Monophyly of Class I Aminoacyl tRNA Synthetase, USPA, ETFP, Photolyase, and PP-ATPase Nucleotide-Binding Domains: Implications for Protein Evolution in the RNA World

L. Aravind,† Vivek Anantharaman, and Eugene V. Koonin
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

ABSTRACT **Protein sequence and structure comparisons show that the catalytic domains of Class I aminoacyl-tRNA synthetases, a related family of nucleotidyltransferases involved primarily in coenzyme biosynthesis, nucleotide-binding domains related to the UspA protein (USPA domains), photolyases, electron transport flavoproteins, and PP-loop-containing ATPases together comprise a distinct class of α/β domains designated the HUP domain after *H*IGH-signature proteins, *U*spA, and *P*P-ATPase. Several lines of evidence are presented to support the monophyly of the HUP domains, to the exclusion of other three-layered α/β folds with the generic "Rossmann-like" topology. Cladistic analysis, with patterns of structural and sequence similarity used as discrete characters, identified three major evolutionary lineages within the HUP domain class: the PP-ATPases; the HIGH superfamily, which includes class I aaRS and related nucleotidyltransferases containing the HIGH signature in their nucleotide-binding loop; and a previously unrecognized USPA-like group, which includes USPA domains, electron transport flavoproteins, and photolyases. Examination of the patterns of phyletic distribution of distinct families within these three major lineages suggests that the Last Universal Common Ancestor of all modern life forms encoded 15–18 distinct α/β ATPases and nucleotide-binding proteins of the HUP class. This points to an extensive radiation of HUP domains before the last universal common ancestor (LUCA), during which the multiple class I aminoacyl-tRNA synthetases emerged only at a late stage. Thus, substantial evolutionary diversification of protein domains occurred well before the modern version of the protein-dependent translation machinery was established, i.e., still in the RNA world. Proteins 2002;48:1–14.**

Key words: HUP domain; last universal common ancestor (LUCA); class I aminoacyl-tRNA synthetases

INTRODUCTION

Genome comparisons that became feasible as a result of the "genomic revolution" made it clear that translation apparatus is the most conserved component of all cellular housekeeping machinery.[1–5] The core of the translation system can be traced back to the last universal common ancestor (LUCA) of all life forms. The universal repertoire of proteins involved in translation consists of 18 aminoacyl-tRNA synthetases (aaRS) of two unrelated classes; approximately 35 ribosomal proteins; translation factors, many of which are GTPases; and some RNA-binding proteins.[6] This is in contrast with other central cellular systems, such as replication and transcription, that show far fewer conserved components that can be traced back to the LUCA, although they are highly conserved within each of the two major divisions of life, the archaeo-eukaryotic and bacterial lineages.[7–9] The simplest interpretation of this pattern of evolutionary conservation is that the translation apparatus was the earliest basic system to "crystallize"[3] into a form close to modern before the divergence of the major lineages of life from the LUCA. In contrast, the replication and transcription systems might not have assumed their modern, complex forms in the LUCA, with some of their central components probably evolving independently in the two principal lines of descent.[9]

Sequence–structure analysis of proteins in the translation machinery points to some ancient duplication and divergence events that occurred before the emergence of the LUCA. In particular, phylogenomic analysis shows that the LUCA encoded most members of the two classes of aaRS that had already attained domain architectures close to those that were extant. Thus, evolution of both classes of aaRS from their respective common ancestors, which included several rounds of duplications, must have predated the LUCA.[10,11] Both classes of aaRS are related to other proteins that do not function as aaRS, but have other diverse, ATP-dependent catalytic activities. Specifically, the HIGH-motif-containing-aaRS (class I) are related to various nucleotidyltransferases, such as glycerol 3-phosphate cytidylyltransferase and pantothenate synthetase, which also preserve an equivalent of the HIGH

†Correspondence to: L. Aravind, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894. E-mail aravind@ncbi.nlm.nih.gov

Received 12 July 2001; Accepted 2 November 2001

motif.[12] Class II aaRS are related to several families of enzymes, including lipoate-protein ligase, biotin-protein ligase, and ammonia-dependent asparagine synthetase.[13] Elucidation of additional homologous relationships and evolutionary analysis of aaRS and their homologs might shed light on early events in the evolution of the translation system.

In the present study, using sequence profile analysis and structural comparisons, we expand the relationships of the class I aaRS and their closest relatives, the HIGH-motif nucleotidyltransferases, by establishing their links with two other superfamilies of NTP-hydrolyzing enzymes and nucleotide-binding proteins. One of these superfamilies includes the UspA family, photolyases, and electron transfer flavoproteins, and the other consists of the PP-loop NTPases. Using sequence-similarity-based clustering, conserved sequence motifs, and structural features, we develop an evolutionary classification for these protein superfamilies. Elucidation of these connections suggests considerable diversification of this class of nucleotide-binding proteins before the emergence of a complete, protein-based translation system.

MATERIALS AND METHODS

The nonredundant protein sequence database, the Expressed Sequence Tags database (National Center for Biotechnology Information, NIH, Bethesda), and individual protein sequence databases of completely and partially sequenced genomes accessible at http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html were searched using the gapped version of the BLAST programs[14,15] (BLASTPGP for proteins and TBLASTNGP for translating searches of nucleotide databases). Sequence profile searches were performed using the PSI-BLAST program; profiles were saved using the -C option and retrieved using the -R option.[15,16] Multiple alignments of amino acid sequences were generated using a combination of PSI-BLAST, T_Coffee,[17] Gibbs sampling,[18,19] and secondary structure predictions that were produced using the PHD program,[20,21] with multiple alignments of individual protein families used as queries. Sequence–structure threading was carried out using the combined fold prediction algorithm.[22] Structural comparisons were performed using the DALI[23] and VAST[24] programs, whereas the FSSP,[25] SCOP,[26,27] and CATH[28] databases were used as guides for structural classification of proteins. Orthologous relationships between proteins were partially derived from the COG database.[6] Cladistic analysis was performed using the PAUP program, with both the exhaustive tree search and branch and bound algorithms.[29,30] The characters were all scored as unordered, i.e., a primitive state (0) could transition independently to multiple derived states (e.g., 1, 2). The character matrix for cladistic analysis was constructed using the NDE program. (See http://www.gwu.edu/~clade/faculty/lipscomb/Cladistics.pdf for a description of cladistic methodology and parsimonious tree searching algorithms.) Three-dimensional (3D) structure visualization, alignment, and modeling were carried out using the SWISS-PDB-Viewer program.[31] Ribbon diagrams were constructed using Molscript.[32]

RESULTS AND DISCUSSION

Detection of Sequence and Structural Relationships of the HIGH and PP-ATPase Superfamilies: the HUP Class of Nucleotide-Binding Domains

The catalytic domain of class I aaRS contains a three-layered α/β domain that has five core strands arranged in the 3–2–1–4–5 order, surrounded by four helices that alternate with these strands[33–35] (Fig. 1). The signature HIGH motif is located in the loop between strand 1 and helix 1 and participates in ATP-binding and the catalysis of the aminoacylation reaction. In addition, class I aaRS have an extension that consists of a loop followed by 2 α-helices, located C-terminal to the core α/β domain. The loop in this extension contains the second characteristic motif, with the consensus sequence KMSKS. The second lysine in the KMSKS motif is indispensable for ATP hydrolysis by class I aaRS[36] and, in addition, the KMSKS motif has been reported to contribute to initial tRNA binding.[37] The detection of candidate HIGH and KMSKS motifs in a diverse family of nucleotidyltransferases, typified by glycerol 3-phosphate cytidylyltransferase and pantothenate synthetase (hereafter referred to as HIGH NTases), by means of sequence analysis, followed by molecular modeling, strongly suggested that they were related to the class I aaRS.[12] This prediction was confirmed when several crystal structures became available for the HIGH NTase family; structural comparisons indicated that class I aaRSs and HIGH NTases formed a distinct superfamily (hereinafter the HIGH superfamily), to the exclusion of all other proteins.[38–40] The structural classification presented in the SCOP database[26,27] reflects an even deeper structural similarity between the HIGH superfamily and the PP (pyrophosphate)-loop ATPases (hereinafter PP-ATPases),[41] a diverse superfamily of domains that catalyze reactions typically involving hydrolysis of the α–β bond in ATP.

