The evolution of sequence specificity in a DNA binding protein family

Meghna Nandy, Mohak Sharda and Aswin Sai Narain Seshasayee*

National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru 560065, India
*For correspondence: aswin@ncbs.res.in

Abstract

Transcriptional regulation enables bacteria to adjust to its environment. This is driven by transcription factors (TFs), which display DNA site recognition specificity with some flexibility built-in. TFs however are not considered essential to a minimal cellular life. How do they evolve? It has been hypothesised that TFs evolve by gaining specificity (and other functions) on a background of non-specific chromosome structuring proteins. We used the IHF/HU family of DNA binding proteins, in which IHF binds DNA in a sequence specific manner whereas HU binds more indiscriminately, to test this hypothesis. We show that neither protein was present in the last common bacterial ancestor, and that HU evolved first in early-branching clades. By reconstructing ancestral sequences *in-silico* on a rooted phylogeny of IHF/HU we show that the common ancestor of this family was probably HU-like and therefore non-specific in binding DNA. Various residues characteristic of IHF and shown to be involved in specific sequence recognition (at least in *E. coli*) evolved gradually with several gains and losses along the phylogeny. Finally, we suggest that the evolution of the IHF/HU from a HU-like ancestor to IHF, and by inference the evolution of DNA site specificity in this family, was driven by selection.

Introduction

Transcriptional regulation enables an organism to react to changing circumstances. In bacteria this is primarily driven by transcription factors (TFs), which bind to sites on the DNA usually upstream or just downstream of transcription start sites and activate or repress the expression of proximal genes (Browning & Busby 2004, 2016; Balleza et al. 2009). The activity of TFs is thus mediated primarily by their ability to recognise and bind to particular nucleotide motifs in DNA, a highly evolvable property (Rk et al. 2012). Unlike restriction modification systems that bind to short sequence motifs with exquisite specificity, TFs binding to their target sites allows for some flexibility in their DNA site recognition.

Though regulation of transcription by TFs is important to adaptation, it is not absolutely required for cellular life. For example, surveys of essential and conserved genes across genomes suggest that TFs are rarely conserved across a large phylogenetic breadth (Gil et al. 2004); for instance, bacterial endosymbionts rarely code for TFs. All in all, increase in bacterial genome size is accompanied by a quadratic increase in the number of TFs (Cases et al. 2003; Molina & van Nimwegen 2009). This suggests that the need for TFs emerges and expands super-linearly with increase in complexity as defined by the number of gene functions required by an organism. Thus, though the ability of a protein to bind to DNA, for example compress it so that it could fit within the dimensions of a cell, is necessary for even the most basal cellular life, the ability to regulate gene expression precisely using TFs is not. Thus, the question arises - where and when did TFs evolve?

Sandhya Visweswariah and Stephen Busby point out that global TFs often bind to a large number of low-affinity sites and bend DNA, not unlike nucleoid-associated proteins, in addition to binding at target sites that mediate regulation of gene expression of proximal genes (Visweswariah & Busby 2015). In fact, the boundary between a global TF and a nucleoid-associated protein is blurry (Dorman et al. 2020). They used this observation to hypothesise that specific DNA site recognition in TFs may have evolved on a background of nonspecific DNA binding observed in nucleoid associated proteins (NAPs). This idea however remains to be tested systematically against data. Such testing would require, at the very least, reliable sequence alignments of both specific and non-specific DNA binding proteins, which can

then permit ancestral sequence reconstructions. However, such diversity in DNA recognition specificity may manifest in wide variation in the sequences of the DNA binding domains of these proteins, which render the task of obtaining good sequence alignments difficult. Doing so would therefore require sequences of proteins belonging to the same sequence family, known to include both sequence-specific and non-specific DNA binding proteins.

Integration host factor (IHF) and Histone-like DNA-binding protein (HU) provide an interesting model system to evaluate this idea (Drlica & Rouviere-Yaniv 1987; Friedman 1988; Boubrik et al. 1991). Both proteins, at least in *E. coli* and several other bacteria, comprise two subunits α and β and bind to DNA as dimers. All the IHF and HU subunits are characterised by the presence of two α helices followed by five β strands forming a sheet and ending at an α helix. HU binds to DNA, but has no known recognition motif and thus has little sequence specificity (Swinger & Rice 2007). On the other hand, IHF is known to bind DNA with high affinity to a consensus sequence motif (Kd=2-20 nM ; consensus: <A/T>ATCA<A/G>NNNNTTR) (Craig & Nash 1984). Despite the variation in their DNA binding specificity, both proteins belong to the same, tightly alignable sequence family. It is believed that IHF evolved from HU (Visweswariah & Busby 2015), but this belief is grounded in learned intuition rather than systematic evidence. In this study, under the assumption that HU and IHF share the same common ancestor, we investigate the evolutionary trajectory of this sequence family, in particular, the evolution of DNA recognition specificity involved in TF activity.

Results

HU and IHF subunits across bacterial phylogeny

We obtained the sequences of the α and β subunits of histone like (HU) and integration host factor (IHF) proteins from the OMA database (Altenhoff et al. 2024). The final dataset used in our study comprised 263 HUα, 626 HUβ, 517 IHFα and 413 IHFβ subunit sequences (Figure 1A). We clustered all these sequences using Principal Component Analysis (Konishi et al. 2019). We found that HU and IHF sequences clustered separately, as did IHFα and IHFβ. However, the HUα and HUβ sequences clustered together without any clear separation between the two (Figure 1B), consistent with findings from earlier work (Dey et al. 2017; Kamashev et al. 2017). Therefore, the superset of HUα and HUβ subunit sequences is referred to as 'HU' in the downstream analyses.

