A highly conserved family of domains related to the DNA-glycosylase fold helps predict multiple novel pathways for RNA modifications

A Maxwell Burroughs* and L Aravind
National Center for Biotechnology Information; National Library of Medicine; National Institutes of Health; Bethesda, MD USA

Keywords: DNA glycosylase, NEMF, Tae2, caliban, FbpA, fibronectin-binding, tRNA 4-thiouridylation, IscS, TusA, base modification

Abbreviations: Fpg, formamidopyrimidine; LUCA, Last Universal Common Ancestor; LECA, Last Eukaryotic Common Ancestor; FMN-DG domain, Formamidopyrimidine, MutM, and Nei/EndoVIII DNA glycosylase; FbpA, fibronectin-binding protein; BER, Base Excision Repair; NFAC domain, NEMF, bacterial FbpA-like proteins, Caliban, and Tae2 domain; NFAC-N domain, NFAC N-terminal domain, NFAC-R domain, NFAC RNA-binding domain; NFAC-C domain; NFAC C-terminal domain; ZnK, zinc knuckle; TGT, transglycosylase; RQC, ribosomal quality control; IRES, Internal Ribosomal Entry Site

A protein family including mammalian NEMF, Drosophila caliban, yeast Tae2, and bacterial FbpA-like proteins was first defined over a decade ago and found to be universally distributed across the three domains/superkingdoms of life. Since its initial characterization, this family of proteins has been tantalizingly linked to a wide range of biochemical functions. Tapping the enormous wealth of genome information that has accumulated since the initial characterization of these proteins, we perform a detailed computational analysis of the family, identifying multiple conserved domains. Domains identified include an enzymatic domain related to the formamidopyrimidine (Fpg), MutM, and Nei/EndoVIII family of DNA glycosylases; a novel, predicted RNA-binding domain, and a domain potentially mediating protein–protein interactions. Through this characterization, we predict that the DNA glycosylase-like domain catalytically operates on double-stranded RNA, as part of a hitherto unknown base modification mechanism that probably targets RNAs. At least in archaea, and possibly eukaryotes, this pathway might additionally include the AMMCR1 family of proteins. The predicted RNA-binding domain associated with this family is also observed in distinct architectural contexts in other proteins across phylogenetically diverse prokaryotes. Here it is predicted to play a key role in a new pathway for tRNA 4-thiouridylation along with TusA-like sulfur transfer proteins.

Introduction

Over 10 y ago, a gene family conserved across all three superkingdoms of life was identified and determined to contain two tandem copies of the nucleic acid-binding helix-hairpin-helix (HhH) domain. Based on its phyletic distribution, this family was traced back to the last universal common ancestor (LUCA) of life. The HhH domain pair and its phyletic pattern suggested a general functional role for the family in a nucleic acid-related role in universally conserved pathways: either RNA-metabolism in the context of translation or DNA repair or recombination. Establishment of a relationship between the HhH-domain pair in these proteins to the one found in ribosomal proteins of the S13/S18 family supported the former function in particular. Since the initial characterization of this family, which includes the Tae2 protein from Saccharomyces cerevisiae, the Caliban (CBo) protein from Drosophila melanogaster, the mammalian NEMF proteins, and the so-called fibronectin-binding (FbpA-like) proteins from bacteria, several studies have resulted in attribution of a wide range of functional roles for these proteins. These include fibronectin binding in certain pathogenic bacteria, a core component of the ribosome-associated, co-translational degradation complex RQC in yeast, regulation of the DNA-damage response in Drosophila, and mediation of nuclear export in Drosophila and human. Given the disparate nature of these findings, we decided to revisit this gene family using state-of-the-art techniques in sequence analysis and comparative genomics while tapping the wealth of new information that has accumulated in the years since its initial characterization. Here we identify and characterize the distinct globular domains conserved across all members of the gene family in addition to the HhH domain pair. One of these domains is predicted to be an

*Correspondence to: A Maxwell Burroughs; Email: max.burroughs@nih.gov
Submitted: 10/28/2013; Revised: 02/04/2014; Accepted: 02/20/2014; Published Online: 03/05/2014
http://dx.doi.org/10.4161/rna.28302
enzymatic domain related to the bifunctional DNA glycosylase/endo nuclease domain involved in Base Excision Repair (BER), commonly referred to as the Formamidopyrimidine, MutM, and Nei/EndoVIII DNA glycosylase (FMN-DG; also referred to in the literature as Fpg/Nei, Fapy DNA glycosylase, glycosylase/AP-lyase, or Endonuclease VIII) domain. We identify shared and distinct features of the active site of these two related domains, implying both similarities and differences in their catalytic mechanisms. Another domain in this gene family is predicted to be a novel RNA-binding domain, with a potential role in a variant of the tRNA 4-thiouridylation pathway present in a subset of prokaryotes. Based on these observations and additional genome contextual evidence, we propose that the fundamental functional role of this ancient gene family is related to processing/modification of double-stranded RNA, perhaps rRNA.

Results
Delineation of the NEMF/FbpA/Caliban/Tae2 gene family and its core architectures
To comprehensively characterize this gene family, we collected all related sequences using known members as seeds to initiate sequence profile searches against the non-redundant (nr) protein database at the National Center for Biotechnology Information. Given the presence of a large coiled-coil domain in the gene family, we applied the low complexity seg filter to these searches to avoid inclusion of genes with spurious similarity. Membership of proteins displaying relationships with borderline significance was confirmed by initiating reverse searches. Sequences obtained were then aligned and potential globular regions shared across the gene family were identified by inspection of these alignments after mapping the location of the known HhH domains and the coiled-coil region characteristic of the NFACT gene family is depicted as a light green circle in architectures and a light green box in gene neighborhoods. Abbreviations: CC, coiled-coil; HHA, helix-hairpin-helix; ZnK, zinc knuckle; ZnR, zinc ribbon.

