ELSEVIER

FEMS Microbiology Reviews 29 (2005) 231–262

FEMS MICROBIOLOGY Reviews

www.fems-microbiology.org

The many faces of the helix-turn-helix domain: Transcription regulation and beyond ☆

L. Aravind *, Vivek Anantharaman, Santhanam Balaji, M. Mohan Babu, Lakshminarayan M. Iyer

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Room 5N50, Bethesda, MD 20894, USA

Received 20 November 2004; received in revised form 22 December 2004; accepted 23 December 2004

First published online 28 January 2005

Abstract

The helix-turn-helix (HTH) domain is a common denominator in basal and specific transcription factors from the three superkingdoms of life. At its core, the domain comprises of an open tri-helical bundle, which typically binds DNA with the 3rd helix. Drawing on the wealth of data that has accumulated over two decades since the discovery of the domain, we present an overview of the natural history of the HTH domain from the viewpoint of structural analysis and comparative genomics. In structural terms, the HTH domains have developed several elaborations on the basic 3-helical core, such as the tetra-helical bundle, the winged-helix and the ribbon-helix–helix type configurations. In functional terms, the HTH domains are present in the most prevalent transcription factors of all prokaryotic genomes and some eukaryotic genomes. They have been recruited to a wide range of functions beyond transcription regulation, which include DNA repair and replication, RNA metabolism and protein–protein interactions in diverse signaling contexts. Beyond their basic role in mediating macromolecular interactions, the HTH domains have also been incorporated into the catalytic domains of diverse enzymes. We discuss the general domain architectural themes that have arisen amongst the HTH domains as a result of their recruitment to these diverse functions. We present a natural classification, higher-order relationships and phyletic pattern analysis of all the major families of HTH domains. This reconstruction suggests that there were at least 6–11 different HTH domains in the last universal common ancestor of all life forms, which covered much of the structural diversity and part of the functional versatility of the extant representatives of this domain. In prokaryotes the total number of HTH domains per genome shows a strong power-equation type scaling with the gene number per genome. However, the HTH domains in two-component signaling pathways show a linear scaling with gene number, in contrast to the non-linear scaling of HTH domains in single-component systems and sigma factors. These observations point to distinct evolutionary forces in the emergence of different signaling systems with HTH transcription factors. The archaea and bacteria share a number of ancient families of specific HTH transcription factors. However, they do not share any orthologous HTH proteins in the basal transcription apparatus. This differential relationship of their basal and specific transcriptional machinery poses an apparent conundrum regarding the origins of their transcription apparatus.
© 2005 Federation of European Microbiological Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Helix-turn-helix; DNA-binding; Transcription regulation; Evolution; Two-component systems; Structure

☆ Edited by Mark J. Pallen.

* Corresponding author. Tel.: +1 301 594 2445; fax: +1 301 435 7794.
E-mail address: aravind@ncbi.nlm.nih.gov (L Aravind).

0168-6445/$22.00 © 2005 Federation of European Microbiological Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.femsre.2004.12.008

Contents

1. Introduction . . . 232
2. Structural scaffold of the HTH domain and its diverse elaborations . . . 233
   2.1. HTH domains with a simple three-helical bundle and its extensions . . . 234
   2.2. The winged HTH domain . . . 235
   2.3. Other highly modified variants of the HTH domain . . . 236
3. General and specific aspects of the domain architectures of HTH proteins . . . 236
   3.1. Simple architectures involving the HTH domain . . . 237
   3.2. Combinations of HTH with other nucleic acid binding domains and protein–protein interaction domains . . . 241
   3.3. Combinations of the HTH domain with catalytic domains . . . 241
   3.4. Architectures related to two-component, PTS and serine/threonine kinase signaling . . . 243
   3.5. Architectures related to single-component signaling . . . 244
   3.6. Unusual functional adaptations of the HTH domain . . . 245
4. The evolutionary classification of HTH domains . . . 246
   4.1. Lineages of basic tri-helical HTH domains . . . 246
   4.2. The tetra-helical HTH superclass and its derivatives . . . 248
   4.3. The wHTH superclass . . . 249
   4.4. Other miscellaneous families of HTH domains . . . 252
5. Proteome-wide demographic trends of HTH domains . . . 252
6. General considerations on the natural history of the HTH fold and implications for the evolution of transcription . . . 254
7. General conclusions . . . 256
8. Supplementary material . . . 257
   References . . . 257

1. Introduction

The general paradigms for the processes of transcription initiation and regulation first emerged from the pioneering studies on gene expression in bacteria and phages [1]. Transcription in bacteria was found to be catalyzed by a single multi-subunit RNA polymerase, which is recruited to promoters by means of a DNA-binding protein, the sigma factor that recognizes specific sequences upstream of genes [2–4]. The sigma factor and the RNA polymerase, together, constitute the basal transcription apparatus that is required for the baseline transcription of all genes. Early studies, especially in the *Bacillus subtilis* sporulation model, suggested that there may be more than one sigma factor that might recruit the catalytic core of the RNA polymerase to alternative sets of genes. This provided a mechanism for regulating the broad changes in gene expression, which correlate with the different developmental or differentiation states of a bacterium [5,6]. A number of early studies on metabolic regulation in bacteria also indicated that there are other regulatory DNA-binding proteins that act as switches to control the expression of specific smaller sets of genes. These sets of genes are often collinear on the chromosome, and encode components of a common pathway for the utilization of a particular environmental metabolite (e.g. lactose), or constitute interacting components of a developmental pathway (e.g. lytic or lysogenic development of phages). These regulatory proteins, termed the specific transcription factors, were found to belong to two distinct functional types: (1) those which negatively regulated transcription of their target gene (repressors) and (2) those which positively regulated transcription of their target genes (activators) [1]. The affinities of the specific transcription factors for their target sequences on DNA were often found to be dependent on their interaction with low-molecular weight compounds (effectors), which bound to them, or phosphorylation and other post-transcriptional modifications [1]. When transcription in eukaryotes was first investigated, several differences in the subunit composition and architecture of the basal transcription machinery and specific transcription factors were noted [7,8]. However, the basic regulatory mechanisms in bacterial transcription, which were brought to light in early studies, remained applicable in a generic sense across the entire Tree of Life [1].

The pioneering investigations of Matthews, Ohlendorf, Sauer, Doolittle and co-workers in 1982 provided the first glimpse of the features unifying diverse transcription regulators [9–13]. They showed that the phage lambda transcription regulators, cro and the cI repressor, and lacI, the lactose operon repressor, shared a similar tri-helical DNA-binding domain. The 2nd and the 3rd helices of this tri-helical domain constituted a Helix-Turn-Helix motif, and this motif was shown to be a critical determinant of their interaction with DNA. Thus, these DNA-binding domains came to be

referred as the helix-turn-helix (HTH) domain. Sequence and secondary structure analysis by these pioneering workers suggested that the HTH domain was a common DNA-binding motif found in several other bacterial repressors as well as activators, such as the cAMP-dependent catabolite activator protein. They perceptively suggested that all these DNA-binding domains have descended from a common ancestor, through duplication and divergence, thereby generating the diversity of transcription regulators that regulate bacterial and phage genes [9,11]. Subsequent sequence analysis revealed that DNA recognition by sigma factors was also mediated by HTH domains, similar to those observed in the specific transcription factors [14–16]. In the second half of the 1980s the conserved domains of transcription factors regulating eukaryotic development and differentiation, namely the homeodomains and Myb domains, were also noted to possess the HTH fold [17,18]. These and other investigations suggested the general significance of this module in DNA–protein interactions across a wide phylogenetic spectrum [18–21].

An explosion of structural studies in the 1990s, while strengthening the basic structure–function relation between the HTH domain and DNA binding, also produced a large amount of data regarding the diversity of DNA–protein interactions mediated by different versions of the HTH domain. A number of sequence and structural analysis studies also uncovered HTH modules in several specific eukaryotic transcription factors, chromatin proteins like histone H1, and basal transcription factors such as TFIIB and TFIIE [22–30]. These findings lead to the idea that the HTH domain is probably one of the most ancient conserved features of transcription apparatus, which was already present in the last universal common ancestor of all extant life forms (LUCA). Studies during this period also showed that although archaea and eukaryotes possess a similar basal transcription machinery, the specific transcription factors of the former are clearly closer to those of the bacteria than the eukaryotes [30–34]. The other major development in the later half of the 1990s was the birth of the genomic era, unleashing the power of comparative genomics. Starting with the earliest comparative genomic studies it became apparent that the HTH domain was a highly prevalent domain in prokaryotes [35]. Comparative analysis also helped in identifying several major monophyletic assemblages of HTH transcription factors, each distinguished by their own distinctive sequence and structure features (for example see Refs. [36–41]). These classes often showed one or more distinctive *domain architectures* – i.e. the fusion of the HTH domain with additional globular domains in the same polypeptide. These globular domains, which are linked to the HTH show a bewildering diversity, and point to the immense variety of functional contexts in which the HTH domain may be deployed. The combined use of genome sequence data and high-throughput expression data also cast light on the multi-level transcriptional regulatory networks in prokaryotes in which these HTH-containing proteins functionally interact to maintain a particular transcriptional state in the cell [42,43].

Over the years, the recruitment of the HTH domains to biological functions beyond transcriptional regulation has also become apparent. Some of these proteins, participating in functions such as DNA repair and RNA metabolism, exploit the nucleic acid-binding properties of the HTH just as in the case of transcriptional regulation (for example see: [44–48]). However, there are other instances where the HTH may be adapted to very different functions, such as mediating specific protein–protein interactions, or as a structural unit of a larger enzymatic domain [49–51].

In this article we aim at providing a synthetic overview of the structural, functional and evolutionary diversity of the HTH domain from the vantage point of over two decades of intense investigation. We currently enjoy unprecedented advantages due to an enormous amount of genomic data, numerous high resolution structures, functional studies and sensitive computational tools to capture the highlights of the natural history of HTH domains. Our focus is principally, but not exclusively, on versions of the domain found in the prokaryotic super-kingdoms. We first discuss the structural diversity seen in this fold, followed by a discussion of the unifying themes in domain architectures of HTH proteins and their significance to different biological functions. We next provide a higher-order natural classification of these domains and discuss their genome-wide demography in light of the general adaptive tendencies that can be inferred from comparative genomics. In this context we also consider the adaptation of the HTH to functional roles beyond transcriptional regulation. Finally, we discuss the relevance of the information gleaned from these diverse areas in reconstructing the origin and evolution of transcription and its regulation.

2. Structural scaffold of the HTH domain and its diverse elaborations

The basic HTH domain is a simple fold comprised of three core helices that form a right-handed helical bundle with a partly open configuration. When it is displayed by placing the third helix in the front and in the horizontal orientation, the 3 helices of the domain form an approximately triangular outline (Fig. 1). We use this as the default orientation for all further illustrations and discussions. The characteristic sharp turn, which is a defining feature of this domain, is situated between the 2nd and the 3rd helix, and typically does not

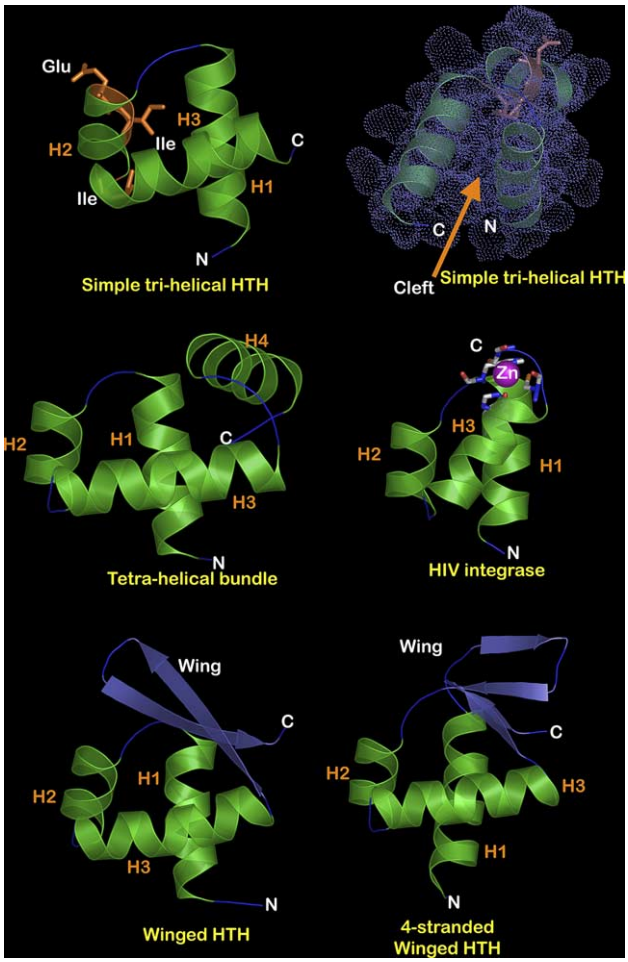

Fig. 1. Salient structural and features of the HTH domains and its chief variants. Helices are shown in green and labeled with an 'H', strands are in blue. The top panel shows a representative of a simple tri-helical HTH domain (PDB: 1k78). Residues that are strongly conserved across all HTH domains are shown in the top panel in stick representation. On the right a surface view of the same domain with the shallow cleft, typical of the partly "open" configuration of the tri-helical bundle, is shown using the surface view. In the middle panel a representative of the tetra-helical bundle (PDB: 1a.4) is shown to the left, while the metal-chelating HTH domain of the retroviral integrases is show to the right (HIV integrase, PDB: 1k6y). In the bottom panel a simple 2-stranded winged HTH (pdb: 1smt) is shown to the left, while a 4-stranded winged HTH (PDB: 1cgp) is shown to the right.

tolerate insertions or distortions. However, the loop between helix-1 and helix-2 shows far greater variability and may accommodate several modifications in the different classes of HTH domains. Furthermore, diverse N- and C-terminal extensions to the core tri-helical bundle are also encountered in several classes of HTH domains. The partly open configuration of the bundle results in a shallow cleft between helix-3 and helix-1 on the side opposite to helix-2. This cleft appears to have acted as a structural niche that favored the evolution of additional structural elements to pack into it via hydrophobic interactions. Thus, most of the extension to the core HTH domain are structural elements that appear to have evolved to generate a more closed configuration by interacting with the cleft (Fig. 1; see below for further discussion). The 3rd helix, known as the recognition helix, typically forms the principal DNA–protein interface by inserting itself into the major groove of the DNA [19,24,52]. Nevertheless, the individual residues involved in DNA contacts may widely vary across the fold. Additional secondary contacts with DNA may be mediated by other parts of the structure or even extensions outside of the core HTH domain, such as a basic patch at the N-terminus of helix-1 [52].

Despite the great sequence diversity observed in this fold, there are a few sequence elements that are widely conserved in members of the fold. The most characteristic of these is the "shs" pattern (where 's' is a small residue, most frequently glycine in the first position, and 'h' is a hydrophobic residue) that lies in the turn between helix-2 and helix-3 of the core HTH structure. The other well-conserved signatures is the "phs" (where 'p' is a charged residue, most frequently glutamate) that is present in helix-2 (Fig. 1). The conserved hydrophobic residues in these motifs, together with at least two other conserved-hydrophobic residues seen in helix-1 and helix-3 localize to the interior and form the characteristic hydrophobic core that stabilizes the domain (Fig. 1). The conservation of these elements across diverse members of this fold from the 3 super-kingdoms of life, taken together with the conserved structure–function associations of this fold strongly supports the monophyletic origin of HTH domains [9,11] from a common ancestor that bore the above-mentioned sequence features.

Based on their distinctive features, members of the HTH fold can be divided into two major structural classes and a few other highly derived structural classes that contain drastic alterations of the core HTH domain.

2.1. HTH domains with a simple three-helical bundle and its extensions

The simplest version of the HTH domain, the *basic tri-helical version*, is comprised entirely of the three core helices with no additional elaborations (Figs. 1 and 2). This configuration appears to be closest to the ancestral state of the domain and is widely seen across the three super-kingdoms of life. This version is represented by the HTH domains such as the region-3 and region-4 of the sigma factors, the RPB10 subunit of the DNA-dependent-RNA polymerases of the archaeo-eukaryotic lineage, bacterial transcription factors of the FIS family, the eukaryotic Homeo, POU and Myb domains, the eukaryotic BRCA2 tower domain, and the paired module and related HTH domains of transposases and resolvases (Fig. 2). A distorted version of the basic tri-helical HTH domain is also seen in the zinc-chelating domain

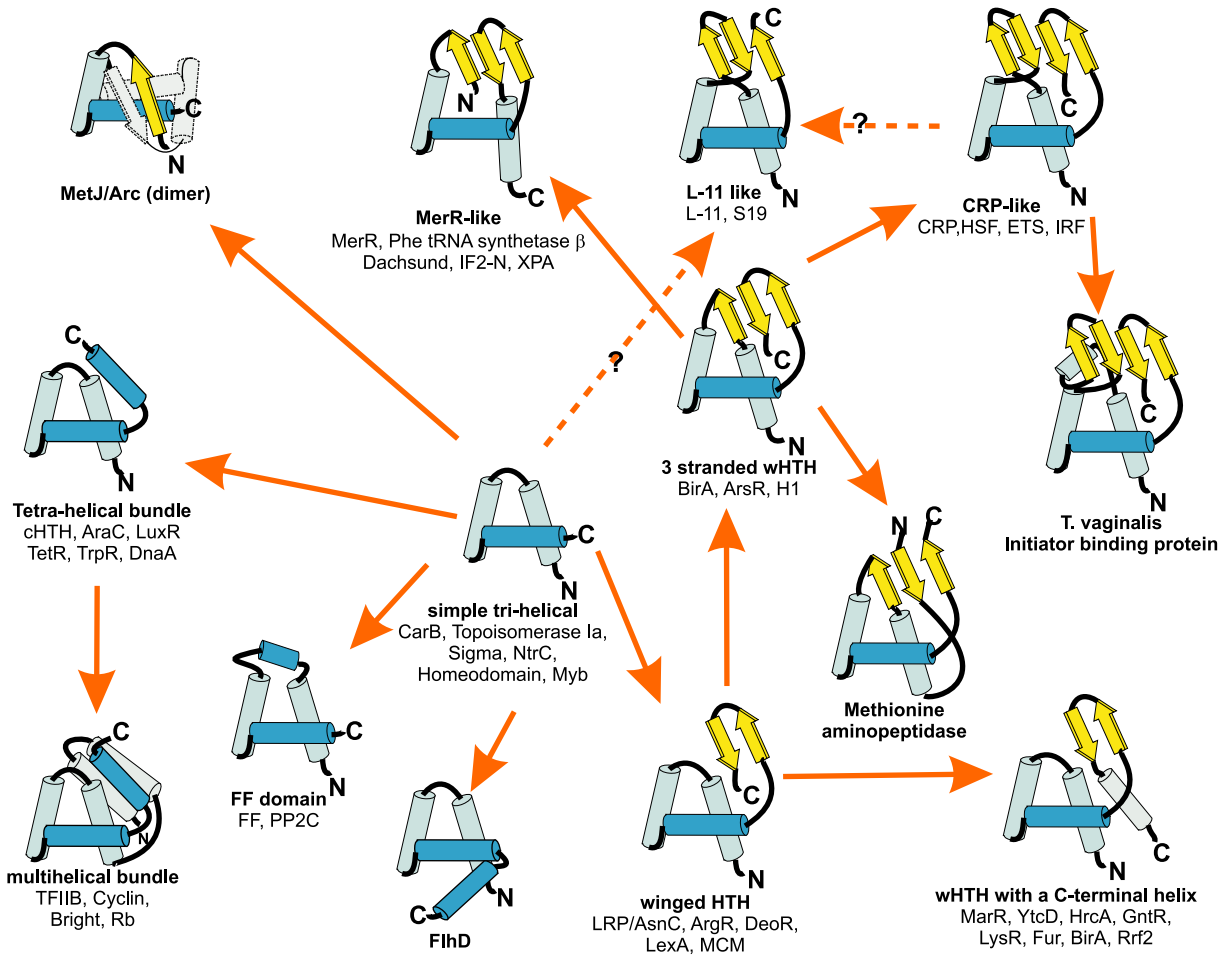

Fig. 2. A pathway showing the structural elaboration of the simple HTH domain into its diverse versions. Strands are shown as yellow arrows with the arrow heads at the C-terminus, helices are shown as blue cylinders. The orange arrows show the probable routes of transformation of the HTH fold. The two possible origins of the L11 like HTHs are shown by dotted arrows with question marks. The topologies have been constructed using the following PDB enteries: (1) simple trihelical bundle: 2hdd (2) FF domain: 1h40 (3) Tetrahelical bundle: 1d5y (4) Multihelical bundle: 1ais (5) MetJ/Arc: 2cpg (6) MerR: 1jbg (7) L11: 1mms (8) CRP-like: 1cgp (9) *T. vaginalis* initiator: 1pp7 (10) 3 stranded wHTH: 2dtr (11) Methionine aminopeptidase: 1xgs (12) winged HTH: 1i1g (13) wHTH with a C-terminal helix: 1jgs.

of several retroviral integrases [53], in which helix-3 is packed more closely against the other helices by means of a Zn ion chelated by conserved cysteines and histidines at the ends of helix-1 and helix-3 (Fig. 1). Likewise, in RPB10, a set of zinc-chelating cysteines help in stabilizing an N-terminal loop against helix-3 [54].