We assembled a 16S rRNA sequence-based species tree for all the bacteria represented in our dataset, and rooted it using the archaeaon *Methanocaldococcus indicus* as the outgroup (Figure 1C). We then mapped the presence of the four subunits on our species tree (Figure 1C). At least one HU subunit is present in 79% of the bacteria represented in our study whereas at least one IHF subunit is found in 64% (Figure 1A and C). This reiterates the widely held idea that HU is much more conserved than IHF (Prieto et al. 2012). Within HU, the HUβ subunit is more abundant across bacteria (found in 64% of bacteria) than HUα (27% of bacteria). Even in bacteria like *E. coli*, that have both HUα and HUβ subunits, both homodimers and heterodimers can form, and each of these forms might have distinct functions (Tanaka et al. 1993). IHF is usually present as a heterodimer in *E. coli* (Drlica & Rouviere-

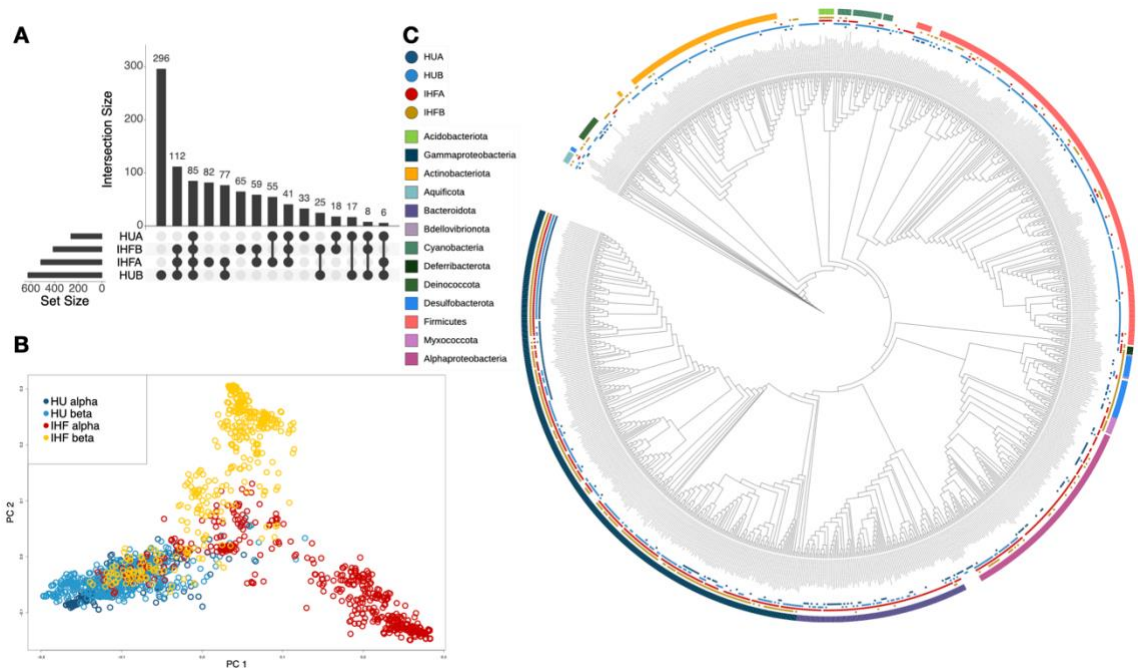

Figure 1: *The IHF/HU sequence family.* (A) An UpSet plot showing the distribution of the four IHF/HU subunits across bacteria. (B) PCA based clustering of IHF and HU sequences. Dark blue circles: HUα, Light blue circles: HUβ, Red circles: IHFα, Yellow circles: IHFβ, Triangles: TraM, an outgroup (See Figure 2A). (C) The bacterial species tree. Outer circle shows various clades of bacteria. Inner circles show the presence of the four subunits of IHF/HU using the same colour scheme as in (B).

Yaniv 1987). However, both IHFα and IHFβ are present together in only 30% of bacteria, whereas 34% of bacteria code for only one of the two subunits (Figure 1A and C). While, IHFα is present in 53% of bacteria, the IHFβ subunit is present in 42%.

Early branching clades, until the emergence of *Firmicutes*, mostly harbour only HU sequences (Figure 1C). Specifically, *Actinobacteria* and *Firmicutes* mostly contain HUβ. The distribution of HUα is sparse across the phylogeny except in the late emerging bacteria in the *Gammaproteobacteria* phylum. An alternating presence of HU α and β was observed in *Bacteroidota*. The phyla emerging after *Firmicutes*, such as *Desulfobacterota*, exhibit a marked increase in the presence of IHFβ in their genomes, while *Myxococcota* show an increase in the presence of IHFα in addition to IHFβ. *Alphaproteobacteria* and *Bacteroidota* mostly contain the IHFα subunit in addition to either HU subunit. Among *Gammaproteobacteria*, the early emerging genomes exhibit the presence of both subunits of IHF but only the β subunit of HU; late emerging subclades of *Gammaproteobacteria* such as the *Enterobacteriaceae* - *Klebsiella*, *Escherichia*, *Shigella* and related bacteria - have all the four subunits (Figure 1C).

We then reconstructed the presence of the subunits across nodes of the species tree to understand how the dynamics of emergence and loss of the subunits has defined the extant distribution of HU and IHF subunits (Figure 2). None of the protein subunits seems to be present at the root, or the last common ancestor of bacteria and archaea, or in the bacterial ancestor. HUβ appears to have emerged the earliest of all the four subunits, in the ancestor of the *Deinococcota* phylum (Figure 2B). It exhibits a major loss in the *Deferribacterota*, *Bdellovibrionota*, *Desulfobacterota* and *Myxococcota* phyla as well as the ancestor of *Alphaproteobacteria*. Within *Alphaproteobacteria*, it is gained and maintained in a smaller clade. The late-branching common ancestor of *Gammaproteobacteria* and *Bacteroidota* harboured HUβ. HUβ was lost several times within *Bacteroidota*; and it is present in most *Gammaproteobacteria*. The first major emergence of HUα was in the common ancestor of *Deferribacterota*, *Bdellovibrionota*,