The HIGH and PP-ATPase superfamilies share a three-layered α/β fold with a central sheet, loosely termed the "Rossmann-like fold" in certain protein classification systems, such as CATH.[28] This generic structural category includes numerous diverse protein families with a broadly similar topology and spatial geometry, such as the dinucleotide-binding domains of NAD/FAD-dependent oxidoreductases (the classic Rossmann fold), and the *S*-adenosyl methionine-binding domain of methyltransferases, the TOPRIM domain of topoisomerases and primases, von Willebrand factor A (vWA) domains, and GTPases of the tubulin-FtsZ family. The hierarchy of relationships between the "Rossmann-like domains" that, in principle, could be established on the basis of the conservation of active sites and specific structural features, such as the number of strands and the angle between them in the core sheet, remains unclear. In an attempt to delineate some of the ancient lines of descent within this broad class of α/β protein domains, we performed a detailed sequence and

TABLE I. Connections Between HUP-Domain Families Detectable Using Sequence and Structure Comparison Methods*

	Class I aaRS	HIGH-nuct	UspA	ETFP	Photolyase	PP-ATPases
Class I aaRS	—	DALI: 2 (Z:8.1) VAST: 10e-11.3 *E*: 0.01 (6)	DALI: 3 (Z:8.3) VAST: 10e-9.1	DALI: 2 (Z:7.0) VAST: 10e-5.9	DALI: 4 (Z:5.5) VAST: 10e-7.2	DALI: 3 (Z:4.4) VAST: 10e-3.4
HIGH-nuct	DALI[a]: 1 (Z:9.3) VAST[b]: 10e-6.7 *E*: 0.001 (3)[c]	—	DALI: 4 (Z:8.2) VAST: 10e-8.2 *E*: 0.3 (12)	DALI: 3 (Z:6.6) VAST: 10e-4.9	DALI: 3 (Z:7.4) VAST: 10e-10.8	10e-2.8 DALI: 1 (Z:6.1)
UspA	DALI: 2 (Z:8.3) VAST: 10e-6.7	DALI: 1 (Z:8.2) VAST: 10e-7.4	—	DALI: 1 (Z:10.4) VAST: 10e-5.9 *E*: 0.01 (6)	DALI: 1 (Z:9.4) VAST: 10e-9.7	DALI: 2 (Z:5.9) VAST: 10e-3.5
ETFP	DALI: 3 (Z:7.3) VAST: 10e-4.7	DALI: 4 (Z:6.8) VAST: 10e-5.1	DALI: 1 (Z:10.4) VAST: 10e-9.8 *E*: 8×10^{-4} (2)	—	DALI: 2 (Z:8.3) VAST: 10e-6.8	DALI: 4 (Z:4.3) VAST: 10e-3.3
Photolyase	DALI: 3 (Z:5.5) VAST: 10e-4.7	DALI: 3 (Z:7.4) VAST: 10e-5.1	DALI: 2 (Z:9.4) VAST: 10e-9.3 *E*: 4×10^{-6} (4)	DALI: 3 (Z:6.6) VAST: 10e-5.2	—	DALI: 5 (Z:4.2) VAST: 10e-3
PP-ATPases	DALI: 3 (Z:5.5) VAST: 10e-2.8	DALI: 5 (Z:6.6) VAST: 10e-4.1	DALI: 5 (Z:6.6) VAST: 10e-6.4 *E*: 5×10^{-4} (8)	DALI: 4 (Z:6.0) VAST: 10e-2.7	DALI: 5 (Z:5.0) VAST: 10e-3.5	—

aaRS, aminoacyl-tRNA synthetases.
*All searches were carried out bidirectionally, with representatives of each family used as queries against the Protein Data Bank (PDB) database with the DALI and VAST programs and against the no. database with the PSI-BLAST program; hence both the lower and upper triangles of the matrix are occupied.
[a]Rank of structural similarity between representatives of the two families (excluding multiples structures from each family), with the *Z*-score (the number of standard deviations above the expectation for unrelated structures) shown in parentheses.
[b]Probability (*P*) value.
[c]Expectation (*E*) value, with the iteration of first detection in PSI-BLAST shown in parentheses.

structural comparison of the HIGH and PP-ATPase superfamilies and other related groups of nucleotide-binding domains. Using the results of these comparisons in conjunction with a cladistic approach, we developed an evolutionary classification of this class of proteins, which has broad implications for the early evolution of the translation system.

Transitive structural comparisons with the PDB database using the DALI and VAST programs[24,25,42] were performed with representatives of the HIGH and PP-ATPase superfamily domains. The DALI searches started with the structures of class I aaRS (e.g., 2ts1, tyrosyl-tRNA synthetase) and HIGH NTases (e.g., 1iho, pantothenate synthetase) detected, among the best hits, and with scores comparable to those of other aaRSs and HIGH NTases, the stand-alone USPA-domain protein MJ0577 (1mjh),[43,44] DNA photolyases (1qnf), and the electron transport flavoprotein (ETFP) (1efv) (Table I). The PP-ATPases, such as GMP synthetase (1gpm), were recovered in these searches with next-best scores. Beyond these hits, a variety of α/β domains were recovered with lower scores and without any discernible preferential ranking. Reciprocal searches started with the MJ0577 and photolyase structures recovered, in addition to each other and ETFP, the HIGH NTases and class I aaRSs as the best hits. Similarly, a search with the PP-ATPase domain of NAD synthetase (2nsy) recovered the MJ0577, ETFP, and tyrosyl- tRNA synthetase as the best hits, ranking above the generic list of hits to α/β domain that did not necessarily share any topological similarity with these above proteins. Furthermore, a VAST search with the USPA domain of MJ0577, which is a compact, stand-alone version of this domain without any inserts or additional domains, produced the longest (105–109 amino acid residues) and best-scoring alignment with tyrosyl-tRNA synthetase and ETFP. In a reciprocal VAST search, the catalytic domain of tyrosyl-tRNA synthetase recovered MJ0577 as the second best hit, ahead of most of the other class I aaRS (Table I). These observations suggested a close structural relationship between USPA domains, ETFP, the α/β domain of photolyases, and the HIGH superfamily. All these domains, in turn, appeared to be related to the PP-ATPases, to the exclusion of all other Rossmann-like domains.

Powerful sequence profile search methods have recently enabled the detection of many distant relationships between proteins that were previously detectable only through structural comparisons.[45,46] This type of analysis complements the relationships detected through structural comparisons, allowing assessment of the statistical significance of the alignments and a more reliable evaluation of potential evolutionary connections.[47] We conducted iterative PSI-BLAST searches,[15] with the USPA and the HIGH NTase domains employed as queries, because they often represent the minimal versions of the nucleotide-binding domain, without any inserts or extensions. A search with the *Escherichia coli* UspA sequence detected PP-ATPases in the 8th iteration, with statistically significant E-values (Table I), without any false positives, and generated a pair-wise alignment compatible with the structural alignment of the USPA and PP-ATPase domains. Similarly, searches against complete proteomes of various organisms with the USPA-domain profiles detected PP-ATPases, in addition to bona fide USPA domains. Likewise, a search initiated with MJ0577 detected

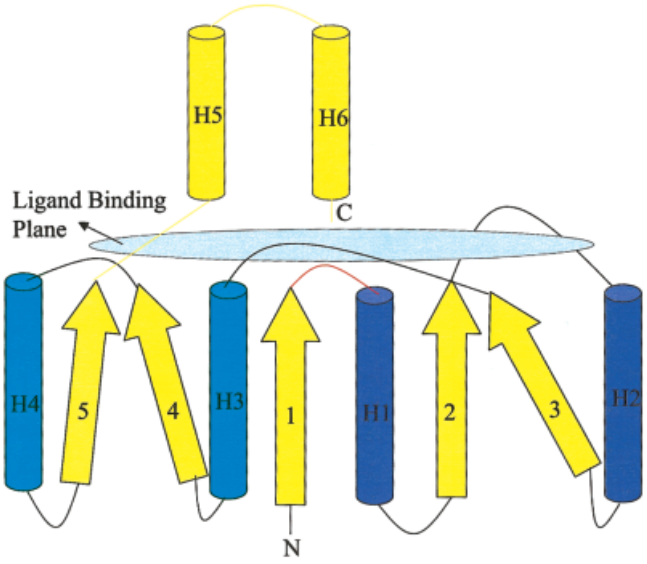


Fig. 1. Topology diagram of the *H*IGH-signature proteins, *U*spA, and *P*P-ATPase—the HUP domain. Yellow arrows, α-strands; blue cylinders, β-helices. Strands and helices are numbered separately in the order of their appearance. The red loop between strand 1 and helix 1 is the nucleotide-binding site, which contains the *H*IGH and PP-loop signatures in the protein superfamilies. Additional α-helical extensions in certain HUP class proteins are shown in orange.

the photolyases in the 4th iteration, whereas a search with a stand-alone USPA-domain protein Vng1536c from *Halobacterium* detected the ETFPs with significant E-values in the second iteration (Table I). Some searches with the USPA domains (e.g., Vng1536c) produced pairwise alignments with HIGH superfamily members comparable to the structural alignments, although these did not have significant E-values (E ~0.3–0.5). Thus, many of the similarities that were initially detected using structural comparisons could also be detected at the sequence level, supporting the hypothesis that these proteins form a distinct class of nucleotide-binding/hydrolyzing domains, which we designated the HUP class, after HIGH, USPA-like and PP-ATPase.