The *tetra-helical version* of HTH domain is an elaboration of the basic tri-helical version and is characterized by an additional C-terminal helix which packs against the shallow cleft formed due to the open configuration of the tri-helical core (Figs. 1 and 2). Several major families of prokaryotic transcription factors display this version of the domain. The *multi-helical version*, typified by the archaeo-eukaryotic basal transcription factor TFIIB, is a further elaboration of the tetra-helical HTH, wherein two additional helices have been added to the N-terminus of the tetra-helical core, resulting in a larger globular helical bundle (Fig. 2). The Bright (ARID) domain, a eukaryotic DNA-binding domain [55], is another more divergent version of the TFIIB-like HTH domains. This version of the fold is very infrequent in the bacteria, and is represented by a circularly permuted version seen in the sporulation regulator Spo0A from spore-forming Gram-positive bacteria [56]. Other versions, which represent relatively infrequent elaborations of the basic tri-helical version, are the KorB-like HTHs [57] and the FlhD-like HTHs [58]. The former version is characterized by an additional N-terminal helix that packs against the basic 3-helix core and the later contains a C-terminal helical extension that packs very differently from the helical extension seen in the above-mentioned classical tetra-helical forms (Fig. 2).

2.2. The winged HTH domain

The winged HTH (wHTH) domains are distinguished by presence of a C-terminal β-strand hairpin unit (the

wing) that packs against the shallow cleft of the partially open tri-helical core [24,25]. The simplest versions of the wHTH domains contain a tight helical core similar to basic tri-helical version followed by the two-strand hairpin (Figs. 1 and 2). However, many wHTH domains display further serial elaborations of the β-sheet. In the 3-stranded version, the loop between helix-1 and helix-2 assumes an extended configuration and is incorporated as the 3rd strand in the sheet, via hydrogen bonding with the basic C-terminal hairpin (Fig. 2). In the 4-stranded version, the linker between helix-1 and helix-2 also forms a hairpin with two β-strands, and along with the C-terminal wing forms an extended β-sheet (Fig. 2). In versions that bind nucleic acids, the wing often provides an additional interface for substrate contact, typically by interacting with the minor groove of DNA through charged residues in the hairpin [24,25,27]. The two- and three-stranded versions of the wHTH are encountered in DNA-binding domains of some of the largest families of prokaryotic transcription factors, as well as several eukaryotic DNA-binding domains. The single-strand RNA-binding La domains also have a version of the wHTH fold with a slightly extended and variable insert between helix-1 and helix-2. An unusual version of the 4-stranded wHTH domain, with an additional small helical insert after helix-1, is observed in the orphan transcription-initiator-sequence-binding protein from the parabasalid protist, *Trichomonas vaginalis* [59] (Fig. 2). The Fur family of bacterial transcription factors, which is involved in metal-responsive transcriptional regulation, shows a regular 2/3-stranded wHTH domain, but the wing is incorporated into a large sheet formed with additional C-terminal strands. A circularly permuted version of the basic wHTH domain, with one of the strands of the wing moved to the N-terminus, is seen in the C-terminal accessory domain of the methionine aminopeptidase-2 (Fig. 2) [60].

2.3. Other highly modified variants of the HTH domain

The *MetJ-Arc family* (also known as ribbon-helix-helix/RHH family) of transcription factors is, to date, known only from the prokaryotes. They are obligate dimers, which pair through a single N-terminal strand, and possess a C-terminal helix-turn-helix unit (Fig. 2). The organization of the C-terminal helical unit is identical to corresponding unit in the classical versions of the HTH domain, and it shows the characteristic conserved sequence features of the HTH domain [61]. The sheet formed by the N-terminal strands of the domain is inserted into the major groove of DNA [61]. Mutagenesis experiments have shown that even single mutations in the N-terminal strand convert the strand of the RHH domain to a helix, and result in a structural packing that is closer to the canonical HTH domain [62]. This result, together with the notable structural and sequence similarities with the HTH domains, suggest that the RHH domain was derived from the HTH domain through conversion of the N-terminal helix to a strand. Concomitant with this modification, the N-terminal strand, which came to lie atop the recognition helix, appears to have taken up the principal DNA-binding role of this protein.

The DNA-binding domain of the bacterial transcription regulator MerR defines an aberrant derivative of a 3-stranded version of the wHTH domain, in which helix-1 has been lost (Fig. 2). In addition to the MerR family of bacterial repressors this version of the fold is seen in a wide variety of phage, bacterial and eukaryotic DNA-binding proteins and translation factors. The topoisomerase II family has two copies of wHTH domains [63]. One of these wHTH domains contains a large insert between helix-1 and helix-2. This insert contains an extended β-sheet structure that forms a brace around double-stranded DNA. Also present in this insert is a second wHTH domain which is circularly permuted with helix-1 occurring to the C-terminus of the wing. The two structurally related ribosomal proteins L11 and S18 represent another distinctive derivative of wHTH domain. While they share a β-strand hairpin between helix-1 and helix-2 with the 4-stranded wHTH, unlike the latter they possess only a single C-terminal strand (Fig. 2). A highly modified HTH domain seen in the catalytic domain of the phage integrases shows a rare insertion between helix-2 and helix-3 of the core domain [63]; however, this insert does not distort the structure of the core. The FF-domain [64] and the C-terminal domain of the PP2C protein phosphatase [65] define another highly modified version of the HTH domain that is currently only known from eukaryotes. This version domain shows the insertion of a helical insert between helix-1 and helix-2 resulting a different packing of the elements and a distortion of the fold (Fig. 2).

3. General and specific aspects of the domain architectures of HTH proteins

HTH domains are combined with other domains in the same protein giving rise to an astonishing array of domain architectures (Table 1 and Fig. 3). Despite the diversity, all the architectures can be classified into a small number of generic architectural classes, the members of each class being unified by certain general organizational and functional principles. These generic architectural classes illustrate how natural selection has convergently engineered similar functional solutions using a relatively small repertoire of domains [66]. Hence, the more highly populated architectural classes

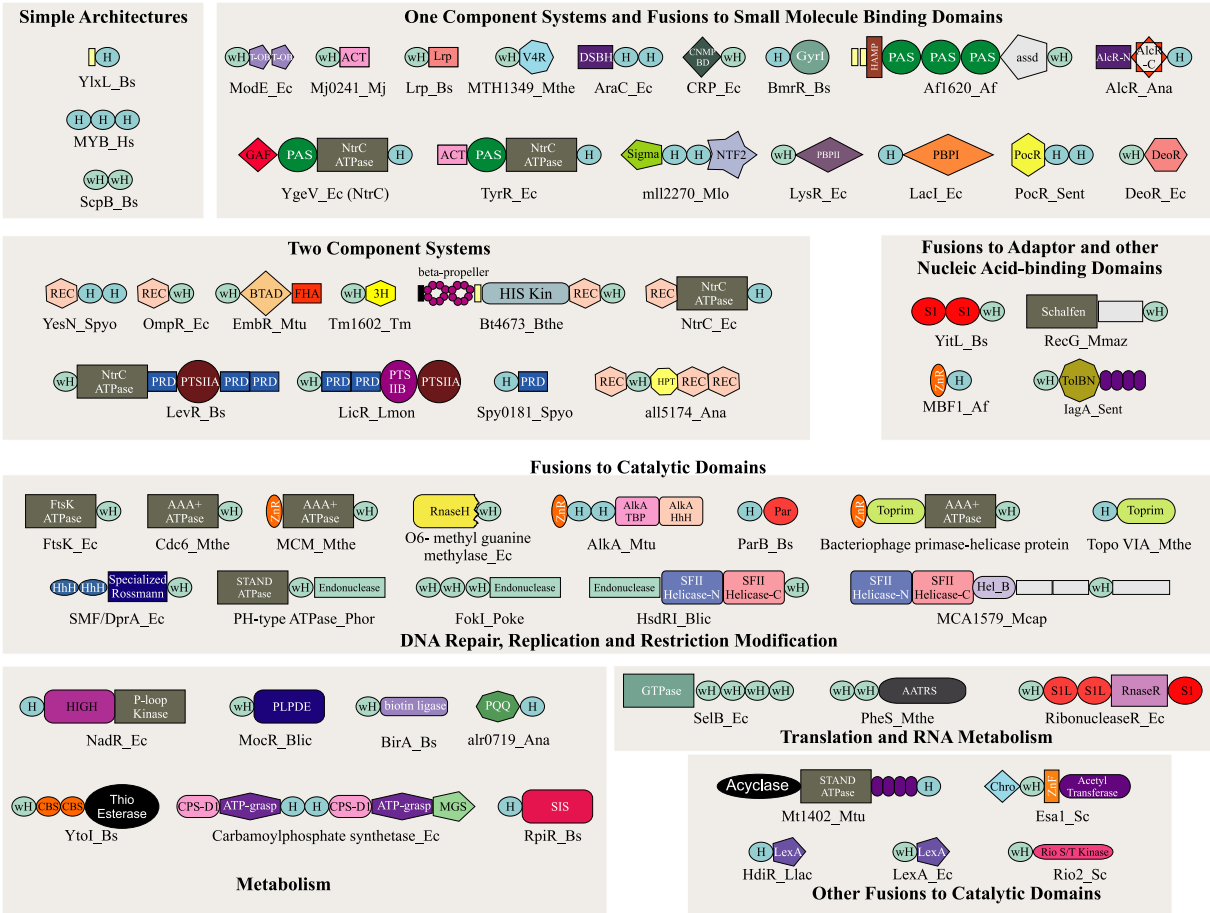

Fig. 3. Examples of domain architectures of proteins with HTH domains. The 3 panels show the HTH domains in one component systems, those in two component systems and those fused to enzymatic domains. Below each protein the name of the protein or its family is indicated. The domain names and abbreviations are as shown in the Table 1. Additional abbreviations are H – any HTH domain; wH – winged HTH; Assd – archaeal specific signaling domain; Acyclase – adenylyl cyclases; AATRS – aminoacyl tRNA synthetase; TPR – tetratricopeptide repeats; HAMP – domain present in histidine kinases, adenylyl cyclases, methyl-accepting proteins and phosphatases, chro-chromodomain BTAD- conserved domain found in bacterial signaling proteins, CPS-D1, the domain 1 of the large subunit of the carbamoylphosphate synthetases, MGS- the methylglyoxal synthase domain that binds ornithine in the carbamoylphosphate synthetases, ParB- the OB-fold nuclease domain seen in ParB proteins, B-hel- a conserved beta-barrel seen in the Lhr helicases. The transmembrane regions are indicated by yellow boxes. The grey boxes represent uncharacterized globular domains that are unique to a particular family of proteins. Af: *Archaeoglobus fulgidus*, Ana: *Nostoc* sp., Blic: *Bacillus licheniformis*, Bs: *Bacillus subtilis*, Bthe: *Bacteroides thetaiotaomicron*, Ec: *Escherichia coli*, Llac: *Lactococcus lactis*, Lmon: *Listeria monocytogenes*, Mjan: *Methanococcus jannaschii*, Mlo: *Mesorhizobium loti*, Mmaz: *Methanosarcina mazei*, Mcap: *Methylococcus capsulatus*, Mthe: *Methanothermobacter thermautotrophicus*, Mtu: *Mycobacterium tuberculosis*, Phor: *Pyrococcus horikoshii*, Poke: *Planomicrobium okeanokoites*, Tm: *Thermotoga maritima*, Sent: *Salmonella enterica*, Spyo: *Streptococcus pyogenes*, Hs: *Homo sapiens*, Sc: *Saccharomyces cerevisiae*.

are likely to represent the more successful functional solutions in which the HTH domain has been involved.

3.1. Simple architectures involving the HTH domain

The simplest architectures involving the HTH domain are seen in certain proteins related to the cI repressors (e.g. the archaeal repressors typified by AF1793), most proteins of the MetJ-Arc superfamily and the Fis proteins. These proteins are almost entirely comprised of just a standalone HTH, and might, at best, have some small extensions that play a role in dimerization or interactions with other components of the transcriptional machinery. The eukaryotic ribosomal protein S10 and the histone H1 respectively contain RNA and DNA binding versions of the wHTH domain, which is fused to low-complexity sequence that forms a non-globular tail [67]. Such non-globular extensions that play a role in non-specific nucleic acid contacts and protein contacts during transcription activation are very common in eukaryotic transcription factors with HTH domains [41]. A family of bacterial proteins typified by the *B. subtilis* sigma D regulator YlxL (SwrB) [68,69] contains a HTH domain fused to a N-terminal transmembrane region (Fig. 3). These HTH proteins might regulate transcription under the influence of signaling events associated with the cell membrane. The next level of

Table 1
Globular domains frequently linked to the HTH domain in the same polypeptide

Domain	Structure	Placement of HTH[a]	Comments
Domains in two-component, phosphorelay and S/T kinase signaling cascades			
REC (receiver domain)	α/β. [Flavodoxin-like topology]. PDB 1NTR	1.19N; 78.35C	Is phosphorylated on an aspartate residue by the histidine kinase dependent phospho-relay system. Typically found fused to OmpR and LuxR-family (in NarL-like proteins) HTH domains
Histidine kinase	α + β. PDB 1BXD	0.10N; 0.03C	Usually occurs as a standalone subunit, but on few instances is fused to the downstream transcription factor with receiver and HTH domains
3H	α + β fold	1.00N	A domain sharing a common fold with the HPr domain of the PTS system. It has 3 conserved histidine residues that are likely to be phosphorylated
FHA	β fold PDB 1LGQ	0.23N; 0.16C	A phosphoserine/threonine peptide-binding domain. Combinations with HTHs are expanded mainly in Actinomyctes, which have numerous serine/threonine kinases
HPT (histidine phosphotransfer domain)	α helical domain PDB 1QSP	–	Receives a phosphate on a conserved histidine in the two-component relay. Found fused to OmpR family wHTH domains along with Receiver domains in cyanobacteria
PRD (PTS Regulatory Domain)	α helical domain PDB 1H99	3.84N	Typically phosphorylated by the EIIB and HPr. It is often found fused to the HTH in the LevR-subfamily with the NtrC-type AAA+ ATPases
PTSIIA	α + β	–	A conserved domain of the PTS system that receives the phosphate in the phosphor-relay and regulates the sugar permeases
PTSIIB	α + β PDB 1BLE	–	A conserved domain of the PTS system that receives the phosphate in the phospho-relay and phosphorylates the PRD domain
SMBDs typically associated with one-component systems			
ACT (aspartokinase, chorismate mutase, TyrA domain)	α + β sandwich [RRM-like fold]. PDB 1PSD	0.19N	Binds various small molecule ligands such as amino acids and purine derivatives
Lrp-C	α + β sandwich [RRM-like fold]	15.55N	This domain is exclusively found at the C-termini of proteins of the Lrp family and is closely related to the ACT fold. It is likely to be an amino-acid-sensing derivative of the ACT fold which is unique to the Lrp proteins.
ArgR-C	α + β. DcoH fold. PDB 1XXA	3.35N	The amino acid-sensing domain exclusively found fused the ArgR-family of HTH domains
DSBH (Double-stranded β-helix domain)	A double-stranded helix of strands. PDB: 2ARC.	6.45N; 14.68C	A widespread fold with numerous enzymes that typically function as oxidases or oxygenases. Most members of this fold contain 3 conserved residues that play a role in ligand-recognition, as well as catalysis in the enzymatic versions of the fold. Binds carbohydrates, oxalate and amino acids
Boa1-N	All β	–	Domain found at the N-terminus of the Boa1-like archaeal HTHs that may be distantly related to the DSBH domain
AlcR-N	Two distinct sub-structures with all β and all α sub-structures	1.65C	This domain shows an N-terminal all β sub-structure that is likely to form a DSBH structure and a C-terminal all-α unit (Fig. 6). A lineage specific expansion is seen in *Nostoc* sp. (17 proteins). Many versions contain a conserved cysteine which may be the site for ligand-interaction or attachment of a prosthetic group
cNMPBD (cNMP-binding domain)	A double-stranded β-barrel domain. PDB: 1RGS	7.65 C	Binds 3′-5′ cyclic nucleotides and typically found fused to the C-terminus of the Crp-like HTH domains
CBS (Cystathionine beta-synthase)	α/β; in nearly all cases, a dimer. PDB 1ZFJ	2.61N	A ligand-binding domain that probably binds a wide range of ligands which may include purine nucleotides and cyclic nucleotides. It is found fused in different proteins to both wHTH and cI-like tetra-helical HTH domains
UTRA (UbiC, transcription regulator associated domain)	α + β. PDB 1FW9	13.39N	Versions of this domain appear to accommodate a wide range of ligands that include amino acids, sugars, fatty acids and alkylphosphonate. Always found fused to a HTH domains of the HutC/FarR subfamily of the GntR family. Shares a fold with the bacterial Chorismate Lyases

DeoR-C	α/β PDB 1LK7	0.48N	This domain shares a common fold with the phosphosugar isomerases such as D-ribose-5-phosphate isomerase. It acts as the sugar-binding domain of the DeoR family of transcription factors
DtxR-N	α-helical domain followed by SH3-like barrel. PDB 2DTR	3.90N	This domain is found exclusively in the iron sensors of the DtxR-Fur family of HTH regulators
FadR-C	α-helical. PDB 1H9G	23.58 N; 0.35C	A unique seven-helical fold found fused to the subfamily of the GntR family typified by the FadR protein
PBP-I (periplasmic binding protein type-I domain)	α/β PDB: 1JWL	29.35N; 1.35C	Members of this fold bind diverse small molecule ligands. Usually found in the N-termini of the LacI family
PBP-II (periplasmic binding protein type-II domain)	α/β PDB: 1AL3	87.97N; 0.61C	Members of this fold bind diverse ligands like N-acetylserine, thiosulphate, amino acids, carbohydrate. It is usually found in the C terminus of LysR family
T-OB	All β PDB: 1B9M	1.23N	A modified version of the OB fold that typically functions as an obligate dimer. Found in ModE family where it binds molybdate
PAS (Per-Arnt-Sim domain)	α/β. PAS-like fold PDB: 2PYP	13.61N; 3.48C	A widely utilized superfamily of ligand-binding domains that bind a range of ligands such as heme, Flavin nucleotides, cinnamate. The ligand-binding domains of the IclR family are divergent versions of the PAS domain
GAF (cGMP phospho-diesterase, Adenylate cyclase, FhlA domain)	α/β. PAS-like fold	0.16N; 2.10C	Another widely utilized superfamily family of ligand-binding domains sharing a common fold with the PAS domain. Binds ligands such as cNMPs, tetrapyrroles and formate. Often found at the N-terminus of NtrC-like proteins
TraR-N	α/β. PAS-like fold. PDB: 1L3L	1.90C	N-terminal domain of *Agrobacterium* TraR transcription factor. It is always found at the C-terminus of several transcription factors of the LuxR family
PocR	Predicted PAS-like fold	0.23C	Domain found fused to the AraC family HTH domains in the 1,2-propanediol-dependent transcription. It also occurs in other contexts fused to diverse signaling domains. Secondary structure predictions suggest that it adopts a PAS-like fold
NTF2	α + β	–	A domain found in several enzymes like Steroid isomerases, Scytalone dehydratase and the carotenoid binding protein. Likely to function as a SMBD fused to certain sigma factors
GyrI	α + β (contains a duplication of SHS2 modules)	4.55N	Frequently found associated with HTH domains of the MerR family. The drugbinding domain of BmrR
Enzymatic domains			
NadR nucleotidyl transferase domain (HIGH)	α/β. HUP fold; PDB 1LW7	0.48N	The catalytic domain functions in the adenylation of the nicotinamide mononucleotide in NAD biosynthesis
mlr6529-C module with metalloprotease-like and metal-chelating domains	Contains two distinct sub-domains An N-terminal all α-unit and a C-terminal α + β unit that might adopt a PAS-like fold	2.16N	A novel conserved module typically found at the C-terminus of HTH domains of the cI-like family. The N-terminal sub-domain possesses the same fold of as the Zn-dependent metalloproteases but many copies of it may be catalytically inactive as they show disruptions of HEXXH signature. The C-terminal sub-domain contains a conserved group of 4 cysteines, which suggests that it chelates metals. Its predicted secondary structure suggests that it may adopt a PAS-like fold
Biotin ligase domain	BirM (Fold: Class II aaRS and biotin synthetases) and BirC (Fold: SH3-like barrel) PDB 1HXD	2.65N	The ligase domain activates biotin to form biotinyl-5′-adenylate which acts as an effector for transcriptional repression. Its regular activity is the transfer of biotin moiety to biotin-accepting proteins
PLPDE	PDB 1DJU	9.48N	Pyridoxal phosphate-dependent aminotransferases. They are typically found fused to GntR family HTH domains
Sugar isomerase (SIS)	α/β PDB 1M3S	2.39N	Usually associated with regulators of polysaccharide metabolism regulons
Uroporphyrinogen-III synthase	α + β	12.65N	An enzyme in Porphyrin biosynthesis pathway. Usually found fused to a wHTH domain of the OmpR family

