Nucleic Acids Research, 2002, Vol. 30, No. 23 ***5229–5243***

Detection of novel members, structure–function analysis and evolutionary classification of the 2H phosphoesterase superfamily

Raja Mazumder, Lakshminarayan M. Iyer, Sona Vasudevan and L. Aravind*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received July 12, 2002; Revised and Accepted September 30, 2002

ABSTRACT

2′,3′ Cyclic nucleotide phosphodiesterases are enzymes that catalyze at least two distinct steps in the splicing of tRNA introns in eukaryotes. Recently, the biochemistry and structure of these enzymes, from yeast and the plant *Arabidopsis thaliana*, have been extensively studied. They were found to share a common active site, characterized by two conserved histidines, with the bacterial tRNA-ligating enzyme LigT and the vertebrate myelin-associated 2′,3′ phosphodiesterases. Using sensitive sequence profile analysis methods, we show that these enzymes define a large superfamily of predicted phosphoesterases with two conserved histidines (hence 2H phosphoesterase superfamily). We identify several new families of 2H phosphoesterases and present a complete evolutionary classification of this superfamily. We also carry out a structure–function analysis of these proteins and present evidence for diverse interactions for different families, within this superfamily, with RNA substrates and protein partners. In particular, we show that eukaryotes contain two ancient families of these proteins that might be involved in RNA processing, transcriptional co-activation and post-transcriptional gene silencing. Another eukaryotic family restricted to vertebrates and insects is combined with UBA and SH3 domains suggesting a role in signal transduction. We detect these phosphoesterase modules in polyproteins of certain retroviruses, rotaviruses and coronaviruses, where they could function in capping and processing of viral RNAs. Furthermore, we present evidence for multiple families of 2H phosphoesterases in bacteria, which might be involved in the processing of small molecules with the 2′,3′ cyclic phosphoester linkages. The evolutionary analysis suggests that the 2H domain emerged through a duplication of a simple structural unit containing a single catalytic histidine prior to the last common ancestor of all life forms. Initially, this domain appears to have been involved in RNA processing and it appears to have been recruited to perform various other functions in later stages of evolution.

INTRODUCTION

Enzymes possessing phosphoesterase activities have emerged in entirely different structural scaffolds on several occasions. Several distinct superfamilies of these enzymes, such as the metallo-β-lactamases, the HD hydrolases, the DHH hydrolases, the calcineurin-fold phosphoesterases, the haloacid dehalogenase superfamily and the HKD superfamily have been identified to date (1–7). Each of these superfamilies is distinguished by a specific constellation of conserved charged residues that define the active site. Some of these superfamilies, such as the HD, metallo-β-lactamase, calcineurin-like and DHH, chelate metal ions, which activate water for hydrolysis of the phosphoester linkage, on very distinct structural scaffolds. Others, such as the haloacid dehalogenase superfamily, hydrolyze the phosphoester linkages through covalent enzyme-phosphate intermediates. The majority of these superfamilies appear to have at least one single representative traceable to the last universal common ancestor (LUCA) of all known life forms and appear to have diverged to occupy numerous other functional niches in the later stages of the evolution of life. Analysis of these phosphoesterase superfamilies has thrown considerable light on the ontology, structure and function of the principal biological systems in which they participate, namely nucleic acid metabolism and signal transduction (8,9).

One superfamily of phosphoesterases, the study whose evolutionary diversification may potentially throw light on multiple biological functions related to nucleic acid metabolism and cellular signaling, is typified by the 2′,3′ cyclic phosphodiesterases (CPDases) (10,11). These enzymes were initially described in yeasts, in studies on the joining of tRNA half molecules generated during the excision of introns by the tRNA splicing endonuclease (12). Action of this endonuclease

*To whom correspondence should be addressed. Tel: +1 301 594 2445; Fax: +1 301 480 9241; Email aravind@ncbi.nlm.nih.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

generates two fragments; a 5′ fragment with a 2′,3′ cyclic phosphate terminus and a 3′ fragment with a free 5′OH. These fragments are joined in a three-step process that includes the processing of a 2′,3′ cyclic phosphate, the phosphorylation of the 5′OH, and the ligation of the two fragments (12,13). All three activities were traced to the yeast tRNA ligase Trl1p, a multidomain protein found only in fungi, with an N-terminal T4-like RNA ligase domain, a polynucleotide kinase domain and a C-terminal CPDase domain (14–16). The CPDase activity of fungal Trl1p involves the cleaving of the 2′,3′ cyclic phosphate to generate a 2′ phosphate and a 3′OH on the 5′ exon junction of the tRNA fragment. The ligation of the exons, respectively, with the 2′ and 5′ phosphate, proceeds via addition of an adenosine phosphate to the 5′ phosphate intermediate and results in the generation of a tRNA with a 2′ phosphate at the junction. The removal of the 2′ phosphate from the ligated tRNA is carried out by transferring it to NAD to generate 1″,2″-cyclic phosphate (Appr>p) and nicotinamide. Subsequently, a second enzyme typified by yeast CPD1p degrades Appr>p to ADP-ribose 1″-phosphate (Appr-1″p) (14,17,18). In plants, a CPDase that cleaves both the 2′,3′ cyclic phosphate and Appr>p has been identified (19,20).

While most bacteria do not possess such introns in their tRNAs, some cyanobacteria and proteobacteria contain type I or type II self-splicing introns inserted into their tRNA genes (21). These introns excise themselves from their parent RNA through consecutive transesterification steps without involving any cyclic phosphate RNA intermediate (22). However, *in vitro*, bacterial extracts have been demonstrated to join eukaryotic tRNA halves by cleaving the 2′,3′ cyclic phosphate in the 5′ fragment to give a 2′ phosphate that is directly ligated to the 5′OH of the 3′ tRNA fragment (23,24). The enzyme responsible for this activity is encoded by the *Escherichia coli* LigT and is conserved across a wide range of bacteria and archaeal species. The brain 2′,3′ phosphodiesterase, which occurs in vertebrates as a structural component of myelin, is another enzyme that hydrolyzes similar phosphodiester bonds in cyclic nucleotides, oligonucleotides and 2′,3′-cyclic NADP (16,25,26).

Direct comparisons of the sequences of these enzymes that share a common biochemistry revealed the presence of two conserved Hh[ST]h motifs (h, a hydrophobic residue) (17,27). Mutational studies in both yeast and plant 1″,2″ CPDases showed that the histidine and the serine/threonine residues (to a lesser extent) are essential for catalytic activity (17,28). Site-directed mutagenesis of the brain phosphodiesterase also supported the role of the conserved His in the catalytic activity of this family (29). Subsequent determination of the crystal structure of the *Arabidopsis* ADP-ribose 1″,2″ CPDase showed that these enzymes are likely to adopt a unique $\alpha + \beta$ fold, displaying an internal dyad symmetry (28,30). The Hh[ST]h signature motifs are located in a cavity where the phosphodiesterase reaction is proposed to occur. The reaction appears to proceed by the interaction of the C-terminal conserved histidine with a water molecule that generates a nucleophilic hydroxide ion, which attacks the cyclic phosphate (17). The N-terminal conserved histidine is proposed to protonate the leaving oxygen (17,27). The structure and the reaction mechanism of these 2′,3′ phosphodiesterases are strikingly different from the classical 3′,5′-cyclic nucleotide phosphodiesterases, which contain an all-α-helical catalytic domain of the HD superfamily (1). The HD domain chelates two zinc ions via several conserved histidine and aspartates, and the metal ions activate water for hydrolysis of the cyclic phosphodiester bond (1,31).