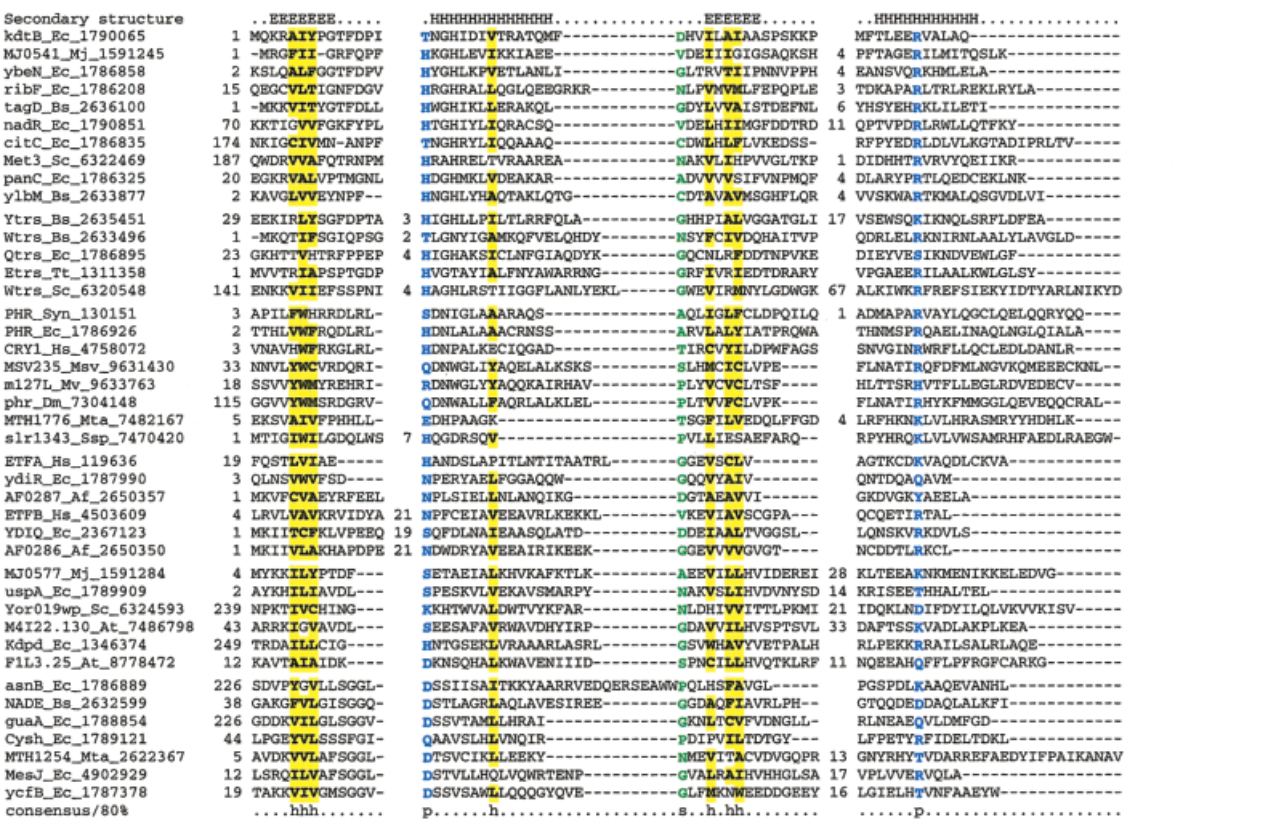
To evaluate in detail the relationships of the domains within the HUP class, we constructed a structure-based multiple sequence alignment of the HIGH superfamily, USPA domains. photolyases, ETFP, and PP-ATPases. This alignment was extended to include, in addition to the proteins whose structures were aligned, homologs identified through sequence conservation (see Fig. 2). All HUP-domain structures share a core domain of five strands, with a conserved 3–2–1–4–5 order, and four helices (Figs. 1, 3). They either end abruptly after the 5th strand (USPA) or continue into an α-helical substructure that forms a cap over the core α/β domain. These features of the core clearly distinguish the HUP domains from other domains with a "Rossmann-fold-like" geometry, such as the TOPRIM domains[48] that have a conserved core with only four strands (2–1–3–4 order), and the classic Rossmann folds of dehydrogenases and methylases that have a core of six strands.[26,27] The HUP class also differs from the classic Rossmann fold in having a well-conserved small residue, mostly glycine, demarcating the end of strand 4 (Fig. 2). The HUP domains lack the signature glycine-rich loop of the Rossmann domains, as well as other, more subtle, features of the latter, namely the conserved polar and charged residues, respectively, at the end of strand 3 and at the beginning and end of strand 4 (most frequently aspartate or asparagine), and the distinct loop with at least one conserved glycine at the base of strand 5. Furthermore, the HUP domains show a unique crossover of the extensions of strands 4 and 5 (Figs. 1, 3), and a strong "horizontal" depression of strand 3 with respect to the rest of the sheet (Fig. 3). Thus, several lines of evidence, including the results of structural and sequence similarity searches, and unique shared structural features suggest a monophyletic origin for the domains of the HUP class, to the exclusion of other "Rossmann-like" folds.

The HUP-class domains share a general similarity in their ligand-binding properties, with the principal ligand bound at the surface formed by the C-termini of the strands in the parallel β-sheet (Fig. 1). The proteins of the HIGH and PP-ATPase superfamilies hydrolyze the α–β bond in ATP,[12,41] whereas the USPA and ETFP-domains bind ATP and AMP, respectively.[43,49] The photolyases bind non-nucleotide ligands, such as methenyltetrahydrofolylpolyglutamate (MTHF) or 8-hydroxy-5-deazariboflavin (8-HDF), which act as chromophores for the light-dependent thymine–dimer reversion catalyzed by these enzymes.[50] Thus, adenosine phosphates are the preferred ligands of the HUP domains, suggesting that their common ancestor bound such a ligand, in contrast to the classic Rossmann-fold domains whose typical, and probably ancestral, ligands are flavin or nicotinamide dinucleotides or S-AdoMet.

Major Divisions Within the HUP Class

Various structural characteristics, distinct sequence features, and biochemical properties of Rossmann-like domains were used for cladistic analysis to verify the monophyly of the HUP class and to discern more precisely the relationships within it (Fig. 4; see figure legend for a description of the character states associated with each lineage and Table II for character matrix). Conventional phylogenetic analysis of proteins using sequence-based parsimony or distance methods cannot be applied for such cases, in which very distant evolutionary relationships are being explored. These conventional phylogenetic methods are generally applicable to more closely related proteins, namely orthologous sets or highly conserved paralogs. While all the domains of the HUP class can be aligned throughout their whole length (Fig. 2), there are not enough conserved amino acid positions to allow finer resolution of the relationships between different superfamilies for the HUP class by using the traditional phylogenetic methods. Cladistic analysis has been successfully applied to derive inferences on evolutionary relationships between organisms by using discrete characters, such as morphological features.[29,30,51] In principle, the same methodology can be applied to the classification of protein domains as a means of formally analyzing structural similarities be-

A


Secondary structure
kdtB_Ec_1790065
MJ0541_Mj_1591245
ybeN_Ec_1786858
ribF_Ec_1786208
tagD_Bs_2636100
nadR_Ec_1790851
citC_Ec_1786835
Met3_Sc_6322469
panC_Ec_1786325
ylbM_Bs_2633877
Ytrs_Bs_2635451
Wtrs_Bs_2633496
Qtrs_Ec_1786895
Etrs_Tt_1311358
Wtrs_Sc_6320548
PHR_Syn_130151
PHR_Ec_1786926
CRY1_Hs_4758072
MSV235_Msv_9631430
m127L_Mv_9633763
phr_Dm_7304148
MTH1776_Mta_7482167
slr1343_Ssp_7470420
ETFA_Hs_119636
ydiR_Ec_1787990
AF0287_Af_2650357
ETFB_Hs_4503609
YDIQ_Ec_2367123
AF0286_Af_2650350
MJ0577_Mj_1591284
uspA_Ec_1789909
Yor019wp_Sc_6324593
M4I22.130_At_7486798
Kdpd_Ec_1346374
F1L3.25_At_8778472
asnB_Ec_1786889
NADE_Bs_2632599
guaA_Ec_1788854
Cysh_Ec_1789121
MTH1254_Mta_2622367
MesJ_Ec_4902929
ycfB_Ec_1787378
consensus/80%


Secondary structure
kdtB_Ec_1790065
MJ0541_Mj_1591245
ybeN_Ec_1786858
ribF_Ec_1786208
tagD_Bs_2636100
nadR_Ec_1790851
citC_Ec_1786835
Met3_Sc_6322469
panC_Ec_1786325
ylbM_Bs_2633877
Ytrs_Bs_2635451
Wtrs_Bs_2633496
Qtrs_Ec_1786895
Etrs_Tt_1311358
Rtrs_Sc_6320548
PHR_Syn_130151
PHR_Ec_1786926
CRY1_Hs_4758072
MSV235_Msv_9631430
m127L_Mv_9633763
phr_Dm_7304148
MTH1776_Mta_7482167
slr1343_Ssp_7470420
ETFA_Hs_119636
ydiR_Ec_1787990
AF0287_Af_2650357
ETFB_Hs_4503609
YDIQ_Ec_2367123
AF0286_Af_2650350
MJ0577_Mj_1591284
uspA_Ec_1789909
Yor019wp_Sc_6324593
M4I22.130_At_7486798
Kdpd_Ec_1346374
F1L3.25_At_8778472
asnB_Ec_1786889
NADE_Bs_2632599
guaA_Ec_1788854
Cysh_Ec_1789121
MTH1254_Mta_2622367
MesJ_Ec_4902929
ycfB_Ec_1787378
consensus/80%
1b6ta
1F9A
1cozA
1g8fa
1ihoA
1ts1
1d2rb
1gtra
1gln
1lbs2A
1qnf
1efv
1mjha
1CT9
1NSY
1GPM
1SUR
HIGH Ntase
Aminoacyl
tRNA
synthetase
Photolyase
ETFP
PP-Loop
ATPase
USPA

Figure 2.