(*continued on next page*)

Table 1 (*continued*)

Domain	Structure	Placement of HTH[a]	Comments
Phosphoribosyltransferase (PRTase)	α/β PDB 1STO	0.77N	These enzymes transfer PRPP an activated form of phosphoribose to orotate and purines in nucleotide biosynthesis. Independent fusions of HTHs to orotate and purine phosphoribosyltransferases are seen
Sugar kinase	α/β. RNAse H fold	6.77N	Sugar kinases involved in polysaccharide metabolism are usually found fused to a distinct subfamily of wHTH domains of the MarR family
β-D-Xylosidase	α/β (TIM barrel) PDB: 1UHVD	0.58N	This enzyme is found fused to the AraC-like HTH domains in regulators of polysaccharide metabolism in low GC Gram positive bacteria
MJ0056	α + β	0.39N	Found fused to HTH domain in well-conserved subfamily of MarR transcription factors that are restricted to the archaea. Genes encoding these proteins are often found in a conserved gene-neighborhood with enzymes of the riboflavin biosynthesis pathway. The conservation pattern suggests that this domain is also enzymatic and may be a novel diaminohydroxyphosphoribosylaminopyrimidine deaminase, because a conventional enzyme has not been identified in most archaea
V4R (Vinyl-4 reductase domain)	α + β	–	This domain might bind intermediates of chlorophyll or porphyrin biosynthesis pathways
Thioesterase domain	α + β	–	A thioesterase domain proto-typed by the 4-hydroxybenzoyl-CoA thioesterase. This domain is found fused to HTH domains of the GntR family in several proteins from low-GC Gram positive bacteria. The proteins usually also possess CBS domains
Cyanate lyase	α + β	–	Always found fused to a cI-like HTH domain. The enzyme detoxifies cyanide by combining it with bicarbonate to produce ammonia and carbon-dioxide
ThiJ_PfpI	α/β PDB: 1OI4	3.97C	Domains of this superfamily possess a diverse range of activities, such as protease, catalase, 5′-phosphoribosylformylglycinamide:L-glutamine amido-ligase, and redox-dependent molecular chaperone activity. Usually they are found fused to the AraC family HTH domains
LexA Protease domain	All β	6.71N	Serine protease domains of the signal peptidase fold. They are found fused to both cI-like and wHTH domains
Carbamoyl phosphate synthatase	α/β fold. PDB 1a9x	–	A dyad of HTH domains are found between the carboxy and carbamate phosphorylating units of the large subunit of enzyme
PQQ biosynthesis protein C domain	α helical. PDB 1RCW	–	Iron-binding redox enzyme involved in biosynthesis of Coenzyme Pyrroloquinoline quinone. Fusions to HTH seen only in cyanobacteria
S-adenosyl-L-methionine-dependent methyltransferase domain	α/β. Rossmann fold	1.55N; 0.26C	There have been multiple independent fusions to methyltransferases with different substrate specificities. These include fusions of the ArsR family (*Pseudomonas* PA0547) and MerR (*Bacillus* BC2672) families with methylases of unknown specificities, plant isoflavone *O*-methyltransferases and restriction methylases
Methyl-DNA protein methyltransferase	α/β. A truncated RNAse H fold	8.52C	These enzymes transfer alkyl groups from *O*-6-methylguanine-DNA to themselves. The methyl acceptor is a conserved cysteine in the wing of the wHTH domain
AlkA	α + β, TBP-like domain followed by helical HhH domain. PDB 1DIZ	0.48N	An α-helical DNA glycosylase involved in DNA repair. The HTH domains in these proteins belong to the AraC family
Ribonuclease R (RnaseR)	α + β	3.06N	A multi-domain enzyme involved in the processing of tmRNA
P-Loop NTPase	α/β. P-loop fold	12.10N; 51.97C	Diverse versions of the P-loop fold are fused to the HTH domain. These include the AAA+ superclass members like Cdc6, MCM and DnaA, STAND NTPases in the AfsR and MalT like regulators and GTPases in the case of the SelB and IF-12 proteins
DOC	Predicted α-helical	1.65C	A metal dependent enzymatic domain predicted to function as a nuclease

ParB	All β, PDB 1VZ0	3.87C	A nuclease domain of the OB-fold involved in plasmid partitioning. Always found in association with HTH domains of the KorB family
TOPRIM	α/β	–	Catalytic domain found in topoisomerases, primases and nucleases with a catalytic DxD motif at the active site
Miscellaneous domains			
Rubredoxin-like ZnR	Metal Chelating	–	These domains are involved in both DNA binding and protein-protein interactions. They are common in the archaeo-eukaryotic lineage in HTH proteins like TFIIE, MBF1 and TFIIB
TolB-N	α + β	–	Domains possessing the same fold as the TolB-N terminal domain are fused to HTH domains like IagA from *S. typhi*. This domain is invariably followed by a further set of TPR domains
S1 and S1-like	All β; OB fold	–	Two related RNA-binding domains typified by the bacterial ribosomal S1 protein and the cold shock protein (CSP)
Schalfen N-terminal and middle domains	N-terminal domain is α + β	–	A domain often found fused to a P-loop ATPase in the animal schalfen proteins and GTPases in some archaeo-eukaryotic proteins. The version associated with the wHTH domain is typically accompanied a second "middle domain" that occurs between the schlafen domain and the wHTH domain. Operon architectures suggest that they may be subunits of restriction-modification or DNA repair systems

[a] The number in this columns gives the frequency (per 1000 HTH proteins in prokaryotic genomes) of HTH domains located immediately adjacent to the globular domain under discussion. The 'N' indicates that the HTH occurs N-terminal to the domain under discussion, whereas C indicates that it occurs to the C-terminus of the domain under discussion. A total set of 31100 proteins from 184 prokaryotic genomes were analyzed. A '–' indicates that the HTH domain is either not found immediately adjacent to the module under discussion or that such combinations are very infrequent.

architectural diversification involves tandem duplications of HTH domains. Such versions are encountered in the sigma factors, where regions 3 and 4 correspond to a tandem duplication of HTH domains [70]. The Myb/SANT family of HTH domains in eukaryotic chromatin proteins and transcription factors also tend to show multiple tandem copies [71]. In certain cases, such duplications of the HTH domain in the same protein are accompanied by functional diversification of the copies. For example, in the sigma factors, the HTH domain corresponding to the region-3 is involved in contacting the core RNA polymerase complex, while the HTH domain associated with region-4 binds DNA associated with the -35 element of the promoter [70].

3.2. Combinations of HTH with other nucleic acid binding domains and protein–protein interaction domains

HTH domains are occasionally combined with other domains that interact with macromolecules resulting in multivalent polypeptides. Examples of such architectures are the basal transcription factors, MBF1 and TFIIE, of the archaeo-eukaryotic lineage, which combine the HTH with a rubredoxin-like Zn-ribbon domain [30]. The Zn-ribbon could either function as a second potential nucleic acid binding domain or mediate specific protein–protein interactions with the basal transcriptional machinery. Similarly, combinations of the HTH with another four-helical DNA binding-domain are seen in the plasmid encoded KorB proteins [57]. In the eukaryotes, the HTH is fused to the single-strand DNA-binding OB-fold domains replication factor RP-A, while in the tumor-suppressor protein, BRCA2, a basic tri-helical HTH is inserted within one of its three OB-folds [72]. In a class of widespread bacterial proteins, prototyped by YitL, a wHTH domain is fused with RNA-binding S1 domains, suggesting that this group of proteins may perform an as yet uncharacterized role in bacterial RNA metabolism (Fig. 3). The domain architectures suggest that these multi-domain proteins usually function as adaptors that may bridge two macromolecular complexes, such as specific transcription factors and the basal transcriptional machinery, by way of the multiple interaction surfaces provided by their distinct domains. An alternative general function for these proteins, which is not mutually exclusive of the previous one, is to induce conformational changes in nucleic acids by binding them at multiple sites.

3.3. Combinations of the HTH domain with catalytic domains

The HTH domain is frequently combined with a diverse set of catalytic domains, and there are several

general functional trends associated with such combinations. One common association of the HTH domain with catalytic domains represents the utilization of the domain as a substrate-recognition or localization domain. The HTH is observed to be linked to catalytic domains in several proteins involved in DNA replication (e.g. the FtsK–HerA superfamily ATPase domain [73] in bacterial chromosome pumping protein FtsK [74], or the AAA+ ATPase domain in MCMs and Cdc6/ORC1, which function in archaeal and eukaryotic replication initiation [75]) and repair (e.g. AlkA-type helical DNA glycosylase domain), certain restriction endonucleases (e.g. FokI [45]) and modification methylases (e.g. ScrFIA methylase) (Fig. 3). In course of this survey we noted that the DprA protein (Smf/Dal/CilB), which is required for protecting DNA during transformation in several bacteria [76], contains a wHTH domain fused to the C-terminus of a large globular domain containing a specialized version of the Rossmann fold (Fig. 3). This wHTH domain might recruit the Rossmann fold domain to DNA, and enable it to catalyze an as yet uncharacterized DNA-modifying activity that is required for efficient transformation. Most transposases and integrases of diverse mobile DNA elements have at least one HTH domains that help their catalytic domains associate with DNA [77]. In these instances the HTH either serves as an additional tether that recruits the catalytic domain to DNA or it participates in substrate recognition. An extreme case of this functional theme is seen in the topoisomerases: the HTH domains have supplied, on at least four independent occasions, the catalytic tyrosine of these enzymes that is covalently linked to DNA ([63,78] and LA unpublished). Similarly, in the methyl-DNA protein methyltransferase (*O*-6-methylguanine-DNA alkyltransferase), a wHTH domain fused to a truncated domain of the RnaseH fold bears the cysteine that receives the alkyl group from damaged DNA [46]. Continuing on this theme, it has been previously noted that the catalytic domain of the lambda integrase family has itself evolved from a duplication of the HTH domains [63]. In some cases, the HTH domain has also been recruited by enzymes involved in RNA metabolism and translation for binding RNA, especially double-stranded structures peculiar to certain RNAs. For example, the GTPase module of the bacterial selenocysteine-specific elongation factor (SelB) is recruited to a specific RNA hairpin of selenoprotein encoding transcripts by a C-terminal extension containing 4 tandem copies of the wHTH domain [44]. We also detected a wHTH domain, similar to those in SelB at the N-terminus of the bacterial RNA-processing enzyme Rnase R, which might bind RNA (Fig. 3). However, in proteobacteria the Rnase R is also known to function as regulator of virulence genes [79], suggesting that this HTH domain also additionally functions as a conventional transcription regulator.

In related examples, the HTH domain recruits a catalytic domain that may act on proteins, rather than nucleic acids. A striking example of this is the wHTH domain fused to the N-terminus of the Rio family of protein kinases from archaea and eukaryotes [30]. The Rio family of kinases function in 40S ribosomal subunit maturation [80,81], and the wHTH domain recruits the linked protein kinase domain to an rRNA processing protein complex. The LexA protein, the repressor of several bacterial DNA repair genes, represents another variation on this general functional theme. It contains a protease domain of the signal peptidase fold fused to a wHTH domain. The protease domain catalyzes an autocatalytic cleavage in response to a DNA damage signal and triggers dissociation of its wHTH domain from target sequences, thereby allowing transcription of DNA repair genes [82]. Architectures analogous to LexA are also seen in the repressors typified by the heat-response transcription factor HdiR from the *Lactococcus lactis* [83], where a LexA-like protease domain is fused to a cI-like HTH instead of the wHTH seen in LexA (Fig. 3). This implies that the mechanism of transcription regulation with a proteolytic processing step was innovated independently on at least two occasions in evolution.

The next major architectural theme involving combinations of HTH and enzymatic domains appears to be related to feedback regulation of metabolic pathways. In such combinations, the HTH domain is fused to an enzymatic domain catalyzing a key step in a biosynthetic pathway, and usually regulates the transcription of genes in that pathway. One of the archetypal representatives of this architectural theme is the biotin operon repressor, BirA, which contains an N-terminal HTH domain fused to a C-terminal biotin ligase domain [23]. In the presence of biotin the enzymatic domain synthesizes the co-repressor, and the HTH domain represses the transcription of the biotin biosynthesis genes. Comparative genomics suggests that architectures involving fusions to a range of enzymes from cofactor, nucleotide, amino acid and carbohydrate metabolism are fairly common in archaea and bacteria [30] (Table 1). Some notable fusions include combination of the HTH with nicotinamide mononucleotide adenylyl transferase and a P-loop kinase in NadR, with the pyridoxal-phosphate dependent aminotransferase domain (in bacterial HTH proteins of the GntR family), the orotate phosphoribosyltransferase (in archaea), sugar kinases (Rok family in bacteria), purine phosphoribosyltransferase (in archaea) and the threonine synthase (restricted to the genus *Pyrococcus*) (Fig. 3). Some of these architectures, like BirA are widely distributed in the prokaryotic genomes and appear to be ancient. Others like the fusion of an OmpR family wHTH with the uroporphyrinogen-III synthase are found only in actinobacteria, while yet others like a fusion to the threonine synthase are restricted to a single genus, and are apparently of more recent prove-

nance. This observation suggests that the combinations of HTHs with enzymatic domains have been repeatedly selected for throughout the span of prokaryotic evolution.

Two other specialized classes of domain architectures arise through fusions of the HTH domains with either of two types of P-loop NTPase domains, namely the NtrC-like AAA+ domains [84] and the STAND (signal transduction ATPases with numerous domains) NTPase domain [85]. Proteins containing the NtrC-like AAA+ domains are found only in those bacteria that contain sigma-54, and they bind a distant enhancer element and activate transcription of sigma-54 bound promoters. The AAA+ ATPase domains of these proteins perform an ATP-dependent chaperone-like activity that converts the "closed" sigma-54-containing transcription complexes to an "open" configuration, which is favorable for transcription initiation [84]. The NtrC-like AAA+ domains are fused to at least two different types of HTH domains. The classical versions like NtrC and TyrR are fused to a C-terminal basic tri-helical HTH domain of the Fis family [86]. The second version typified by the *Bacillus* levanase operon regulator, LevR, instead contains an N-terminal wHTH domain, suggesting that there have been two independent fusions of the HTH domain with NtrC-like AAA+ ATPases (Fig. 3). The STAND P-loop NTPases are, as a rule, large multi-domain proteins that appear to catalyze the ATP-dependent assembly of complexes in variety of signaling contexts [85]. They typically contain repetitive super-structure-forming domains, such as the WD and TPR domains, which may serve as surfaces for the assembly of multi-protein complexes [85]. The archetypal members of the architectural class combining the HTH and STAND domains are the *E. coli* MalT [87], *Bacillus* GutR [88] and *Streptomyces* AfsR proteins [89]. A recent analysis of the STAND superfamily revealed that the HTH domains have been fused to them on several independent occasions [85]. The fusions involving the OmpR family of wHTH domains (e.g. in AfsR) usually link the HTH to the N-terminus of the STAND NTPase domain. In contrast, fusions involving the LuxR family of HTH link it to the C-terminus of the STAND module, with a set of α-helical repeats occurring between these two modules (e.g. GutR and MalT) (Fig. 3). The STAND-domain-containing transcription regulators are likely to integrate multiple signaling inputs via interactions of their STAND and super-structure forming domains, and are particularly prevalent in the developmentally or organizationally more complex bacteria. Another version of the HTH domain, associated with a restriction endonuclease fold and STAND NTPases domains (Fig. 3), is found in the PH-type ATPases that are expanded in Pyrococci. These domains have been predicted to localize the endonuclease domains of these proteins to their target sequences [30].

3.4. Architectures related to two-component, PTS and serine/threonine kinase signaling

The two component phospho-relay system, involving the histidine kinase and the receiver domain, which is phosphorylated on a conserved aspartate, comprise one of the most common signaling systems in bacteria and certain archaea [90,91]. The fusions of the receiver domain with the HTH are typical of transcriptional regulators responding to histidine kinase-dependent signaling. Two of the most common architectures, seen in the majority of bacteria, involve combinations of a single N-terminal receiver domain to either a LuxR-like tetra-helical HTH domain (e.g. UhpA and NarL) or wHTH domain (e.g. OmpR and PhoB, Fig. 4). Less frequent fusions involving HTH domains of the AraC- and the CitB families are seen in certain bacteria. Other than these simple architectures, several more complicated architectures involving multiple receiver domains or even fusions to additional histidine kinase (e.g. *B. cereus* protein BC3207) and NtrC-like AAA+ ATPase (e.g. *E. coli* NtrC) domains are also observed (Fig. 3) [90,91].

The PTS sugar-transport systems [92] use a phospho-relay cascade to transfer a phosphate from phosphoenol pyruvate to a histidine on the PTS regulatory domain (PRD), which often co-occurs in the same polypeptide with HTH domains [93,94]. The PRDs receive the phosphates from the HPr and EIIB proteins of the PTS system, and depending on their phosphorylation state regulate transcription. Architectures involving the PRD domain are analogous to those involving the receiver domain of the two-component system. The simplest versions contain an N-terminal wHTH domain fused to a C-terminal PRD domain. The more complex forms contain more than one PRD domains, or fusions to NtrC-like AAA+ domains and PTS system EIIB domains, which determine sugar specificity [95,96] (Fig. 3). The *B. subtilis* LicR protein contains an N-terminal HTH fused to two PRDs and both EIIB and EIIA components of the PTS system [95,96], indicating that it is a multi-functional protein that directly regulates both sugar uptake and transcription of sugar-utilization genes. The 3H domain, which is related to the HPr domain of the PTS system, is also found fused to BirA-related wHTH domain in several bacterial proteins typified by Tm1602 from *Thermotoga maritima* [97]. The 3H domain may represent another novel domain that may be regulated by phosphorylation on its conserved histidines, perhaps via a PTS-like system.