While these enzymes are extremely divergent at the sequence level, they share a common catalytic site suggesting that they define a novel superfamily of enzymes with several, distinct biological roles. In order to understand the provenance of this superfamily and the entire extent of its diversity, we undertook an exhaustive analysis of these proteins using sensitive sequence analysis methods. As a result of this analysis, we were able to detect several new members of this superfamily, including previously entirely undetected families of these enzymes: three from bacteria and two from eukaryotes. Furthermore, we were able to identify several predicted enzymes of this superfamily from type C rotaviruses, coronaviruses, piscine retroviruses, large double stranded DNA viruses, and additional divergent versions in eukaryotic, archaeal and bacterial proteomes. This superfamily is hereinafter referred to as the 2H phosphoesterase superfamily after the presence of the two conserved histidines that characterize its active site. By analyzing contextual information, in the form of phyletic profiles, domain fusions and predicted operon organization, we were able to get evidence for the association of these enzymes with different tRNA processing events, phosphonate metabolism and different signal transduction pathways. Based on phylogenetic analysis of these proteins we present evidence that they were probably involved in RNA metabolism in the LUCA itself and have subsequently been recruited to several distinct functions on independent occasions in evolution.

MATERIALS AND METHODS

Sequence searches of the non-redundant and unfinished genome databases at the NCBI were conducted using the BLASTP program and PSI-BLAST programs (32). Iterative profile searches using the PSI-BLAST program were done using a profile inclusion cutoff of 0.01. Hidden Markov models (HMMs) built from the multiple sequence alignments of previously detected members were used to search the database for previously undetected members using the HMMER2 package (33). Further, in a parallel approach, position-specific score matrices (PSSMs) were constructed using alignments of the confirmed members and used with the RPS-BLAST program to detect all significant matches to the PSSM (34). The conserved motifs were detected and evaluated for statistical significance using the Gibbs sampling procedure (implemented in the MACAW program) (35). Sequence-structure threading of the PDB database was done using the 3D-PSSM (36) and hybrid-fold recognition methods (37). Multiple sequence alignments were generated using the T-Coffee program (38) and further manually refined using the output from PSI-BLAST searches, the predicted secondary structure of the individual families and the crystal structure of the plant CPDase. Secondary structure predictions were obtained using the PHD and Jpred programs (39,40).

Clustering of families was done using the BLASTCLUST program with empirically determined length and score threshold cutoff values (for documentation see

ftp://ftp.ncbi.nih.gov/blast/documents/README.bcl). Phylogenetic analysis was done using the neighbor joining or least squares method with subsequent local rearrangements using the maximum-likelihood algorithm to determine the most likely tree (41). The robustness of tree topology was assessed with 10 000 Resampling of Estimated Log Likelihoods (RELL) bootstrap replicates. The MOLPHY and Phylip software packages were used for the phylogenetic analyses (42,43). Structural manipulations and modeling were carried out using the SWISS-PDB Viewer (44), Molscript (45) (http://www.avatar.se/molscript/) and Pymol (http://www.pymol.org) programs. In order to construct the surface conservation diagrams the residue positions, in the multiple alignment, that were conserved in each family, but not universally conserved in the entire 2H superfamily, were first extracted using the Consensus program (http://www.bork.embl-heidelberg.de/Alignment/consensus.html). The residues corresponding to these positions on the template structure, the Plant 1″,2″ CPDase (PDB id: 1FSI), were mapped based on the alignment and highlighted in red on the surface representation of this protein.

RESULTS AND DISCUSSION

Detection of new members of the 2H phosphoesterase superfamily

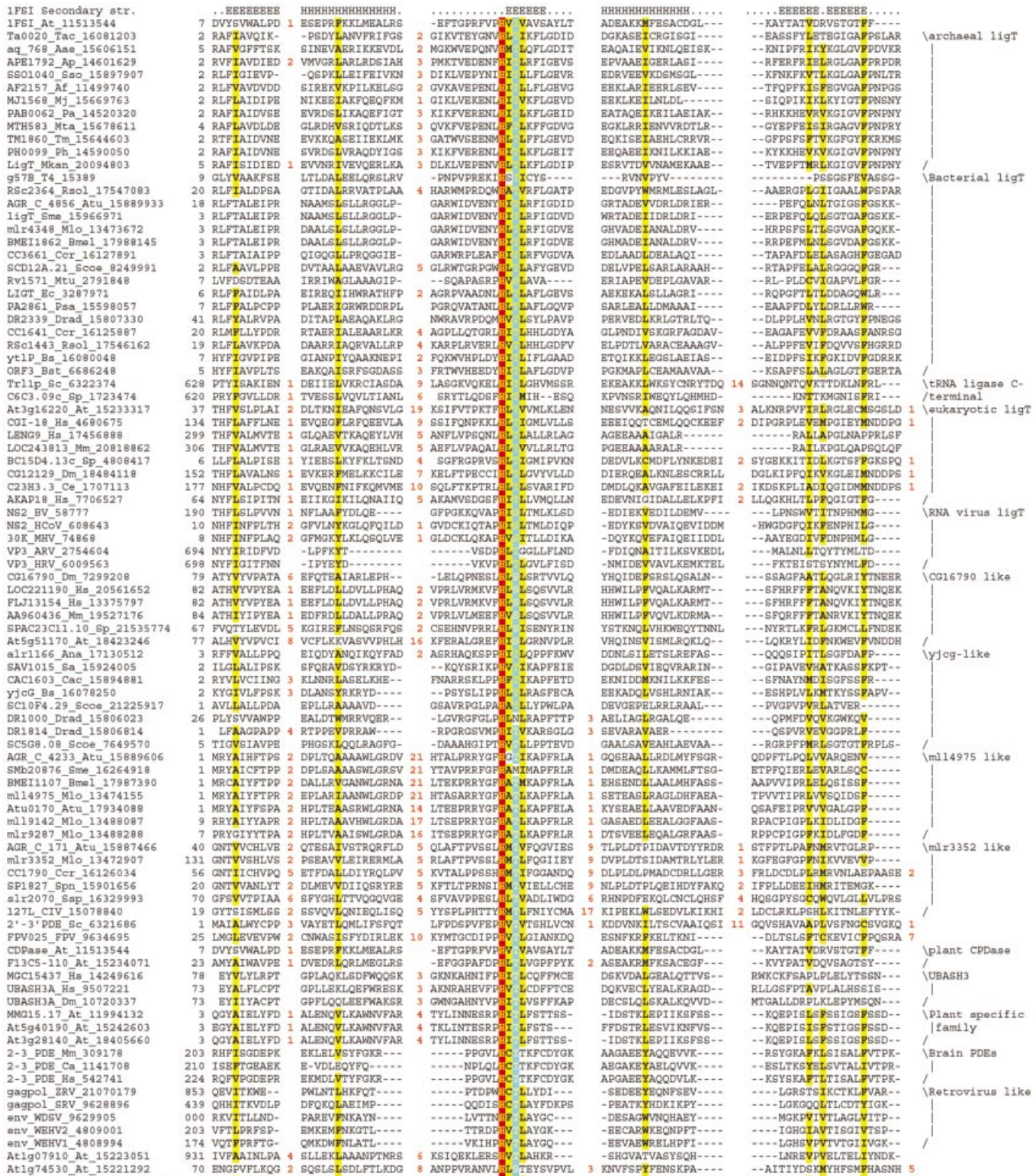
A hallmark of the 2H superfamily is their extreme sequence divergence despite the conservation of the active site motifs. This presents a challenge for their identification via sequence analysis and calls for a combination of a variety of sensitive techniques. In order to perform an exhaustive search of the sequence databases, we conducted a series of PSI-BLAST searches seeded with representatives of all previously identified members. We clustered all the members detected at convergence in these searches and transitively searched the databases with representatives of any new versions that were detected. Searches initiated with 2H domains corresponding to the CPDase module of the yeast tRNA ligase, the yeast and plant ADP-ribose 1″,2″ CPDases and the CPDase domain of the brain phosphodiesterases usually recovered their close relatives and orthologs, but did not detect much else. Searches initiated using the LigT-like proteins fared better and recovered several new families and divergent members. For example, searches initiated from the *Pyrococcus abyssi* LigT ortholog, PAB0062 (gi: 14520320), recovered the archaeal and the bacterial LigT-like proteins in the first iteration of the PSI-BLAST search at significant expect (e)-values (*Methanococcus jannaschii* LigT, e-value = 7×10^{-40}; *Agrobacterium* LigT, e-value = 5×10^{-10}). The same search recovered several eukaryotic LigT-like proteins (e.g. AKAP18, gi: 7706527) and the related YjcG-type bacterial proteins were recovered between the third to the fifth iterations with e-values between 8×10^{-10} and 10^{-3}. The coronavirus non-structural protein 2 (NS2), the rotavirus VP3 capping protein, and *Drosophila* CG16790 were recovered in subsequent iterations (e-value = ~10^{-5} to 10^{-3}). At convergence, with e-values of borderline significance (~0.1) several proteins that had a similar conservation pattern, including the Hh(S/T)h motifs, were detected. These included proteins typified by *Agrobacterium* AGR_C_171, Fowlpox virus FPV025, the UBASH3-group of proteins from animals and *Mesorhizobium* mlr3352. A transitive search with the *Bacillus anthracis* YjcG protein recovered not only the proteins detected in the above search, but also the proteins like FPV025, mlr3352, UBASH3 with significant e-values (e-value = ~10^{-8} to 10^{-3}) at the point of first detection, and confirmed their membership to the 2H superfamily.