Figure 1. Domain architectures and conserved gene neighborhoods relating to NFACT proteins. Domains architectures are depicted by adjoining polygonal shapes labeled with individual domain names, sizes of domains are not drawn precisely to scale. Individual genes within conserved gene neighborhoods are depicted with boxed arrows with the arrowhead pointing toward the 3′ end of the gene. Each architecture/neighborhood is labeled with gene name, GenBank gene identifier (gi) number, and organism name separated by semicolons. Architectures/neighborhoods relating to the NFACT gene family are boxed in purple; those relating specifically to independent contexts of the NFACT-R domain are boxed in orange. The coiled-coil region characteristic of the NFACT gene family is depicted as a light green circle in architectures and a light green box in gene neighborhoods. Abbreviations: CC, coiled-coil; HHA, helix-hairpin-helix; ZnK, zinc knuckle; ZnR, zinc ribbon.

orthologs of the gene family across all three superkingdoms of life were identified, including the NEMF, bacterial FbpA-like proteins, Caliban, and Tae2; accordingly, we termed this family NFACT. Representatives of the family were found across all major archael lineages including the euryarchaeota, crenarchaeota, korarchaeota, and thaumarchaeota. The NFACT family is also found across most major bacterial lineages, although it is notably absent in the α- and γ-proteobacterial lineages (despite being present in β- and ε-proteobacterial and actinobacteria. In eukaryotes, the NFACT family is again present in all major lineages including the diplomonads, parabasalids, hetrolobosystems, kinetoplastids, chromosalvelates, apicomplexa, and the crown-group eukaryotes encompassing the plant, amoebozoa, animal, and fungal lineages (with a notable absence in the basidomyocyte fungi). Taken as a whole, despite losses in certain terminal lineages, this phylogenetic spread unquestionably points to presence of the NFACT family in the LUCA (see Supplemental Material for complete sequence and phylic distribution).

The conserved core of the NFACT family found across all members is formed by four domains interrupted by the coiled-coil region (Fig. 1, Supplemental Material): from N terminus to C terminus...
these entail an uncharacterized N-terminal domain, the two HhH domains, the coiled-coil region, and a domain currently annotated as DUF814 (Domain of Unknown Function 814) in Pfam.20 The first three domains from the N-terminus are currently incorrectly annotated as a single domain in Pfam: the FbpA domain. We propose renaming the N-terminal domain the NFACT-N (for NEMF, FbpA, Caliban, Tae2, N-terminal) domain and separating it from the downstream HhH domains (Fig. 1). In archaea and eukaryotes, an additional C-terminal domain annotated in Pfam as DUF3441 is present, clearly establishing the archaeal version of the family as the one inherited by eukaryotes. The core (NFACT-N+HhH+HhH+coiled-coil+DUF814+DUF3441) has proven resistant to domain accretion during evolution, although a few limited elaborations are observed in eukaryotes. In plants, Entamoeba, and Giardia, a zinc knuckle domain (ZnK) insertion is present between the DUF814 and DUF3441 domains. The most parsimonious explanation for this unusual phyletic distribution is independent, secondary acquisition of the ZnK in these three distant lineages, a scenario supported by the lack of specific sequence similarity across the different ZnKs. While the ZnK is not universally present in chlorophyte algae, its presence in Ostreococcus suggests that it was acquired early in the evolution of Viridiplantae. One additional potential domain fusion of note is the C-terminal fusion to four copies of the RNA-binding RRM domain in the roundworm Loa loa (Fig. 1).

To better understand the roles of the distinct domains in the NFACT proteins, we investigated in detail the previously uncharacterized domains in NFACT proteins using sensitive sequence-profile searches.

**NFACT-N domain**

Iterative profile searches initiated with NFACT-N domain sequences and their downstream HhH domains against the
nr database recovered significant matches extending along the length of the DNA glycosylase (FMN-DG) domain and its characteristic C-terminal HhH domain pair. For example, a sequence from the *H. sapiens* NEMF protein recovered the FMN-DG and HhH domains from the fungus *Togninia minima* (gi: 500259929, iteration: 4, e-value: $2 \times 10^{-4}$) and the acidobacterium *Terriglobus roseus* (gi: 390956257, iteration: 4, e-value: $7 \times 10^{-4}$) in PSI-BLAST. Reciprocal searches confirmed these relationships: a search initiated with the same sequence from *Terriglobus* yielded the NFACT-N from *Methanocella arvoryzae* (gi: 147920849, PSI-BLAST iteration: 1, e-value: $4 \times 10^{-4}$). As an independent means of confirmation, a HMM profile constructed from the multiple sequence alignment of the NFACT-N and HhH domains was searched against a database of HMMs constructed for individual pdb entries using the HHpred program. In addition to detecting PDB: 3doa (structure of a N-terminal NFACT fragment), this search again revealed a significant relationship between the NFACT-N and HhH domains and cognate domains in FMN-DG; for example, significant matches are retrieved for FMN-DGs from *Arabidopsis thaliana* (PDB: 3twl, p-value: $1 \times 10^{-5}$, probability: 95.9%) and *Geobacillus stearothermophilus* (PDB: 3u6p, p-value: $6.2 \times 10^{-5}$, probability: 94.0%).

The FMN-DG family is well-distributed across bacteria, absent in archaea, and only present in scattered eukaryotic lineages. The family can be divided into two distinct subfamilies, the Fpg/MutM-like subfamily found in bacteria, plants, and fungi and the eukaryotic Nei subfamily primarily observed in animals. Previous analysis indicated that eukaryotic versions of the Fpg/MutM-like subfamily likely emerged via horizontal gene transfer (HGT) from bacteria relatively early in eukaryotic evolution, while the Nei subfamily emerged later following HGT from a bacterial source to the stem of the animal lineage.
analysis identified Nfe homologs in the early-branching eukaryote diplomonad *Giardia*, suggesting that the possibility that Nfe also might have been acquired earlier in eukaryotic evolution (AMB, LA, personal observations).