The serine–threonine kinases are over-represented in certain organizationally complex bacteria, like the cyanobacteria and the actinobacteria. In the latter group there is class of proteins, typified by the protein EmbR, containing a fusion of the HTH domain with the FHA domain [98]. The FHA domain in this protein binds phosphoserine peptides, and mediates its interaction

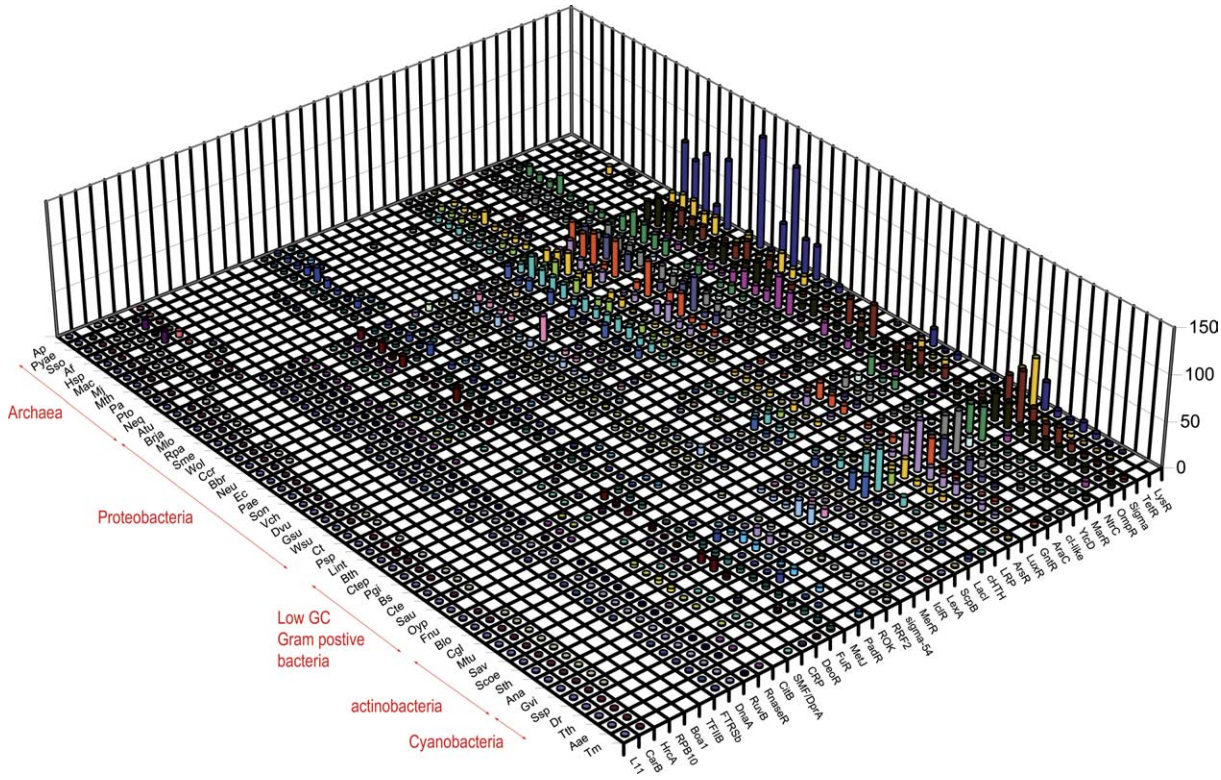

Fig. 4. Phyletic patterns and demography of selected families of HTH domains in selected prokaryotic proteomes. The bar graph depicts actual counts of the number of HTHs in each family per genome. Species abbreviations are as follows: Ap: *Aeropyrum pernix*, Atu_w:*Agrobacterium tumefaciens* C58 (U. Washington), Aae: *Aquifex aeolicus*, Af: *Archaeoglobus fulgidus* DSM 4304, Bs: *Bacillus subtilis*, Bth_V:*Bacteroides thetaiotaomicron* VPI-5482, Blo: *Bifidobacterium longum*, Bbr: *Bordetella bronchiseptica*, Brja: *Bradyrhizobium japonicum*, Ccr: *Caulobacter crescentus*, Ct: *Chlamydia trachomatis*, Ctep: *Chlorobium tepidum*, Ctet: *Clostridium tetani* E88, Cgl: *Corynebacterium glutamicum*, Dr: *Deinococcus radiodurans*, Tth: *Thermus thermophilus*, Dvu: *Desulfovibrio vulgaris*, Ec_C:*Escherichia coli* CFT073, Fnu: *Fusobacterium nucleatum*, Gsu: *Geobacter sulfurreducens*, Gvi: *Gloeobacter violaceus*, Hasp: *Halobacterium* sp. NRC-1, Lint: *Leptospira interrogans* serovar lai 56601, Mlo: *Mesorhizobium loti*, Mj: *Methanococcus jannaschii*, Mac: *Methanosarcina acetivorans*, Mth: *Methanothermobacter thermautotrophicus*, Mtu_H:*Mycobacterium tuberculosis H37Rv*, Neq: *Nanoarchaeum equitans*, Neu: *Nitrosomonas europaea*, Ana: *Anabaena* sp. PCC 7120, Oyp: *Onion yellows phytoplasma*, Pto: *Picrophilus torridus*, Psp: *Pirellula* sp., Pgi: *Porphyromonas gingivalis*, Pae: *Pseudomonas aeruginosa*, Pyae: *Pyrobaculum aerophilum*, Pa: *Pyrococcus abyssi*, Rpa: *Rhodopseudomonas palustris*, Son: *Shewanella oneidensis*, Sme: *Sinorhizobium meliloti*, Sau_MR:*Staphylococcus aureus* subsp. aureus MRSA252, Sav_MA:*Streptomyces avermitilis* MA-4680, Scoe: *Streptomyces coelicolor*, Sso: *Sulfolobus solfataricus*, Sth: *Symbiobacterium thermophilum*, Ssp: *Synechocystis* sp. PCC 6803, Tm: *Thermotoga maritima*, Vch: *Vibrio cholerae*, Wsu: *Wolinella succinogenes*.

with the upstream protein kinase in regulating the biogenesis of the mycobacterial cell wall [99]. Taken together, HTH domains fused to the receiver, PRD, 3H and FHA domains represent a distinct class of architectures that are typical of proteins responding to environmental and physiological stimuli downstream of signaling cascades.

3.5. Architectures related to single-component signaling

In contrast to the above-discussed signaling cascades, the single-component systems are defined as those signaling systems in which the transcription regulatory domain and the stimulus sensor domain are combined in a single protein. These architectures, which are functionally analogous to the fusions of the HTHs with the metabolic enzymes, are by far the most prevalent architectural category in prokaryotes (Table 1 and Figs. 3 and 4). In their simplest form they combine a HTH domain with a small molecule binding domain (SMBDs) [97]. More complex architectures may involve multiple SMBDs or even additional domains such as the NtrC-like AAA+. The same SMBDs found in the single component systems may also occasionally be found fused to two-component regulators, where they may supply secondary allosteric inputs (Fig. 3).

The most common SMBDs fused to HTHs in the single component systems are drawn from a relative small set of ancient protein folds (Table 1): (1) The PAS-like fold, with representatives such as the PAS domain, the GAF domain, and the ligand binding domains of the IclR-type transcription factors [66,100]. (2) The periplasmic binding protein types I and II domains, which include the ligand-binding domains of the LysR family [101–103]. (3) The ferredoxin-like fold, which includes the ACT domain and related ligand-sensing domains

of the Lrp-like transcription factors and the classic ferredoxins, which are fused to HTH domains in archaeal and cyanobacterial proteins [104–106]. (4) The double-stranded β-helix (cupin) fold, which contains the AraC-type ligand-binding domains, as well as the cNMP binding domains [97,107]. (5) The CBS domain that occurs as an obligate dyad [108]. (6) The GyrI domain, which contains two copies of the SHS2 structural module, appears to be one of the principal ligand-binding domains of the MerR family [109].

Some other SMBDs share a common fold with enzymatic domains, but appear to be catalytically inactive versions that merely bind low-molecular weight substrates. Examples of these are: (1) the UTRA domain, which is found in the HutC/FarR group of GntR family transcription factors and possesses the same fold as chorismate lyase [110] and (2) The DeoR ligand-binding domain, which shares a common α/β fold, which is also present in the enzymes of the phosphosugar isomerase family such as ribose phosphate isomerase [111]. The enormous genomic information has resulted in the availability of proteins from numerous prokaryotes, displaying several new domain architectures. In many of these proteins, uncharacterized globular domains are fused with the HTH and other signaling domains, in architectures analogous to those of known sensory domains. Thus, these analogous architectures enable the prediction of novel sensory domains of one-component systems. By this procedure several new candidate sensory domains, with somewhat lower abundance than the previously described domains, were uncovered (Table 1). An example of such a domain is suggested by the PocR protein, from *Salmonella*, with an AraC-like HTH domain, which binds the effector 1,2-propanediol, and regulates the propanediol regulon [112]. It contains a distinct globular N-terminal domain that is also found fused to histidine kinases, chemotaxis receptors, and diguanylate phosphodiesterases of the HD-GYP family in other bacterial proteins (VA and LA unpublished). These domain combinations suggest that it a novel evolutionarily mobile, small-molecule-sensing domain, which probably initiates responses through the domains with which it is linked.

We also found a conserved domain in the *Salmonella typhi* invasion regulator IagA [113], which occurs C-terminal to the OmpR-like wHTH domain domain (Fig. 3). Additionally, it occurs independently in other bacterial proteins fused to adenylyl cyclase and histidine kinase domains (data not shown). Iterative sequence profile searches suggest that this domain shares a common fold with the TolB N-terminal domain [114,115], and is typically found at the N-termini of super-structure forming repeats such as TPR and WD40 repeats (Fig. 3). These architectures, and the interactions of the TolB protein [116], suggest that this domain probably acts in unison with the super-structure forming proteins as a potential sensor for the assembly state of certain multi-protein complexes. Hence, transcription factors like IagA with the TolB-N-related domains might regulate transcription in response to the dynamics of multi-protein complexes.

3.6. Unusual functional adaptations of the HTH domain

Beyond its usual DNA binding role the HTH domain appears to have been exapted for a variety of functions, where it is utilized as a molecular adaptor. For example the permuted version of the wHTH in the N-termini of the methionine aminopeptidases appears to represent an ancient recruitment to a protein–protein interaction function [60]. Several such instances of recruitment of the HTH domain to protein–protein interactions are seen in the eukaryotes. One such example is the PINT domain, which forms the structural scaffold of the proteasomal lid, the signalosome and the eukaryotic initiation factor eIF3 [117,118]. It appears to have been derived from a prokaryotic wHTH precursor which secondarily lost its DNA-binding properties. The Snf8 family of proteins in eukaryotes contains two tandem copies of a wHTH domain related to the PINT domain. This family includes the Vps22, Vps25 and Vps36 proteins, which are required for sorting of transmembrane proteins and lipids into the multivesicular-bodies in the eukaryotic vesicular transport system [119–121]. Just as in the case of the PINT domain, the duplicate wHTH domains of the Snf8 family proteins provide a scaffold for the formation of multi-protein ESCRT complexes required for vesicular trafficking. Additionally, the same complex of Snf8 family proteins are also implicated in transcriptional elongation [119], suggesting that they might have been secondarily recruited for a eukaryote-specific role in vesicular transport from an original role in transcriptional regulation. The cyclins and Retinoblastoma are derivatives of the ancestral TFIIB protein which were utilized for specific protein–protein interactions, respectively involved in regulating the eukaryotic cell-cycle controlling kinases and the E2F/DP1 proteins [122]. Likewise, the DEP domain, which is found in several signaling proteins [49], the cullin C-terminal domain (found in cullins, which are the adaptors for the anaphase-promoting ubiquitin ligases) [50], and the C-terminal domain of Esa1-like histone acetylases [123] are other notable examples in which the wHTH domain has been recruited to mediate specific protein–protein interactions in diverse eukaryote-specific signaling contexts. In most of these proteins the HTH domain sticks out as a distinct domain in a complex modular architectural context.

The plant isoflavone *O*-methyltransferases contain a N-terminal wHTH domain which is related to that in bacterial transcription factors, and appears to function as a dimerization domain [124]. The closest relatives of

these plant methyltransferases are seen in bacteria (e.g. McmR from *Streptomyces lavendulae*) suggesting that the plant lineage probably acquired these enzymes through lateral transfer from a bacterial source. Hence, it is possible that the wHTH domain in the bacterial precursor originally functioned as a transcription regulator that was fused to the methyltransferase domain (see above for discussion on analogous fusions) but was subsequently reused in the plants in a structural role. A similar case is presented by the carbamoylphosphate synthetase (CPS), which contains a tandem duplication of HTH domains between the carboxyphosphate and carbamoyl phosphate synthetic modules of the enzyme (Fig. 3). These HTH domains are of the simple tri-helical versions, like those encountered in bacterial transcription factors such as Fis and LuxR. However, in CPS, rather than binding DNA, they apparently function as protein–protein interaction domains that convert the enzyme to its oligomeric form in the presence of uridine [125].

4. The evolutionary classification of HTH domains

As the HTH is a small domain, which exhibits extreme sequence divergence, reconstruction of its higher-order natural classification is fraught with problems arising from the erosion of evolutionary signal. Nevertheless, the availability of numerous high-resolution structures and extensive sequence information allows us at least to reconstruct the major evolutionary radiations of this domain. This reconstruction is based on three distinct sources of information: (1) Structural features (see above for discussion) help in establishing the relationships at the highest level. (2) Sequence information can be used for clustering based on similarity scores, conventional phylogenetic analysis and cladistic analysis with discrete sequence characters. These sequence-based procedures help in resolving the relationships at a lower level, such as defining the principal sequence families, the relationships within them, and some of the higher-level groupings between sequence families. (3) Phyletic patterns (Fig. 4) help in reconstructing the temporal aspects of the evolutionary history of these domains and also help in constraining the directions of derivation of particular versions from others. The combined scenario gleaned from these directions is presented schematically in Fig. 5 (for details of the combined approach see [85]).

This higher-order evolutionary scheme is characterized by the presence of several basal lineages that retain the primitive basic tri-helical version HTH fold, but have no other shared-derived character that groups them together. These basal lineages are followed by the two great monophyletic lineages, namely the tetra-helical superclass and its derivatives and the wHTH and its derivatives. Beyond these, there are few other highly derived versions whose provenance is hard to establish on account of their extreme sequence and structure divergence. Below, we briefly describe the major evolutionary lineages of the HTH along with their phyletic patterns and functional diversification.

4.1. Lineages of basic tri-helical HTH domains

Several distinct families that retain the primitive simple tri-helical HTH domain are represented in one or more of the major divisions of life. The duplicate HTH domains found in the *carbamoyl phosphate synthetase* represent a distinctive lineage of simple trihelical domains present in all the 3 super-kingdoms of life. Phylogenetic trees show that these proteins follow the "standard model topology", with a distinct archaeo-eukaryotic branch and a bacterial branch [126]. This topology and their phyletic pattern suggest that this lineage most probably, goes back to the LUCA. The HTH domains bearing the catalytic tyrosine of the topoisomerase I family that is found in all the 3 super-kingdoms and the archaeal topoisomerase VI are also distinctive lineages of tri-helical HTH domains with no specific relationship to any other HTH domains. The phyletic pattern of the former lineage suggests that it was probably present in the common ancestor of the 3 super-kingdoms [127].

Rbp10 family, which is defined by the eponymous RNA polymerase core subunit [54,128], is universal in both archaea and eukaryotes and appears to have been part of the shared vertical inheritance of the archaeo-eukaryotic lineage. Likewise the *sigma factor family* [129] is conserved throughout the bacteria but *bona fide* representatives of this group are absent in the archaea and eukaryotes. Members of this group are characterized by a tandem duplication of the HTH domain. sigma-70, which is the basal transcription factor of the bacteria, is usually present in single copy in all bacterial lineages and shows a noticeable phylogenetic signal suggestive of a largely vertical inheritance since the last common ancestor of all the extant bacteria [129,130]. In contrast other sigma factor subfamilies show evidence for lateral transfer, gene loss and lineage specific expansions (Fig. 4). In particular, the ECF subfamily appears to have been widely expanded in numerous bacterial lineages, especially in those with complex metabolic and developmental capabilities [129]. The lineage specific diversification of the ECF subfamily of sigma appears to have been a major contributor to the evolution of niche-specific adaptations in bacteria by allowing diverse patterns of differential gene expression (see below). The *sigma-54 family* appears to have been derived independently from the remaining sigma factors and is another distinct lineage of tri-helical HTHs that is sporadically distributed in bacteria.

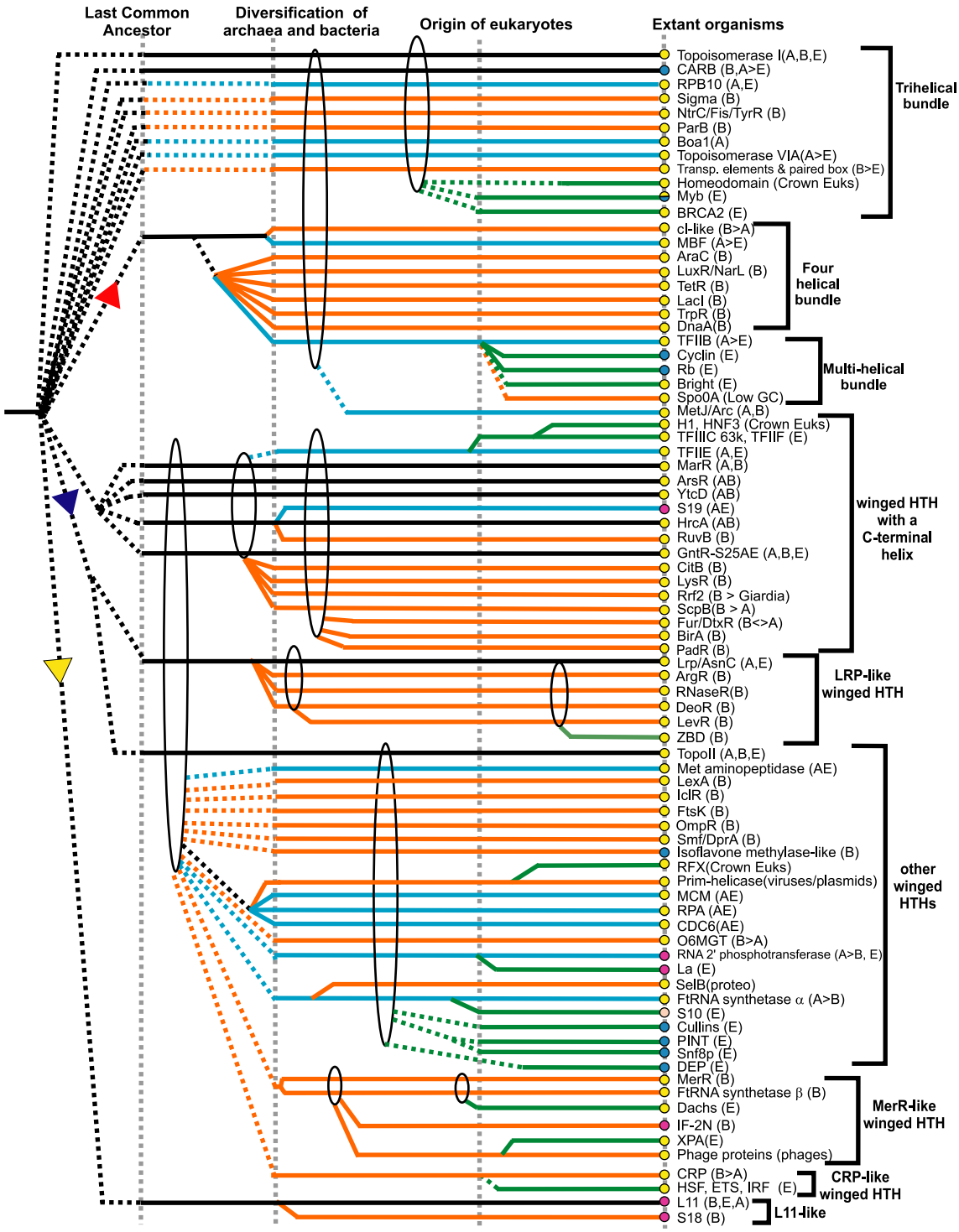

Fig. 5. Higher order evolutionary relationships of HTH domains. The horizontal lines show temporal epochs corresponding to three major transitions in evolution, the last Universal common ancestor, the diversification of archaea and bacteria and the evolution of the eukaryotes. Solid lines reflect the maximum depth of time to which a particular family can be traced. Broken lines indicate an uncertainty with respect to the exact point of origin of a lineage. Colored circles at the termini of the lines represent broad functional classes: where yellow represents DNA binding, pink represents RNA binding and blue, interaction with proteins. The ellipses encompass groups of lineages from which a new lineage with relatively limited distribution could have potentially emerged. Lineages of archaeal origin are colored blue, those of bacterial origin are colored orange and those present in archaea and bacteria are colored black. Lineages only detected in the eukaryotes are colored green. The yellow triangle reflects the origin of the L11 family of proteins, the blue triangle reflects the origin of the winged HTHs and the red triangle reflects the origin of the tetra-helical version. The phyletic distribution of the lineages are also shown in brackets, where A: Archaea; B: bacteria, E: eukaryotes, proteo: Proteobacteria and Crown: Crown group eukaryotes. The '>' reflects lateral transfer with the arrow head pointing to the potential direction of transfer.