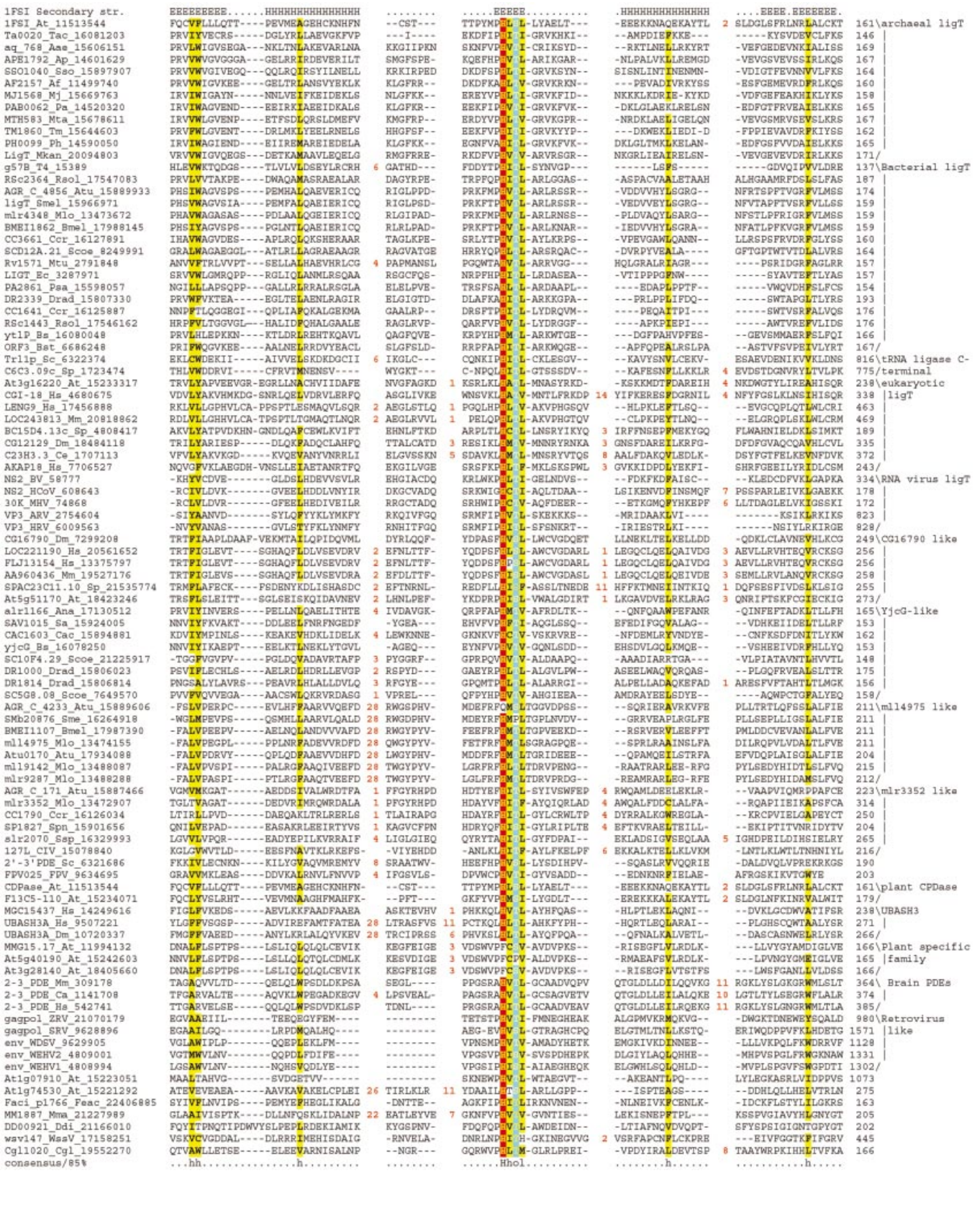
A multiple alignment of all the detected proteins was prepared using the T-Coffee program and used to prepare a HMM that was used to search the complete or nearly complete proteomes (NCBI Genome database: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). These searches additionally recovered other members not detected in the above PSI-BLAST searches mentioned above such as the yeast tRNA ligase, the ADP-ribose 1″,2″ CPDases, gene 57B from phage T4 and Chilo iridescent virus (CIV) protein 127L. Additionally, pattern searches were initiated with the conserved motifs shared by all these 2H proteins and all the thus recovered proteins were queried against a library of PSSMs, including one for the 2H family, using the RPS-BLAST program. This not only confirmed the above-detected proteins, but also revealed a new group of 2H proteins such as those embedded in the poly-proteins of piscine retroviruses such as the Zebrafish endogenous retrovirus (e-value for matching 2H domain PSSM = 10^{-4}), and other divergent proteins such as Cgl1020 from *Corynebacterium glutamicum* (e-value = 5.2×10^{-8}), the DD00921 protein from *Dictyostelium discoideum* (e-value = 4.8×10^{-4}), MM1887 from *Methanosarcina mazei* (e-value = 10^{-5}), and WSV147 from Shrimp white spot syndrome virus (e-value = 10^{-3}).

A search for conserved motifs across the entire set of 2H proteins, detected above, using Gibbs sampling revealed the presence of two conserved motifs of length approximately 16 and 18 amino acids, with a probability of chance occurrence in this protein set being less than 10^{-15}. Secondary structures were predicted individually for each of the newly detected families as well as divergent un-clustered members and compared to the previously known structure of the plant 1″,2″ CPDase (PDB id: 1FSI). The excellent correlation between the predicted structural elements for these proteins and that known for 1″,2″ CPDase further supported their membership to the 2H superfamily. This result was also supported by the sequence-structure threading, performed for a subset of these proteins, using 3D-PSSM or the hybrid-fold recognition method that suggested 1FSI to be the most likely template.