Thus, the phylectic patterns of NFACT-N point to an origin in the ancestor of the core elements of FMN-DGs that have emerged first in bacteria. This suggests that the latter are likely to have been derived from the earlier in bacterial evolution. To better understand the relationship between NFACT-N and FMN-DNA glycosylases, we constructed a structure-guided super-alignment, first aligning known FMN-DG structures with the 3DOA structure and then adding further NFACT-N/FHh sequences (see Materials and Methods). At this point it became evident that despite the clear homology between the domains, the shared core scaffold had undergone a multi-step structural reorganization via duplication during divergence (Fig. 2). Both domains feature a core containing eight β-strands and two α-helices leading into the dyad of HH motifs. The NFACT-N domain, which was inferred to represent the ancestral condition, has two repeats of a basic structural element, each containing an α-Helix leading into a 4-stranded β-meadow, yielding a sandwich-like fold with the two stacking β-sheets, is retained but the connectivity between the helix/meander units has been substantially altered (Fig. 2). The most parsimonious explanation for this “re-writing” of the connectivity in FMN-DG entails the following steps, in some ways reminiscent of the recently elucidated steps underlying the derivation of the FYVE domain from the canonical bimucin treble clef domain core22. (Fig. 2B): (1) duplication of one of the repeats in the original 2-repeat structure yielding a 3-repeat intermediate. (2) Given the packing of the sheets against each other to form a sandwich, the 3-repeat state leads to an inherently unstable condition with competition between alternative repeats to reconstitute the original sandwich. (3) This instability was resolved by natural selection through partial loss of a subset of the elements to reconstitute the original two-sandwich. This reconstitution, rather than proceeding via the loss of a complete superfluous repeat, resulted from complementary, partial loss of elements from repeat 1 and 3, while via the loss of a complete superfluous repeat, resulted from complementary, partial loss of elements from repeat 1 and 3, while the loss of a full repeat as in the case of repeat 2 may be the cause of the isolated FMN-DG+ domains with selected FMN-DG+ domains. (Fig. 2). The column corresponding to the conserved glutamate/histidine residue in the HH domains is shaded in red, colored in yellow; c, small (shaded in green); α, alkaline (shaded in yellow); , negatively charged (shaded in purple); p, polar (shaded in blue); +, positively charged (shaded in purple); a, aromatic (shaded in yellow); b, big (shaded in gray); u, tiny (shaded in green); c, charged (shaded in purple). Columns corresponding to active site residue positions are shaded in red, colored in yellow, and marked with “.”. The column corresponding to the conserved glutamate/histidine residue specific to NFACT-N is marked with a “.”. (Orgnisms abbreviations as follows: Aabo, Alcaligenes faecalis; Aaor, Aporrhais spinulosa; Atca, Aspergillus oryzae; Atla, Atriplaspha thalassiae; Bcor, Bacillus cereus; Bthu, Bacillus thuringiensis; CCA, Candidatus Caldispepsium; CCH, Candidatus Chloridobacteriaceae; Ckit, Candidatus Kiritakeaeae; Cale, Caldothrix elegans; Cerr, Coriopsis intermidii; Cere, Caspofungus novemarum; Cjas, Chlamydia psittaci; Cphi, Chlorella phaeocystisi; Cpyr, Cynodegus sylvestris; Dpym, Diphtheria pneumoniae; Dpne, Fusobacterium necrophorum; Gla2, Giardia lamblia; Gap, Geobacillus sp.; GSt, Geobacillus stearothermophilus; Hmac, Hyphomicrobiurn macrospallatum; Hnaa, Homo sapiens; Hter, Hydrobacter termitis; Hvul, Heliobacter volcanii; Klac, Klyuyveromyces lactis; Lbd, Leucobacter lucii; Llue, Leucobacter lucii; Lnmo, Listeria monocytogenes; Mnbe, Micrococcus bercovieri; Mmmu, Mmuseus musculus; Mrys, Methanobrevibacter; Mny1, Methylophana yellowstonensis; Mns5, Mycoplasma sputorum; Mmyd, Metalliphana yellowstonensis; Mspe, Mesoplasma sulzeneri; Mapt, Mycoplasma parvum; Pnor, Parachlorobacterium novum; Ptpp, Pseudococcis sp.; Punt, Pananurium trimalum; Rnor, Rattus norvegicus; Saci, Saccharomyces cerevisiae; Sacc, Schizosaccharomyces; Secc, Streptomyces coelicolor; Spor, Streptococcus pneumoniae; Tbru, Thermoplasma brucei; Ttyc, Thermostoxys cruent; Tlie, Thermostoxys lain; Ttr, Theileria orvitalis; Vip, Thermus sp.; Uarv, uncultured archaeon.

The transition from the core enzymatic domain to the β-depicted as follows: extended loop regions are represented by black lines, sequences from solved crystal structures are at the top of the alignment followed by the NFA...
Currently holds that the proline residue forms a Schiff base intermediate with the C1' atom of the damaged base and that at least one of the two conserved glutamate residues is involved in initiating the base removal step. The role of the lysine has been debated but appears to also be involved in the base removal step while possibly also being a key player in the backbone cleavage reaction. Helix-1 in NFACT-N also contains a well-conserved aspartate found in one or the other of the two positions where conserved glutamates are found in the FMN-DG domain (Figs. 2A and 3). Strikingly, NFACT-N does not contain any absolutely conserved prolines, either at the N terminus or elsewhere in the domain (Fig. 3). NFACT-N also features an additional absolutely conserved aspartate in the predicted active site pocket, one residue downstream of the aforementioned arginine, which could compensate for the second aspartate seen in the N terminus of H1 in the cognate FMN-DG domain (Figs. 2A and 3). Two additional well-conserved residues outside of the active site pocket with conserved cognates in FMN-DG include a glutamate near the C terminus of strand-3 and a well-conserved asparagine at the N terminus of strand-4. The glutamate residue appears equivalent to a highly conserved histidine residue in FMN-DG, which plays a role in substrate nucleic acid binding; however, in NFACT-N, this residue appears to form a conserved salt bridge with a positively charged residue in the linker region found between the NFACT-N domain and the HhH domains. The conserved NFACT-N asparagine appears positionally equivalent to a polar residue typically taking the form of a serine or threonine in FMN-DG domains that contributes to nucleic acid recognition (Figs. 2A and 3); in NFACT-N, this residue might help position a well-conserved NFACT-N-specific arginine/lysine, which points out from the predicted catalytic core and could be involved in nucleic acid recognition (Fig. 3). Crosslinking experiments capturing FMN-DG active site intermediates clearly indicate that proline-mediated Schiff base intermediates clearly indicate that proline-mediated Schiff base