The *Fis family* of basic tri-helical HTH domains is also found solely in the bacteria and typically appears at the C-termini of the NtrC-like AAA+ domains [131]. It appears to have emerged early in bacterial evolution and spread widely via lateral gene transfer along with the spread of sigma factors of the sigma-54 family. The Fis protein itself appears to have been secondarily derived in the proteobacteria through the gene fission of an NtrC-like regulator [131]. Likewise, the trihelical HTH domains of the *Rok family* are restricted to the bacteria and always found in combination with sugar kinase domains of Hsp70/actin fold. The archaeal *Boa1 family* of tri-helical HTH domains contains a unique all β-strand domain at the N-terminus that is likely to bind its effectors (Table 1, Fig. 4). The *Myb family* is a pan-eukaryotic family of simple tri-helical domains that appears to have diversified into multiple members prior to the diversification of all extant eukaryotic lineages (Fig. 5). Some versions of the Myb family, the SANT subfamily, appear to have been recruited secondarily for protein–protein interactions in the eukaryotic chromatin [132]. In bacteria, the Myb domain is only seen in the RsfA-related pre-spore transcription factors of the low-GC Gram-positive bacteria [133], suggesting that it was acquired relatively late through lateral transfer from the eukaryotes. In addition to the Myb domain, the *homeodomain and POU domain families* [41], which are also basic tri-helical HTH domain, are respectively widespread in the crown-group eukaryotes and metazoans. The HTH domain of the eukaryotic tumor suppressor BRCA2 [72] is another simple trihelical version that was derived early in eukaryotes. However, it shows no specific relationships to any other HTH domains with a similar structural configuration.

Beyond these classical families there are several tri-helical HTH domains associated with diverse transposases and resolvases, such as the gamma–delta resolvase. The exact point of origin of the HTH domains associated with these mobile elements is difficult to ascertain, but they appear to have given rise to families of HTH domains found in cellular transcription factors on multiple occasions. Particularly striking examples of these include the *Paired box and Pipsqueak families* involved in metazoan developmental gene expression, and the *CENBP family* (centromere binding protein) in the crown group eukaryotes [134–137]. Likewise, the HTH domain of the bacterial *YlxL(SwrB) family* (Fig. 3) is also related to the HTH domains of the gamma–delta resolvase [138]. It is possible that other transcription factors families with a relatively restricted phyletic distribution, like the eukaryotic homeodomain, and the bacterial sigma-54 family have also ultimately been derived from the HTH domains of transposases. The metal-binding domain of the retroviral integrase also appears to have been derived from the HTHs of transposase/resolvase class through acquisition of metal-chelating residues. The *KorB–ParB family* also contains a basic tri-helical version of the HTH domain, and functions as the partitioning protein for diverse bacterial plasmids. The KorB subfamily contains an additional C-terminal 4-helical DNA-binding domain [57] fused to the HTH domain, while the ParB subfamily contains a fusion to a nuclease domain of the OB-fold [139].

The MetJ-Arc (RHH) family of transcription factors appears to have been derived from the basic tri-helical bundle [61]. They are most frequently found as transcriptional regulators of the mobile toxin–antitoxin operons [140]. Hence, it is possible that they were originally derived in such toxin–antitoxin systems, through rapid divergence from a conventional HTH. This appears to have happened early in the evolution of one of the prokaryotic lineages, after which they were widely disseminated across the prokaryotes through horizontal mobility.

4.2. The tetra-helical HTH superclass and its derivatives

The first major monophyletic clade of HTH domains is defined by the unifying structural feature of the tetra-helical bundle. Sequence similarities help in identifying several major lineages within this group. The *cI-like family*, typified by the phage lambda cI protein, contains representatives from across the 3 super-kingdoms of life (Fig. 4). Several distinct subfamilies can be recognized within this family. The largest of these is the bacterial repressor subfamily typified by the protein PbsX (Xre) from the *B. subtilis* prophage 168 [141]. Another notable subfamily is the MBF1 subfamily, which is nearly universally conserved in the archaeo-eukaryotic lineage functions, and is an adaptor that appears to bridge the specific transcriptional regulators to the basal transcription machinery MBF1 [142].

The next major assemblage within the tetra-helical superclass comprises of 6 major families that are exclusively prokaryotic in their distribution (Figs. 4 and 5). This assemblage includes the *AraC, LuxR, LacI, DnaA, TrpR and TetR families*, which are predominantly bacterial with several independent lateral transfers to archaea (Fig. 4). The first four of these families are nearly pan-bacterial in their distribution suggesting that these HTH families had probably diverged from each other even in the common ancestor of all bacteria (Fig. 5). The latter two lineages are more limited in their distribution, but are found in most proteobacteria, and low GC Gram positive bacteria (Fig. 4). Within some of these families several distinct lineages, often defined by specific architectural themes can be identified. However, most of these families contain a dominant architectural theme suggesting these might have been the earliest versions of these families. The AraC family contains a duplication of the tetra-helical version of the HTH domain [143], and typically occurs fused to a sugar binding

domain of the DSBH fold [97,107] suggesting that they predominantly function as sugar-sensing transcription factors. However, in complex cyanobacteria, like *Nostoc*, the majority of AraC family HTH domains occur fused to a novel sensor domain that is also found in the siderophore alcalgin sensing transcriptional activator, AlcR, from *Bordetella* [144] (Fig. 6 and Table 1). The most common subfamily of the LuxR family contains fusions to the receiver domain, as seen in the case of the *E. coli* protein NarL. In the case of the LacI family the dominant architecture features a fusion of the HTH domain with a small-molecule binding domain of the periplasmic solute-binding protein fold. DnaA is usually found only in a single copy in all bacterial genomes, with the HTH occuring at the C-terminus of the AAA+ domain [145]. The DnaA protein functions in replication initiation, and also as a transcription factor [146]. Additionally, sporadic versions of the tetrahelical HTH superclass are also seen in several phage transposases related to the Mu transposase.

TFIIB, a basal transcription factor in the archaeo-eukaryotic lineage, defines the *TFIIB family*, a derivative of the tetra-helical class. In the archaea there is typically a single version of this family (Fig. 4). However, in the eukaryotes not only did TFIIB undergo duplication, but it also spawned two more divergent families, namely the cyclins and the Rb proteins [122]. Structural comparisons suggest that the eukaryote-specific *Bright domain family* [55,147], which includes DNA-binding proteins involved in chromatin dynamics was also derived from a TFIIB-like precursor prior to the radiation of eukaryotes from their common ancestor. The provenance of the only bacterial relatives of this version, namely the Spo0A protein from endospore-forming bacteria [56] remains unclear.

4.3. The wHTH superclass

The second major monophyletic clade of HTH domains are unified by the presence of a striking derived feature, the "wing". Several major assemblages with varying distributions can be identified within the wHTH superclass. The wHTH superclass includes the majority of prokaryotic transcription factors. Thirteen major families of prokaryotic wHTH domains, namely the *BirA, ArsR, GntR, DtxR-FurR, CitB, LysR, ModE, MarR, PadR, YtcD, Rrf2, ScpB and HrcA-RuvB families*, are unified by the presence of a characteristic helix after the wing, and comprise the largest monophyletic assemblage within the wHTH superclass (Fig. 5). Of these, representatives of the ArsR, MarR, YctD, GntR, HrcA-RuvB are seen in both archaea and bacteria (Fig. 4) with phylogenetic trees suggesting distinct pan-archaeal and pan-bacterial branches within them (data not shown). This would imply that these families possibly even go back to the LUCA (Fig. 5). The members of these families, barring some striking exceptions (see below) function as transcription factors suggesting that they could have descended from an ancestral protein that had some role in transcriptional regulation. The MarR family has vastly proliferated in the archaea to give rise to several archaeal subfamilies and includes most of the major archaea-specific wHTH transcription factors. In bacteria the RuvB sub-family, which is derived from the ancestral HrcA lineage, appears to have secondarily acquired a role in DNA recombination after a fusion with the AAA+ domain in the ATPase subunit of the Holliday junction resolvase [148]. Another subfamily of the HrcA–RuvB family, S19AE is a small subunit ribosomal protein in the archaeo-eukaryotic lineage, with a potential RNA-binding function (Fig. 5). A similar case is observed in the GntR family, which has vastly proliferated in bacteria giving rise to many of the major bacterial one-component transcription factors. However, in the archaeo-eukaryotic clade it is represented by a single lineage, the small subunit ribosomal protein S25AE (Figs. 4 and 5).

In contrast to the above families, the BirA, ModE, PadR, ScpB and DtxR-Fur families appear to have had a bacterial origin followed by sporadic lateral transfers to the archaea (Fig. 4). The DtxR–Fur family appears to have specialized early on in regulating metal-dependent transcription of genes. The ScpB family is typically represented in most bacteria by a single protein, which is encoded in the same operon as a kleisin and a SMC-type ABC ATPase [149,150]. The ScpB protein contains a tandem duplication of the wHTH with a C-terminal helix after the wing (Fig. 3), and has been shown to regulate the activity of the chromosome reorganizing SMC ATPases [149,150]. The archaeal representatives of this family appear to have been acquired through a lateral transfer from the cyanobacteria (LMI and LA unpublished). The pan-bacterial families, namely CitB, LysR and Rrf2, are largely absent in the archaea (Fig. 4). Interestingly, a single member of the Rrf2 family (GLP_14_27362_29578) is seen in the early branching eukaryote *Giardia lamblia*. A number of wHTH domains with distinct sequence conservation profiles, and occurring in large multi-domain proteins also belong to this assemblage of wHTH domains with a C-terminal helix after the wing. These include (1) two HTH domains of the topoisomerase II family, (2) the HTH domain of the enigmatic topoisomerase V, which is currently found only in *Methanopyrus kandleri* [78], (3) The ESA1 family of eukaryotic histone acetyltransferases and (4) the N-terminal HTH domains of the Rio kinases from archaea and eukaryotes. Two related wHTH domains from *Giardia* (Genbank Gis: 29245940, 29247865) also appear to have been derived from within the above-discussed assemblage of wHTH domains, but their precise affinities are obscured due to rapid sequence divergence.

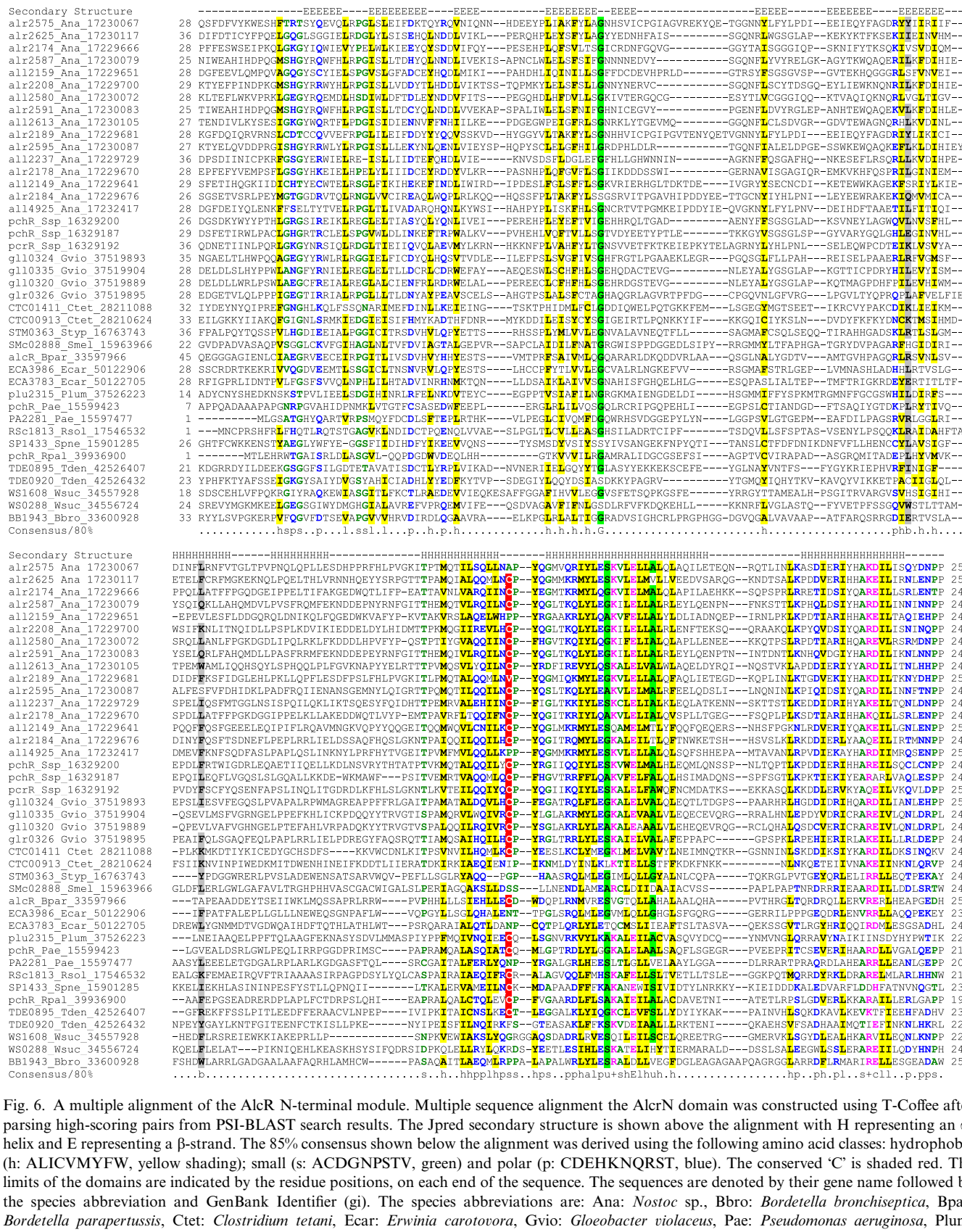

Fig. 6. A multiple alignment of the AlcR N-terminal module. Multiple sequence alignment the AlcrN domain was constructed using T-Coffee after parsing high-scoring pairs from PSI-BLAST search results. The Jpred secondary structure is shown above the alignment with H representing an α-helix and E representing a β-strand. The 85% consensus shown below the alignment was derived using the following amino acid classes: hydrophobic (h: ALICVMYFW, yellow shading); small (s: ACDGNPSTV, green) and polar (p: CDEHKNQRST, blue). The conserved 'C' is shaded red. The limits of the domains are indicated by the residue positions, on each end of the sequence. The sequences are denoted by their gene name followed by the species abbreviation and GenBank Identifier (gi). The species abbreviations are: Ana: *Nostoc* sp., Bbro: *Bordetella bronchiseptica*, Bpar: *Bordetella parapertussis*, Ctet: *Clostridium tetani*, Ecar: *Erwinia carotovora*, Gvio: *Gloeobacter violaceus*, Pae: *Pseudomonas aeruginosa*, Plum: *Photorhabdus luminescens*, Psyr: *Pseudomonas syringae*, Rpal: *Rhodopseudomonas palustris*, Rsol: *Ralstonia solanacearum*, Smel: *Sinorhizobium meliloti*, Spne: *Streptococcus pneumoniae*, Ssp: *Synechocystis* sp., Styp: *Salmonella typhimurium*, Tden: *Treponema denticola*, Wsuc: *Wolinella succinogenes*.

The next major monophyletic assemblage of wHTH superclass includes *the DeoR, ArgR, LevR, YitL, Lrp-AsnC, ZBD (Z-DNA binding domain), and RNase R families*. These families are unified by overall sequence similarity, and a conserved pattern with a conserved glutamine or arginine residue between helix-1 and helix-2 of the HTH domain. Of these the Lrp-AsnC family is widely conserved in both bacteria and archaea (Figs. 4 and 5) and in phylogenetic trees displays distinct branches separating the majority of archaeal and bacterial members. Hence, it is possible that the Lrp-AsnC family goes back to the LUCA. The ArgR and DeoR are predominantly bacterial families, whereas the LevR group is sporadically found, mainly in low GC Gram positive bacteria. The RNase R family is a limited group that is represented by just a single pan-bacterial orthologous lineage in the form of the wHTH domain at the extreme N-terminus of the Rnase R protein (Fig. 3). Its widespread distribution in the bacteria suggests that it emerged early in the evolution of this lineage. The ZBD family is restricted to the crown group eukaryotes and in animals it is fused to the deaminase domain involved in hyper-mutation of the immunoglobulin genes [151,152]. The restricted phyletic pattern of the ZBD family suggests that it may have evolved from one of the prokaryotic families after lateral transfer to the crown group eukaryotes (Fig. 5).

The wHTH domains found in the archaeo-eukaryotic proteins involved in replication initiation, namely the MCM proteins, the CDC6–Orc1 proteins (C-terminal to the AAA+ ATPase domains in both these cases) and the RP-A protein comprise the *replication initiation family* of wHTH domains. Both the MCM and the CDC6 versions appear to have been present right from the base of the archaeo-eukaryotic lineage [30,153,154]. The version associated with RP-A occurs as a standalone protein in the archaea, while it appears to have been fused to the single strand DNA binding OB fold domains in the eukaryotes. Also belonging to this family are the C-terminal wHTH domains from the replicative helicase-primase enzymes of various viruses such as P4, plasmid Rep proteins and the eukaryotic RFX-type DNA binding domain seen in transcription factors like the MHC class II transcription factor HRFX1 [75,155,156]. This family might have originally evolved to recognize specific DNA features associated with the replication initiation sites, and recruiting the catalytic activities involved in pre-initiation and initiation to these sites [75]. The RFX domains are specifically related to certain phage helicase-primase wHTH domains belonging to this family. This suggests that the crown group eukaryotes may have acquired the RFX domains from such a viral replication protein and reused them as a transcription regulator (Fig. 5).

The wHTH domains of the *archaeal phenylalanyl tRNA synthetase α-subunits*, the *eukaryotic ribosomal protein S10* and the bacterial selenocysteine-specific elongation factor *SelB* appear to comprise a family principally associated with translation and RNA metabolism proteins (Figs. 3 and 5). Their phyletic patterns suggest that their recruitment to RNA-specific functions appears to have occurred after the separation of the major superkingdoms of life, though it is possible that a standalone precursor of this family was already present in the LUCA. Another distinct family of wHTH domains with an exclusive single-stranded RNA-binding function is the *La domain family* [47,48,157]. The La domain has previously only been reported from eukaryotes; however, using sequence profile analysis we show that it is homologous to the N-terminal domain of the NAD-dependent RNA 2′-phosphotransferase [158], which removes the phosphate from the 2′ ends of RNA. In the RNA 2′-phosphotransferase the La domain bears two of the four absolutely conserved catalytic histidines (LA unpublished), suggesting that it is another case of recruitment of the HTH domain for a catalytic role. The RNA 2′-phosphotransferases are highly conserved in the archaeo-eukaryotic and sporadically observed in the bacteria. This suggests that the La family of HTH domains emerged early in the archaeo-eukaryotic lineage and were subsequently laterally transferred to bacteria. In the eukaryotes the non-catalytic versions (the classical La domains) were recruited for binding 3′ poly(U)-rich elements of nascent RNA polymerase III transcripts and translation regulation [47,48]. The fungal protein frequency [157] contains an as yet functionally uncharacterized version of the La domain, which may regulate the circadian clock via RNA metabolism.

There are other distinct families of wHTH transcription factors in prokaryotes with related 2- or 3- stranded wHTH domains, but they do not appear to belong to any of the aforementioned assemblages. These include the *LexA, OmpR, and IclR families* that appear to be pan-bacterial families, with at best a rare presence in the archaea (Figs. 4 and 5). The classical representatives of the LexA family appear to be involved in regulating responses to DNA damage in diverse bacteria [82]. However, a highly divergent, potential offshoot of the LexA family is seen fused to the C-termini of large DNA helicases (lhr) that are found sporadically in several bacteria (Fig. 3). The eukaryotes display their own families of specific transcription factors, such as the *Forkhead-histone H1* [25,26] and *E2F-DP1 families* [159], and the basal transcription factors, such as the *TFIIF(Rap30)-TFIIIC-63K family* [160] that show structurally similar wHTH domains, but lack specific sequence relationships with any of the prokaryotic families. Within the TFIIF-TFIIIC-63K family, the RNA polymerase III transcription factor, TFIIIC-63K, is conserved throughout eukaryotes, but TFIIF appears to be restricted to the crown group and the apicomplexans.

The wHTH domains of the TFIIF and TFIIIC-63K are functionally similar to those of the Forkhead-histone H1 family suggesting that the latter were probably derived from more widespread family of basal transcription factors [160]. The remaining eukaryotic families appear to be chiefly represented in the eukaryotic crown group, implying that they arose relatively late from pre-existing eukaryotic wHTH domains found in the basal transcription machinery or via rapid divergence from laterally transferred prokaryotic transcription factors. A similar scenario appears to be applicable for the four eukaryotic wHTH families that are not involved in DNA-binding, namely *the PINT domain, the SNF8 (with two tandem copies of the wHTH domain), cullin and the DEP domain families*.