Figure 2. (Continued.) Multiple alignment of the *HI*GH-signature proteins, *U*spA, and *PP*-ATPase—HUP domain. The alignments for each major lineage were generated using the T_Coffee[17] program, followed by adjustments based on PSI-BLAST-generated pairwise alignments. These multiple alignments were, in turn, further aligned on the basis of structural alignments of representatives of each major lineage, for which 3D structures were available. The structural alignments were derived using FSSP.[25] The core secondary structure shared by all the families is shown above the alignment, with E representing a β-strand, and H an α-helix. The 80% consensus shown below the alignment was derived using the after amino acid classes: polar (p: KRHEDQNST,blue), hydrophobic (h: ALICVMYFW, yellow), the aliphatic subset of the former (l; ALIVMC, yellow), small (s: ACDGNPSTV, green), and its tiny subset (u: GAS, green shading). Limits of the domains are indicated by position numbers on each side of the alignment. Numbers within the alignment are inserts that are not shown. Sequences are denoted by their gene name followed by the species abbreviation, and GenBank Identifier. The different classes, along with their functions, are listed on the right. Protein Data Bank codes of structures, where they exist for the protein, are also shown on the right. Species abbreviations: Af, *Archeoglobus fulgidus*; Mj, *Methanococcus jannaschii*; Mta, *Methanothermobacter thermoautotrophicus*; Ec, *Escherichia coli*; Bs, *Bacillus subtilis*; Ssp, *Synechocystis* sp.; Syn, *Synechococcus leopoliensis*; Tt , *Thermus thermophilus*; Mv, *Myxoma virus*; Msv, *Melanoplus sanguinipes entomopoxvirus*; At, *Arabidopsis thaliana*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Sc, *Saccharomyces cerevisiae*.

tween proteins, such as those described in the present study, or those used in the classification presented in the SCOP database.[26] In addition to structural similarity patterns, sequence motifs and biochemical features of protein domains can also be scored as discrete characters in cladistic analysis.

In applying this methodology to protein domains, we defined various discrete characters (as mentioned above) that are likely to provide distinct evolutionarily derived states and scored them across all the domains (taxa) considered in this analysis (Table II). The nine distinct superfamilies of domains with a basic "Rossmann-like" topology and geometry were considered as individual taxa and 21 informative characters were scored. The characters were chosen objectively, with the only criterion being that they provided useful evolutionary information in the range appropriate for the examined relationships. Features that were conserved only within a particular superfamily or those that were universal across all the considered superfamilies were not considered because they do not provide any information to resolve inter-superfamily affinities. The availability of large amounts of sequence data for each of the superfamilies additionally helped us to select only robust characters that were not merely peculiarities of a particular lineage, for which a representative structure was available, but appeared to be ancestral features of the respective superfamily. This selection procedure also helped to minimize the effects of the diverse functional adaptations seen within each superfamily that could potentially result in certain unusual features that could obscure the actual evolutionary relationships. The resulting matrix (Table II) was then used to search for the tree requiring the shortest sequence of steps (most parsimonious tree) that fully accounts for the character states seen in the terminal taxa. The most parsimonious trees were determined with exhaustive search and branch-and-bound algorithms of the PAUP program package.[29,30] This cladistic analysis was followed by a detailed analysis of each individual lineage of the HUP class through delineation of distinct families within each major line of descent by similarity-based clustering (Table III).

Cladistic analysis using the character matrix shown in Table II generated three most parsimonious trees (see the consensus tree in Fig. 4), with the monophyly of the HUP class supported by all of them. The apparent shared derived characters (synapomorphies) that supported this clade included the structure and sequence features described above. Within the HUP class, the PP-ATPases were set aside from the rest on the basis of a poorly stabilized helix 4 (which in some of the PP-ATPases even adopts a loop-like conformation) and strand 5 that is shorter than in other HUP domains and is displaced forward with respect to the base of the parallel β-sheet (Fig. 3). PP-ATPases also have a unique sequence signature in the phosphate-binding loop between strand-1 and helix-1, with the consensus SXGXDS (Fig. 2). The HIGH superfamily is another distinct lineage within the HUP class, which is supported by the HIGH motif and the presence of a C-terminal extension with a long loop and two helices (Figs. 1, 3). The KMSKS motif or its equivalent is situated in the extension, within the loop, immediately N-terminal to the first helix. The second lysine in this

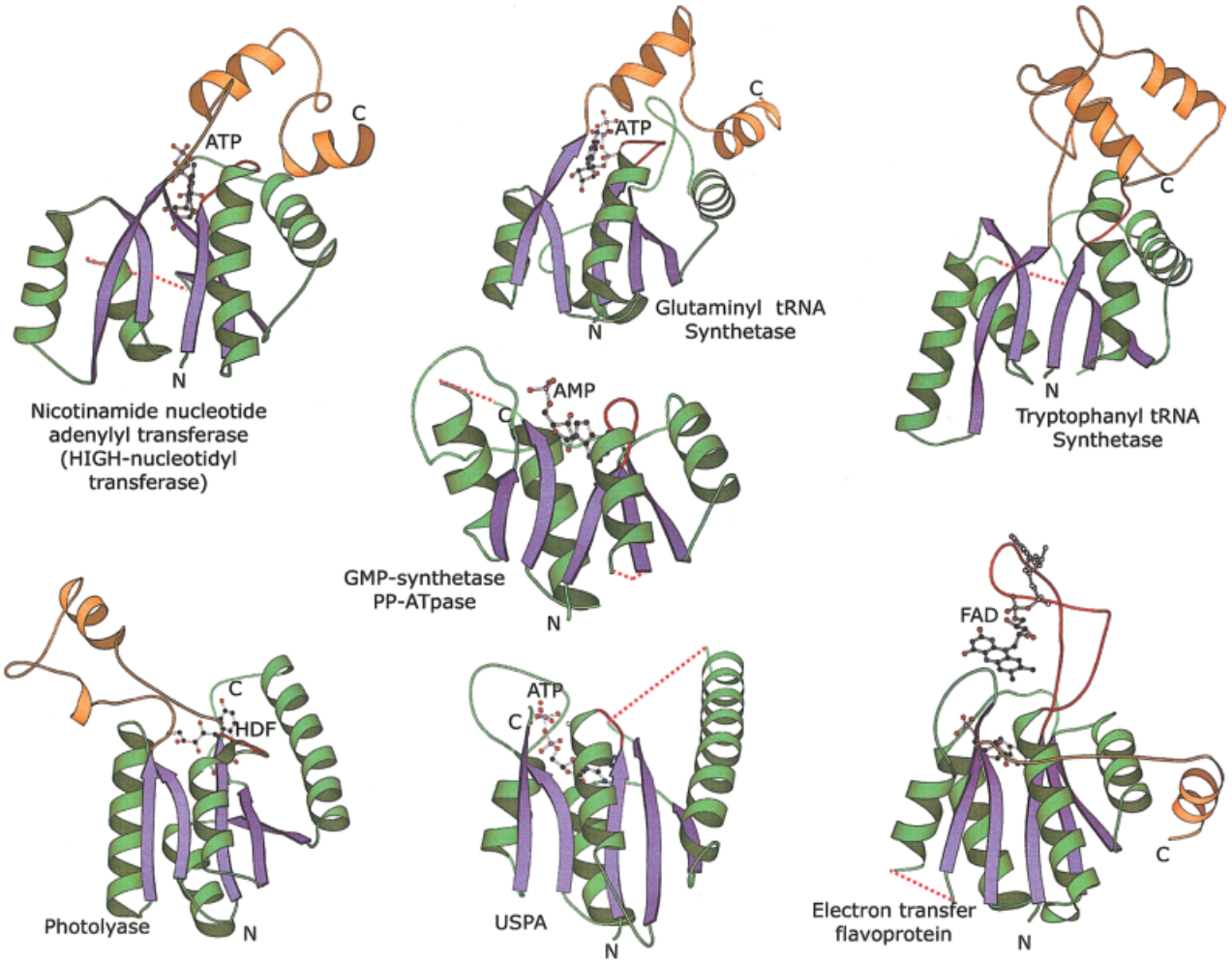


Fig. 3. HUP domain structures (*HI*GH-signature proteins, *U*spA, and *PP*-ATPase). PDB codes of the structures: nicotinamide nucleotide adenylyl transferase, 1ej2; GlnTRS, 1gtr; tryptophanyl tRNA synthetase, 1d2r; GMP synthetase, 1gpm; photolyase, 1qnf; USPA, 1MJH; ETFP, 1efv. The loop between strand 1 and helix 1, where the HIGH and PP-loop are located, is rendered in red, and the α-helical extensions present in some of these proteins, with the loop N-terminal to the helices containing the KMSKS motif in the HIGH superfamily, are rendered in orange.

motif is almost always present in the aaRS, but typically absent in the other NTases of the HIGH superfamily, indicating different modes of interaction with the NTP phosphates.