Figure 4. Known and predicted reactions. (A) Base removal and ring-opening steps catalyzed by FMN-DNA glycosylases (top) and predicted analogous steps catalyzed by NFACT-N during a potential base-exchange reaction (bottom). The introduced free base in the NFACT-N reaction is labeled with a red “M,” indicating the base is potentially modified in some way despite being shown here as a uridine. (B) 2-thiouridylation, canonical 4-thiouridylation, and the novel, predicted 4-thiouridylation sulfur relay pathways. Potential intermediate step in the novel 4-thiouridylation pathway involving transfer to a conserved cysteine on IscU is shown in dotted lines, reflecting uncertainty in whether this step is present in all organisms containing the Thi+NFACT-R fusion or whether it is restricted to a subset of them.
formation occurs following base excision and deoxyribose ring-opening but prior to nicking of the phosphatidylethanolamine backbone (Fig. 4A). One of the two conserved His aspartates is consistently implicated in base removal upstream of the Schiff base formation, although which aspartate is involved may depend on the substrate. Conflicting views on the lysine indicate a role in either base-removal upstream of Schiff base formation or in initiating the DNA cleavage reaction. The lack of a conserved proline or any compensatory amine-hearing residue in the cognate NFAC-T-N active site, but lacking NFAC-T-N as incapable of forming the Schiff base intermediate and thus is unlikely to be involved in DNA cleavage as catalyzed by FMN-DG. However, like FMN-DG, NFAC-T-N displays an arginine equivalent to lysine in the former and two acidic active site residues. Hence, based on the above-outlined spatial position- and residue- conservation between the domains, we predict NFAC-T-N potentially catalyzes base-removal as observed in FMN-DG (Fig. 4A, see below). These observations, combined with the knowledge that the FMN-DG domains were derived from the ancestral NFAC-T-N domain, also potentially assists in distinguishing between proposed roles for the lysine in the FMN-DG reaction mechanism. As the conserved proline, Schiff base formation, and nuclease activity are unique to FMN-DG, they must necessarily be catalytic innovations secondary to the roles of the ancestral aspartate glutamate and arginine/lysine. Thus, a role for the lysine/arginine in base removal during glycosylase activity is likely to be the ancestral role (Fig. 4A), although we cannot rule out that the lysine has secondarily acquired an additional role in DNA cleavage in FMN-DG. This role for the lysine is also consistent with mutational studies finding base removal, as opposed to DNA cleavage, to be most affected by lysine substitution. One additional residue conserved in both NFAC-T-N and FMN-DG-fused HHH domains is noteworthy: a well-conserved glutamate residue (sometimes replaced by histidine in NFAC-T-N) found in the first helix of the second HHH domain, which mediates a backbone contact with a distinctive, conserved loop structure immediately C-terminal to the HHH domains in NFAC-T-N and C-terminal to an inserted zinc ribbon domain found downstream of the HHH domains in FMN-DG. This glutamate-backbone contact positions the HHH domains for interaction with the nucleic acid substrate on the side opposite to the NFAC-T-N and FMN-DG active sites (Figs. 2A and 3). Thus, this structural constraint appears to have maintained similar active site clefts for the two families with the HHH motifs forming a nucleic acid binding “cap” in both cases to accommodate a double-stranded substrate.

DEUF814 domain

Searches with individual DEUF814 sequences as seeds failed to recover any remote relationships with other known domains. A multiple sequence alignment constructed for the domain indicates the DEUF814 domain consists of an α/β structure with at least seven β-strands and three α-helices. DEUF814 domains of the NFAC gene family contain a well-conserved DxxCxxH motif with two conserved cysteines at either end of the third predicted strand in addition to a conserved downstream serine residue found in the second helix of this domain (Supplemental Material). These domains are present in two additional contexts: (1) N-terminal fusion to a PP-loop domain in a diverse range of bacteria including E. coli, nitrobacter, fusobacteria, synergistes, spiru- ntes, delta- and epsilon-, and some gamma-proteobacteria, aquificae, dacryoglo- mi, planktomycetes as well as few thauvarchaeota and (2) as a solo domain C-terminally fused to a small coiled-coil region present across all eukaryotic lineages, typified by the CCDC25 protein in humans and the Ipj2 protein in Saccharomyces cerevisiae (Supplemental Material). An alignment constructed of only members of the PP-loop-fused DEUF814 family revealed the same core secondary structure that is maintained between the DEUF814 family, instead featuring two nearly-absolutely conserved arginines N-terminal to the first strand and the first helix, respectively (Supplemental Material). The PP-loop domain belongs to the Thi-like PP-loop family, which catalyzes 4-thiouridylation at nucleotide 8 of bacterial and archaeal tRNAs. Thl-like domains were previously thought to be universally fused to an N-terminal RNA-binding THUMP domain34 with versions from many bacteria and a few archaea additionally fused C-terminally to a Rhodanese (RHOD) domain containing an absolutely conserved cysteine residue. Bacteria and archaea lacking the C-terminal RHOD domain likely interact with a stand-alone RHOD domain during thiouridylation. Canonical Thl-mediated 4-thiouridylation of tRNA begins with mobilization of sulfite from free cysteine via the thsC desulfurase, followed by transfer of the sulfite to the conserved cysteine residue found in the RHOD domain (Fig. 4B). In parallel, the THUMP domain binds and positions the tRNA and the PP-loop domain activates the C-terminal RNA-binding THUMP domain44 with versions from many bacteria and a few archaea that play dual roles in thiouridylation. The ThiI+DUF814 proteins lack both THUMP and RHOD fusions, ThiI-like enzymes (Supplemental Material) and (2) while the ThiI+DUF814 proteins lack both THUMP and RHOD fusions, they retain the conserved residues required for ATP utilization/adenylation found across all PP-loop domains in addition to conserving the internal cysteine residue and DxxCxxH motif characteristic and specific to the Thi-like PP-loop domains (Supplemental Material). It is also worth noting the variable flexible loop region covering the Thl active site appears to be more elaborate in the ThiI+DUF814 fusion proteins than canonical ThiI-like domains and houses several unique, strongly conserved motifs (GrxRxQ and TxsE and a glutamate (Supplemental Material). In Francisellaphilomadegia, the ThiI+DUF814 protein is additionally fused to a TisA/Siru-like (TisA) domain at the C terminus (Fig. 1). Inspection of the gene neighborhoods surrounding the ThiI+DUF814 fusion proteins revealed further association with a TisA domain, the only conserved gene neighborhood association found across phylogenetically diverse bacteria. Within these neighborhoods, a subset of bacteria including several...
domains in humans, consistent with a role for the domain in associations with several genes encoding diverse RNA-binding genes; henceforth, we refer to this domain as the NFACT-R dict DUF814 acts as an RNA-binding domain in the NFACT protein family. Given these observations and previous connections between the THUMP domain and FeS cluster biogenesis, the high degree of sequence similarity between the ThiI domain and the THUMP domain, positioning it for ThiI-catalyzed sulfur transfer (Supplemental Material). The web of contextual information presents several reasons strongly supporting the ThiI+DUF814 proteins being part of a distinct pathway catalyzing 4-thiouridylation in a subset of prokaryotes: (1) mutual exclusivity in phyletic distributions between ThiI+DUF814 and other Thi-like enzymes indicates functional equivalence of the two (Supplemental Material). (2) The presence of the ThiI domain, which can compensate for the absence of the RHOD domain and act as a sulfur acceptor prior to transfer to the tRNA. This proposed 4-thiouridylation pathway is predicted to proceed as follows (Fig. 4B): (1) Analogous to its role in 2-thiouridylation of tRNA, the ThiI domain likely accepts sulfur abstracted from the free cysteine pool in the cell by the IscS-like desulfurase domain. (3) The sulfur relay could also include transfer through IscU/NifU based on its scattered presence in gene neighborhoods (Fig. 1), potentially representing the first known involvement of IscU/NifU in tRNA thiolation as opposed to its standard role in FeS cluster biogenesis. (2) ThiI then functions similar to the RHOD domain in standard 4-thiouridylation pathways by interacting with ThiI and positioning the donor sulfur near the active site of ThiI. (3) The DUF814 domain binds the tRNA substrate in lieu of the THUMP domain, positioning it for ThiI-catalyzed adenylation of the target uridine and ultimately sulfur transfer. Given these observations and previous connections between the NFACT gene family and nucleic acid-binding, we can also predict DUF814 acts as an RNA-binding domain in the NFACT family gene; henceforth, we refer to this domain as the NFACT-R domain. The DUF814-R domain protein family of eukaryotic CCDC25/Jan2-like proteins have been previously experimentally characterized; however, functional coupling networks detect strong associations with several genes encoding diverse RNA-binding domains in humans, consistent with a role for the domain in RNA-binding. CCDC25 shares interactions with proteins which also interact with NFACT proteins, though the two have yet to be directly linked. The CCDC25/Jan2-like family of NFACT-R domains exhibits strong conservation in eukaryotes suggesting purifying selection, a feature of RNA-associated domains functioning in core biological processes like base modification and translation. The close relationship between the two NFACT-R domains and the version in the NFACT proteins along with its pan-eukaryotic distribution suggests that it likely emerged through partial duplication from a NFACT precursor prior to the Last Eukaryotic Common Ancestor (LECA). Hence, we predict that it is likely to function in an RNA-binding role in a core cellular function, perhaps even in the same complex as the NFACT proteins (see below).