Beyond these multi-member families there are highly conserved lineages of 2- or 3- stranded wHTH domains that are typically found in single orthologous groups of proteins, and cannot be linked to any of the larger assemblages. Such lineages include wHTHs found in the bacterial proteins such as FtsK, DprA and *O*-6-methylguanine-DNA alkyltransferase, and the archaeo-eukaryotic proteins like TFIIE (Fig. 3) [30,161].

Distinct from the 2- and the 3-stranded HTH domains are the 4-stranded HTH domains that appear to form a separate monophyletic assemblage within the wHTH clade (Fig. 5). The main prokaryotic family in this assemblage is the *Crp family* [36] that has a pan-bacterial distribution and sporadic presence in few archaea. Thus, it seems to represent a bacterial innovation that was disseminated to the archaea via lateral transfer. Members of this family are typically fused to a C-terminal cNMP-binding domain, and appear to have specialized early on as cyclic nucleotide dependent regulators. The eukaryotes contain a single major family of this assemblage, the *HSF family* (heat shock transcription factor), which is present only in the crown group eukaryotes. In the animals this protein appears to have spawned two distinct sub-families that are prototyped by the ETS domain and the IRF domain (interferon regulator factors) [162,163]. Given their relatively restricted presence in eukaryotes, it is possible that they have originated through rapid divergence from laterally transferred prokaryotic versions. A similar scenario could be envisaged for the origin of the orphan initiator binding protein from *Trichomonas*, which also contains a 4-stranded version of the HTH domain.

The *MerR-like assemblage* of truncated wHTH domains are derivatives of 3-stranded wHTH domains (Fig. 2). The MerR family proper [37] and the related wHTH domains present in the DNA-binding region of the bacterial phenyalanyl tRNA synthetase β-subunit [164] show a pan-bacterial distribution. In bacteria the MerR family has vastly proliferated into several distinct subfamilies, like the *SoxR and CueR subfamilies* [37]. However, most other versions show a more restricted distribution. A potential RNA-binding version of this domain is the N-terminal domain of translation initiation factor IF2 from certain bacteria [165]. The remaining versions, observed in the phage lambda excisionase and terminase proteins, the phage Mu-repressor family and the eukaryotic DNA repair protein XP-A and animal transcription factors of the Dachshund family, appear to be DNA binding [166–168]. Eukaryotic Xp-A family is involved in nucleotide excision repair [169], and appears to have been derived at some point in eukaryotic evolution from the functionally similar phage excisionases. The principal diversification of this assemblage appears to have happened early in bacterial evolution resulting in the two ancient families, MerR and the FTRS β-subunit N-terminal domain. Given that regular 3-standed wHTHs are also found in association with other translation proteins, like the archaeal FTRS-α subunit and SelB, it possible that the prototype of the MerR-like version was derived from such a form through loss of the initial helix.

4.4. Other miscellaneous families of HTH domains

In addition to the above-described major assemblages, there are a few ancient HTH families with uncertain affinities. The chief amongst these are the related ribosomal proteins L11 and S18 (Fig. 2). The former is conserved in all the three super-kingdoms of life and binds RNA in the L11-stalk structure, which appears to go back to the shared ancestral core of the ribosome [170]. S18 appears to have been derived from L11 in the bacteria. Despite its apparent ancient origins L11 appears to be a highly derived version of the HTH. Hence, notwithstanding certain general similarities with the 4-stranded wHTHs, it is more likely that L11 has convergently acquired these features early in evolution.

5. Proteome-wide demographic trends of HTH domains

The availability of a large number and phyletic diversity of complete genome sequences allows robust estimation of the general trends in the proteome-wide distribution of HTH domains. In order to detect the occurrences of HTH domains in proteomes, position-specific score matrices or sequence profiles were constructed for the various distinct families of these domains using seed alignments with diverse representatives. These sequence profiles were then used to iteratively search the target proteomes with the PSI-BLAST program [42]. Alternatively, the alignments were used to generate hidden Markov models, which were similarly used to search the proteomes with the HMMER program [42]. Using a combination of these procedures we determined the total number of proteins containing HTH domains encoded by all completely sequenced organisms that were available at

the time of the analysis. For prokaryotes, the plot of the total number of HTH domains against gene number per genome is best fitted by the power equation of the form $y = k \times x^c$ (where k and c are constants; $r^2 = 0.89$; see Fig. 7(a)). The r^2 of 0.89 for this fit suggests that this tendency is indeed strongly maintained across a wide diversity of genomes. This non-linear scaling of HTH domain numbers with gene number is consistent with recent studies that have suggested that the transcription factor counts follow a power equation with respect to gene number [171]. Given that the HTH domains are the main transcription factors in most of the prokaryotic genomes it is clear that the trend observed for transcription factors is principally a reflection of the distribution of HTH domains (Fig. 7). This distribution function suggests that as gene number increases, a greater than linear number of HTH domain regulators are required per gene.

Examination of major architectural classes of HTH indicates that there is an interesting differential class-wise partitioning of the trends. HTH proteins belonging to two-component, serine/threonine kinase and PTS signaling cascades are almost entirely missing in archaea. In the bacteria, where they are abundantly present, their numbers show a linear scaling with respect to gene count per genome ($r^2 = 0.8$) (Fig. 7(c)) [42]. This suggests that each HTH in two-component and related phosphorylation cascades regulates a fixed number of target genes, and as the gene numbers grow larger, the regulators increase in direct proportion to their increase in targets. This is consistent with a model of evolution in which the two-component systems and their target genes undergo duplication with approximately the same probability as the size of a bacterial genome increases in evolution. In contrast to the two-component systems, the one component systems, which chiefly comprise of those HTH domains fused to specific SMBDs or metabolic enzymes, scaled non-linearly with increase in gene count per genome. The best fit for the one component systems was obtained with the power equation of the form $y = k \times x^c$ ($r^2 = 0.76$, Fig. 7(b)). This equation suggests that the HTH domains belonging to the one-component system are likely to be a significant contributor to the over-all power equation-type distribution of the HTH domains. This observation implies that as genome size increases a greater than proportional increase in the numbers of one-component transcription factors is required for controlling the newly added genes. This tendency might correlate with the need to regulate specialized groups of their genes, by combining the distinct inputs sensed by the effector-binding domains of multiple sets of one-component transcription factors in the metabolically or organizationally complex bacteria with large genomes.

The counts of sigma factors, too on an average, are positively correlated with gene count per genome (Fig. 7(d)), and scale non-linearly with it (the best fit being provided by a quadratic curve of the form $y = a \times x^2 + b \times x$; where '$a$' and '$b$' are constants; $r^2 = 0.76$). The non-linear scaling of the sigma factors suggests that in the more complex genomes the additional genes are distributed amongst several functionally specialized gene batteries, which are under the regulation of devoted sigma factors responding to specific situations. Interestingly, a few genomes show a significantly greater than expected number of sigma factors (Fig. 7(d)). The most striking example is seen in the case of *Phytoplasma asteris*, which, like other Mycoplasmas, has a highly reduced genome with just over 700 genes [172]. Whereas, the other Mycoplasmas have only a basal sigma-54, *P. asteris* has, in addition to sigma-54, a recent lineage-specific expansion of 11 sigma factors that are related to the *Bacillus* sigma F. Likewise, *Bacteroides thetaiotaomicron* and *Nitrosomonas* show recent lineage-specific expansions of ECF-type sigma factors that have given rise to at least 10 closely related paralogous members in their proteomes. In the case of *P. asteris* there is evidence that the sigma factors may constitute a novel transposon [173]. While this possibility also exists in the case of the other bacteria that show a greater than expected number of sigma factors, it is likely that some of them might have been utilized as transcriptional regulators. This potential link between sigma factors and transposon is consistent with the repeated recruitment of HTH domains from transposases as specific transcriptional regulators.

The genomic data from eukaryotes is still not adequate to discuss general genome-wide trends. However, the available data suggests that eukaryotes display different tendencies from the prokaryotes. The HTH domains of the homeo and Myb proteins are amongst the most prominent DNA-binding domains of transcriptional regulators in the plant and animal lineages. However their numbers are relatively low in the other eukaryotic lineages. This suggests that the rise in prominence of HTH transcription factors may have been a relatively late phenomenon that occurred on multiple occasions in the eukaryotic crown group. The parasitic apicomplexa, including those forms that have genome sizes comparable to some free-living fungi, have far fewer transcription factors in general [174]. The early branching eukaryote *Giardia lamblia* has at least 13 Myb domains and 2 Bright domains, but apparently no other representatives of the HTH domains found in the eukaryotic specific transcription factors (LA unpublished). Furthermore, the scaling of the total number of transcription factors in eukaryotes with gene counts per genome is not equivalent with what is observed in the prokaryotes. This difference may arise due to two main reasons: (1) the emergence of a complex apparatus for chromatin-structure regulation might have changed the nature of transcriptional control in eukaryotes. (2) Most eukaryotic transcription factors are effectively

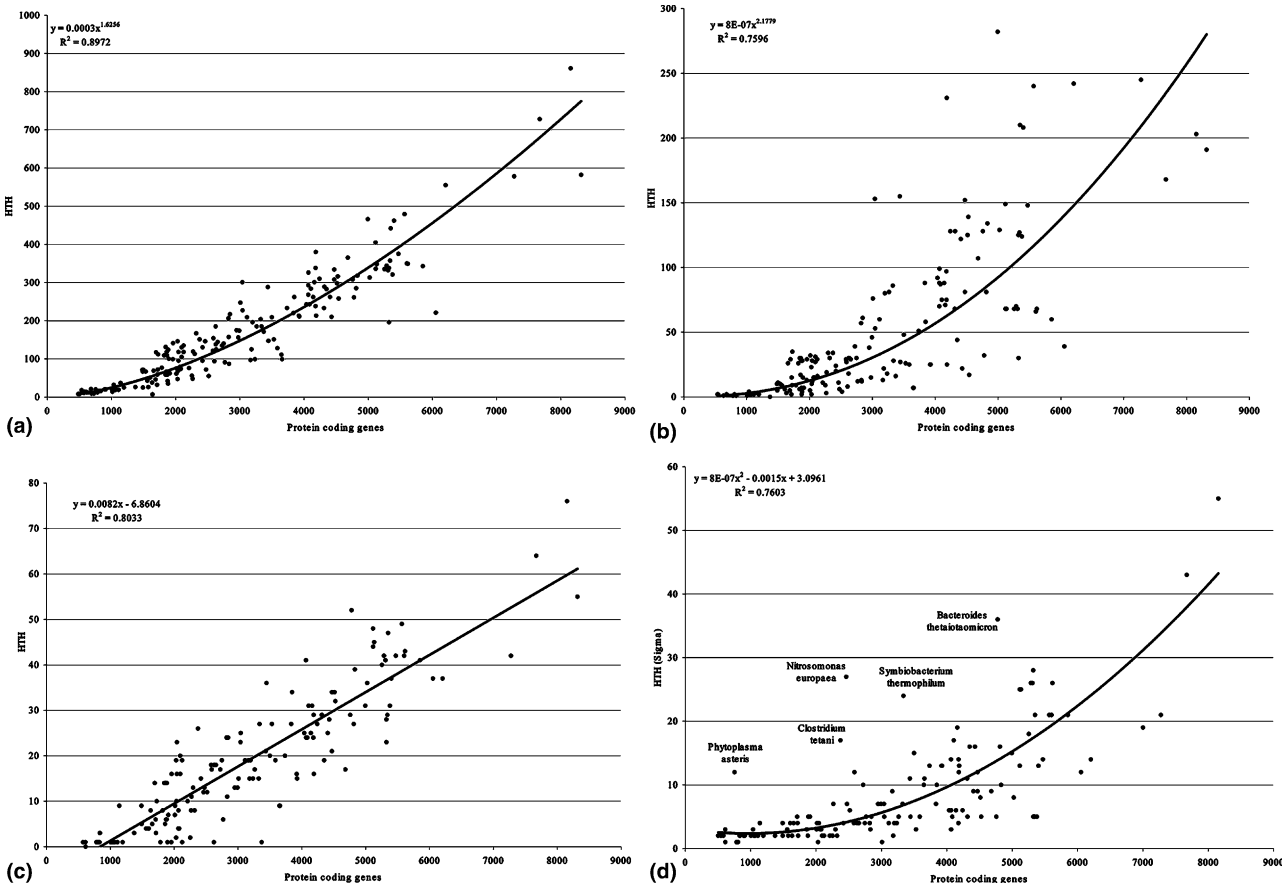

Fig. 7. Scaling of HTH domains with gene number per genome. All graphs show a scatter plot of number of proteins with HTH domains in a given proteome (*Y*-axis) versus the number of protein-coding genes in that organism (*X*-axis). (a) The *Y*-axis is the overall number of proteins with HTH domains. (b) The *Y*-axis is the number of predicted one-component system proteins with HTH domains. (c) The *Y*-axis is number of two-component system and other phospho-relay system proteins with HTH domains. (d) The *Y*-axis shows the number of sigma factors. For each graph the best-fitting trend line along with its r^2 value is shown.

down-stream of signaling cascades that communicate from the cell membrane, or the cytoplasm to the nucleus. Hence, there are there are few equivalents of the genuine prokaryote-type two-component systems in the eukaryotes. Additionally, the eukaryotes might also extensively employ post-transcriptional control mechanisms, involving regulatory RNAs, resulting in a lower dependence on transcriptional regulators [175,176]. A combination of these factors might account for the difference in the average number of genes controlled by per transcription factor in eukaryotes and prokaryotes.

6. General considerations on the natural history of the HTH fold and implications for the evolution of transcription

With the exception of the ribosomal protein L11 no bona fide HTH domains with an ancestral RNA-binding role can be confidently traced back to the LUCA. In the HrcA and GntR families the representatives from the archaeo-eukaryotic clade are ribosomal proteins (S19AE and S25AE); whereas all the known bacterial representatives of this family are specific transcription factors (Fig. 5). All other versions of the HTH associated with translation and RNA metabolism, such as those found in La/RNA 2′ phosphatase, SelB and ribosomal protein S10E, appear to have been derived after the separation of the archaeo-eukaryotic and bacterial clades. The simplest interpretation of these observations is that the majority of HTH domains associated with RNA metabolism settled into their extant functional niches only after the divergence of the major lineages from LUCA. Hence, excluding L11, it is likely that other HTH domains associated with RNA metabolism in the LUCA performed more generic functions compared to their extant counterparts. The DNA-binding property is strongly preserved across diverse lineages and structural variants of the HTH fold, and involves helix-3, despite the several variations in the details of the interactions of individual versions of the domain. This observation, taken together with the more sporadic

distribution of the versions of the domain associated with translation and RNA metabolism, suggests that the ancestor of most of the extant versions of the HTH, excluding L11, was a DNA-binding protein. Furthermore, the diversification of this domain was potentially associated with the emergence of DNA as genetic material [75,153].

While there are HTH domains in the basal transcription factors of both the bacterial (sigma factors) and archaeo-eukaryotic (TFIIB, TFIIE, and MBF1) lineages, none of these can be considered as being truly orthologous [30,32,33]. In contrast, several families of the HTH domains in specific transcription factors appear to be extensively shared by the bacteria and archaea (Fig. 4). Though several of the prokaryotic families shared by bacteria and archaea can be easily explained as arising from relatively recent lateral transfer between the prokaryotic super-kingdoms, some others like the MarR, ArsR, YctD, Lrp, HrcA and GntR families appear to show distinct pan-archaeal and pan-bacterial groups suggesting that they were present in the earliest representatives of each of the super-kingdoms, hence potentially go back to the LUCA (Fig. 5). Despite the basal transcription machinery shared with the archaea, the specific transcription factors that are traceable to the last eukaryotic common ancestor do not belong to the same families as the specific transcription factors seen in the archaea [30]. These patterns raise a profound conundrum regarding the origin of the phyletic patterns of HTH domains in the basal and specific transcriptional regulators of extant organisms.

Even though several scenarios could in principle explain these patterns, there are only a few parsimonious alternatives that account for the currently available data. Given that both the basal transcription machinery and the domain architecture of the RNA polymerase catalytic subunits are very different in the bacterial and the archaeo-eukaryotic lineages [130], one could extrapolate that the basal transcription factors arose only after the two great lineages had separated from the LUCA. In this situation, the sharing of the specific transcription factors by the archaea and the bacteria could be explained in two possible ways: (1) The LUCA had several specific transcription factors but no basal transcription factors. (2) Alternatively, there were neither specific nor basal transcription factors in the LUCA and both types emerged after the lineages separated. However, multiple very early lateral transfer events resulted in prokaryotic lineages sharing a common set of specific transcription factors. This latter scenario is the consistent with the evidence for extensive lateral transfer between the two prokaryotic super-kingdoms throughout their evolution [32,177]. At least in the case of the HrcA and GntR families, the striking difference in functions of the extant bacterial and archaeo-eukaryotic representatives suggests that the ancestral versions of these families in LUCA performed a generic nucleic-acid-binding function. They appear to have been secondarily recruited as specific transcription factors only in the bacteria, thereby supporting the second scenario. The basal transcription factors and ribosomal proteins tend to show a stronger signal of vertical inheritance as compared to specific transcriptions factors which are prone to rampant gene loss and lateral transfer [32,177]. This is not unexpected, given the functional constraints acting on the basal transcription factors, and might confound the reconstruction of the actual evolutionary scenario for the specific transcription factors.

Although the first scenario of basal transcription factors emerging after the specific transcription factors might appear counter-intuitive, detailed analysis of the reconstructed house-keeping functions in the LUCA suggest that it is hardly implausible. Earlier studies on the DNA replication and chromosome partitioning systems suggest that the central enzymes of the DNA replication apparatus appear to have emerged only after the split of the archaeo-eukaryotic and bacterial lineages [32,127,178]. Thus, the DNA replication system resembles the basal transcription apparatus in its origins. More specifically, the absence of an ancestral DNA polymerase and associated replication enzymes in the LUCA suggest that it probably had a system of replication involving reverse transcription [127]. However, simpler but completely functional DNA-dependent RNApolymerase (DdRP) subunits have been reconstructed for the LUCA [130,179]. Hence, it is possible that the LUCA did not use basal transcription factors and the DdRPs chiefly functioned as enzymes that supplied the RNA template for the replication process involving a reverse transcription step. The precursors of the specific transcription factors might have still functioned under these circumstances, primarily acting as general repressors that regulated the synthesis of the RNA template. The basal transcription factors probably arose only when the genome got organized into multiple tandem operons, each needing its own transcription initiation signal. This suggestion is consistent with the fact that the prokaryotic specific transcription factors can function equally well with both archaeo-eukaryotic- and bacterial-type basal transcription machinery and RNA polymerases, despite the numerous differences [180].

Irrespective of the scenarios, the HTH domains of both the simple tri-helical type (e.g. carbamoylphosphate synthetase and topoisomerase I) and the wHTH type (e.g. topoisomerase II family) are present in proteins that can be confidently traced back to the LUCA. Depending on the scenario, there were at least 6–11 different HTH domains in the cellular genomes, suggesting that the fold had undergone structural and functional differentiation even before the period of the LUCA. Subsequently, in course of prokaryotic evolution, the

HTH domains rapidly expanded in several prokaryotic genomes to rank amongst the folds with highest representation [32,35]. The currently available evidence suggests that the common ancestor of extant eukaryotes arose later, via an endosymbiotic event involving an archaeal precursor for the nucleus, translation and secretion apparatus and an α-proteobacterial precursor for the mitochondrion. However, neither the mitochondrial, nor the nuclear genome, appears to have retained the specific transcription factors might have been inherited from the respective prokaryotic precursors. Instead, the principal pan-eukaryotic HTH transcription factor families, like Bright and MYB domains, are only distantly related to the prokaryotic counterparts. The origin of the eukaryotes saw the emergence of a distinct nuclear compartment, extensive RNA-dependent post-transcriptional gene regulation, a complex chromatin structure and the proliferation of enzymatic complexes involved in chromatin dynamics, such as the Swi2 ATPases, acetylases and deacetylases [181]. The compartmentalization of the cell probably rendered the prokaryote-type one-component systems ineffective in the eukaryotes. Furthermore, the regulators of chromatin dynamics, such as the histone deacetylases and associated Swi2 ATPases took up the role of transcriptional repressors [181], and probably resulted in the ancestral prokaryote-type repressors becoming superfluous. Hence, the origin of the eukaryotes was probably accompanied by a massive loss of the prokaryote-type transcription factors, along with the innovation through rapid sequence divergence of new versions that suited the eukaryotic milieu. Some of the HTH domains inherited from the ancestral prokaryotic genomes were also reused by the eukaryotes as adaptor modules in signaling systems un-related to DNA-binding or transcription regulation. However, many versions inherited from the archaea appear to have persisted in the basal transcription machinery, where they were indispensable for transcription of the nuclear genome.