The USPA domains, photolyases, and ETFPs form another well-defined clade within the HUP class. These domains show the greatest similarity to each other in both sequence and structure searches (Table I) and also share a distinct arrangement of the last two strands of the core sheet (Fig. 3). The USPA-domain superfamily has no conserved extension and represents the minimal HUP domain that could resemble the ancestral form (Fig. 3). The photolyase superfamily has a distinct, C-terminal α-helical extension (Fig. 3) that, at least in some cases, cooperates with the HUP domain to bind the ligand.[50] The ETFP superfamily is distinguished by a large insert, comprising of a β-strand hairpin, between helix-4 and strand-5 (Fig. 3). In structural comparisons, the entire USPA-like assemblage (the USPA, photolyase, and ETFP superfamilies) shows closer similarity to the HIGH superfamily than to the PP-ATPases (Table I). The primary feature that supports this relationship is the specific shared conformation of the region including helix-4 and strand-5 in the former group (see above). Thus, the evolutionary scenario for the HUP class has the PP-ATPases branching off first, with subsequent radiation of the USPA-like group and the HIGH superfamily (Fig. 4). The domains of the HUP class, with the exception of ETFP, have undergone extensive radiation to colonize several functional niches (see below). Nevertheless, the basic features that support their monophyly and internal relationship are well preserved across proteins performing diverse functions, suggesting that the above analysis has not been affected by influences of secondary functional convergence or divergence.

Phyletic Distribution of the Principal Families and Groups of Orthologs in Different HUP Lineages

To define the ancient lineages within each superfamily of the HUP class, the phyletic distribution of the principal

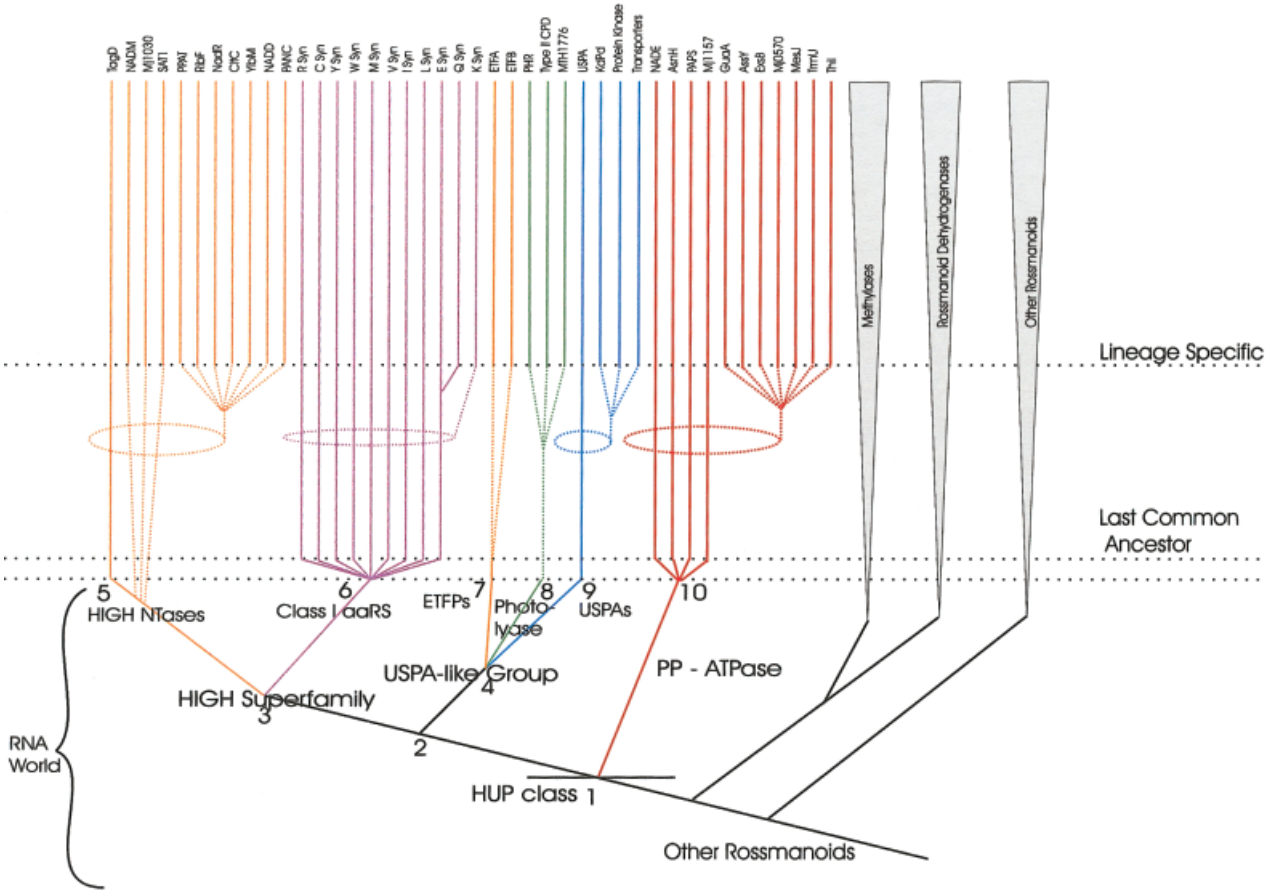


Fig. 4. Evolutionary scenario (cladogram) for the HUP domain class (*H*IGH-signature proteins, *U*spA, and *P*P-ATPase). The specific characters associated with each of the nodes as derived through the cladistic analysis are as follows. Node 1 (the HUP domain): presence of a core 5-strand sheet in the 3–2–1–4–5 order, horizontal depression of strand 3 with respect to the rest of the sheet, tendency of crossover of strand 4 and 5 or their extensions, a sequence motif corresponding to strand 4, with a conserved small residue at its C-terminus, adenosine phosphate ligands. Node 2: configuration of the region between the last helix and strand of the HUP domain, strand 4 and 5 hydrogen-bonded through most of their lengths. Node 3: HIGH motif in the loop between strand 1 and helix 1, bihelical extension C-terminal to the core HUP domain, with a secondary role in nucleotide-binding, Node 4: loop between strand 3 and helix 4 tends to face outward, helix 4 tends to be placed behind the central sheet, no strongly conserved motif between strand-1 and helix-1, possible loss of ability to hydrolyze α–β bond in ATP. Node 5: nucleotidyl transferase activity, two small residues in the KMSKS loop. Node 6: aaRS activity, classical KMSKS motif. Node 7: insertion of a β-hairpin between helix 4 and strand 5. Node 8: fusion with large, α-helical C-terminal domain. Node 9: distinct loop between strand 4 and helix 4. Node 10: SXGXDS motif between strand-1 and helix-1, conserved helix N-terminal to strand 1. The conserved families (see Table III) and their probable temporal points of origin are shown for each of the six major lineages. Dotted ellipses and lines leading to a particular lineage indicate uncertainty regarding its emergence.

families and orthologous groups identified within each family was examined (Table III). As documented previously, nine class I aaRS are traceable to the LUCA.[10,11] The NTases of the HIGH superfamily are typically involved in metabolic functions, such as biosynthesis of cofactors (pantothenate, nicotinamide, and riboflavin) or assimilation of sulfur for cysteine synthesis. Involvement in these central metabolic functions suggests that at least some of these enzymes might be of ancient provenance. However, the phyletic pattern of all orthologous groups of HIGH NTases is patchy (Table III), which indicates loss of these proteins in several lineages of heterotrophs and displacement of the ancestral versions after horizontal gene transfer.[52] A comparison of the families widely represented among organisms with extensive metabolic capabilities (including autotrophs) suggests that at least the TagD-Pct1-like family, the NadD family, the MJ1030 family, and the sulfate adenylyltransferase family were probably represented in the LUCA. Most of these enzymes are small proteins that consist almost completely of the core HUP domain, with a bihelical extension equivalent to the KMSKS-containing extension of class I aaRS. They lack the specialized accessory domains or inserts found in aaRS[10] and probably resemble the ancestral state of the HIGH superfamily.

The photolyases, which function in DNA repair or as photoreceptors,[50] form three distinct but closely related families, including a new family identified during the present analysis (Table III). Each of these families shows sporadic distribution in bacteria, archaea, and eukaryotes (Table III). The photo-dependency of these proteins might explain their absence in organisms that are not exposed to light. However, the lack of diversity and limited phyletic distribution suggests that photolyases evolved at a relatively late stage of evolution and spread across the microbial world through horizontal gene transfers.

TABLE II. Character Matrix Used for Cladistic Analysis of the Proteins With Rossmann-like Topology and Geometry

	Characters[a]																				
Protein families	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
HIGH NTases	1	1	1	1	2	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0
aaRS	1	1	1	1	2	1	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0
UspA	1	1	1	1	2	1	1	1	?	0	0	0	0	0	0	0	0	0	0	1	1
Photolyase	1	1	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0	1	1
ETFP	1	1	1	1	2	1	1	1	0	1	0	0	0	0	1	0	0	0	0	1	1
PP-ATPases	1	1	1	1	2	1	0	2	1	0	0	1	0	0	0	0	0	0	0	0	0
Rossmann methylases	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Rossmann oxidoreductases	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Toprim domains	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

aaRS, aminoacyl-tRNA synthetase.