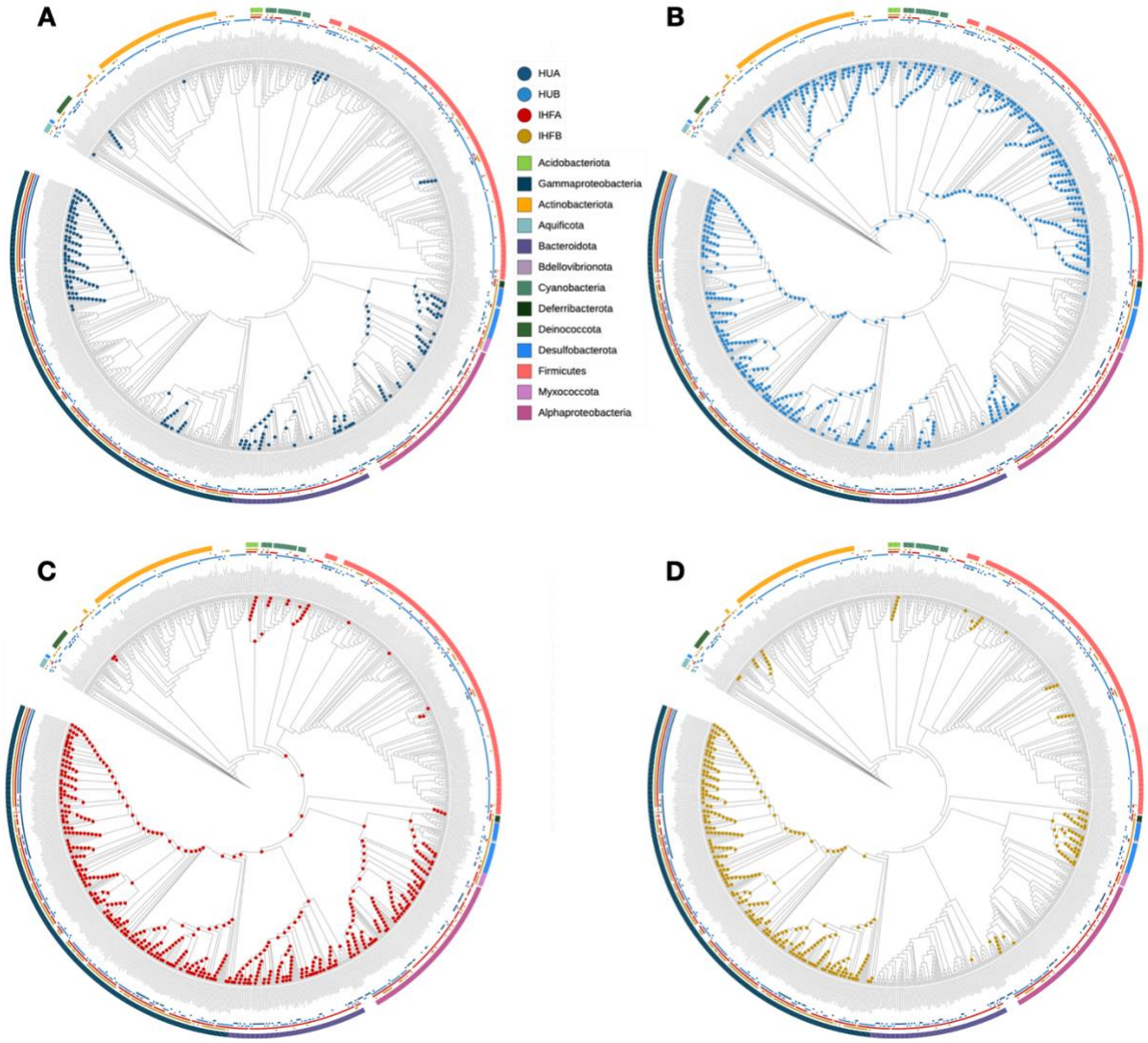

Figure 2: *Gains and losses of IHF/HU subunits across bacteria*. Species trees are as in Figure 1C. Here, the presence of the four IHF/HU subunits in the ancestral / internal nodes is shown for (A) HUA (B) HUB (C) IHFA and (D) IHFB. The presence of a coloured circle indicates the presence of the subunit.

Myxococcota and *Alphaproteobacterota* (Figure 2A). This subunit is however present sporadically in a few terminal clades of earlier-branching bacteria. It is lost in the ancestor of the clade that includes *Myxococcota* and *Alphaproteobacteria*. HUα was gained in the common ancestor of *Bacteroidota*, but replaced sporadically by HUβ in the terminal clades (Figure 2A and B). Finally, the most distinct gain of HUα was in late emerging clades of *Gammaproteobacteria* including *E. coli*, where it is consistently maintained (Figure 2A). Here we re-emphasise the fact that the discrimination of HUα from HUβ is purely based on orthology annotations in the OMA database and that in our alignments (and in other published works), the sequences of the two subunits cluster together without any clear demarcation between the two. The common ancestor of *Acidobacteria* and *Cyanobacteria* is the earliest node with IHFα (Figure 2C). Downstream, it is lost in the ancestor of *Firmicutes*. It is then maintained throughout the subsequent clades with the exception of *Deferribacterota* and *Desulfobacterota*. IHFβ emerged in the common ancestor of *Deferribacterota*, *Desulfobacterota* and *Myxococcota*, where it is maintained in almost all of the genomes. Another major emergence of IHFβ occurred as late as in a *Gammaproteobacteria* ancestor (Figure 2D). IHFβ is largely maintained downstream but for the occasional loss in a few bacteria.

Thus, HUβ, among all four subunits of HU and IHF, emerged first, with HUα, being rarer, emerging much later. The two subunits of IHF do not always occur together, with IHFα being somewhat more conserved than IHFβ (Figure 2).

IHF and HU gene tree

We constructed a rooted gene phylogeny to conduct subsequent evolutionary analyses on the IHF/HU sequence family. We first made an amino acid sequence alignment, consisting of all the 1819 IHF and HU protein subunits in our dataset, which we then converted to a codon alignment. An appropriately rooted phylogeny is essential for estimating the origins of this family. To do this we searched for a protein sequence that could be used as an outgroup for this sequence family. Some published phylogenies of this sequence family have identified *Bacillus subtilis* phage protein transcription factor 1 (TF1) family of proteins as a close relative of IHF/HU (Muñoz et al. 2010). We found, however, that members of this family clustered within a larger HU clade in addition to forming an external cluster as depicted in the published phylogeny. To identify a different outgroup, we looked for other closely related sequence families as candidates and identified relaxosome protein TraM, a protein involved in F-plasmid transfer. The prokaryotic DNA-binding protein family containing IHF and HU sequences and the TraM protein family belong to the same broader protein superfamily, implying the presence of a common ancestor from which the two protein families may have diverged. TraM also aligns with IHF and HU protein subunit sequences while clustering separately from both proteins (Figure 3A). Using this as an outgroup we built a rooted phylogeny of the IHF-HU sequence family (Figure 3B).

The resulting gene tree exhibited the clustering of a large majority of HU, IHFα and IHFβ sequences into distinct clades (Figure 3B, 3C). While IHFα and IHFβ form separate clusters consisting purely of the respective sequence type, the HU sequences form two major clades - henceforth referred to as early and late branching - but without clearly separating HUα and HUβ (Figure 3B). In other words, both early and late-branching HU clades consisted of both HUα and HUβ sequences. The late branching HU clade shares an ancestor with the common ancestor of all IHF sequences. From the common ancestor of all IHF, four early branching clades of mixed sequence types diverge, which are labelled 'IHF_trns'1-4 (Figure 3C); 'trns' here refers to the presence of a 'transition zone' that exists between major clades. The node ancestral to IHF_trns4 gives rise to the common ancestor of 'IHFβ majority' and 'IHFα majority' clades, which primarily comprise IHFβ and IHFα sequences. Even within these majority clades, early branching subclades harbour sequences of other subtypes. Eventually, an IHFβ-only clade divides into an 'eco-like' clade, which contains the IHFβ sequence for *E. coli*, and a 'non-eco-like' clade that does not (Figure 3C). Similarly, the IHFα-only clade bifurcates into a 'non-eco-like' and finally an 'eco-like' clade. We noticed that 8 eukaryotic members of this family had been inadvertently added to our dataset. These are mostly from protists and one from plant. They cluster mostly in the IHF_trns region, with one sequence being part of an early-branching clade. These are probably horizontally acquired if the possibility of contamination in these eukaryotic genome sequences could be ruled out. We did not remove these sequences subsequently from the analysis.