Finally, the rise of organizational complexity in plant, animal and fungal lineages went hand-in-hand with the emergence of new specific transcriptional regulators. Some were drawn from pre-existing HTH families, like the Myb domain, while others like the HSF, Homeo, Pou, Pipsqueak and Paired families arose through rapid divergence from different sources. The HTH domains of transposons provided the source material for some of these domains, whereas other might have diverged rapidly from laterally transferred prokaryotic transcription factors. Thus, on one hand prokaryotes appear to share a sizeable common set of transcription factors, whose phyletic patterns are chiefly governed by the lateral transfers and gene losses acting over and above a basic signal of vertical inheritance. On the other hand, the eukaryotes share only a few unique ancient DNA-binding domains, and their transcription factors have chiefly emerged through multiple lineage-specific expansions.

7. General conclusions

The HTH domain, one of the best-studied of the double-stranded DNA-binding domains, is one of the key protein domains in the transcriptional apparatus of all extant organisms. With the "hindsight" of over two decades of investigations since the discovery of the domain we attempt to provide a synthetic overview of the natural history of the HTH domain from the viewpoint of comparative genomics. Despite the HTH being a rather simple structural scaffold, it is observed to be capable of considerable structural variety and functional versatility, while still preserving a core set of correlated structure–function features. Most HTH domains, despite their structural diversity, participate in a variety of functions that depend on their DNA-binding properties. These include their central role in mediating the substrate interactions of various enzymes that operate on DNA, and their role as both basal and specific transcription regulators. Thus, the HTH domains are the predominant transcription factors in all prokaryotic organisms and the more complex eukaryotes, such as the plants, animals and fungi. Beyond these DNA-binding functions, the HTH domains have been recruited on multiple occasions in a RNA-binding capacity and as mediators of protein–protein interactions. The last universal common ancestor already had anywhere between 6 and 11 distinct versions of the HTH domain, which covered much of the structural diversity, and at least some of the functional diversity seen in the extant versions of the domain. Though several families of specific transcription factors are shared by the two prokaryotic kingdoms and may even go back to their common ancestor, the HTH proteins in the basal transcription factors do not appear to be orthologous. This presents an interesting evolutionary conundrum, whose solution might emerge from new data on alternative transcription and replication systems, like those in viruses and other selfish elements [75,156].

The HTH domain occurs frequently in modular proteins, whose domain architectures are often correlated with the general functional properties of the protein. In prokaryotes the dominant domain architecture is the one-component system that combines the HTH with a sensor domain. While many of the sensor domains of commonly known one-component systems have been characterized previously, several others remain structurally and functionally unexplored, and suggest a new direction for exploring the intricacies of biological sensors in prokaryotic systems. The extensive use of divergent HTH domains in protein–protein interactions, especially in eukaryotes, is another area that might

develop further in the future as the actual mechanistic details of such interactions become clearer. In prokaryotic systems the wealth of sequence and structure data might finally allow us to investigate some of the more difficult problems such as, the overall transcriptional regulatory network of organisms, and the details of how target DNA sequence and ligand specificity are achieved by transcriptional regulators. We hope that the overview presented by us will provide a framework for such future investigations.

8. Supplementary material

A complete list of gis of the HTH domains detected in 183 completely sequenced prokaryotic genomes, and alignments of major families will be available by ftp. ftp.ncbi.nih.gov/pub/aravind/.

References

[1] Ptashne, M. (2004) Genetic Switch: Phage Lambda Revisited. Cold Spring Harbor Laboratory Press, New York.

[2] Tahirov, T.H., Temiakov, D., Anikin, M., Patlan, V., McAllister, W.T., Vassylyev, D.G. and Yokoyama, S. (2002) Structure of a T7 RNA polymerase elongation complex at 2.9 Å resolution. Nature 420, 43–50.

[3] Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K. and Darst, S.A. (1999) Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. Cell 98, 811–824.

[4] Darst, S.A., Polyakov, A., Richter, C. and Zhang, G. (1998) Structural studies of *Escherichia coli* RNA polymerase. Cold Spring Harb. Symp. Quant. Biol. 63, 269–276.

[5] Haldenwang, W.G. (1995) The sigma factors of *Bacillus subtilis*. Microbiol. Rev. 59, 1–30.

[6] Stragier, P. and Losick, R. (1990) Cascades of sigma factors revisited. Mol. Microbiol. 4, 1801–1806.

[7] Kornberg, R.D. (1999) Eukaryotic transcriptional control. Trends Cell Biol. 9, 46–49.

[8] Kornberg, R.D. (1998) Mechanism and regulation of yeast RNA polymerase II transcription. Cold Spring Harb. Symp. Quant. Biol. 63, 229–232.

[9] Ohlendorf, D.H., Anderson, W.F. and Matthews, B.W. (1983) Many gene-regulatory proteins appear to have a similar α-helical fold that binds DNA and evolved from a common precursor. J. Mol. Evol. 19, 109–114.

[10] Ohlendorf, D.H., Anderson, W.F., Fisher, R.G., Takeda, Y. and Matthews, B.W. (1982) The molecular basis of DNA–protein recognition inferred from the structure of cro repressor. Nature 298, 718–723.

[11] Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature 298, 447–451.

[12] Steitz, T.A., Ohlendorf, D.H., McKay, D.B., Anderson, W.F. and Matthews, B.W. (1982) Structural similarity in the DNA-binding domains of catabolite gene activator and cro repressor proteins. Proc. Natl. Acad. Sci. USA 79, 3097–3100.

[13] Matthews, B.W., Ohlendorf, D.H., Anderson, W.F. and Takeda, Y. (1982) Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. Proc. Natl. Acad. Sci. USA 79, 1428–1432.

[14] Gribskov, M. and Burgess, R.R. (1986) Sigma factors from *E. coli*, *B. subtilis*, phage SP01, and phage T4 are homologous proteins. Nucl. Acids Res. 14, 6745–6763.

[15] Yura, T., Tobe, T., Ito, K. and Osawa, T. (1984) Heat shock regulatory gene (htpR) of *Escherichia coli* is required for growth at high temperature but is dispensable at low temperature. Proc. Natl. Acad. Sci. USA 81, 6803–6807.

[16] Landick, R., Vaughn, V., Lau, E.T., VanBogelen, R.A., Erickson, J.W. and Neidhardt, F.C. (1984) Nucleotide sequence of the heat shock regulatory gene of *E. coli* suggests its protein product may be a transcription factor. Cell 38, 175–182.

[17] Frampton, J., Leutz, A., Gibson, T. and Graf, T. (1989) DNA-binding domain ancestry. Nature 342, 134.

[18] Otting, G., Qian, Y.Q., Muller, M., Affolter, M., Gehring, W. and Wuthrich, K. (1988) Secondary structure determination for the Antennapedia homeodomain by nuclear magnetic resonance and evidence for a helix-turn-helix motif. EMBO J. 7, 4305–4309.

[19] Brennan, R.G. and Matthews, B.W. (1989) The helix-turn-helix DNA binding motif. J. Biol. Chem. 264, 1903–1906.

[20] Dodd, I.B. and Egan, J.B. (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. Nucl. Acids Res. 18, 5019–5026.

[21] Dodd, I.B. and Egan, J.B. (1987) Systematic method for the detection of potential lambda Cro-like DNA-binding regions in proteins. J. Mol. Biol. 194, 557–564.

[22] Schultz, S.C., Shields, G.C. and Steitz, T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90°. Science 253, 1001–1007.

[23] Wilson, K.P., Shewchuk, L.M., Brennan, R.G., Otsuka, A.J. and Matthews, B.W. (1992) *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. Proc. Natl. Acad. Sci. USA 89, 9257–9261.

[24] Brennan, R.G. (1993) The winged-helix DNA-binding motif: another helix-turn-helix takeoff. Cell 74, 773–776.

[25] Clark, K.L., Halay, E.D., Lai, E. and Burley, S.K. (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature 364, 412–420.

[26] Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L. and Sweet, R.M. (1993) Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. Nature 362, 219–223.

[27] Swindells, M.B. (1995) Identification of a common fold in the replication terminator protein suggests a possible mode for DNA binding. Trends Biochem. Sci. 20, 300–302.

[28] Kodandapani, R., Pio, F., Ni, C.Z., Piccialli, G., Klemsz, M., McKercher, S., Maki, R.A. and Ely, K.R. (1996) A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex. Nature 380, 456–460.

[29] Gajiwala, K.S. and Burley, S.K. (2000) Winged helix proteins. Curr. Opin. Struct. Biol. 10, 110–116.

[30] Aravind, L. and Koonin, E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. Nucl. Acids Res. 27, 4658–4670.

[31] Bell, S.D. and Jackson, S.P. (2001) Mechanism and regulation of transcription in archaea. Curr. Opin. Microbiol. 4, 208–213.

[32] Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I. and Koonin, E.V. (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res. 9, 608–628.

[33] Bell, S.D. and Jackson, S.P. (1998) Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. Trends Microbiol. 6, 222–228.

[34] Rivera, M.C., Jain, R., Moore, J.E. and Lake, J.A. (1998) Genomic evidence for two functionally distinct gene classes. Proc. Natl. Acad. Sci. USA 95, 6239–6244.
[35] Koonin, E.V., Tatusov, R.L. and Rudd, K.E. (1995) Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. Proc. Natl. Acad. Sci. USA 92, 11921–11925.
[36] Korner, H., Sofia, H.J. and Zumft, W.G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. FEMS Microbiol. Rev. 27, 559–592.
[37] Brown, N.L., Stoyanov, J.V., Kidd, S.P. and Hobman, J.L. (2003) The MerR family of transcriptional regulators. FEMS Microbiol. Rev. 27, 145–163.
[38] Rigali, S.b., Derouaux, A., Giannotta, F. and Dusart, J. (2002) Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. J. Biol. Chem. 277, 12507–12515.
[39] Gallegos, M.T., Schleif, R., Bairoch, A., Hofmann, K. and Ramos, J.L. (1997) Arac/XylS family of transcriptional regulators. Microbiol. Mol. Biol. Rev. 61, 393–410.
[40] Weickert, M.J. and Adhya, S. (1992) A family of bacterial regulators homologous to Gal and Lac repressors. J. Biol. Chem. 267, 15869–15874.
[41] Gehring, W.J., Affolter, M. and Burglin, T. (1994) Homeodomain proteins. Annu. Rev. Biochem. 63, 487–526.
[42] Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. and Teichmann, S.A. (2004) Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14, 283–291.
[43] Teichmann, S.A. and Babu, M.M. (2004) Gene regulatory network growth by duplication. Nat. Genet. 36, 492–496.
[44] Selmer, M. and Su, X.-D. (2002) Crystal structure of an mRNA-binding fragment of *Moorella thermoacetica* elongation factor SelB. EMBO J. 21, 4145–4153.
[45] Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I. and Aggarwal, A.K. (1997) Structure of the multimodular endonuclease FokI bound to DNA. Nature 388, 97–100.
[46] Moore, M.H., Gulbis, J.M., Dodson, E.J., Demple, B. and Moody, P.C. (1994) Crystal structure of a suicidal DNA repair protein: the Ada O6-methylguanine-DNA methyltransferase from *E. coli*. EMBO J. 13, 1495–1501.
[47] Alfano, C., Sanfelice, D., Babon, J., Kelly, G., Jacks, A., Curry, S. and Conte, M.R. (2004) Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. Nat. Struct. Mol. Biol. 11, 323–329.
[48] Dong, G., Chakshusmathi, G., Wolin, S.L. and Reinisch, K.M. (2004) Structure of the La motif: a winged helix domain mediates RNA binding via a conserved aromatic patch. EMBO J. 23, 1000–1007.
[49] Wong, H.C., Mao, J., Nguyen, J.T., Srinivas, S., Zhang, W., Liu, B., Li, L., Wu, D. and Zheng, J. (2000) Structural basis of the recognition of the dishevelled DEP domain in the Wnt signaling pathway. Nat. Struct. Biol. 7, 1178–1184.
[50] Zheng, N., Schulman, B.A., Song, L., Miller, J.J., Jeffrey, P.D., Wang, P., Chu, C., Koepp, D.M., Elledge, S.J., Pagano, M., Conaway, R.C., Conaway, J.W., Harper, J.W. and Pavletich, N.P. (2002) Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex. Nature 416, 703–709.
[51] Guo, F., Gopaul, D.N. and van Duyne, G.D. (1997) Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature 389, 40–46.
[52] Otting, G., Qian, Y.Q., Billeter, M., Muller, M., Affolter, M., Gehring, W.J. and Wuthrich, K. (1990) Protein–DNA contacts in the structure of a homeodomain–DNA complex determined by nuclear magnetic resonance spectroscopy in solution. EMBO J. 9, 3085–3092.
[53] Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G.M. and Gronenborn, A.M. (1997) Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. Nat. Struct. Biol. 4, 567–577.
[54] Mackereth, C.D., Arrowsmith, C.H., Edwards, A.M. and McIntosh, L.P. (2000) Zinc-bundle structure of the essential RNA polymerase subunit RPB10 from *Methanobacterium thermoautotrophicum*. Proc. Natl. Acad. Sci. USA 97, 6316–6321.
[55] Iwahara, J. and Clubb, R.T. (1999) Solution structure of the DNA binding domain from Dead ringer, a sequence-specific AT-rich interaction domain (ARID). EMBO J. 18, 6084–6094.
[56] Zhao, H., Msadek, T., Zapf, J., Madhusudan, Hoch, J.A. and Varughese, K.I. (2002) DNA complexed structure of the key transcription factor initiating development in sporulating bacteria. Structure (Camb) 10, 1041–1050.
[57] Khare, D., Ziegelin, G.n., Lanka, E. and Heinemann, U. (2004) Sequence-specific DNA binding determined by contacts outside the helix-turn-helix motif of the ParB homolog KorB. Nat. Struct. Mol. Biol. 11, 656–663.
[58] Campos, A., Zhang, R.G., Alkire, R.W., Matsumura, P. and Westbrook, E.M. (2001) Crystal structure of the global regulator FlhD from *Escherichia coli* at 1.8 Å resolution. Mol. Microbiol. 39, 567–580.
[59] Schumacher, M.A., Lau, A.O.T. and Johnson, P.J. (2003) Structural basis of core promoter recognition in a primitive eukaryote. Cell 115, 413–424.
[60] Liu, S., Widom, J., Kemp, C.W., Crews, C.M. and Clardy, J. (1998) Structure of human methionine aminopeptidase-2 complexed with fumagillin. Science 282, 1324–1327.
[61] Gomis-Rueth, F.X., Solã¡, M., Acebo, P., Pãrraga, A., Guasch, A., Eritja, R., Gonzalez, A., Espinosa, M., del Solar, G. and Coll, M. (1998) The structure of plasmid-encoded transcriptional repressor CopG unliganded and bound to its operator. EMBO J. 17, 7404–7415.
[62] Cordes, M.H., Walsh, N.P., McKnight, C.J. and Sauer, R.T. (1999) Evolution of a protein fold in vitro. Science 284, 325–328.
[63] Grishin, N.V. (2000) Two tricks in one bundle: helix-turn-helix gains enzymatic activity. Nucl. Acids Res. 28, 2229–2233.
[64] Allen, M., Friedler, A., Schon, O. and Bycroft, M. (2002) The structure of an FF domain from human HYPA/FBP11. J. Mol. Biol. 323, 411–416.
[65] Das, A.K., Helps, N.R., Cohen, P.T. and Barford, D. (1996) Crystal structure of the protein serine/threonine phosphatase 2C at 2.0 Å resolution. EMBO J. 15, 6798–6809.
[66] Aravind, L., Mazumder, R., Vasudevan, S. and Koonin, E.V. (2002) Trends in protein evolution inferred from sequence and structure analysis. Curr. Opin. Struct. Biol. 12, 392–399.
[67] Lewis, J.D., Saperas, N.R., Song, Y., Zamora, M.J., Chiva, M. and Ausió, J. (2004) Histone H1 and the origin of protamines. Proc. Natl. Acad. Sci. USA 101, 4148–4152.
[68] Werhane, H., Lopez, P., Mendel, M., Zimmer, M., Ordal, G.W. and Márquez-Magaña, L.M. (2004) The last gene of the fla/che operon in *Bacillus subtilis*, ylxL, is required for maximal sigmaD function. J. Bacteriol. 186, 4025–4029.
[69] Kearns, D.B., Chu, F., Rudner, R. and Losick, R. (2004) Genes governing swarming in *Bacillus subtilis* and evidence for a phase variation mechanism controlling surface motility. Mol. Microbiol. 52, 357–369.
[70] Campbell, E.A., Muzzin, O., Chlenov, M., Sun, J.L., Olson, C.A., Weinman, O., Trester-Zedlitz, M.L. and Darst, S.A. (2002) Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Mol. Cell. 9, 527–539.
[71] Aasland, R., Stewart, A.F. and Gibson, T. (1996) The SANT domain: a putative DNA-binding domain in the SWI-SNF and ADA complexes, the transcriptional co-repressor N-CoR and TFIIIB. Trends Biochem. Sci. 21, 87–88.

[72] Yang, H., Jeffrey, P.D., Miller, J., Kinnucan, E., Sun, Y., Thoma, N.H., Zheng, N., Chen, P.-L., Lee, W.-H. and Pavletich, N.P. (2002) BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure. Science 297, 1837–1848.

[73] Iyer, L.M., Makarova, K.S., Koonin, E.V. and Aravind, L. (2004) Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. Nucl. Acids Res. 32, 5260–5279.

[74] Aussel, L., Barre, F.X., Aroyo, M., Stasiak, A., Stasiak, A.Z. and Sherratt, D. (2002) FtsK Is a DNA motor protein that activates chromosome dimer resolution by switching the catalytic state of the XerC and XerD recombinases. Cell 108, 195–205.

[75] Giraldo, R. and FernÃ¡ ndez-Tresguerres, M.E. (2004) Twenty years of the pPS10 replicon: insights on the molecular mechanism for the activation of DNA replication in iteron-containing bacterial plasmids. Plasmid 52, 69–83.

[76] Berge, M., Mortier-Barriére, I., Martin, B. and Claverys, J.-P. (2003) Transformation of *Streptococcus pneumoniae* relies on DprA- and RecA-dependent protection of incoming DNA single strands. Mol. Microbiol. 50, 527–536.

[77] Pietrokovski, S. and Henikoff, S. (1997) A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. Mol. Gen. Genet. 254, 689–695.

[78] Belova, G.I., Prasad, R., Nazimov, I.V., Wilson, S.H. and Slesarev, A.I. (2002) The domain organization and properties of individual domains of DNA topoisomerase V, a type 1B topoisomerase with DNA repair activities. J. Biol. Chem. 277, 4959–4965.

[79] Tobe, T., Sasakawa, C., Okada, N., Honma, Y. and Yoshikawa, M. (1992) vacB, a novel chromosomal gene required for expression of virulence genes on the large plasmid of *Shigella flexneri*. J. Bacteriol. 174, 6359–6367.

[80] Angermayr, M., Roidl, A. and Bandlow, W. (2002) Yeast Rio1p is the founding member of a novel subfamily of protein serine kinases involved in the control of cell cycle progression. Mol. Microbiol. 44, 309–324.

[81] LaRonde-LeBlanc, N. and Wlodawer, A. (2004) Crystal structure of *A. fulgidus* Rio2 defines a new family of serine protein kinases. Structure (Cambridge) 12, 1585–1594.

[82] Peat, T.S., Frank, E.G., McDonald, J.P., Levine, A.S., Woodgate, R. and Hendrickson, W.A. (1996) Structure of the UmuD′ protein and its regulation in response to DNA damage. Nature 380, 727–730.

[83] Savijoki, K., Ingmer, H., Frees, D., Vogensen, F.K., Palva, A. and Varmanen, P. (2003) Heat and DNA damage induction of the LexA-like regulator HdiR from *Lactococcus lactis* is mediated by RecA and ClpP. Mol. Microbiol. 50, 609–621.

[84] Zhang, X., Chaney, M., Wigneshweraraj, S.R., Schumacher, J., Bordes, P., Cannon, W. and Buck, M. (2002) Mechanochemical ATPases and transcriptional activation. Mol. Microbiol. 45, 895–903.

[85] Leipe, D.D., Koonin, E.V. and Aravind, L. (2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. J. Mol. Biol. 343, 1–28.