**DUF3441 domain**

In the Pfam database, DUF3441 is presented as an eukaryotic-specific domain. However, we recovered divergent yet clearly homologous versions at the C-terminus of archaeal NFACT proteins. For example, a PSI-BLAST search with the C-terminal domain from the archaeon *Aquaspirillum nifidum* recovered the entirety of the eukaryotic NFACT DUF3441 domain from the annelid *Capitella teleta* (gi: 443707385, iteration: 4, e-value: 2e-15). Given its presence as a core NFACT domain, we rename this domain the NFACT-C domain (NFACT-C terminal domain). Secondary structure predictions based on a multiple sequence alignment suggest that it adopts an α/β fold. In contrast to other domains in NFACT proteins, there is little absolute conservation of residues outside of a strongly-conserved PG motif (Supplemental Material); however, several positions in the domain are retained as different polar or charged residues. This pattern suggests NFACT-C could mediate protein–protein contacts within a larger complex rather than playing a catalytic role.

**Contextual analysis of the NFACT gene family**

We then investigated the NFACT family itself for potential conserved gene neighborhoods and functional interaction networks, as this information can provide insight into function of uncharacterized domains. Gene neighborhoods extracted for bacterial NFACT genes did not recover any conserved associations; however, archaeal NFACT genes formed a conserved gene neighborhood across all archaeal lineages, including the euryarchaeota, crenarchaeota, nanarchaeota, thaumarchaeota, and the caldarchaeota with two genes coding for: (1) An active Mut7-C RNase domain of the PIN nuclease fold with its accompanying C-terminal Zn-ribbon domain and (2) the AMMCR1 domain containing the RAGNYA fold (Supplemental Material). The conserved connection to the Mut7-C RNase is again strongly suggestive of an RNA-related role for NFACT. Additionally, AMMCR1 has previously been linked to involvement in an as-yet-uncharacterized RNA base modification, potentially entailing the transfer of a modifying group onto an RNA base via a conserved cysteine. Notably, most AMMCR1 domains found in this conserved neighborhood retains all previously predicted enzymatic residues (except the crenarchaeotum *Vulcanisaeta* where the predicted catalytic cysteine residue is replaced by a serine; see Supplemental Material). Functional interaction networks constructed primarily by co-expression patterns and protein–protein interaction data were also suggestive of an RNA-related role for the NFACT family. The yeast Tic2 protein is most strongly linked to a cluster of proteins involved in ribosomal subunit biogenesis. Consistent with this, recent studies identified it as a part of the large (60S) ribosomal subunit interacting Ribosomal Quality Control Complex (RQC).
Cognate NEMF proteins from mammals were predominantly linked to proteins harboring diverse RNA-binding domains, many with demonstrated roles in splicing and RNA biogenesis.

Discussion
Emerging picture of NFACT as a potential RNA-modifying enzyme
Multiple independent lines of evidence support a RNA-related role for NFACT proteins: (1) the previously determined evolutionary relationship between the NFACT HHH domains and those observed in the S13/S18 ribosomal proteins, (2) the evolutionary relationship between the NFACT-N domain and the FMN-DG catalytic domain suggesting NFACT might function as an enzyme operating on bases in double stranded nucleic acids, (3) conserved operonic associations in archaea with the AMMECR1 and the Mut7-C RNase domain pointing in the direction of RNA processing and modification, and (4) experimental evidence from yeast13,15 and functional network associations suggesting one or more rRNAs as probable substrates.

In terms of RNA-modification reactions, parallels can be drawn between the glycosylase reaction catalyzed by FMN-DG and base-modification mechanisms catalyzed by structurally unrelated RNA-modifying enzymes: the pseudouridine synthases catalyzing formation of pseudouridine and the RNA-guanine trianglycylases (TGTs) catalyzing the "base-swapping" mechanism, which inserts the preQ0 base precursor of archaeosomes in archaea and queuosine in bacteria and eukaryotes.54,55 Members of the pseudouridine synthase fold, which includes the Trub, RhuA, RuaA, and Trua-like families, similar to FMN-DG require an initial base-flipping step to position a base in the active site prior to modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine modification. Pseudouridine synthases utilize an absolutely conserved aspartate as the primary catalytic residue during the uridine modification.