[a]1, basic sheet topology (0: 2–1–3–4, 1: 3–2–1–4–5, 2: 3–2–1–4–5–6); 2, crossover state of strand 4 and strand 5; 3, angle of strand 3 with respect to the sheet; 4, presence of conserved motif associated with strand 4; 5, ligand-binding properties (0: generic nucleotide or nucleic acid; 1: SAM or dinucleotide; 2: adenosine phosphates); 6, curve at C-terminal end of strand 4; 7, state of helix 4 (0: very irregular helix or poorly structured helix; 1: regular), 8, state of strand 5 (0: ancestral Rossmann-like state with no particular alignment, 1: N-terminal strand of strand aligned with the base of the sheet and hydrogen bond throughout, 2: strand 5 weakly formed and "raised" with respect to the base of sheet and hydrogen bonds only part of the length), 9, tendency to cleave phosphate bonds in nucleotide; 10, tendency to lead to α-helical extensions; 11, HIGH motif; 12, PP-motif; 13, KMSKS-like motif; 14, distal K of the KMSKS motif; 15, Insertion of a β-hairpin between helix 4 and strand 5; 16, glycine/small amino acid-rich loop preceding strand 5; 17, motif with a polar position (typical N or D) at the termini of strand 4; 18, motif with a polar residues at the end of strand 3; 19, toprim-specific motifs (conserved E and DXD); 20, structure of junction between strand 3 and helix 4 (0: loop in a "depressed" conformation; 1: characteristic exposed loop); 21, position of helix 4 with respect to central sheet (0: "laterally placed," 1—placed more or less behind the sheet).

The ETFPs are intermediaries in electron transfer from diverse dehydrogenases to the main respiratory redox chain.[53,54] The two distantly related subunits, ETFA and ETFB, are widely, but not universally, represented in bacteria, eukaryotic mitochondria, and archaea (Table III). ETFA contains a unique long insert in the loop between strand-1 and helix-1 that developed into a FAD-binding site specific to this subunit (Fig. 3). The most parsimonious evolutionary inference that can be drawn from the phyletic pattern of the ETFP proteins is a duplication of an ancestral ETFP giving rise to ETFA and ETFB before the radiation of the bacterial and archaeo-eukaryotic lineages from the LUCA. However, the current absence of a nonmitochondrial form of the ETFPs in eukaryotes, along with its absence in some archaea, could also support a more complex, alternative history, with an origin at the base of the bacterial lineage followed by an early horizontal transfer to the archaea.

The USPA superfamily is widely represented in archaea and bacteria; among the eukaryotes with extensive genomic sequence availability, these proteins were detected in fungi and plants (Table III).[55] The most common architecture seen in the USPA superfamily is the USPA-only form, with a single or two USPA domains. Distinct multidomain architectures of USPA-containing proteins, such as fusions to serine/threonine kinases, histidine kinases, or cation antiporters, are lineage-specific (Table III). This phyletic distribution suggests that the stand-alone USPA domains were represented in the LUCA. Two USPA domains, *E. coli* UspA and MJ0577, have been shown to interact with ATP.[43,56] The proteins containing the USPA domain in a stand-alone or duplicated form might function as ATP-dependent switches undergoing a conformational change upon ATP binding and relaying the signal to their interaction partners. Such a model appears particularly plausible given the fusions to different types of protein kinase domains: a histidine kinase in the bacterial KdpD proteins and serine-threonine kinases in several plant proteins. It seems likely that, at least in these proteins, ligand-binding by the USPA domain results in a conformational change that affects the kinase activity. Several of the plant proteins, in addition to the kinase-fusion, also show a fusion to the U-box domain, which functions as an ubiquitin ligase,[57] suggesting that the USPA domains might also regulate ubiquitin-mediated signaling in plants. Additionally, like many other small-molecule-binding domains, the USPA domains probably also regulate ion transport across the membrane in prokaryotes.[55] Thus, the available information and domain architectures seem to point to an ancient role for the USPA domain, ATP-dependent signaling.[44,55]

The PP-ATPases are a large superfamily of ATP-using enzymes,[41,58] which consist of at least four families with a widespread or nearly universal distribution in the three superkingdoms of life. These ancient families include glutamine-dependent asparagine synthetases, NAD synthetases, phosphoadenosine phosphosulfate reductases, and MJ1157-like, predicted, tRNA-thiouridine synthases (Table III). Several other families, such as GMP synthetases, ThiI-like thiouridine synthases, argininosuccinate synthetases, and the ExsB/YbaX-like PP-ATPases are widespread in archaea and bacteria, with the eukaryotic members, if any, apparently acquired by horizontal transfer from bacteria, probably through the mitochondrial precursor. Thus, the LUCA can be inferred to have encoded at least four, and possibly more, distinct members of the PP-ATPase superfamily, with a biochemical diversity spanning the

TABLE III. Phyletic Distribution of HUP-Domain Protein Families

Families[a]	Phyletic distribution[b]	Comments
	HIGH superfamily	
HIGH nucleotidyl transferases		
Glycerol-3-phosphate cytidyltransferase (tagD) 2636100_Bs	AB(Aae,Bs,Nm,Ec,Hi, Pa,Hp,Cj)E[c]	TagD is involved in polyglycerol phosphate teichoic acid biosynthesis; includes RFAE_Ec (N-terminal ADP Heptose/Ribokinase, Ec,Hi,Pa,Hp,Cj), TagD-like bacterial proteins (Aae,Bs,Nm), Pct1_Sc and Muq1_Sc (eukaryotes have two Muq1-like nucleotidyl transferases)
Riboflavin biosynthesis protein RIBF 1786208_Ec	B,At	N-Terminal riboflavin kinase fusion; involved in riboflavin biosynthesis
Nicotinate nucleotide adenylyltransferase NADD 1786858_Ec.	BE	Catalyzes reversible adenylylation of nicotinate mononucleotide (the yeast ortholog Ylr328w is called NADM); *Mycoplasma* have a C-terminal HD hydrolase
Nicotinamide nucleotide adenylyltransferase NADM 1591245_Mj	AB(Dr,Ssp)	Catalyzes the synthesis of NAD^+ and nicotinic acid adenine dinucleotide
Phosphopantetheine adenylyltransferase PPAT 1790065_Ec	B	Catalyzes reversible transfer of an adenylyl group from ATP to 4′ phosphopantetheine yielding dephospho-COA and pyrophosphate
MJ1030 1591685_Mj	AE	Uncharacterized nucleotidyl transferases; Dm and Ce have a C-terminal dephospho-CoA-like P-loop kinase
Transcription regulator NadR 1790851_Ec	B(Ec,Hi,Mt)	NadR is bifunctional; at low NAD levels, it permits NMN transport into the cell; at high NAD levels, it represses NADA, NADB, and PNCB genes; C-terminal ABC–ATPase fusion
Citrate–lyase ligase CitC 1786835_Ec	B(Ec,Hi)	Involved in acetylation of prosthetic group [2-(5″-phosphoribosyl)-3′-dephospho-CoA] of the γ-subunit of citrate lyase; *N*-terminal *N*-acetyl transferase fusion
ylbM 2633877_Bs	A(Mj,Mta)B(Bs,Tm)	Uncharacterized nucleotidyl transferases
Sulfate adenylyltransferase SAT1 2633932_Bs	A(Ap,Af,Pab)[BE]	Involved in adenylylsulfate synthesis in cysteine biosynthesis pathway; fusion to P-loop kinase in animals and Aae
Pantoate-β-alanine ligase PanC 1786325_Ec	BE(At,Sc,Sp)	Involved in pantothenate biosynthesis
	Class I Aminoacyl-tRNA synthetase	
Arginyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Cysteinyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Lysyl-tRNA synthetase	AB(Tp,Bb,Rp)	Charging of tRNA with cognate amino acid
Glutamyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Glutaminyl-tRNA synthetase	BE	Charging of tRNA with cognate amino acid
Tyrosyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Tryptophanyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Methionyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Valyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Isoleucyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
Leucyl-tRNA synthetase	ABE[c]	Charging of tRNA with cognate amino acid
	USPA-like group	
USPA (only major orthologous groups listed)		
USPA 1789909_Ec	ABE(At,Sc,Sp)[c]	A nucleotide-binding domain widely represented in prokaryotes and plants; implicated in signal transduction and transport regulation; small expansion in plants
KdpD (histidine kinase) 1786912_Ec	B(Ec,Pa,Mt,Ssp,Dr)	Involved in signal transduction; fusion to histidine kinase, a P-loop ATPase, and a GAF domain; Ssp and Dr do not have histidine kinase and GAF
Protein kinase 7488259_At	At only	C-terminal STY kinase with or without a modified ring-finger domain/U-BOX
Cationic amino acid transporter 2648945_Af	A(Af,Hsp)B(Ssp)	Other than the N-terminal cation transporter fusion in archaea and Ssp, there is an N-terminal chloride ion transporter, followed by 2 CBS fusion in 1653666_Ssp
Photolyase		
Deoxyribodipyrimidine photolyase PHR 1786926_Ec	BE, Hsp	Catalyzes light-dependent monomerization of UV-induced cyclobutyl pyrimidine dimers that are formed between adjacent bases on the same DNA strand