We next used the gene tree to reconstruct sequences at each ancestral node of the phylogeny. The reconstructed sequences for the ancestral nodes of the IHF-HU sequence family as well as the ancestral sequences to all of the early branching clades, when aligned with the extant sequences, cluster among

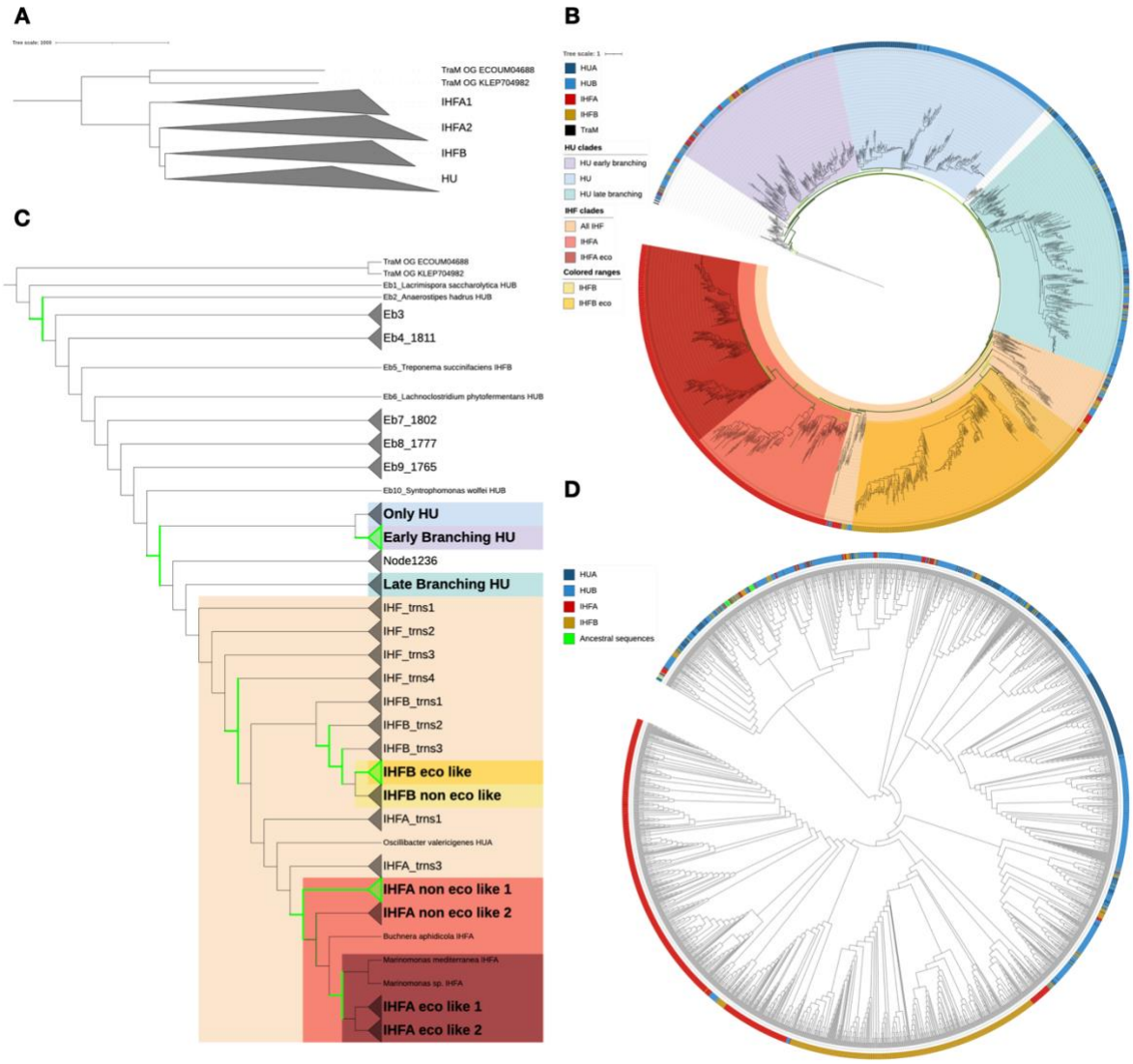

Figure 3: *Gene trees*. (A) A collapsed hierarchical clustering showing TraM as an outgroup to IHF/HU. (B) The IHF/HU gene tree. (C) A simplified version of the IHF/HU gene tree showing major clades and the branching pattern. Branches in green show signatures of positive selection (dN/dS > 1). (D) Hierarchical clustering of IHF/HU with ancestors of early-branching nodes shown in green (clustered near the 11-o-clock region of the tree). HUA is in dark blue, HUB in light blue, IHFA in red and IHFB in yellow.

HU subunit sequences (Figure 3D). The ancestor of the IHF-HU sequence family was therefore likely HU-like, and hence probably exhibited non-specific DNA binding.

In summary, the gene tree shows that HU protein clades were the first to emerge from the ancestor of the IHF-HU sequence family (Figure 3B and C). Ancestral sequence reconstruction also suggests that the last common ancestor of HU and IHF was HU-like (Figure 3D). The node ancestral to the late branching clade of HU sequences also gives rise to the common ancestor of all IHF sequences (Figure 3C). The ancestor of all IHF is separated from the node that further branches off into IHFα-majority and IHFβ-majority by a 'transition zone' of clustered sequences of different types. The nodes giving rise to IHFα and β majority clades also give rise to such transition zones before the emergence of clades consisting entirely of IHFα and IHFβ sequences (Figure 3B and C). In both IHFα and IHFβ, the *E. coli* protein forms a separate clade with its close relatives.

Characteristic amino acid residues

In order to determine how the properties of the proteins have evolved across the gene tree we determined the residues that would be characteristic for the structure and function of the subunits. We identified residues that are conserved within the combined HUα-β group in addition to IHFα and IHFβ. We calculated Shannon's entropy for each amino acid residue position across multiple sequence alignments of each group and selected sites exhibiting entropy below the 25th percentile of the entropy distribution, as 'conserved residues.'