Despite similarities, the aspartate in FMN-DG plays only a transient role in the catalytic mechanism, sharply contrasting the centrality of the aspartate in the pseudouridine synthase mechanism, which is thought to directly form a Michael adduct with a carbon in the pyrimidine ring of the uridine as a reaction intermediate.56,57 Thus, it seems unlikely NFACT would follow the mechanism template established in pseudouridine synthases. TGTs, members of the TIM-barrel fold, also require an absolutely conserved aspartate residue when catalyzing the swapping of a rRNA guanine base for the preQ0 precursor base. In contrast to the pseudouridine synthase mechanism and closer to the FMN-DG glycosylase mechanism, the aspartate is involved in the initial step of the reaction facilitating the removal of the base from the sugar backbone.58 A further parallel with FMN-DG is observed in the base-exchange step wherein N9 of the amidoamine ring of the preQ0 base directly attacks the C1' of the "base-less" sugar, resulting in attachment of the preQ0 to the rRNA sugar backbone (Fig. 4A). This step strikingly resembles the attack by the nitrogen atom belonging to the conserved FMN-DG proline residue on the C1' of the DNA substrate backbone following base removal. It is possible that the secondary emergence of the proline and the Schiff base observed in FMN-DG activity displaced a nitrogen-based attack from an incoming base similar to the elucidated TGT mechanism in the ancestral NFACT-N domain (Fig. 4A). This reasoning leads to a proposal that the NFACT proteins could possibly catalyze a base-exchange reaction in which a regular RNA base is initially removed and replaced by a new base, similar to the action of the TGTs.

While more speculative in nature, it is possible to obtain certain clues regarding the nature of the potential target of the predicted NFACT-N catalytic domain: if NFACT-N were to act similar to FMN-DG, then among the endogenous bases in RNA uridine or cytidine contain spatially comparable carbonyl groups to FMN-DG substrates.59-61 In addition to formation of pseudouridine and several pseudouridine derivatives, a range of rRNA base modifications have been previously characterized and include various forms of methylated, hydroxylated, and acetylated bases.62,63 However, while pseudouridine and deazaguanine (e.g., preQ0 formation shares some kinship with base-exchange reactions and at least some modified rRNA bases do exist as free bases in the cell,61 most well-studied base modifications typically proceed via direct enzymatic attachment of a chemical group on an existing base and not through a base-exchange reaction. Additionally, it is not clear that NFACT catalyzes a terminal reaction on a base: a potential exchanged base could, similar to preQ0 attachment via TGT, be followed by further modifications catalyzed by distinct enzymes. Another possibility is that NFACT could act as an rRNA repair enzyme, replacing damaged bases with normal bases. Oxidation of RNA bases, particularly in conjunction with cell death, is currently an area of emerging research interest,57,64 given the connections between NFACT-N and FMN-DG, this possibility could warrant additional investigation.

This proposed role for NFACT in RNA base-modification potentially unifies some of the disparate experimental findings on these proteins. The yeast NFACT protein Tac2 appears to activate the transcription factor Hsf1 in response to translational stress detected by the RQC, ultimately leading to degradation of peptides derived from mRNAs lacking a stop codon.65 It is conceivable that this proceeds via an effect of the proposed modification on translation fidelity consistent with the observed role of certain rRNA modifications, such as pseudouridylation, on ribosomal stability and translation fidelity.66 Translation of Internal Ribosomal Entry Sites (IRES)-bearing mRNAs, which play a role in responding to stress conditions across eukaryotes,62 also depends on rRNA base modifications like pseudouridylation.60-61 Hence, it is conceivable that the proposed modification mediated by NFACT proteins might also interact with IRES-mediated translation under stress. The Drosophila NFACT protein Caliban has been proposed to be part of a network including p53, caspase, and Hid proteins during DNA damage-induced apoptosis.62 Here again the modification could play a role as part of a stress response ensuing from DNA damage, as many key oncomers and apoptosis factors are translated via IRES.

Certain studies have suggested that members of the NFACT family are virulence factors of pathogenic bacteria involved in
mediating adhesion and invasion in light of its capacity to bind fibronectin/fibrinogen.7,73,74 These studies have also sought to demonstrate its presence in the extracellular space around some of these bacteria.7,75,76 From an evolutionary perspective, however, demonstra...and invasion during infection.69,77-80 Thus, the proposed extracellular NFACT proteins are byproducts of cell lysis during biofilm formation;77 consistent with this, these proteins have been identified in extracellular space occupied by biofilm with no apparent direct attachment to the cell membrane.77,79 This would also be consistent with recent results questioning various aspects of potential direct NFACT roles in mediating adhesion and invasion during infection.80-77,80 Thus, the proposed extracellular fibronectin/fibrinogen-binding role is unlikely to be a general one for this family and might merely reflect promiscuous interactions mediated by the extensive coiled-coil regions and facilitated by biofilm formation. However, we cannot rule out the possibility that in certain pathogenic bacteria NFACT might contribute specifically to biofilm formation or adhesion.

General conclusions

Through synthesis of sequence, genome, functional interaction, and structural data, we propose a potential role for the NFACT proteins conserved across the three superkingdoms of life in RNA-base modification in the context of translation. We present evidence that this function is mediated by the catalytic activity of the NFACT-N domain, shown here to be related to DNA glycosylases. We also predict an RNA-binding role for the NFACT-R (formerly DUF814) domain. In a diverse subset of bacteria lacking the canonical 4-thiouridylation pathway,81 the NFACT-R domain. In a diverse subset of bacteria lacking the canonical 4-thiouridylation pathway,81 the NFACT-R domain. In a diverse subset of...demonstrate its presence in the extracellular space around some of these bacteria.7,73,74 From an evolutionary perspective, however, demonstra...and invasion during infection.69,77-80 Thus, the proposed extracellular NFACT proteins are byproducts of cell lysis during biofilm formation;77 consistent with this, these proteins have been identified in extracellular space occupied by biofilm with no apparent direct attachment to the cell membrane.77,79 This would also be consistent with recent results questioning various aspects of potential direct NFACT roles in mediating adhesion and invasion during infection.80-77,80 Thus, the proposed extracellular fibronectin/fibrinogen-binding role is unlikely to be a general one for this family and might merely reflect promiscuous interactions mediated by the extensive coiled-coil regions and facilitated by biofilm formation. However, we cannot rule out the possibility that in certain pathogenic bacteria NFACT might contribute specifically to biofilm formation or adhesion.