Conserved sequence features and the structural core of the 2H superfamily

An examination of the multiple alignment of the entire 2H superfamily reveals that both the Hh[ST]h motifs are almost absolutely conserved with threonine present in 86% of the motifs (Fig. 1). The only disruptions of the Hh[ST]h motifs that we were able to detect were observed in the C-terminal copy of the *Arabidopsis* At5g40190-like family and *Agrobacterium* AGR_C_4233 protein, where the histidine is replaced by phenylalanine and glutamine, respectively. The rest of the sequence conservation corresponds mainly to the hydrophobic residues that stabilize the β-strand elements (Fig. 1). Comparison of the common core conserved across the 2H superfamily with the structure of the plant 1″,2″ CPDase (PDB id: 1FSI) reveals that it comprises of two topologically


1FSI Secondary str.
1FSI_At_11513544
Ta0020_Tac_16081203
aq_768_Aae_15606151
APE1792_Ap_14601629
SSO1040_Sso_15897907
AF2157_Af_11499740
MJ1568_Mj_15669763
PAB0062_Pa_14520320
MTH583_Mta_15678611
TM1860_Tm_15644603
PH0099_Ph_14590050
LigT_Mkan_20094803
g57B_T4_15389
RSc2364_Rsol_17547083
AGR_C_4856_Atu_15889933
ligT_Sme_15966971
mlr4348_Mlo_13473672
BMEI1862_Bmel_17988145
CC3661_Ccr_16127891
SCD12A.21_Scoe_8249991
Rv1571_Mtu_2791848
LIGT_Ec_3287971
PA2861_Psa_15598057
DR2339_Drad_15807330
CC1641_Ccr_16125887
RSc1443_Rsol_17546162
ytlP_Bs_16080048
ORF3_Bst_6686248
Trl1p_Sc_6322374
C6C3.09c_Sp_1723474
At3g16220_At_15233317
CGI-18_Hs_4680675
LENG9_Hs_17456888
LOC243813_Mm_20818862
BC15D4.13c_Sp_4808417
CG12129_Dm_18484118
C23H3.3_Ce_1707113
AKAP18_Hs_7706527
NS2_BV_58777
NS2_HCoV_608643
30K_MHV_74868
VP3_ARV_2754604
VP3_HRV_6009563
CG16790_Dm_7299208
LOC221190_Hs_20561652
FLJ13154_Hs_13375797
AA960436_Mm_19527176
SPAC23C11.10_Sp_21535774
At5g51170_At_18423246
alr1166_Ana_17130512
SAV1015_Sa_15924005
CAC1603_Cac_15894881
yjcG_Bs_16078250
SC10F4.29_Scoe_21225917
DR1000_Drad_15806023
DR1814_Drad_15806814
SC5G8.08_Scoe_7649570
AGR_C_4233_Atu_15889606
SMb20876_Sme_16264918
BMEI1107_Bmel_17987390
mll4975_Mlo_13474155
Atu0170_Atu_17934088
mll9142_Mlo_13488087
mlr9287_Mlo_13488288
AGR_C_171_Atu_15887466
mlr3352_Mlo_13472907
CC1790_Ccr_16126034
SP1827_Spn_15901656
slr2070_Ssp_16329993
127L_CIV_15078840
2'-3'PDE_Sc_6321686
FPV025_FPV_9634695
CDPase_At_11513544
F13C5-110_At_15234071
MGC15437_Hs_14249616
UBASH3A_Hs_9507221
UBASH3A_Dm_10720337
MMG15.17_At_11994132
At5g40190_At_15242603
At3g28140_At_18405660
2-3_PDE_Mm_309178
2-3_PDE_Ca_1141708
2-3_PDE_Hs_542741
gagpol_ZRV_21070179
gagpol_SRV_9628896
env_WDSV_9629905
env_WEHV2_4809001
env_WEHV1_4808994
At1g07910_At_15223051
At1g74530_At_15221292
Faci_p1766_Feac_22406885
MM1887_Mma_21227989
DD00921_Ddi_21166010
wsv147_WssV_17158251
Cgl1020_Cgl_19552270
consensus/85%
\archaeal ligT
\Bacterial ligT
\tRNA ligase C-terminal
\eukaryotic ligT
\RNA virus ligT
\CG16790 like
\yjcg-like
\mll4975 like
\mlr3352 like
\Plant specific family
\Brain PDEs
\Retrovirus like
\plant CPDase
\UBASH3



equivalent, closely interacting repeats (Fig. 2A and B). Each individual repeat comprises of a β-α-β-α-β unit, with the last β strand of each repeat unit being structurally associated with the strands 1 and 2 of the opposite unit to form a three-stranded sheet. The penultimate strand of each repeat associate with each other to form a two-stranded sheet that is in a different plane relative to the two three-stranded sheets formed by the rest of the two repeats (Fig. 2A and B). This results in an incomplete barrel-like structure with a large water-filled cavity that houses the active site of these enzymes (Fig. 2A) (27,46). The histidine and alcoholic residues (serine or threonine) are associated with strand 2 of each repeat unit and are in proximity with each other in the space within the cavity, with the histidines at opposite ends (Fig. 2A and B) (27,46). Additionally, certain members of this superfamily appear to have a C-terminal strand that stacks against the basic three-stranded sheet of the first repeat. The distribution of the inserts with respect to the above-described structural elements, which are seen in different members of this superfamily, suggest that they are unlikely in any way to distort the catalytic cavity (Fig. 2B).

This dyad symmetry in the sequence conservation pattern and the structure suggest that this module emerged through the ancestral duplication of a single β-α-β-α-β unit. Originally, two such units, each contributing a Hh[ST]h motif to the active site, are likely to have functioned as a dimer that was stabilized by the interaction of their terminal strand 3. Upon duplication these terminal strands appear to have grown further giving rise to a β-strand extension that was incorporated into the sheet formed by the strands 1 and 2 of the opposite structure. This is likely to have been fixed as it would have contributed to the stability of these sheets and thereby maintained the integrity of the catalytic pocket.

RNase A, RNase T1 of the barnase superfamily and RNase T2 possess phosphodiesterase activities similar to the 2H superfamily (30,47). A direct comparison of the structures of these enzymes with that of the 2H superfamily reveals that they are very distinct from each other in their overall topology. However, both the RNase A and RNase T2 superfamilies contain active sites with two catalytic histidines, and are likely to exhibit similarities to the 2H superfamily in catalytic mechanisms. Furthermore, the symmetric spatial placement of these histidines in a pocket formed by a curved β-sheet scaffold is reminiscent of the arrangement of the histidines in the 2H superfamily. Nevertheless, the topology of the secondary structure elements that constitute the catalytic pocket of these RNases differs completely from that of the 2H superfamily active site. Hence, it is likely that the superficially similar active sites in these enzyme superfamilies have arisen through convergent evolution.

Classification of the 2H phosphoesterase superfamily

The preliminary clustering of the 2H superfamily proteins was carried out using single linkage clustering across a range of (bit score)/(sequence length) thresholds using the BLASTCLUST program. This analysis provided the major groups of proteins that were further classified through phylogenetic tree reconstruction, using the maximum-likelihood method, to identify orthologous groups. The classification above the level of the orthologous groups was reinforced by using information from phyletic patterns and presence of shared derived characters (synapomorphies) specific to a given group, in addition to the clustering done with BLASTCLUST. As a result a total of 12 families and a few extremely divergent members that could not be confidently included within any specific family were identified within the 2H superfamily (Table 1). Most of these families could further be accommodated within four major clades namely, the archeo-bacterial LigT clade, the eukaryotic–viral LigT clade, the YjcG-like clade and the mlr3352 clade (Table 1). Generally, 2H superfamily members show good sequence conservation within a family but little conservation beyond the core residues between families. In this aspect, the 2H superfamily resembles other phosphoesterase superfamilies such as HD superfamily (1) and the calcineurin-like phosphoesterase superfamily (3). This suggests that there has been extensive diversification in regions of the domain beyond the catalytic core probably in response to adaptation to different functional niches specific to each family (see below).

We briefly describe below the various major divisions of 2H superfamily.