A total of 23 residues are conserved across the entirety of the IHF-HU sequence family. However, our primary interest was in residues that are characteristic of each protein and therefore we focused on residues conserved in one subunit but not the others. Asn2, Ala11, Lys18, Ala73 and Lys86 (residue positions are from structure of *Anabaena* HU-DNA co-crystal, PDB: 1P51, (Swinger et al. 2003)) are specifically conserved in the HUα-β group but not in IHFα or IHFβ. Thr4 and Val27 (residue positions are from *E. coli* IHFα-β-DNA complex structure, PDB: 1IHF, (Rice et al. 1996)) are specifically conserved in IHFα. We identified residues conserved within the 'eco-like' clade that harbours the IHFα sequence from *E. coli* clade. These are Lys45, Ser47, Arg76, Val78, Lys86. Within the IHFβ subunit, Ile7, Pro78, Lys81 and Lys84 are conserved. Ser4, Arg42, Glu44, Arg46 and Lys84 are conserved within the IHFβ '*eco*-like' clade.

We then used our reconstruction of HU/IHF sequences at various internal, ancestral nodes to ask when, along the evolution of the extant IHF and HU subunits, the conserved residues were gained or lost (Figure 4A). The ancestral nodes of the early-branching clades of the phylogeny, including those that give rise to HU-majority clades, harbour all the HU-specific conserved residues in addition to some IHFβ residues. These IHFβ residues happen to be conserved in a substantial proportion of HU proteins as well, though not within the top 25 percentile (refer to methods for conservation threshold), but their conservation increases in IHFβ. The ancestor of all IHF sequences has some IHFα residues, while the nodes descending into the IHFβ clade eventually lose the HU residues. Within the IHFα-majority clade, we see eventual gain of IHFα residues and IHFα *eco*-type residues and the loss of IHFβ and HU residues, which are almost completely lost in the IHFα *eco*-type clade.

We examined the literature to identify residues that are important for DNA binding in each subunit. Residues contributing to DNA binding of HU are Lys13 from *Acinetobacter baumannii*, (Liao et al. 2017); Phe47 from *E. coli* (Goshima et al. 1990); Arg55, Arg58, Arg61 from *Bacillus sterothermophilus, Bacillus subtilis and Francisella tularensis* (Saitoh et al. 1999; Köhler & Marahiel 1998; Pavlik & Spidlova 2022); Lys62, Val63 and Pro64 in *Helicobacter pylori* (Chen et al. 2004); Ser 74 in *F. tularensis* (Pavlik & Spidlova 2022); and Lys80 and Lys86 in *F. tularensis* and *B. subtilis* (Köhler & Marahiel 1998). DNA recognition specificity in IHF appears to arise from a sum of small contributions from many residues (Rice et al. 1996). Some of these, highlighted in the literature, are listed here. Within the IHF heterodimer (in *E. coli*), Ser47, Arg60, Pro61, Pro65, Lys66, Ile71 and Ile73 in IHFα interact with DNA (Rice et al. 1996; Lee et al. 1992). In a recent work, IHF was shown to bind to the DNA in multiple modes of which the so-called 'fully-wrapped' mode was considered to represent specific DNA binding (Yoshua et al. 2021). The following residues present in the IHFα subunit are involved in making DNA contacts specifically in this fully-wrapped form: Thr4, Lys5, Lys20, Arg21, Lys24. In the IHFβ subunit, residues linked to DNA recognition are: Arg42, Glu44, Arg46, Pro64, Val70, Leu72, Arg87, Arg89 and Ala90 (Rice et al. 1996; Read et al. 2000; Mengeritsky et al. 1993).

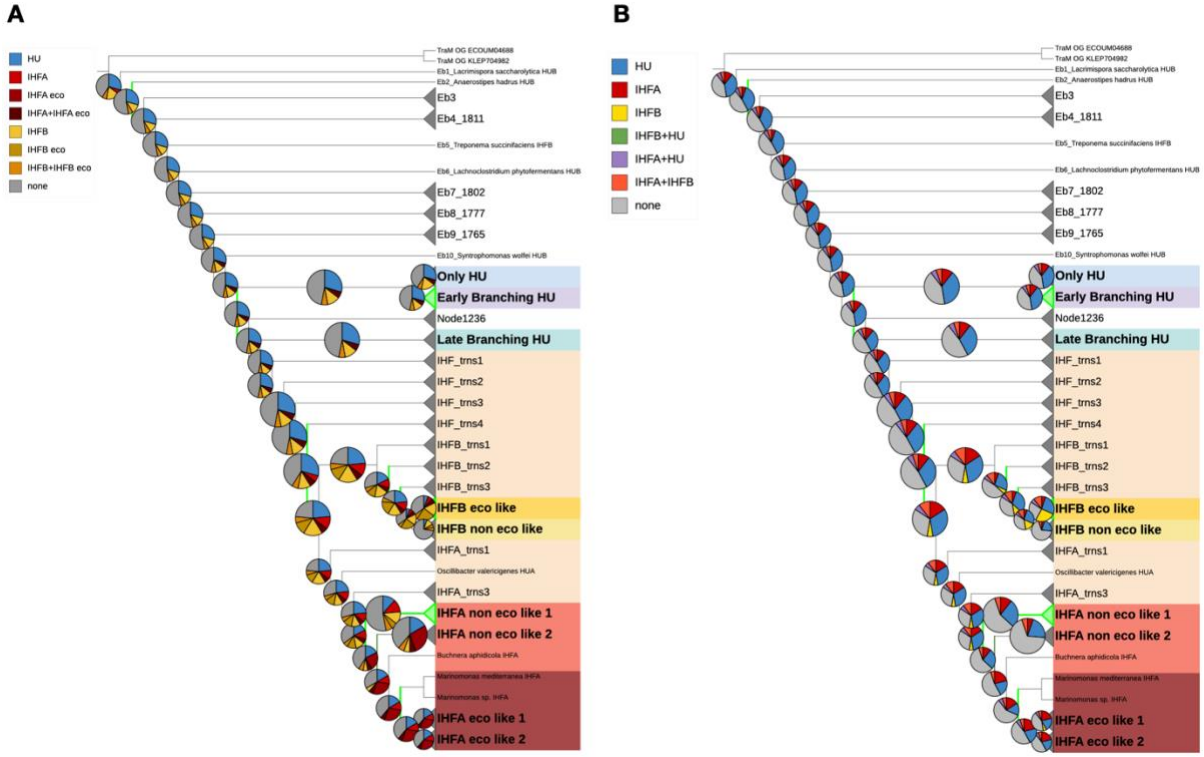

Figure 4. *Characteristic residues*. Shows the prevalence of residues characteristic of various clades. (A) residues conserved in particular clades of IHF/HU. (b) residues annotated in the literature as involved in DNA recognition of IHF/HU. In these pie charts grey represents residues (A) not conserved anywhere (B) not known to be involved in DNA binding. Blue represents HU residues; red shades stand for IHFA residues and; yellow is for IHFB residues.