Materials and Methods

Iterative sequence profile searches were performed using the web implementation of the PSI-BLAST program87 (with the following listed parameters different from default: -num_descriptions 200,000, -evalue 20, -comp_based_stats 1, -pseudo_count 30, -psiblast_threshold 0.01) and web version 1.5 of the jackhmmer (http://jhmmer.janelia.org/search/jackhmmer) program run with default parameters against the non-redundant (NR) protein database at the National Center for Biotechnology Information (NCBI). Multiple sequence alignments were built by the Kalign28 and MUSCLE88 programs with default parameters, followed by manual adjustments on the basis of profile-profile and structural alignments. Similarity-based clustering for both classification and culling of nearly identical sequences was performed using the BLASTCLUST program with empirically determined length and score threshold parameters (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The web-based implementation of the HHpred program89 based on the HHsuite-2.0.15 software package with default parameters was used for profile-profile comparisons, searching against the PDB70 and PfamA_27.0 pre-configured databases. Structure similarity searches were performed using the DALI Lite v. 3 program.90 Secondary structures were predicted using the JPred 3 program with default parameters.90 For previously known domains, the Pfam database release 27.0 was used as a guide and augmented by addition of newly detected divergent members. Structural visualization and manipulations were performed using the Open-Source PyMOL 1.5.0.3 (http://www.pymol.org) program. Funcoup3.0 was used to analyze contextual information based on interaction and expression data.89 Gene neighborhoods were determined using either the PTT file (downloadable from the NCBI ftp site) or the GenBank file in the case of whole genome shotgun sequence data.89 The protein sequences of all neighbors were clustered using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/docs/documents/blastclust.html) to identify related sequences in gene neighborhoods. Each cluster of homologous proteins were then assigned an annotation based on the domain architecture or conserved shared domain. Neighborhoods were further refined by ensuring that genes are unidirectional on the same strand of DNA and shared a putative common promoter to be counted as a single operon. If they were on opposite strands they were examined for potential bidirectional promoter sharing patterns. In-house Perl scripts were used to automate this analysis of genome context.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
Burroughs AM and Aravind LV’s research is supported by the funds of the Intramural Research Program of the NIH, National Library of Medicine.

Supplemental Material
Supplemental material may be found here: www.landesbioscience.com/journals/rnabiology/article/28302
The tumor suppressor Caliban regulates DNA damage. Defeucanu C, Jacquier A, et al. Cdc48-associated complex in Streptococcus gordonii. Microbiology 2002; 148:639-49; PMID:15228540; http://dx.doi.org/10.1099/0076-6879-148-3-639.

Shiba K, Arai Y, Matsuda M, et al. Regulation of mitotic chromosome separation by a novel WD-repeat protein. Cell 2000; 100:659-69; PMID:10765157; http://dx.doi.org/10.1016/S0092-8674(00)00127-8.

Henderson B, Nair S, Fullam J, Williams MA. The crystallographic solution of a E. coli chromosome has been determined by bacterial histone-fluorescent proteins. FEMS Microbiol Rev 2001; 25:479-88; PMID:11399295; http://dx.doi.org/10.1111/j.1574-6976.2001.tb00545.x.

Bosman F, Steeghs D, van der Does van Dijck F, et al. The tRNA modification enzyme Thiomodifying enzyme ThiK from Bacillus subtilis is an Escherichia coli thiouridine thiolase. J Bacteriol 2000; 182:620-9; PMID:10628583; http://dx.doi.org/10.1128/JB.182.3.620-629.2000.

DeMott J, Lee V, Meinicke J, Namba M, Galas J, Ducruix A, Leenstra M, Fekete A, et al. CD8-associated protein bound to folic acid is required for the clearance of aberrant transitional precursors. Proc Natl Acad Sci U S A 2013; 110:15664-9; PMID:23847576; http://dx.doi.org/10.1073/pnas.1305203110.