The archaeo-bacterial LigT-like group

The principal family in the archaeo-bacterial LigT-like group is prototyped by the *E.coli* LigT protein. Orthologs of this protein are sporadically distributed across most major bacterial lineages and are encoded by all euryarchaeal and crenarchaeal genomes, available to date, with the exception of *Halobacterium* sp. where the gene appears to have recently degenerated (Vng2440h, gi:15791217) (Table 1). Additionally, a very divergent member of this family is encoded by the gene 57B of phage T4. This family is typified by a very characteristic set of sequence features that include an

Figure 1. (Previous two pages) Multiple alignment of a selected set of 2H domains. Proteins are represented by their gene names, species abbreviations and gi numbers. The 85% consensus shown below the alignment was based on the following amino acid classes: h, hydrophobic residues (L,I,Y,F,M,W,A,C,V) and l, aliphatic (L,I,A,V) residues shaded yellow; o, alcohol (S,T) group containing residues, shaded blue. The secondary structure of the *Arabidopsis* Appr>p cyclic phosphodiesterase is shown above the alignment, where H denotes residues present in helices and E (extended) in strands. Family specific groupings are shown to the right of the alignment. Species abbreviations are as follows: Aae, *Aquifex aeolicus*; Af, *A.fulgidus*; Ana, *Anabaena* sp.PCC 7120; Ap, *A.pernix*; ARV, Avian rotavirus; At, *Arabidopsis thaliana*; Atu, *Agrobacterium tumefaciens*; Bmel, *Brucella melitensis*; Bs, *B.subtilis*; Bst, *Bacillus stearothermophilus*; BV, Berne virus; Ca, *Carassius auratus*; Cac, *Clostridium acetobutylicum*; Ccr, *Caulobacter crescentus*; Ce, *Caenorhabditis elegans*; Cgl, *C.glutamicum*; CIV, Chilo iridescent virus; Ddi, *D.discoideum*; Dm, *Drosophila melanogaster*; Drad, *Deinococcus radiodurans*; Ec, *E.coli*; Feac, *Ferroplasma acidarmanus*; FPV, Fowlpox virus; HCoV, Human coronavirus; HRV, Human rotavirus; Hs, *Homo sapiens*; MHV, Mouse hepatitis virus; Mj, *M.jannaschii*; Mkan, *M.kandleri*; Mlo, *M.loti*; Mm, *Mus musculus*; Mma, *M.mazei* Goe1; Mta, *M.thermautotrophicus*; Mtu, *Mycobacterium tuberculosis*; Pa, *P.abyssi*; Ph, *Pyrococcus horikoshii*; Psa, *P.aeruginosa*; Rsol, *Ralstonia solanacearum*; Sa, *Staphylococcus aureus*; Sc, *S.cerevisiae*; Scoe, *Streptomyces coelicolor*; Sme, *Sinorhizobium meliloti*; Sp, *S.pombe*; Spn, *Streptococcus pneumoniae*; SRV, Snakehead retrovirus; Sso, *S.solfataricus*; Ssp, *Synechocystis* sp. PCC 6803; T4, Bacteriophage T4; Tac, *Thermoplasma acidophilum*; Tm, *Thermotoga maritima*; WDSV, Walleye dermal sarcoma virus; WEHV1, Walleye epidermal hyperplasia virus type 1; WEHV2, Walleye epidermal hyperplasia virus type 2; WssV, Shrimp white spot syndrome virus; ZRV, Zebrafish endogenous retrovirus.

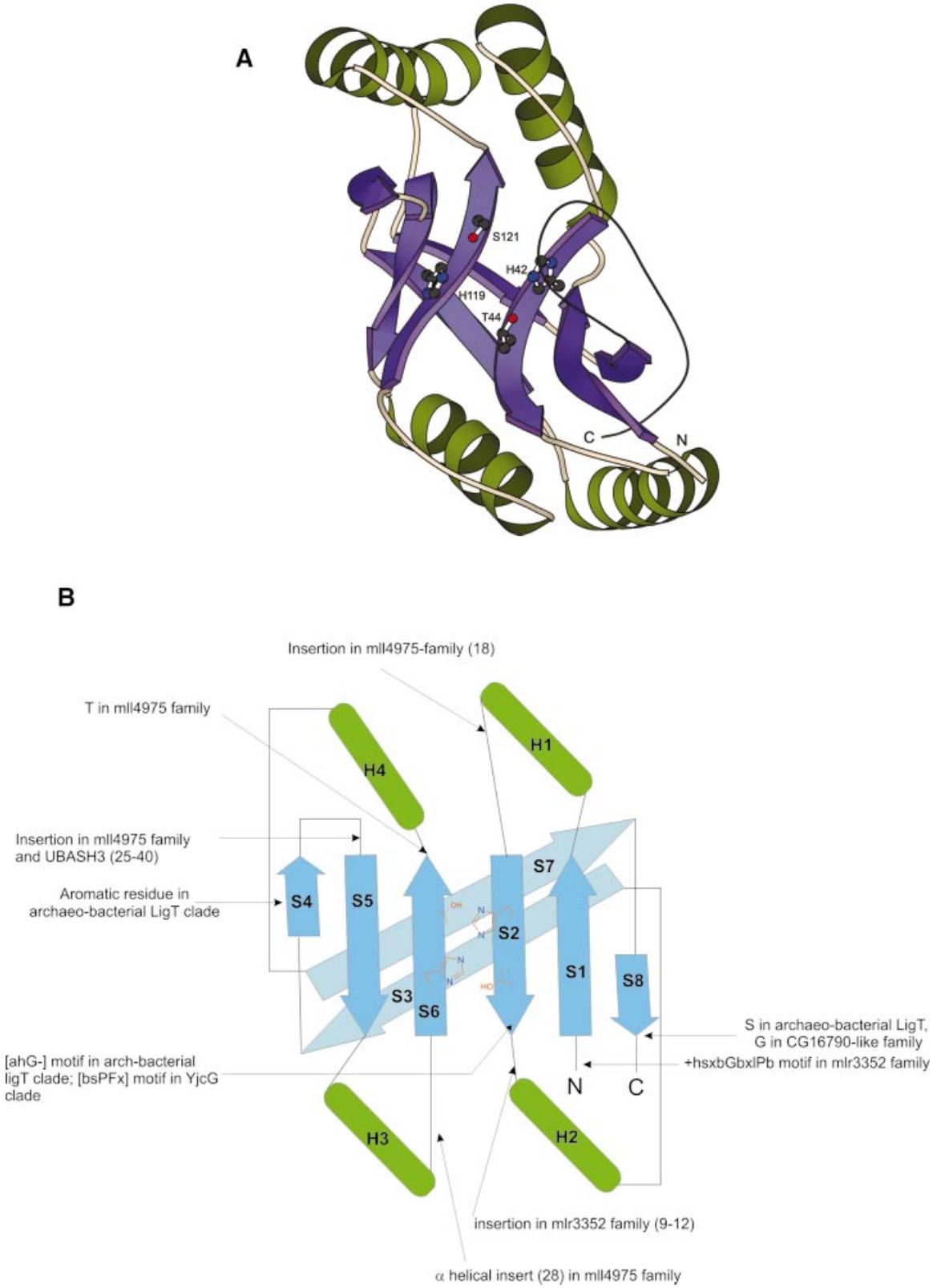


Figure 2. Evolutionarily conserved structure of the 2H phosphoesterase domain. (**A**) Structure of the plant CPDase (PDB id: 1FSI) showing the secondary structure elements conserved across the superfamily. The residues involved in catalysis are shown in the ball-and-stick representation. (**B**) Schematic representation of the secondary structure topology of the 2H phophoesterase domain. β strands are represented as arrows, while the α helices are rods. Secondary structural element numbering is based on ascending order from the N-terminal end. Side chains comprising the catalytic core are shown in greater detail. Inserts and sequence synapomorphies are shown with the number of residues in inserts given in brackets. Note the two topologically similar and equivalent structural units.