We compared residues conserved within HU, IHFα and IHFβ subunit sequences with residues annotated in published literature as important for DNA binding. None of the annotated HU residues are conserved, as defined by our entropy thresholds. Among residues annotated as involved in DNA binding in IHFα: (a) Thr4 is conserved in all IHFα (b) Ser47 is conserved in IHFα *eco*-clade (c) Lys5 conserved in HU and (d) Lys20 in all protein subunits. We consider Lys5 and Lys20 as conferring specificity in the IHF context purely based on their ability to form contacts primarily in the fully-wrapped IHF-DNA structure despite their broader conservation. Residues Arg42, Glu44, Arg46, Pro64 annotated in published literature to be important for IHFβ-DNA interaction for specific binding are conserved in *eco*-like clade of IHFβ.

We once again used the reconstructed ancestral sequences across nodes to estimate when these residues were lost and gained along the phylogeny (Figure 4B). Early branching clades until the ancestor of the late branching HU exhibit the presence of all HU residues as well as a few IHFα residues. Some, but not all, HU residues are lost and sometimes regained rather sporadically as IHFs evolve. This probably explains why HU residues are not highly conserved. The common ancestor of IHFα and β majority clades is characterised by the first gain of IHFβ residues and the gain of an IHFα residue. Within the IHFβ clade the accumulated IHFα residues are however lost steadily until only one such residue is retained. This loss is predictably accompanied with the eventual gain of the IHFβ residues. However, the non-*eco*-type clade of IHFβ losses all of the IHFβ residues, indicating that residues involved in specific DNA site recognition in IHFβ are restricted to *E. coli* and its close relatives. Similarly, in the IHFα clade, the IHFβ residues are lost accompanied by the gain of IHFα residues until all the IHFβ residues are lost. However, nearly half of HU residues are retained in the ancestors of both late emerging

IHF clades, suggesting that residues known to be involved in DNA binding in HU may still have a role in recognition, if not specificity, in IHF.

Signatures of selection

Using the reconstructed ancestral sequences at all nodes, we further looked for signatures of positive selection by estimating dN/dS ratios of sequences across specific branches in the gene tree (Figure 3C). The branch leading to the common ancestor of the early and late branching HU clades exhibits signatures of positive selection. The branch leading from early-branching HU to the ancestor of the HU clade also harbouring multiple subunit types also shows high dN/dS. On the other hand, the branch leading to the clade comprising just HU sequences does not. This reflects the fact that the early branching HU ancestor is similar to extant HU, and the evolution of a more mixed clade is driven by selection. Several branches downstream all show signatures of positive selection. These include: (a) within the IHF transition zone, the branch leading from the node giving rise to the IHF_trns4 sequences to the common ancestor of IHFα and IHFβ majority clades; (b) from the common ancestor of IHFβ majority clade, the branches leading to nodes that give rise to IHFβ_trns2 and 3; (c) the branch leading to the ancestral node of the IHFβ *eco*-type clade; (d) the branch leading to the ancestor of the clade consisting of IHFα sequences; (b) branches leading to IHFα non *eco*-type clade 1 and of IHFα *eco*-type clade. Thus, the generation of sequence diversity leading from the HU-like common ancestor of all HU and IHF all the way down to IHFα appears to be driven by selection.

Conclusions

In this study we have traced the evolutionary history of the IHF/HU sequence family. The four subunits comprising this family emerged / evolved at different stages in the bacterial phylogeny, with HUβ, the most widely distributed subunit, emerging early though not in the last bacterial common ancestor. Its emergence is followed by IHFα, IHFβ and HUα respectively. All four subunits are present together most prominently in late emerging *Gammaproteobacteria*, the phylum most distant from the bacterial common ancestor. The ancestral sequence of the IHF/HU family appears to be similar to a HU, and might therefore have been a non-specific DNA binding protein. Specific recognition of DNA sites by IHF may arise from the sum total of small-effect base-specific interactions of multiple residues (Rice et al. 1996), and these residues seem to have emerged in a gradual manner along the phylogeny of the IHF/HU family. This supports the view that sequence-specific transcription regulators might have evolved from non-specific nucleoid-shaping proteins. We also suggest that this evolution of increased specificity in this protein family is under selection.

In an earlier work, Dey *et al* (Dey et al. 2017) had used an amino acid residue-based phylogeny of IHF/HU, and evolutionary trace analysis to identify sequence and structural features contributing to DNA recognition specificity. In a book chapter, Munoz et al. also used a HU/IHF tree to identify residues associated with major clade diversifications (Muñoz et al. 2010). We complement these by assembling a codon-based phylogeny of IHF/HU carefully rooted using an appropriate outgroup, and by mapping gains and losses of HU and IHF on a species tree, to infer the family's selection-driven evolutionary trajectory from a non-specific HU-like ancestor towards IHF.

Finally, sequence-specific DNA recognition by IHF has been characterised in detail primarily, if not exclusively, in *E. coli* and close relatives. We note that IHF sequences from *E. coli* form a separate clade diverging from a distinct set of other IHFs, and several residues known to be involved in DNA

recognition in *E. coli* IHF are not conserved widely. If IHFs across the board show similar DNA recognition specificity, then it stands to reason that the two clades of each subunit of IHF might have discovered their own independent solutions to the problem of acquiring specificity. On the other hand, if the *E. coli*-like IHF sequences in fact shows greater DNA recognition specificity, with IHFα predominantly making sequence-specific contacts as suggested by the fully-wrapped conformation in (Yoshua et al. 2021), then one can argue that specificity has possibly increased over the course of this protein family's evolution.

Methods

Dataset curation

Orthologous groups for IHF alpha (IHFα), IHF beta (IHFβ), HU alpha (HUα) and HU beta (HUβ) were obtained from the July 2023 release of the Orthologous MAtrix (OMA) database (Altenhoff et al. 2024). Out of a total of 2757 protein sequences, 389 belonged to HUA subunit, 995 to HUB subunit, 764 to IHFA and 609 to IHFB. A few proteins found in eukaryotic genomes such as those of protozoans and algae were removed. The combination of proteins present within each bacterial strain was determined. A single strain representative for each species was selected at random. A total of 1819 proteins were found across 971 bacterial species. The final alignment consisted of 263 HUA subunit sequences, 626 HUB sequences, 517 IHFA sequences and 413 IHFB sequences.