TABLE III. (Continued)

Families[a]	Phyletic distribution[b]	Comments
Photolyase (Continued)		
Type II CPD photolyase 7304148_Dm	DNA viruses, Dm,At,Mta	Type II photolyase
MTH1776 7482167_Mta	A(Mta,Hsp)B(Ssp, Vc, Cc, Scoe)	A previously uncharacterized photolyase family detected as a part of this study.
ETFP		
Electron transfer flavoprotein ETFA 119636_Hs	A[BE]	ETFP acts as an electron acceptor for several dehydrogenases, including five acyl-CoA dehydrogenases, glutaryl-CoA, and sarcosine dehydrogenase; the electrons are transferred to the main mitochondrial respiratory chain via ETF dehydrogenase; ETFA/FixB is the alpha subunit
Electron transfer flavoprotein ETFB 4503609_Hs	A[BE]	ETFB/FixA is the β-subunit; 10639599_Ta has a FixA and FixB fusion in addition to stand-alone ETFA and ETFB proteins
	PP-LOOP superfamily	
PP-loop ATPases (only major orthologous groups listed)		
NAD^+ synthetase NadE 1788036_Ec	ABE[c]	NH_3^+-dependent NAD^+ synthase; N-terminal amidohydrolase in eukaryotes and some bacteria
Glutamine-dependent asparagine synthetase AsnH 2636538_Bs	A[BE]E[c]	Glutamine-dependent asparagine synthase; one eukaryotic set is a horizontal transfer from bacteria, the other an ancient conserved group
Phosphoadenosine phosphosulfate (PAPS) reductase CysH 1789121_Ec	ABE[c]	Reduction of sulfate into sulfite; in bacteria and eukaryotes, involved in cysteine biosynthesis (two subunits, CysH and CysD); in archaea, predicted to be involved in RNA modification (fusion to PUA and Zn ribbon RNA binding domains in Mj and Ph) (unpublished observation VA, EVK, LA).
GMP synthase GuaA 1788854_Ec	A[BE]	Glutamine-dependent GMP synthase; N-terminal amidohydrolase fusion; a related archaeal group (MJ0830-like) is predicted to be involved in RNA modification (V. Anantharaman, E.V. Koonin, and L. Aravind, unpublished observation)
Argininosuccinate synthase AssY 1789563_Ec	A[BE]	Involved in arginine biosynthesis
Succinoglycan biosynthesis regulator (exsB) 1786648_Ec	AB	Some members function in succinoglycan biosynthesis
MJ0570 1591277_Mj	AE	Uncharacterized PP-loop ATPase; eukaryotes have C-terminal fusion to two YabJ-like chorismutase-fold domain
MesJ 1786386_Ec	BE (At,Sp)	Involved in RNA modification; Dr has a C-terminal deaminase fusion also predicted to be involved in RNA modification (V. Anantharaman, E.V. Koonin, and L. Aravind, unpublished observation)
tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase TrmU 1787378_Ec	BE	RNA modification—thiouridylate formation
ThiI 1786625_Ec	AB	Thiouridylation; N-terminal RNA-binding THUMP domain and a C-terminal rhodanese-like domain involved in sulfur transfer
MJ1157 1592323_Mj	ABE[c]	Predicted to be involved in RNA modification as a thiouridine synthetase. N- and C-terminal fusions to Zn ribbon (V. Anantharaman, E.V. Koonin, and L. Aravind, unpublished observation)
MJ0690 1591405_Mj	A (Mj,Mta,Ph,Hsp)	Predicted to be involved in RNA modification as a thiouridine synthetase (found in an operon with ribosomal proteins) (V. Anantharaman, E.V. Koonin, and L. Aravind, unpublished observation)

aaRS, aminoacyl tRNA synthetase.

[a]For each family, except for the aaRS, a gene identification (GI) number and the gene name of a representative sequence are given to ensure unambiguous identification in databases.

[b]A, archaea; B, bacteria; E, eukaryotes. In cases where the eukaryotic homologs appeared to be of bacterial (primarily mitochondrial) origin, [BE] is indicated in brackets. When the distribution of the respective orthologous group in the given superkingdom is sporadic, the species are indicated in parentheses.

[c]Ancient conserved groups that show nearly universal distribution, largely attributable to vertical descent, and by inference, are thought to originate from the LUCA. Species abbreviations are as shown in Figure 3 and Hsp, *Halobacterium* sp.; Ph, *Pyrococcus horikoshii;* Pab, *Pyrococcus abyssi;* Ta, *Thermoplasma acidophilum;* Aae, *Aquifex aeolicus;* Bb, *Borrelia burgdorferi;* Cc, *Caulobacter crescentus;* Cj, *Campylobacter jejuni;* Dr, *Deinococcus radiodurans;* Hi, *Haemophilus influenzae;* Hp, *Helicobacter pylori;* Nm, *Neisseria meningitidis;* Mt, *Mycobacterium tuberculosis;* Pa, *Pseudomonas aeruginosa;* Rp, *Rickettsia prowazekii;* Scoe, *Streptomyces coelicolor;* Tm, *Thermotoga maritima;* Vc, *Vibrio cholerae;* Sp, *Schizosaccharomyces pombe.*

range from metabolic reactions and cofactor biosynthesis to tRNA modification crucial for translation.

Implications of the Evolutionary Relationships of the HUP Domains for Reconstruction of Ancient Biological Systems

The inferred evolutionary relationships within the HUP-domain class have notable implications for reconstruction of ancient biochemical systems that preceded the LUCA. Since most of the HUP-domain proteins bind adenosine phosphates, it appears most likely that the ancestral HUP domain also did so, primarily via the loop between strand-1 and helix-1 (Figs. 1,3). Furthermore, the ability to hydrolyze the α–β bond of ATP that is seen in both the PP-ATPase and HIGH branches (Fig. 4) suggests that the ancestral domain also possessed an α–β bond ATPase activity, perhaps an inefficient one. During subsequent evolution, this ancestral activity independently diversified in different lineages through acquisition of more specific sequence features in the nucleotide-binding loop and structural elaborations. Such lineage-specific developments included the evolution of the SXGXDS signature in the nucleotide-binding loop of the PP-ATPases and of the corresponding HIGH signature, and the helical extension that harbors the KMSKS motif, in the HIGH branch. The HUP domains of the USPA–like assemblage apparently retained only the ancestral nucleotide-binding properties, whereas the phosphohydrolase activity was probably lost in most if not all members of this group.

Examination of the phyletic patterns of the major families within each superfamily of HUP domains indicates that 15–18 members of the HUP class, including 9 aaRS, were already present in the LUCA (Table III and Fig. 4). An inevitable corollary of this conclusion is that several rounds of duplication and divergence of the HUP domains antedate the LUCA. Thus, it seems most likely that, during this phase of early evolution, the ancestral undifferentiated HUP proteins performed multiple nucleotide-dependent functions that are currently performed by their distinct descendents. The most notable aspect of the evolution of the HUP class is that multiple HUP-domain-containing proteins, namely aaRS of 9 specificities and tRNA thiouridine synthases (members of the PP-ATPase superfamily), are indispensable for translation in its modern form. As considerable diversification of the HUP domains preceded the radiation of the aaRS (Fig. 4), one has to conclude that ancestors of modern proteins with well-defined structural and sequence features, such as the ancestral PP-ATPase, HIGH, and USPA domains, antedate the modern translation apparatus.

The common ancestor of the HUP class, an ATP-binding domain with generic ATP-pyrophosphatase and/or nucleotidyltransferase activity, could not perform the multiple, specific functions assumed by its descendants. In particular, it was impossible for this generic ATP-hydrolase to catalyze reactions that require highly specific molecular interactions, such as tRNA aminoacylation or thiouridylation of specific bases in tRNA. The most plausible solution to this conundrum lies in the well-known hypothesis that, in the primitive translation system, many functions currently performed by proteins relied on RNA molecules, including the ancestors of rRNA and tRNA.[59] The above reconstruction shows that, at this early stage of evolution, proteins already had catalytic and ligand-binding capabilities, suggesting that the RNAs were mainly responsible for the specificity. In particular, the ancestor of the HUP class could interact nonspecifically with proto-tRNAs facilitating aminoacylation, whereas the specificity of these reactions came from the same tRNA, or possibly other, accessory RNA molecules. The recent demonstration of specific self-aminoacylation catalyzed by an RNA molecule[60] is compatible with this scenario. The same ancestral protein could have functioned, in cooperation with other RNAs, to facilitate other reactions that require ATP hydrolysis or nucleotide transfer, such as RNA modification and cofactor biosynthesis. After the first duplication leading to the radiation of the two main branches of the HUP class (Fig. 4), the ancestor of the HIGH-USPA lineage could have functioned as a generic nucleotidyltransferase involved in translation and cofactor biosynthesis, whereas the ancestor of the PP-loop ATPases might have become a generic ATPase involved in RNA modification and some metabolic functions (Table III).