Wang Y, Wang Z, Ichiki RH, Pali RK, Stel P, Yuan Q, Wei Z, Hu Y, Wang Y, Dong X, et al. The tumor suppressor Caliban regulates DNA damage by promoting the recruitment of p53 to independent activity. Oncogene. 2013; 32:8577-8586; PMID:24166672; http://dx.doi.org/10.1038/onc.2013.395.
43. Aloyzas S, Schmitz T, Tjärnberg A, Guala D, Ferrer O, Sommarin EL. Comparative inter- 
action with Pancreatic Z-B, Nuclear Acids Res 2012; 40:682-68. PMCID: 3218854; h 
44. Aravind L. Gap as association: contextual information in 
45. Hayman M, Stad L, Bock W, 3rd, Bock P. Predicting protein--protein interactions: 
46. Nacura T, Takah K, Kukita J, Nishi S, Tagahara K, Saito M, Ogasawara H, et al. Inter 
47. Asahi S, Nishimura S, Katze JR, Hashizume T, Crain PF. The RNA Modification Database, 
48. Rama C. Critical aspartic acid residues for the action of a guanine insertion enzyme, a specific 
49. Garcia GA. tRNA-guanine transglycosylase from 
50. Kitt O. Guilt by association: contextual informa-
51. Aravind L, Aravindyan L, Zhang D, de Sousa RT, Jey LM. Gene flow and biological conflict 
52. Jey LM, Aravindyan L, Zhang D, de Sousa RT, Jey LM. Gene flow and biological conflict 
53. Spedding CJ, Ginter JM, Johnston MV, Mueller EG. The pseudouridylation synthetase 
54. Xu G, Liu Y, Sanyu DW. The mechanism of post-
55. Barkh S, Aravind L. The RNAMbase: a novel tool with multiple epigenetic variants found in function 
56. Watanabe M, Marato M, Tanioka S, Akahane H, Hino K, Hishika K, et al. Analysis of 
57. Sumi Y, Li Z, Wu J, Dalois CJ. tRNA Gene Damma and 
58. Yoon Y, Choe HY, Alsberg E, Wu J, et al. Insights into the evolution of Archaea and 
59. Pracikovics B, Szafiendi N, Erdélyi E, Kurkowska M, Olchowik A, Januszewski K, et al. 
60. Ohara-Wada N, Kitaura H, Nishiyama M, Ohara O, Do-Okada N, Okada N, Ohgi T, 
61. Wang J, Saito H, Hattori M, Kikuchi S, et al. Structural and functional comparison of the Ty 
62. Lanfear D, Elm C, Gerber J, Bergmann S, Rohde M, Hammerschmidt S. PavA of Str 
63. Chiang WC, Chee GJ, Hattori M, Kanai A, Atomi E, Nakanishi S, Nagata M, et al. A 
64. Natori M, Takai J, Ando H, Yamada K, Itakura S, Ikehara S, Yoshida T, et al. Gene expression 
65. Tjärnberg A, Guala D, Ferrer O, Sommarin EL. Comparative interaction with Pancreatic 
66. Miura K, Tsuda K, Yamagata H, et al. Prediction of multiple topological variants found in function 
67. Spedding CJ, Ginter JM, Johnston MV, Mueller EG. The pseudouridylation synthetase 
68. Huan C J, Mueller EG. Critical aspartic acid residues for the action of a guanine insertion 
69. Treff R, Sere P, Bagl I, Parrot Y, Ishikawa H, et al. KRAS gene amplification in colorectal 
70. Prabhu P, Verma SK, Chaudhry A, Misra B, Bora S, et al. Predicting protein--protein interactions 
71. Ferrer O, Sommarin EL. Comparative interaction with Pancreatic Z-B, Nuclear Acids Res 
72. Nacura T, Takah K, Kukita J, Nishi S, Tagahara K, Saito M, Ogasawara H, et al. Insights into the evolution of Archaea and eukaryotic proteins: motifs revealed by the genome of a novel archaeal group. Nuclear Acids Res 2011; 39:3204-23. PMCID: 3189985; h 
73. Asahina A, Ono K, Ono S, Ito T, Nishimura A, Wada M, et al. Identification of gene 
74. Nacura T, Takah K, Kukita J, Nishi S, Tagahara K, Saito M, Ogasawara H, et al. Insights into the evolution of Archaea and eukaryotic proteins: motifs revealed by the genome of a novel archaeal group. Nuclear Acids Res 2011; 39:3204-23. PMCID: 3189985; h 
75. Asahi S, Nishimura S, Katze JR, Hashizume T, Crain PF. The RNA Modification Database, 
76. Garcia GA. tRNA-guanine transglycosylase from 
77. Sumi Y, Li Z, Wu J, Dalois CJ. tRNA Gene Damma and 
78. Yoon Y, Choe HY, Alsberg E, Wu J, et al. Insights into the evolution of Archaea and 
79. Pracikovics B, Szafiendi N, Erdélyi E, Kurkowska M, Olchowik A, Januszewski K, et al. 
80. Ohara-Wada N, Kitaura H, Nishiyama M, Ohara O, Do-Okada N, Okada N, Ohgi T, 
81. Wang J, Saito H, Hattori M, Kikuchi S, et al. Structural and functional comparison of the Ty 
82. Lanfear D, Elm C, Gerber J, Bergmann S, Rohde M, Hammerschmidt S. PavA of Str 
83. Chiang WC, Chee GJ, Hattori M, Kanai A, Atomi E, Nakanishi S, Nagata M, et al. A 
84. Natori M, Takai J, Ando H, Yamada K, Itakura S, Ikehara S, Yoshida T, et al. Gene expression 
85. Tjärnberg A, Guala D, Ferrer O, Sommarin EL. Comparative interaction with Pancreatic 
86. Nacura T, Takah K, Kukita J, Nishi S, Tagahara K, Saito M, Ogasawara H, et al. Insights into the evolution of Archaea and eukaryotic proteins: motifs revealed by the genome of a novel archaeal group. Nuclear Acids Res 2011; 39:3204-23. PMCID: 3189985; h 
87. Asahina A, Ono K, Ono S, Ito T, Nishimura A, Wada M, et al. Identification of gene 
88. Nacura T, Takah K, Kukita J, Nishi S, Tagahara K, Saito M, Ogasawara H, et al. Insights into the evolution of Archaea and eukaryotic proteins: motifs revealed by the genome of a novel archaeal group. Nuclear Acids Res 2011; 39:3204-23. PMCID: 3189985; h 
89. Asahi S, Nishimura S, Katze JR, Hashizume T, Crain PF. The RNA Modification Database, 
90. Garcia GA. tRNA-guanine transglycosylase from 
91. Sumi Y, Li Z, Wu J, Dalois CJ. tRNA Gene Damma and 
92. Yoon Y, Choe HY, Alsberg E, Wu J, et al. Insights into the evolution of Archaea and 
93. Pracikovics B, Szafiendi N, Erdélyi E, Kurkowska M, Olchowik A, Januszewski K, et al. 
94. Ohara-Wada N, Kitaura H, Nishiyama M, Ohara O, Do-Okada N, Okada N, Ohgi T, 
95. Wang J, Saito H, Hattori M, Kikuchi S, et al. Structural and functional comparison of the Ty 
96. Lanfear D, Elm C, Gerber J, Bergmann S, Rohde M, Hammerschmidt S. PavA of Str 
97. Chiang WC, Chee GJ, Hattori M, Kanai A, Atomi E, Nakanishi S, Nagata M, et al. A 
98. Natori M, Takai J, Ando H, Yamada K, Itakura S, Ikehara S, Yoshida T, et al. Gene expression 
99. Tjärnberg A, Guala D, Ferrer O, Sommarin EL. Comparative interaction with Pancreatic 
100. Nacura T, Takah K, Kukita J, Nishi S, Tagahara K, Saito M, Ogasawara H, et al. Insights into the evolution of Archaea and eukaryotic proteins: motifs revealed by the genome of a novel archaeal group. Nuclear Acids Res 2011; 39:3204-23. PMCID: 3189985; h
84. Lassmann T, Frings O, Sonnhammer EL. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 2009; 37:958-65; PMID:19301465; http://dx.doi.org/10.1093/nar/gkn1006

85. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004; 5:113; PMID:15516955; http://dx.doi.org/10.1186/1471-2105-5-113

86. Hildebrand A, Rotmer M, Biegert A, Stößing J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009; 75(Suppl 1):S20-32; PMID:19626712; http://dx.doi.org/10.1002/prot.22499

87. Holm L, Kääriäinen S, Rosenström P, Schenkel A. Searching protein structure databases with DaliLite v3. Bioinformatics 2008; 24:2780-1; PMID:18818215; http://dx.doi.org/10.1093/bioinformatics/btn507

88. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 2008; 36:W200-5; PMID:18463136; http://dx.doi.org/10.1093/nar/gkn238

89. Schütt T, Opio C, Sonnhammer EL. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res 2014; 42:D380-8; PMID:24185702; http://dx.doi.org/10.1093/nar/gkt984