Species based phylogeny

A 16S rRNA based species tree was constructed for all the 971 bacterial species. The bacterial species names as annotated in the OMA database were used to obtain 16SrRNA sequences from the SILVA database (Glöckner et al. 2017). Alternative name annotations (heterotypic synonyms) were obtained from NCBI website for 40 bacteria because the species annotations as present in the OMA database were not compatible with the names of bacterial species as mentioned in the SILVA database. The Archaeon *Methanococcus indicus* used as an outgroup for the bacterial species phylogeny.

The sequences were aligned and a phylogenetic tree was subsequently constructed using the Silva Alignment Classification Tree (ACT) website. ACT uses SILVA Incremental Aligner (SINA version 1.2.12) to align the rRNA sequences (Pruesse et al. 2012). The phylogeny was computed using RAxML with a general time reversible model (GTR) of sequence evolution (Stamatakis 2014). The rate model for likelihoods was obtained/sampled from a gamma distribution.

Gene based phylogeny

Amino acid sequences of HU and IHF were used for the primary alignment. The unaligned sequences were visualised using Jalview (version 2.11.3.2) (Waterhouse et al. 2009) and alignment was done using MUSCLE. The JAva Bioinformatics Analysis Web Services system (JABAWS) (version 2.2) (Troshin et al. 2018) and Jalview were used for this. Different alignment sets were made for each of the four subunit sequences. For each of these alignment sets, the sequences that introduced a gap of 5 or more amino acids in the alignment were removed and the remaining sequences were realigned. This process was continued iteratively till all such outlier sequences were eliminated. All the subunit alignment sets were combined and outliers were eliminated, thus producing a final alignment consisting of all of the

IHF and HU subunits . The relationship among the HU and IHF protein sequences was visualised using the Principal Component Analysis approach described by (Konishi et al. 2019).

PAL2NAL (version 14) (Suyama et al. 2006)was then used to convert the protein sequence alignment into a codon alignment. The combined IHF and HU subunit codon alignment was used to make a gene-based phylogeny.

The 2.2.5 version of IQ-TREE software was used to perform model testing on the codon alignment and thereby determine a substitution model to define the sequence evolution (Nguyen et al. 2015). Model selection in IQ-TREE allows for testing across a wide variety of sequence evolution models in addition to detection of variability in evolutionary rates across sites in the input multiple sequence alignment. The 'modelomatic' option in IQ-TREE was used to determine the best model for sequence evolution given the alignment (Kalyaanamoorthy et al. 2017). KOSI07, which is an empirical codon model (Kosiol et al. 2007), was chosen as the optimal model for sequence evolution for the IHF-HU gene alignment using the Bayesian Information Criterion (BIC) value. KOSI07 allows incorporation of simultaneous di- and tri-two to three nucleotide changes in addition to single nucleotide substitutions. The physicochemical properties of the encoded amino acids are also considered while estimating the sequence evolution model by KOSI07.

Translocating membrane associated chain protein (TraM) from E. coli and Klebsiella pneumoniae, containing to the Integration Host Factor (IHF)- like DNA binding domain superfamily, used as outgroup to root the gene tree.

The Codon based IHF-HU gene tree rooted using a pair of TraM protein sequences from E. coli and Klebsiella pneumoniae was used for ancestral sequence reconstruction. Sequences at each node of this gene tree were reconstructed, using the same model used for phylogeny construction. IQTREE was used for ancestral sequence reconstruction and it provides posterior probability for all possible codons for each site at each node. The ancestral sequence at each node consists of the most probable codon at each site for every node. Sites with posterior probability greater than 0.5 were included in the analysis.

Conserved sites

To identify conserved sites in the protein sequence alignment, Shannon's entropy over 20 amino acid categories (normalised between 0 and 1) was calculated for each site. Sites with greater than 50% of gaps were not considered. At other sites, gaps were not included as a category for entropy calculation. Sites with entropy in the lower quartile were considered as conserved and the amino acid most prevalent at that site noted. These were identified separately for HU, IHFA, IHFB and the subset of proteins belonging to the clade containing the E. coli IHFA. In addition, residues described in the literature as being involved in sequence-specific binding of E. coli IHF to its target binding site were identified from the following papers: (Rice et al. 1996; Lee et al. 1992; Read et al. 2000; Mengeritsky et al. 1993; Yoshua et al. 2021).

Signatures of selection

dN/dS ratios, as calculated using the Yang-Nielsen 2000 model as implemented in yn00 (PAML, version 4.9j, Feb 2020) (Yang & Nielsen 2000) were used to identify pairs of sequences showing signs of positive selection. This was done for pairs of reconstructed ancestral sequences leading from the common ancestor of all IHF and HU sequences to the major clades

References

Altenhoff AM et al. 2024. OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. Nucleic Acids Res. 52:D513–D521. doi: 10.1093/nar/gkad1020.

Balleza E et al. 2009. Regulation by transcription factors in bacteria: beyond description. FEMS Microbiol Rev. 33:133–151. doi: 10.1111/j.1574-6976.2008.00145.x.

Boubrik F, Bonnefoy E, Rouvière-Yaniv J. 1991. HU and IHF: similarities and differences. In Escherichia coli, the lack of HU is not compensated for by IHF. Res Microbiol. 142:239–247. doi: 10.1016/0923-2508(91)90036-a.

Browning DF, Busby SJ. 2004. The regulation of bacterial transcription initiation. Nat. Rev. Microbiol. 2:57–65. doi: 10.1038/nrmicro787.

Browning DF, Busby SJW. 2016. Local and global regulation of transcription initiation in bacteria. Nat Rev Microbiol. 14:638–650. doi: 10.1038/nrmicro.2016.103.

Cases I, de Lorenzo V, Ouzounis CA. 2003. Transcription regulation and environmental adaptation in bacteria. Trends Microbiol. 11:248–253. doi: 10.1016/s0966-842x(03)00103-3.

Chen C, Ghosh S, Grove A. 2004. Substrate specificity of Helicobacter pylori histone-like HU protein is determined by insufficient stabilization of DNA flexure points. Biochem J. 383:343–351. doi: 10.1042/BJ20040938.

Craig NL, Nash HA. 1984. E. coli integration host factor binds to specific sites in DNA. Cell. 39:707–716. doi: 10.1016/0092-8674(84)90478-1.

Dey D, Nagaraja V, Ramakumar S. 2017. Structural and evolutionary analyses reveal determinants of DNA binding specificities of nucleoid-associated proteins HU and IHF. Mol Phylogenet Evol. 107:356–366. doi: 10.1016/j.ympev.2016.11.014.