Table 1. Classification and phyletic distribution of 2H-phosphoesterases

	Bacteria	Archaea	Eukaryotes and viruses
		Group 1: Archaeo-bacterial LigT	
Family 1 LigT/2′-5′ RNA ligase	Atu[a], Aae[b], Bmel, Bs, Bst, Ccr, Cte, Cgl, Drad, Ec, St, Psa, Rsol, Mlo, Mle, Mtu, Scoe, Sme, Tm[b], Thte[b], Xax, Xca, Yp	Ap, Sso, Pyae, Af, Mta, Mac, Mkan, Mma, Pa, Ph, Tac, Tvo	–
Family 2 tRNA-ligase-C-terminal domain-like	–	–	Cal, Sp, Sc
		Divergent members in Group I	
Bacteriophage T4-like	Bacteriophage T4		
		Group II: Eukaryotic ligT	
Family 1 Eukaryotic LigT-like family	–	–	Tc, Ag, Cpa, Mm, Nc, Ce, Os, At, Hs, Dm, Sp
Family 2 RNA virus LigT-like family	–	–	BV, HCoV, MHV, ARV, HRV
Family 3 *Drosophila* CG16790-like family	–	–	At, Dm, Ehi, Fr, Hs, Sp
		Group III (YjcG-like)	
Family 1 YjcG-like	Ana, Bs, Cac, Drad, Sa, Scoe, Tel	–	–
Family 2 mll4975-like α-proteobacterial family	Atu, Bmel, Mlo, Sme	–	–
Family 3 2H domains in UBASH3A proteins	–	–	Dm, Hs
Family 4 At5g40190-like plant-specific family			At
		Group IV (mlr3352 like)	
Family 1 mlr3352-like family	Atu, Ccr, Mlo, Spn, Ssp	–	CIV
		Divergent members of the 2H superfamily	
Family 1 Brain phosphodiesterase	–	–	Vertebrates
Family 2 Piscine retrovirus-polyprotein associated	–	–	SRV, WDSV, WEHV1, WEHV2, ZRV
Plant CPDases	–	–	At
DNA virus	–	–	FPV
DNA virus			WssV
Cpd1p	–	–	Sc
Faci_p1766-like	–	Feac	–
DD00921-like	–	–	Ddi
Cgl1020-like	Cgl	–	–
MM1887-like	–	Mma	–

[a]Organism abbreviations: Aae, *A.aeolicus*; Af, *A.fulgidus*; Ag, *Anopheles gambiae*; Ana, *Anabaena* sp. PCC 7120; Ap, *A.pernix*; ARV, Avian rotavirus; At, *A.thaliana*; Atu, *A.tumefaciens*; Bmel, *B.melitensis*; Bs, *B.subtilis*; Bst, *B.stearothermophilus*; BV, Berne virus; Ca, *C.auratus*; Cal, *Candida albicans*; Cac, *C.acetobutylicum*; Ccr, *C.crescentus*; Ce, *C.elegans*; Cgl, *C.glutamicum*; CIV, Chilo iridescent virus; Cpa, *C.parvum*; Cte, *Chlorobium tepidum*; Ddi, *D.discoideum*; Dm, *D.melanogaster*; Drad, *D.radiodurans*; Ec, *E.coli*; Ehi, *E.histolytica*; Feac, *F.acidarmanus*; Fr, *Fugu rubripes*; FPV, Fowlpox virus; HCoV, Human coronavirus; HRV, Human rotavirus; Hs, *H.sapiens*; MHV, Mouse hepatitis virus; Mac, *Methanosarcina acetivorans*; Mma, *M.mazei* Goe1; Mj, *M.jannaschii*; Mkan, *M.kandleri*; Mle, *Mycobacterium leprae*; Mlo, *M.loti*; Mm, *M.musculus*; Mta, *M.thermautotrophicus*; Mtu, *M.tuberculosis*; Nc, *Neurospora crassa*; Os, *Oryza sativa*; Pa, *P.abyssi*; Ph, *P.horikoshii*; Psa, *P.aeruginosa*; Pyae, *P.aerophilum*; Rsol, *R.solanacearum*; Sa, *S.aureus*; Sc, *S.cerevisiae*; Scoe, *S.coelicolor*; Sme, *S.meliloti*; Sp, *S.pombe*; Spn, *S.pneumoniae*; SRV, Snakehead retrovirus; Sso, *S.solfataricus*; ZRV, Zebrafish endogenous retrovirus; Ssp, *Synechocystis* sp. PCC 6803; St, *Salmonella typhi*; T4, Bacteriophage T4; Tac, *T.acidophilum*; Tc, *Trypanosoma cruzi*; Tel, *Thermosynechococcus elongates*; Tm, *T.maritima*; Thte, *Thermoanaerobacter tengcongensis*; Tvo, *Thermoplasma volcanium*; WDSV, Walleye dermal sarcoma virus; WEHV1, Walleye epidermal hyperplasia virus type 1; WEHV2, Walleye epidermal hyperplasia virus type 2; WssV, Shrimp white spot syndrome virus; Xax, *Xanthomonas axonopodis*; Xca, *Xanthomonas campestris*; Yp, *Yersinia pestis*, ZRV, Zebrafish endogenous retrovirus.
[b]2H phosphodiesterase of the archaeal LigT family.