Evolutionary analysis of several other classes of proteins, such as class II aaRS,[13] GTPases, nucleic acid polymerases, and RNA-binding domains in RNA-modification enzymes and ribosomal proteins, similarly indicates that substantial diversification of proteins within each of these classes has already occurred before the emergence of universal and currently indispensable components of the translation system (D. Leipe, V. Anantharaman, E.V. Koonin, and L. Aravind, unpublished observations). These proteins, for which ancient diversification could be specifically demonstrated, are involved in all aspects of RNA metabolism and translation ranging from modifications of tRNAs and rRNA, such as pseudouridylation and methylation, to translation initiation factors and core protein components of the ribosomes. Thus, it appears that a fairly efficient and accurate translation system existed with the assistance from proteins that had only generic biochemical properties and were ancestral to multiple proteins with distinct, essential functions in all modern cells. From a complementary perspective, the evolutionary reconstruction presented here and similar analyses for several other protein classes show that much of the known diversity of protein domains has evolved within the primordial "RNA world."

CONCLUSIONS

Sequence and structure comparisons described here point to the monophyly of class I aaRS, HIGH NTases, USPA domains, photolyases, ETFPs, and PP-ATPases, which together comprise a distinct class of α/β domains designated the HUP domain. Several lines of evidence support the distinctness of the HUP class with respect to all other three-layered α/β folds with generic "Rossmann-like" topology and geometry. Cladistic analysis of the HUP class, which scored patterns of structural and sequence

similarity, identified three major evolutionary lineages, the previously well-known HIGH and PP-ATPase superfamilies, and the previously unrecognized USPA-like group, which includes USPA domains, ETFP, and photolyases. Examination of the patterns of phyletic distribution of distinct families within these three major lineages suggests that LUCA encoded 15-18 distinct members of the HUP class, which points to extensive pre-LUCA evolution, during which the common ancestor of the class I aaRS arose relatively late. Thus, substantial diversification of protein domains within the HUP class of ATPases and nucleotidyltransferases occurred before the modern-type, protein-based translation machinery was established or, in other words, still in the RNA world.

NOTE ADDED IN PROOF

After this paper was accepted for publication, the crystal structure of the UspA protein from *Haemophilus influenzae* has become available (Sousa MC, McKay DB. Structure of the universal stress protein of Haemophilus influenzae. Structure (Camb) 2001 Dec; 9(12):1135–41). Analysis of this structure indicates that, in contrast to MJ0577, UspA does not contain an ATP-binding pocket and does not seem to bind ATP. Thus, the evolution of a distinct family of the USPA protein superfamily might have gone even further than previously suspected, resulting in the loss of not only ATPase activity, but also the nucleotide-binding capacity.

REFERENCES

1. Koonin EV, Tatusov RL, Rudd KE. Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. Proc Natl Acad Sci USA 1995;92:11921–11925.
2. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 1996;93:10268–10273.
3. Woese C. The universal ancestor. Proc Natl Acad Sci USA 1998;95:6854–6859.
4. Woese CR. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA 2000;97:8392–8396.
5. Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA 1998;95: 6239–6244.
6. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22–28.
7. Edgell DR, Doolittle WF. Archaea and the origin(s) of DNA replication proteins. Cell 1997;89:995–998.
8. Forterre P. Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins. Mol Microbiol 1999;33: 457–465.
9. Leipe DD, Aravind L, Koonin EV. Did DNA replication evolve twice independently? Nucleic Acids Res 1999;27:3389–3401.
10. Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 1999;9:689–710.
11. Woese CR, Olsen GJ, Ibba M, Soll D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 2000;64:202–236.
12. Bork P, Holm L, Koonin EV, Sander C. The cytidylyltransferase superfamily: identification of the nucleotide-binding site and fold prediction. Proteins 1995;22:259–266.
13. Artymiuk PJ, Rice DW, Poirrette AR, Willet P. A tale of two synthetases. Nat Struct Biol 1994;1:758–760.
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
15. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
16. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999;15:1000–1011.
17. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
18. Schuler GD, Altschul SF, Lipman DJ. A workbench for multiple alignment construction and analysis. Proteins 1991;9:180–190.
19. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. Extracting protein alignment models from the sequence database. Nucleic Acids Res 1997;25:1665–1677.
20. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
21. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.
22. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pacif Symp Biocomput 2000:119–130.
23. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res 2001;29:55–57.
24. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
25. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
26. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
27. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
28. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
29. Swofford DL. Phylogenetic Analysis Using Parsimony (PAUP), version 3.0s. Champaign, Illinois, USA: Illinois Natural History Survey; 1991.
30. Swofford DL. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland, MA: Sinauer Associates; 2000.
31. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
32. Kraulis PJ. Molscript. J Appl Crystallogr 1991;24:946–950.
33. Cusack S. Aminoacyl-tRNA synthetases. Curr Opin Struct Biol 1997;7:881–889.
34. Burbaum JJ, Schimmel P. Structural relationships and the classification of aminoacyl-tRNA synthetases. J Biol Chem 1991;266: 16965–16968.
35. Nagel GM, Doolittle RF. Evolution and relatedness in two aminoacyl-tRNA synthetase families. Proc Natl Acad Sci USA 1991;88: 8121–8125.
36. Chan KW, Koeppe RE II. Role of lysine-195 in the KMSKS sequence of *E. coli* tryptophanyl-tRNA synthetase. FEBS Lett 1995;363:33–36.
37. Xin Y, Li W, First EA. The "KMSKS" motif in tyrosyl-tRNA synthetase participates in the initial binding of tRNA(Tyr). Biochemistry 2000;39:340–347.
38. von Delft F, Lewendon A, Dhanaraj V, Blundell TL, Abell C, Smith AG. The crystal structure of *E. coli* pantothenate synthetase confirms it as a member of the cytidylyltransferase superfamily. Structure 2001;9:439–450.
39. Izard T, Geerlof A. The crystal structure of a novel bacterial adenylyltransferase reveals half of sites reactivity. EMBO J 1999;18:2021–2030.
40. D'Angelo I, Raffaelli N, Dabusti V, Lorenzi T, Magni G, Rizzi M. Structure of nicotinamide mononucleotide adenylyltransferase: a key enzyme in NAD^+ biosynthesis. Struct Fold Des 2000;8:993–1004.
41. Bork P, Koonin EV. A P-loop-like motif in a widespread ATP

pyrophosphatase domain: implications for the evolution of sequence motifs and enzyme activity. Proteins 1994;20:347–355.
42. Wang Y, Addess KJ, Geer L, Madej T, Marchler-Bauer A, Zimmerman D, Bryant SH. MMDB: 3D structure data in Entrez. Nucleic Acids Res 2000;28:243–245.
43. Zarembinski TI, Hung LW, Mueller-Dieckmann HJ, Kim KK, Yokota H, Kim R, Kim SH. Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc Natl Acad Sci USA 1998;95:15189–15193.
44. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 1999;9:608–628.
45. Aravind L, Koonin EV. Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. J Mol Biol 1999;287:1023–1040.
46. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
47. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998;23:444–447.
48. Aravind L, Leipe DD, Koonin EV. Toprim—a conserved catalytic domain in type IA and II topoisomerases, DNAG-type primases, OLD family nucleases and RecR proteins. Nucleic Acids Res 1998;26:4205–4213.
49. Sato K, Nishina Y, Shiga K. Electron-transferring flavoprotein has an AMP-binding site in addition to the FAD-binding site. J Biochem (Tokyo) 1993;114:215–222.
50. Deisenhofer J. DNA photolyases and cryptochromes. Mutat Res 2000;460:143–149.
51. Harvey HP, Pagel MD. The comparative method in evolutionary biology. Oxford: Oxford University Press; 1991. 239 p.
52. Galperin MY, Walker DR, Koonin EV. Analogous enzymes: independent inventions in enzyme evolution. Genome Res 1998;8:779–790.
53. Roberts DL, Salazar D, Fulmer JP, Frerman FE, Kim JJ. Crystal structure of Paracoccus denitrificans electron transfer flavoprotein: structural and electrostatic analysis of a conserved flavin binding domain. Biochemistry 1999;38:1977–1989.
54. Roberts DL, Frerman FE, Kim JJ. Three-dimensional structure of human electron transfer flavoprotein to 2.1-A resolution. Proc Natl Acad Sci USA 1996;93:14355–14360.
55. Anantharaman V, Koonin EV, Aravind L. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. J Mol Biol 2001;307:1271–1292.
56. Freestone P, Nystrom T, Trinei M, Norris V. The universal stress protein, UspA, of Escherichia coli is phosphorylated in response to stasis. J Mol Biol 1997;274:318–324.
57. Pringa E, Martinez-Noel G, Muller U, Harbers K. Interaction of the ring finger-related U-box motif of a nuclear dot protein with ubiquitin-conjugating enzymes. J Biol Chem 2001;276:19617–19623.
58. Tesmer JJ, Klem TJ, Deras ML, Davisson VJ, Smith JL. The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. Nat Struct Biol 1996;3:74–86.
59. Crick FH. The origin of the genetic code. J Mol Biol 1968;38:367–379.
60. Illangasekare M, Yarus M. Specific, rapid synthesis of Phe-RNA by RNA. Proc Natl Acad Sci USA 1999;96:5470–5475.