, **18**, 4971–4976.
70. Takemaru,K., Li,F.Q., Ueda,H. and Hirose,S. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 7251–7256.
71. Aravind,L. and Landsman,D. (1998) *Nucleic Acids Res.*, **26**, 4413–4421.
72. Knight,K.L., Bowie,J.U., Vershon,A.K., Kelley,R.D. and Sauer,R.T. (1989) *J. Biol. Chem.*, **264**, 3639–3642.
73. Gomis-Ruth,F.X., Sola,M., Acebo,P., Parraga,A., Guasch,A., Eritja,R., Gonzalez,A., Espinosa,M., del Solar,G. and Coll,M. (1998) *EMBO J.*, **17**, 7404–7415.
74. Min,Y.N., Tabuchi,A., Womble,D.D. and Rownd,R.H. (1991) *J. Bacteriol.*, **173**, 2378–2384.
75. Qian,X., Gozani,S.N., Yoon,H., Jeon,C.J., Agarwal,K. and Weiss,M.A. (1993) *Biochemistry*, **32**, 9944–9959.
76. Fourmy,D., Dardel,F. and Blanquet,S. (1993) *J. Mol. Biol.*, **231**, 1078–1089.
77. Dauter,Z., Wilson,K.S., Sieker,L.C., Moulis,J.M. and Meyer,J. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 8836–8840.
78. Glaser,P., Presecan,E., Delepierre,M., Surewicz,W.K., Mantsch,H.H., Barzu,O. and Gilles,A.M. (1992) *Biochemistry*, **31**, 3038–3043.
79. Frey,M., Sieker,L., Payan,F., Haser,R., Bruschi,M., Pepe,G. and LeGall,J. (1987) *J. Mol. Biol.*, **197**, 525–541.
80. Vargas,M., Kashefi,K., Blunt-Harris,E.L. and Lovley,D.R. (1998) *Nature*, **395**, 65–67.
81. Holstege,F.C., van der Vliet,P.C. and Timmers,H.T. (1996) *EMBO J.*, **15**, 1666–1677.
82. Tan,S. and Richmond,T.J. (1998) *Curr. Opin. Struct. Biol.*, **8**, 41–48.
83. Gibson,T.J., Thompson,J.D., Blocker,A. and Kouzarides,T. (1994) *Nucleic Acids Res.*, **22**, 946–952.
84. Rodriguez-Monge,L., Ouzounis,C.A. and Kyrpides,N.C. (1998) *Trends Biochem. Sci.*, **23**, 169–170.
85. Chervitz,S.A., Aravind,L., Sherlock,G., Ball,C.A., Koonin,E.V., Dwight,S.S., Harris,M.A., Dolinski,K., Mohr,S., Smith,T., Weng,S., Cherry,J.M. and Botstein,D. (1998) *Science*, **282**, 2022–2028.
86. Kawarabayasi,Y., Hino,Y., Horikawa,H., Yamazaki,S., Haikawa,Y., Jin-no,K., Takahashi,M., Sekine,M., Baba,S., Ankai,A., Kosugi,H., Hosoyama,A., Fukui,S., Nagai,Y., Nishijima,K., Nakazawa,H., Takamiya,M., Masuda,S., Funahashi,T., Tanaka,T., Kudoh,Y., Yamazaki,J., Kushida,N., Oguchi,A., Kikuchi,H. *et al.* (1999) *DNA Res.*, **6**, 83–101, 145–152.
87. Kyrpides,N.C. and Ouzounis,C.A. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 8545–8550.