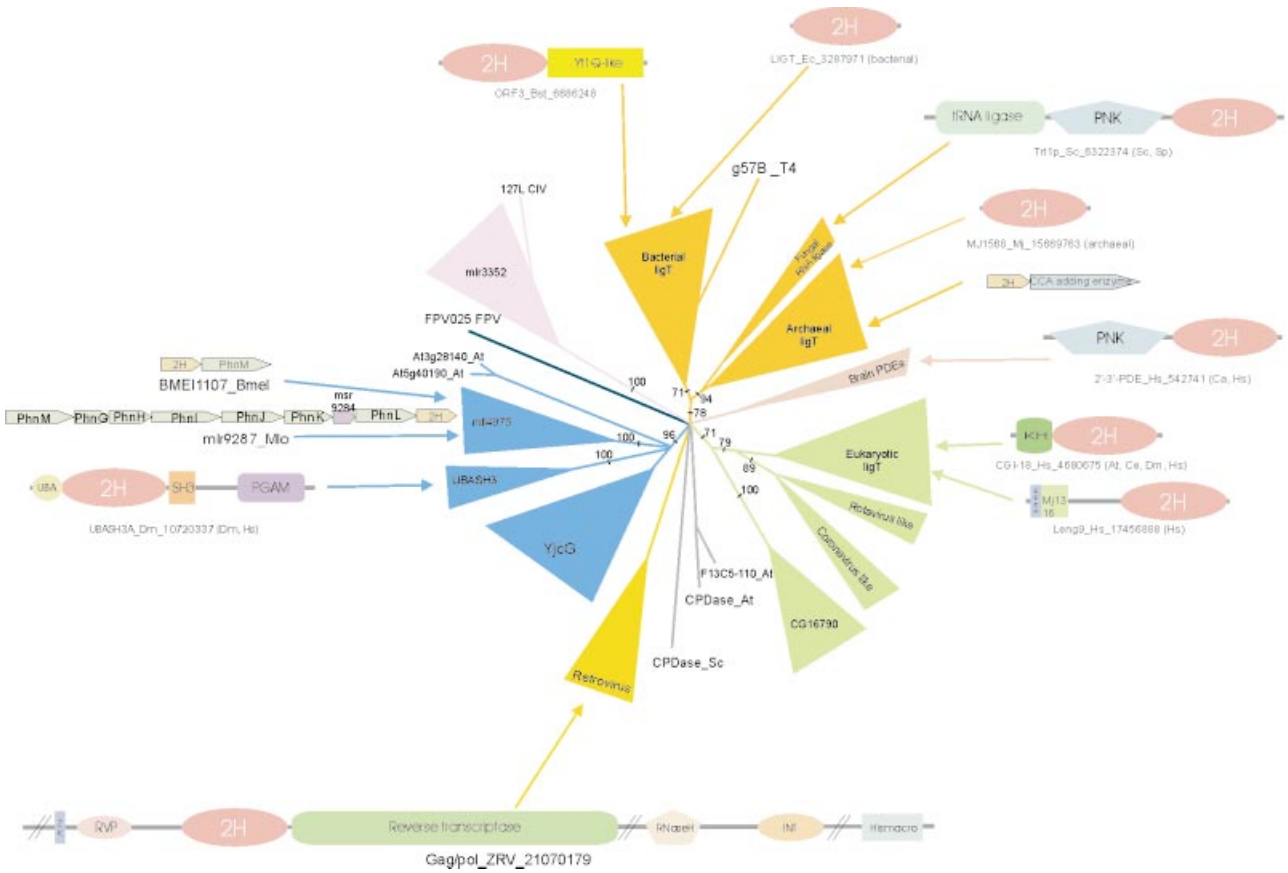


Figure 3. Maximum-likelihood phylogenetic tree, domain architectures and operon organization of 2H proteins. All branches with RELL bootstrap support <50% are collapsed and the values that support a node are shown in the remaining cases. Conserved gene neighborhoods (operons) that are discussed in the text are represented by boxed arrows with the gene names written within. Domain abbreviations are as follows: 2H, 2H phosphoesterase; KH, K homology; UBA, ubiquitin associated; SH3, Src homology 3; PGAM, phosphoglycerate mutase; PNK, P-loop nucleotide kinase; ZK, zinc knuckle; rvp, retroviral aspartyl protease; INT, integrase; Hismacro, phosphoesterase domain found in Macro histone 2 and the Appr-1″-p processing enzyme. Mj1316 domain is a predicted RNA binding domain typified by the protein MJ1316 protein of *Methanococcus*. The species abbreviations are as in Figure 1.

ahG- motif (a, aromatic; h, hydrophobic; -, acidic residues) present at the end of strand 2, an aromatic position in the middle of strand 4 and a serine residue at the end of strand 8 (Fig. 1). The maximum-likelihood phylogenetic tree (Fig. 3) strongly supported a monophyletic group that included the archaeal, bacterial and phage T4 LigT-like proteins (RELL bootstrap: 78%). All the archaeal LigTs strongly grouped together, with the orthologs from *Aquifex*, *Thermoanaerobacter* and *Thermotoga* grouping with the archaeal versions, consistent with the previous observations regarding acquisition of archaeal genes by thermophilic bacteria (48,49). The rest of the bacterial proteins grouped together into a coherent group to the exclusion of the archaeal and thermophile LigTs (Fig. 3). This analysis also placed 57B protein from phage T4 within the bacterial clade (RELL bootstrap: 76%), albeit on a long branch, suggesting that it was acquired from the hosts, and subsequently diverged extensively. The above phyletic pattern and topology of the phylogenetic tree suggests that the LigT-like protein was probably present in the LUCA. While there was a strong selection for its maintenance in the archaeal lineage, the bacteria appear to have lost it on multiple occasions.

In order to gain insight into the potential functions of this family we analyzed the gene neighborhoods of the LigT homologs (Fig. 3). In prokaryotes, genes are often present in co-transcriptional units called operons. Evolutionary conservation of operon organization correlates well with the physical interactions of the gene products with each other or sequential function in a particular pathway (50). Thus, the contextual information present in conserved operons can be used to make functional predictions for poorly characterized genes on the basis of their association with functionally characterized genes (51–53). The ligT orthologs co-occurred in the same predicted operon as the archaeal tRNA-CCA adding enzymes in at least seven archaeal genomes that include crenarchaea (*Aeropyrum pernix*, *Sulfolobus solfataricus* and *Pyrobaculum aerophilum*) and euryarchaea (*Archaeoglobus fulgidus*, *Pyrococcus fulgidus*, *Methanothermobacter thermautotrophicus* and *Methanopyrus kandleri*) (Fig. 3). As few operons are so strongly conserved across such a wide phylogenetic range, this suggests a strong association of these proteins with tRNA metabolism. Given that the archaea follow a mechanism of tRNA splicing that, just as in eukaryotes (54,55), involves 2′,3′ cyclic phosphate intermediates, it is very likely that these enzymes act as 2′,3′ phosphodiesterases. Furthermore, the strong operonic association of the CCA adding enzyme and a predicted 2′,3′ phosphodiesterase suggests that in archaea the process of tRNA splicing and terminal CCA addition is likely

to occur successively or concomitantly in a tightly linked pathway. Similarly in the phage T4, the 2H phosphodiesterase encoding gene 57B occurs in the vicinity of a cluster of eight tRNA genes (56). Interestingly, the phage T4 also encodes tRNA ligases that function similar to the yeast tRNA ligase (57). This observation, together with the earlier finding that phage genes are usually present in temporally co-regulated and functionally related gene clusters (58), suggests that gene 57B may be involved in processing phage tRNA intermediates or repairing nicked RNAs.

The LigT protein from *E.coli* has a strong affinity for tRNA molecules *in vitro* and is capable of joining tRNA halves by forming an unusual 2′-5′ linkage, as well as cleaving this linkage (24). However, bacterial tRNAs do not contain introns as in archaea or eukaryotes and the LigT gene has been shown to be non-essential in *E.coli* (24). Some bacteria like *E.coli* and *Pseudomonas aeruginosa* possess enzymes such as the RNA 2′,3′ cyclase (59) and the NAD utilizing phosphotransferase (60) that generates 1″,2″-Appr>p suggesting that they could generate potential substrates for the LigT enzyme. Nevertheless, the phyletic spread of LigT orthologs in bacteria is far greater than these two enzymes, suggesting that they could function entirely independently of each other. This observation, coupled with the affinity of the bacterial LigT for tRNAs, suggests that it may be involved in an alternative tRNA processing step, where the action of some other processing nuclease might be generating 2′,3′ cyclic phosphate structures.

The spatial distribution of the residues restricted to a given family, within a superfamily, often correlate with the locations that specifically contact substrates and cofactors (61,62). Plotting these family-specific residues on the molecular surface view of the representative tertiary structure of the superfamily helps in predicting potential interaction surfaces for that family (61,62). When we plotted the residues that are specific to the archaeo-bacterial LigT family on the molecular surface of the template structure (PDB id: 1FSI) we observed that they are almost entirely distributed on two broad patches flanking the active site-containing cavity (Fig. 4A). This is consistent with a fairly large substrate like a RNA molecule(s) binding along the surface of the protein that exposes the active site. The presence of two broad patches on either side of the active site may provide the explanation for these proteins binding two RNA fragments in the 2′-5′ ligation reaction carried out by them.

The maximum-likelihood phylogenetic tree suggests that the small family comprising of the fungal tRNA ligase-linked 2H phosphoesterase domains also group with the archaeo-bacterial LigT-like family (RELL bootstrap: 94%) (Fig. 3). These domains are extremely divergent forms, but contain at least some features of this clade such as the aromatic residue in the middle of strand 4 (Fig. 1). Orthologs or even close homologs of these proteins are absent in all other eukaryotes sampled to date, suggesting they are likely to be a fungi-specific feature. However, the ligase domain of these fungal RNA ligases are most closely related to the RNA ligase domains of bacteriophages such as T4 and some baculoviruses. They also share the linkage with the P-loop kinase domain and certain unique sequence signatures with the viral forms (L.M.Iyer and L.Aravind, unpublished data). Given the presence of 2H domains in viruses, it is likely that the fusion with the RNA ligase occurred in a virus and was secondarily transferred to the fungal lineage. Evidence for such acquisitions from viruses have been recently demonstrated for a few other fungal proteins (63,64).