Dorman CJ, Schumacher MA, Bush MJ, Brennan RG, Buttner MJ. 2020. When is a transcription factor a NAP? Curr Opin Microbiol. 55:26–33. doi: 10.1016/j.mib.2020.01.019.

Drlica K, Rouviere-Yaniv J. 1987. Histonelike proteins of bacteria. Microbiol Rev. 51:301–319. doi: 10.1128/mr.51.3.301-319.1987.

Friedman DI. 1988. Integration host factor: a protein for all reasons. Cell. 55:545–554. doi: 10.1016/0092-8674(88)90213-9.

Gil R, Silva FJ, Peretó J, Moya A. 2004. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev. 68:518–537, table of contents. doi: 10.1128/MMBR.68.3.518-537.2004.

Glöckner FO et al. 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotechnol. 261:169–176. doi: 10.1016/j.jbiotec.2017.06.1198.

Goshima N, Kohno K, Imamoto F, Kano Y. 1990. HU-1 mutants of Escherichia coli deficient in DNA binding. Gene. 96:141–145. doi: 10.1016/0378-1119(90)90355-u.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14:587–589. doi: 10.1038/nmeth.4285.

Kamashev D et al. 2017. Comparison of histone-like HU protein DNA-binding properties and HU/IHF protein sequence alignment. PLoS One. 12:e0188037. doi: 10.1371/journal.pone.0188037.

Köhler P, Marahiel MA. 1998. Mutational analysis of the nucleoid-associated protein HBsu of Bacillus subtilis. Mol Gen Genet. 260:487–491. doi: 10.1007/s004380050921.

Konishi T et al. 2019. Principal Component Analysis applied directly to Sequence Matrix. Sci Rep. 9:19297. doi: 10.1038/s41598-019-55253-0.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Mol Biol Evol. 24:1464–1479. doi: 10.1093/molbev/msm064.

Lee EC, Hales LM, Gumport RI, Gardner JF. 1992. The isolation and characterization of mutants of the integration host factor (IHF) of Escherichia coli with altered, expanded DNA-binding specificities. EMBO J. 11:305–313. doi: 10.1002/j.1460-2075.1992.tb05053.x.

Liao J-H et al. 2017. Acetylome of Acinetobacter baumannii SK17 Reveals a Highly-Conserved Modification of Histone-Like Protein HU. Front Mol Biosci. 4:77. doi: 10.3389/fmolb.2017.00077.

Mengeritsky G, Goldenberg D, Mendelson I, Giladi H, Oppenheim AB. 1993. Genetic and biochemical analysis of the integration host factor of Escherichia coli. J Mol Biol. 231:646–657. doi: 10.1006/jmbi.1993.1316.

Molina N, van Nimwegen E. 2009. Scaling laws in functional genome content across prokaryotic clades and lifestyles. Trends Genet. 25:243–247. doi: 10.1016/j.tig.2009.04.004.

Muñoz A, Valls M, Lorenzo V. 2010. Extreme DNA Bending: Molecular Basis of the Regulatory Breadth of IHF. In: Bacterial Chromatin. pp. 365–393. doi: 10.1007/978-90-481-3473-1_16.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32:268–274. doi: 10.1093/molbev/msu300.

Pavlik P, Spidlova P. 2022. Arginine 58 is indispensable for proper function of the Francisella tularensis subsp. holarctica FSC200 HU protein, and its substitution alters virulence and mediates immunity against wild-type strain. Virulence. 13:1790–1809. doi: 10.1080/21505594.2022.2132729.

Prieto AI et al. 2012. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. Nucleic acids research. 40:3524–37. doi: 10.1093/nar/gkr1236.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 28:1823–1829. doi: 10.1093/bioinformatics/bts252.

Read EK, Gumport RI, Gardner JF. 2000. Specific recognition of DNA by integration host factor. Glutamic acid 44 of the beta-subunit specifies the discrimination of a T:A from an A:T base pair without directly contacting the DNA. J Biol Chem. 275:33759–33764. doi: 10.1074/jbc.M910381199.

Rice PA, Yang S, Mizuuchi K, Nash HA. 1996. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell. 87:1295–306.

Rk S, Sj M, Jf K, Mb E. 2012. Probing the informational and regulatory plasticity of a transcription factor DNA-binding domain. PLoS genetics. 8. doi: 10.1371/journal.pgen.1002614.

Saitoh F, Kawamura S, Yamasaki N, Tanaka I, Kimura M. 1999. Arginine-55 in the beta-arm is essential for the activity of DNA-binding protein HU from Bacillus stearothermophilus. Biosci Biotechnol Biochem. 63:2232–2235. doi: 10.1271/bbb.63.2232.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–1313. doi: 10.1093/bioinformatics/btu033.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609-612. doi: 10.1093/nar/gkl315.

Swinger KK, Lemberg KM, Zhang Y, Rice PA. 2003. Flexible DNA bending in HU-DNA cocrystal structures. EMBO J. 22:3749–3760. doi: 10.1093/emboj/cdg351.

Swinger KK, Rice PA. 2007. Structure-based analysis of HU-DNA binding. Journal of molecular biology. 365:1005–16. doi: 10.1016/j.jmb.2006.10.024.

Tanaka H, Goshima N, Kohno K, Kano Y, Imamoto F. 1993. Properties of DNA-binding of HU heterotypic and homotypic dimers from Escherichia coli. J Biochem. 113:568–572. doi: 10.1093/oxfordjournals.jbchem.a124084.

Troshin PV et al. 2018. JABAWS 2.2 distributed web services for Bioinformatics: protein disorder, conservation and RNA secondary structure. Bioinformatics. 34:1939–1940. doi: 10.1093/bioinformatics/bty045.

Visweswariah SS, Busby SJW. 2015. Evolution of bacterial transcription factors: how proteins take on new tasks, but do not always stop doing the old ones. Trends Microbiol. 23:463–467. doi: 10.1016/j.tim.2015.04.009.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25:1189–1191. doi: 10.1093/bioinformatics/btp033.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43. doi: 10.1093/oxfordjournals.molbev.a026236.

Yoshua SB et al. 2021. Integration host factor bends and bridges DNA in a multiplicity of binding modes with varying specificity. Nucleic Acids Res. 49:8684–8698. doi: 10.1093/nar/gkab641.

Acknowledgment

This work was funded by a DBT/Wellcome Trust India Alliance Intermediate Fellowship (IA/I/16/2/502711 to A.S.N.S.) and core funding from the Department of Atomic Energy, Government of India (project number 12-R&d-TFR-5.04-0800).