[86] Wang, Y., Zhao, S., Somerville, R.L. and Jardetzky, O. (2001) Solution structure of the DNA-binding domain of the TyrR protein of *Haemophilus influenzae*. Protein Sci. 10, 592–598.

[87] Larquet, E., Schreiber, V.r., Boisset, N. and Richet, E. (2004) Oligomeric assemblies of the *Escherichia coli* MalT transcriptional activator revealed by cryo-electron microscopy and image processing. J. Mol. Biol. 343, 1159–1169.

[88] Poon, K.K., Chu, J.C. and Wong, S.L. (2001) Roles of glucitol in the GutR-mediated transcription activation process in *Bacillus subtilis*: glucitol induces GutR to change its conformation and to bind ATP. J. Biol. Chem. 276, 29819–29825.

[89] Lee, P.-C., Umeyama, T. and Horinouchi, S. (2002) afsS is a target of AfsR, a transcriptional factor with ATPase activity that globally controls secondary metabolism in *Streptomyces coelicolor* A(32). Mol. Microbiol. 43, 1413–1430.

[90] Pao, G.M. and Saier, M.H. (1995) Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J. Mol. Evol. 40, 136–154.

[91] West, A.H. and Stock, A.M. (2001) Histidine kinases and response regulator proteins in two-component signaling systems. Trends Biochem. Sci. 26, 369–376.

[92] Stulke, J. and Hillen, W. (1998) Coupling physiology and gene regulation in bacteria: the phosphotransferase sugar uptake system delivers the signals. Naturwissenschaften 85, 583–592.

[93] Stulke, J., Arnaud, M., Rapoport, G. and Martin-Verstraete, I. (1998) PRD–a protein domain involved in PTS-dependent induction and carbon catabolite repression of catabolic operons in bacteria. Mol. Microbiol. 28, 865–874.

[94] Hu, K.-Y. and Saier, M.H. (2002) Phylogeny of phosphoryl transfer proteins of the phosphoenolpyruvate-dependent sugar-transporting phosphotransferase system. Res. Microbiol. 153, 405–415.

[95] Reizer, J. and Saier, M.H. (1997) Modular multidomain phosphoryl transfer proteins of bacteria. Curr. Opin. Struct. Biol. 7, 407–415.

[96] Tobisch, S., Stülke, J. and Hecker, M. (1999) Regulation of the lic operon of *Bacillus subtilis* and characterization of potential phosphorylation sites of the LicR regulator protein by site-directed mutagenesis. J. Bacteriol. 181, 4995–5003.

[97] Anantharaman, V., Koonin, E.V. and Aravind, L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. J. Mol. Biol. 307, 1271–1292.

[98] Hofmann, K. and Bucher, P. (1995) The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. Trends Biochem. Sci. 20, 347–349.

[99] Molle, V., Kremer, L., Girard-Blanc, C., Besra, G.S., Cozzone, A.J. and Prost, J.-F.O. (2003) An FHA phosphoprotein recognition domain mediates protein EmbR phosphorylation by PknH, a Ser/Thr protein kinase from *Mycobacterium tuberculosis*. Biochemistry 42, 15300–15309.

[100] Taylor, B.L., Zhulin, I.B. and Johnson, M.S. (1999) Aerotaxis and other energy-sensing behavior in bacteria. Annu. Rev. Microbiol. 53, 103–128.

[101] Tyrrell, R., Verschueren, K.H., Dodson, E.J., Murshudov, G.N., Addy, C. and Wilkinson, A.J. (1997) The structure of the cofactor-binding fragment of the LysR family member, CysB: a familiar fold with a surprising subunit arrangement. Structure 5, 1017–1032.

[102] Tam, R. and Saier, M.H. (1993) Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. Microbiol. Rev. 57, 320–346.

[103] Vartak, N.B., Reizer, J., Reizer, A., Gripp, J.T., Groisman, E.A., Wu, L.F., Tomich, J.M. and Saier, M.H. (1991) Sequence and evolution of the FruR protein of *Salmonella typhimurium*: a pleiotropic transcriptional regulatory protein possessing both activator and repressor functions which is homologous to the periplasmic ribose-binding protein. Res. Microbiol. 142, 951–963.

[104] Ettema, T.J.G., Brinkman, A.B., Tani, T.H., Rafferty, J.B. and Van Der Oost, J. (2002) A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes. J. Biol. Chem. 277, 37464–37468.

[105] Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J. Mol. Biol. 287, 1023–1040.
[106] Bull, P.C. and Cox, D.W. (1994) Wilson disease and Menkes disease: new handles on heavy-metal transport. Trends Genet. 10, 246–252.
[107] Dunwell, J.M., Culham, A., Carter, C.E., Sosa-Aguirre, C.R. and Goodenough, P.W. (2001) Evolution of functional diversity in the cupin superfamily. Trends Biochem. Sci. 26, 740–746.
[108] Bateman, A. (1997) The structure of a domain common to archaebacteria and the homocystinuria disease protein. Trends Biochem. Sci. 22, 12–13.
[109] Heldwein, E.E. and Brennan, R.G. (2001) Crystal structure of the transcription activator BmrR bound to DNA and a drug. Nature 409, 378–382.
[110] Aravind, L. and Anantharaman, V. (2003) HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. FEMS Microbiol. Lett. 222, 17–23.
[111] Zhang, R.G., Andersson, C.E., Savchenko, A., Skarina, T., Evdokimova, E., Beasley, S., Arrowsmith, C.H., Edwards, A.M., Joachimiak, A. and Mowbray, S.L. (2003) Structure of *Escherichia coli* ribose-5-phosphate isomerase: a ubiquitous enzyme of the pentose phosphate pathway and the Calvin cycle. Structure (Cambridge) 11, 31–42.
[112] Ailion, M., Bobik, T.A. and Roth, J.R. (1993) Two global regulatory systems (Crp and Arc) control the cobalamin/propanediol regulon of *Salmonella typhimurium*. J. Bacteriol. 175, 7200–7208.
[113] Miras, I., Hermant, D., Arricau, N. and Popoff, M.Y. (1995) Nucleotide sequence of iagA and iagB genes involved in invasion of HeLa cells by *Salmonella enterica* subsp. *enterica* ser. Typhi. Res. Microbiol. 146, 17–20.
[114] Abergel, C., Bouveret, E., Claverie, J.M., Brown, K., Rigal, A., Lazdunski, C. and Benedetti, H. (1999) Structure of the *Escherichia coli* TolB protein determined by MAD methods at 1.95 Å resolution. Struct. Fold Des. 7, 1291–1300.
[115] Carr, S., Penfold, C.N., Bamford, V., James, R. and Hemmings, A.M. (2000) The structure of TolB, an essential component of the tol-dependent translocation system, and its protein–protein interaction with the translocation domain of colicin E9. Struct. Fold Des. 8, 57–66.
[116] Walburger, A., Lazdunski, C. and Corda, Y. (2002) The Tol/Pal system function requires an interaction between the C-terminal domain of TolA and the N-terminal domain of TolB. Mol. Microbiol. 44, 695–708.
[117] Hofmann, K. and Bucher, P. (1998) The PCI domain: a common theme in three multiprotein complexes. Trends Biochem. Sci. 23, 204–205.
[118] Aravind, L. and Ponting, C.P. (1998) Homologues of 26S proteasome subunits are regulators of transcription and translation. Protein Sci. 7, 1250–1254.
[119] Wernimont, A. and Weissenhorn, W. (2004) Crystal structure of subunit VPS25 of the endosomal trafficking complex ESCRT-II. BMC Struct. Biol. 4, 10.
[120] Teo, H., Perisic, O., GonzÃ¡lez, B. and Williams, R.L. (2004) ESCRT-II, an endosome-associated complex required for protein sorting: crystal structure and interactions with ESCRT-III and membranes. Dev. Cell 7, 559–569.
[121] Hierro, A., Sun, J., Rusnak, A.S., Kim, J., Prag, G., Emr, S.D. and Hurley, J.H. (2004) Structure of the ESCRT-II endosomal trafficking complex. Nature 431, 221–225.
[122] Gibson, T.J., Thompson, J.D., Blocker, A. and Kouzarides, T. (1994) Evidence for a protein domain superfamily shared by the cyclins, TFIIB and RB/p107. Nucl. Acids Res. 22, 946–952.
[123] Yan, Y., Barlev, N.A., Haley, R.H., Berger, S.L. and Marmorstein, R. (2000) Crystal structure of yeast Esa1 suggests a unified mechanism for catalysis and substrate binding by histone acetyltransferases. Mol. Cell 6, 1195–1205.
[124] Zubieta, C., He, X.Z., Dixon, R.A. and Noel, J.P. (2001) Structures of two natural product methyltransferases reveal the basis for substrate specificity in plant *O*-methyltransferases. Nat. Struct. Biol. 8, 271–279.
[125] Kim, J. and Raushel, F.M. (2001) Allosteric control of the oligomerization of carbamoyl phosphate synthetase from *Escherichia coli*. Biochemistry 40, 11030–11036.
[126] Lawson, F.S., Charlebois, R.L. and Dillon, J.A. (1996) Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. Mol. Biol. Evol. 13, 970–977.
[127] Leipe, D.D., Aravind, L. and Koonin, E.V. (1999) Did DNA replication evolve twice independently. Nucl. Acids Res. 27, 3389–3401.
[128] Rubbi, L., Labarre-Mariotte, S., Chéldin, S. and Thuriaux, P. (1999) Functional characterization of ABC10alpha, an essential polypeptide shared by all three forms of eukaryotic DNA-dependent RNA polymerases. J. Biol. Chem. 274, 31485–31492.
[129] Gruber, T.M. and Gross, C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. Annu. Rev. Microbiol. 57, 441–466.
[130] Iyer, L.M., Koonin, E.V. and Aravind, L. (2004) Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene 335, 73–88.
[131] Morett, E. and Bork, P. (1998) Evolution of new protein function: recombinational enhancer Fis originated by horizontal gene transfer from the transcriptional regulator NtrC. FEBS Lett. 433, 108–112.
[132] Gruene, T., Brzeski, J., Eberharter, A., Clapier, C.R., Corona, D.F.V., Becker, P.B. and Mueller, C.W. (2003) Crystal structure and functional analysis of a nucleosome recognition module of the remodeling factor ISWI. Mol. Cell. 12, 449–460.
[133] Juan Wu, L. and Errington, J. (2000) Identification and characterization of a new prespore-specific regulatory gene, rsfA, of *Bacillus subtilis*. J. Bacteriol. 182, 418–424.
[134] Izsak, Z., Khare, D., Behlke, J., Heinemann, U., Plasterk, R.H. and Ivics, Z.n. (2002) Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in sleeping beauty transposition. J. Biol. Chem. 277, 34581–34588.
[135] Czerny, T., Schaffner, G. and Busslinger, M. (1993) DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. Genes Dev. 7, 2048–2061.
[136] Tanaka, Y., Nureki, O., Kurumizaka, H., Fukai, S., Kawaguchi, S., Ikuta, M., Iwahara, J., Okazaki, T. and Yokoyama, S. (2001) Crystal structure of the CENP-B protein–DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA. EMBO J. 20, 6612–6618.
[137] Smit, A.F. and Riggs, A.D. (1996) Tiggers and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA 93, 1443–1448.
[138] Subramanian, G., Koonin, E.V. and Aravind, L. (2000) Comparative genome analysis of the pathogenic spirochetes *Borrelia burgdorferi* and *Treponema pallidum*. Infect. Immun. 68, 1633–1648.
[139] Close, S.M. and Kado, C.I. (1992) A gene near the plasmid pSa origin of replication encodes a nuclease. Mol. Microbiol. 6, 521–527.
[140] Anantharaman, V. and Aravind, L. (2003) New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. Genome Biol., 4.
[141] Wood, H.E., Devine, K.M. and McConnell, D.J. (1990) Characterisation of a repressor gene (xre) and a temperature-

sensitive allele from the *Bacillus subtilis* prophage, PBSX. Gene 96, 83–88.

[142] Takemaru, K.i., Li, F.Q., Ueda, H. and Hirose, S. (1997) Multiprotein bridging factor 1 (MBF1) is an evolutionarily conserved transcriptional coactivator that connects a regulatory factor and TATA element-binding protein. Proc. Natl. Acad. Sci. USA 94, 7251–7256.

[143] Rhee, S., Martin, R.G., Rosner, J.L. and Davies, D.R. (1998) A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. Proc. Natl. Acad. Sci. USA 95, 10413–10418.

[144] Brickman, T.J., Kang, H.Y. and Armstrong, S.K. (2001) Transcriptional activation of *Bordetella alcaligin* siderophore genes requires the AlcR regulator with alcaligin as inducer. J. Bacteriol. 183, 483–489.

[145] Fujikawa, N., Kurumizaka, H., Nureki, O., Terada, T., Shirouzu, M., Katayama, T. and Yokoyama, S. (2003) Structural basis of replication origin recognition by the DnaA protein. Nucl. Acids Res. 31, 2077–2086.

[146] Messer, W. and Weigel, C. (2003) DnaA as a transcription regulator. Meth. Enzymol. 370, 338–349.

[147] Wilsker, D., Patsialou, A., Dallas, P.B. and Moran, E. (2002) ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. Cell Growth Differ. 13, 95–106.

[148] Yamada, K., Miyata, T., Tsuchiya, D., Oyama, T., Fujiwara, Y., Ohnishi, T., Iwasaki, H., Shinagawa, H., Ariyoshi, M., Mayanagi, K. and Morikawa, K. (2002) Crystal structure of the RuvA–RuvB complex: a structural basis for the Holliday junction migrating motor machinery. Mol. Cell 10, 671–681.

[149] Soppa, J.r., Kobayashi, K., Noirot-Gros, M.-F.o., Oesterhelt, D., Ehrlich, S.D., Dervyn, E., Ogasawara, N. and Moriya, S. (2002) Discovery of two novel families of proteins that are proposed to interact with prokaryotic SMC proteins, and characterization of the *Bacillus subtilis* family members ScpA and ScpB. Mol. Microbiol. 45, 59–71.

[150] Mascarenhas, J., Soppa, J.R., Strunnikov, A.V. and Graumann, P.L. (2002) Cell cycle-dependent localization of two novel prokaryotic chromosome segregation and condensation proteins in *Bacillus subtilis* that interact with SMC protein. EMBO J. 21, 3108–3118.

[151] Schwartz, T., Rould, M.A., Lowenhaupt, K., Herbert, A. and Rich, A. (1999) Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. Science 284, 1841–1845.

[152] Schade, M., Turner, C.J., Lowenhaupt, K., Rich, A. and Herbert, A. (1999) Structure–function analysis of the Z-DNA-binding domain Zalpha of dsRNA adenosine deaminase type I reveals similarity to the $(\alpha + \beta)$ family of helix-turn-helix proteins. EMBO J. 18, 470–479.

[153] Giraldo, R. and Diaz-Orejas, R. (2001) Similarities between the DNA replication initiators of Gram-negative bacteria plasmids (RepA) and eukaryotes (Orc4p)/archaea (Cdc6p). Proc. Natl. Acad. Sci. USA 98, 4938–4943.

[154] Liu, J., Smith, C.L., DeRyckere, D., DeAngelis, K., Martin, G.S. and Berger, J.M. (2000) Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. Mol. Cell 6, 637–648.

[155] Emery, P., Strubin, M., Hofmann, K., Bucher, P., Mach, B. and Reith, W. (1996) A consensus motif in the RFX DNA binding domain and binding domain mutants with altered specificity. Mol. Cell Biol. 16, 4486–4494.

[156] Yeo, H.-J., Ziegelin, G.N., Korolev, S., Calendar, R., Lanka, E. and Waksman, G. (2002) Phage P4 origin-binding domain structure reveals a mechanism for regulation of DNA-binding activity by homo- and heterodimerization of winged helix proteins. Mol. Microbiol. 43, 855–867.

[157] Anantharaman, V., Koonin, E.V. and Aravind, L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucl. Acids Res. 30, 1427–1464.

[158] Spinelli, S.L., Kierzek, R., Turner, D.H. and Phizicky, E.M. (1999) Transient ADP-ribosylation of a 2′-phosphate implicated in its removal from ligated tRNA during splicing in yeast. J. Biol. Chem. 274, 2637–2644.

[159] Zheng, N., Fraenkel, E., Pabo, C.O. and Pavletich, N.P. (1999) Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev. 13, 666–674.

[160] Groft, C.M., Uljon, S.N., Wang, R. and Werner, M.H. (1998) Structural homology between the Rap30 DNA-binding domain and linker histone H5: implications for preinitiation complex assembly. Proc. Natl. Acad. Sci. USA 95, 9117–9122.

[161] Meinhart, A., Blobel, J. and Cramer, P. (2003) An extended winged helix domain in general transcription factor E/IIE alpha. J. Biol. Chem. 278, 48267–48274.

[162] Landsman, D. and Wolffe, A.P. (1995) Common sequence and structural features in the heat-shock factor and Ets families of DNA-binding domains. Trends Biochem. Sci. 20, 225–226.

[163] Mackereth, C.D., Scharpf, M., Gentile, L.N., MacIntosh, S.E., Slupsky, C.M. and McIntosh, L.P. (2004) Diversity in structure and function of the Ets family PNT domains. J. Mol. Biol. 342, 1249–1264.

[164] Dou, X., Limmer, S. and Kreutzer, R. (2001) DNA-binding of phenylalanyl-tRNA synthetase is accompanied by loop formation of the double-stranded DNA. J. Mol. Biol. 305, 451–458.

[165] Laursen, B.S.g., Mortensen, K.K., Sperling-Petersen, H.U. and Hoffman, D.W. (2003) A conserved structural motif at the N terminus of bacterial translation initiation factor IF2. J. Biol. Chem. 278, 16320–16328.

[166] Sam, M.D., Cascio, D., Johnson, R.C. and Clubb, R.T. (2004) Crystal structure of the excisionase–DNA complex from bacteriophage lambda. J. Mol. Biol. 338, 229–240.

[167] Wojciak, J.M., Iwahara, J. and Clubb, R.T. (2001) The Mu repressor–DNA complex contains an immobilized 'wing' within the minor groove. Nat. Struct. Biol. 8, 84–90.

[168] de Beer, T., Fang, J., Ortega, M., Yang, Q., Maes, L., Duffy, C., Berton, N., Sippy, J., Overduin, M., Feiss, M. and Catalano, C.E. (2002) Insights into specific DNA recognition during the assembly of a viral genome packaging machine. Mol. Cell 9, 981–991.

[169] Ikegami, T., Kuraoka, I., Saijo, M., Kodo, N., Kyogoku, Y., Morikawa, K., Tanaka, K. and Shirakawa, M. (1998) Solution structure of the DNA- and RPA-binding domain of the human repair factor XPA. Nat. Struct. Biol. 5, 701–706.

[170] Agrawal, R.K., Linde, J., Sengupta, J., Nierhaus, K.H. and Frank, J. (2001) Localization of L11 protein on the ribosome and elucidation of its involvement in EF-G-dependent translocation. J. Mol. Biol. 311, 777–787.

[171] van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. Trends Genet. 19, 479–484.

[172] Oshima, K., Kakizawa, S., Nishigawa, H., Jung, H.-Y., Wei, W., Suzuki, S., Arashida, R., Nakata, D., Miyata, S.-i., Ugaki, M. and Namba, S. (2004) Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nat. Genet. 36, 27–29.

[173] Lee, I.-M., Zhao, Y. and Bottner, K.D. (2003) Identification of putative insertion sequence (IS) associated with members of the aster yellows (AY) phytoplasma group. Phytopathology, 93.

[174] Templeton, T.J., Iyer, L.M., Anantharaman, V., Enomoto, S., Abrahante, J.E., Subramanian, G.M., Hoffman, S.L., Abrahamsen, M.S. and Aravind, L. (2004) Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Res. 14, 1686–1695.

[175] Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116, 281–297.

[176] Mattick, J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2, 986–991.
[177] Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E.R., NesbÃ¸, C.L., Case, R.J. and Doolittle, W.F. (2003) Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. 37, 283–328.
[178] Edgell, D.R. and Doolittle, W.F. (1997) Archaea and the origin(s) of DNA replication proteins. Cell 89, 995–998.
[179] Iyer, L.M., Koonin, E.V. and Aravind, L. (2003) Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. BMC Struct. Biol. 3, 1.
[180] Brinkman, A.B., Bell, S.D., Lebbink, R.J., de Vos, W.M. and van der Oost, J. (2002) The *Sulfolobus solfataricus* Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. J. Biol. Chem. 277, 29537–29549.
[181] Grewal, S.I.S. and Moazed, D. (2003) Heterochromatin and epigenetic control of gene expression. Science 301, 798–802.