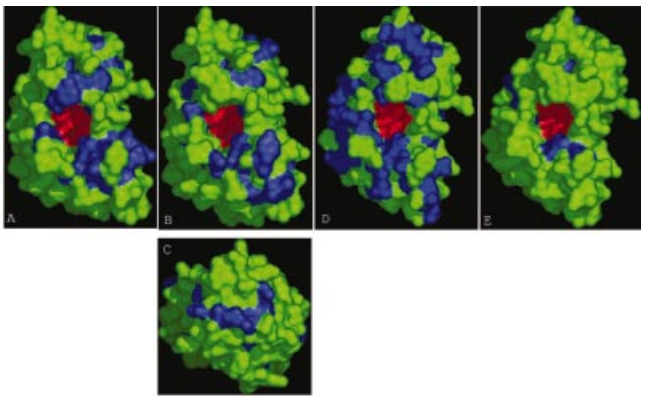


Figure 4. Surface view of family-specific conserved residues in different 2H phosphoesterase families. The family-specific conserved areas are shown in blue and the catalytic residues are shown in red. Note the pocket forming the active site. (**A**) Archaeo-bacterial LigT-like group. (**B**) CGI-18-like eukaryotic LigT proteins. (**C**) Top view of the same. (**D**) CG16790 family. (**E**) YjcB family.

Eukaryotic-viral LigT-like group

A well conserved family of LigT homologs could be identified in the proteomes of most eukaryotes and was found to be distinct from the archaeo-bacterial LigT-like clade. The eukaryotic LigT-like family is typified by the human CGI-18 gene and is represented in plants, animals, fungi (except *Saccharomyces cerevisiae*) and *Cryptosporidium parvum*. The maximum-likelihood tree revealed that the 2H phosphoesterases encoded by some coronaviruses and rotaviruses are the closest relatives of this eukaryotic LigT-like family (RELL bootstrap: 79%) (Fig. 3). Another distinct eukaryotic family that had a similar phyletic profile to the CGI-18, typified by the *Drosophila* CG16790 gene, with members in animals, plants, fungi (except *S.cerevisiae*) and the protist *Entamoeba histolytica*, forms a sister lineage to the clade comprising of the CGI-18-like and viral 2H proteins (RELL bootstrap: 71%) (Fig. 3). This suggests that early in eukaryotic evolution the 2H phosphoesterases split in the two distinct families that have been conserved ever since. The viral family contains proteins from two evolutionarily unrelated viruses: the type C rotaviruses (VP3 protein) that are double stranded multipartite RNA viruses and the coronaviruses (NS2 protein) that are positive strand RNA viruses. Given that these viruses are found only in the vertebrates, it is likely that the viral 2H phosphoesterases were derived in one of these viral groups from a host protein of the CGI-18-like family protein followed by rapid sequence divergence. Subsequently, it appears to have been exchanged between the viral families. Although the direction of the exchange is not clear, it is possible that a double stranded replicative form of a subgenomic RNA transcript of the coronavirus NS2 was stabilized by a rotavirus and incorporated into its multiple double stranded RNA genome.

The N-termini of the animal and plant orthologs of CGI-18 contain a RNA binding KH domain, whereas one of its mammal-specific paralogs, LENG9, contains another RNA binding domain, the zinc chelating CCCH finger at its N-terminus (Fig. 3) (8). Additionally, between the CCCH and 2H domains, LENG9 contains an ancient, predicted RNA binding domain typified by the archaeal protein MJ1316. These domain architectures strongly suggest that the eukaryotic LigT-like family participates in RNA metabolism, just as the archaeo-bacterial LigT-like family. However, plotting the conserved family specific residues on the surface of the structural template (PDB id: 1FSI) reveals that, while the 'lower' substrate binding site is similar to that seen in the archaeo-bacterial LigT-like family, there is no equivalent surface on the 'upper' side (Fig. 4B). This suggests that the potential interaction with RNA occurs only via the lower surface (Fig. 4B). CGI-18 is a component of the Asc-1 (Activating signal co-integrator) transcription co-activator complex, which facilitates the interaction of transcription factors with components of the basal transcription machinery and causes transcriptional activation (65,66). In addition to Asc-1 and the CGI-18 (p50), this co-activator complex has the human ortholog of Brr2p (p200), which has two superfamily II helicase domains and four Sec63 domains (8,67), and p100, which contains a Cue domain (65,67). A variety of transcription factors, such as the nuclear hormone receptors, NFκB, AP-1 and SRF, CBP/p300 and components of the transcription machinery such as TBP and TFIIA are known to bind the conserved Zn-chelating finger of the Asc-1 protein (65,66). The Brr2p protein is a component of the U5 snRNP splicing machinery, and may additionally be involved in cytoplasmic RNA processing (8,67,68). Thus, the Asc-1 complex is likely to be a ribonucleoprotein complex that participates in transcriptional co-activation and in RNA processing events related to maturation of spliceosomal particle RNAs, pre-mRNA splicing or cytoplasmic RNA degradation. In particular, this suggests that the CGI-18 protein could process uncharacterized cyclic phosphates in any of these functional contexts.

The spatial mapping of conserved residues specific to the eukaryotic proteins reveals the presence of another potential interaction surface situated on the 'upper' side (Fig. 4C). As this is located on the side facing away from the catalytic pocket it is likely to form a site for interaction with another protein, rather than the RNA substrate. This observation is consistent with the eukaryotic LigT-like proteins, such as CGI-18 and AKAP18, existing in multi-protein complexes. AKAP18 has been shown to be an activated protein kinase A (PKA) anchoring protein (69). Interestingly, another activated PKA anchoring protein, AKAP149, is a RNA binding domain that contains KH and tudor domains (70). This suggests that activated PKA may be targeted to cytoplasmic RNA processing complexes, which contain 2H phosphoesterases and other RNA binding proteins, and regulate these complexes through phosphorylation. The spatial mapping of residues specific to the CG16790 family reveals a large, conserved tract, which could potentially bind RNA, on the lower side of the catalytic surface, similar to that seen in the CGI-18-like proteins. However, these proteins also have two other, distinct, conserved surfaces on the top and side of the protein that could potentially interact with other protein partners (Fig. 4D). Many RNA metabolism proteins, which are parts of complexes or functional networks involved in post-transcription gene silencing or spliceosomal functions, are well conserved in animals, plants and *Schizosaccharomyces pombe*, but are entirely missing or extremely divergent in *S.cerevisiae* (8,71). Given that the eukaryotic LigT-like and CG16790 families display the above phyletic distribution, they could belong to such RNA processing complexes or networks that have been lost in *S.cerevisiae*.

As capping of mRNAs in eukaryotes takes place in the nucleus during transcription, cytoplasmic RNA viruses have evolved diverse independent methods to circumvent this problem (72). The VP3 protein of rotaviruses has been shown to catalyze the generation of the mRNAs cap (73). This protein is a multidomain protein and the 2H phosphoesterase domain occurs at the extreme C-terminus. Given the consistent demonstration of phosphoesterase activity in this superfamily, it is possible that the viral 2H proteins possess the phosphohydrolase activity required in the first step of capping. Alternatively, at least in the coronaviruses, it is possible that these proteins participate in the ligation process that has been proposed to give rise to the subgenomic RNAs with the same 5′ terminal portion from the genomic RNA (74). Experimental investigation of these viral enzymes may uncover as yet undiscovered aspects of viral RNA metabolism.

The YjcG-like group

Besides the LigT-like family, the bacteria possess a few other fairly widespread families of 2H phosphoesterases. The *Bacillus subtilis* YjcG protein, whose orthologs are found in Gram-positive bacteria, some actinomycetes and cyanobacteria, and *Deinococcus*, is the archetype of one of these families. The other family is restricted to the α-proteobacteria and is typified by the protein mll4975 from *Mesorhizobium loti*. This family is represented by multiple copies in *M.loti* and is characterized by the presence of an insert after helix 1, a large helical insert after helix 3, and a conserved threonine after strand 6 (Figs 1 and 3). The maximum-likelihood tree strongly supported (RELL bootstrap: 96%) a higher order relationship between the YjcG family and the mll4975 family. This clade is also supported by a unique motif of the form bsPFxl (where b, big; s, small; l, aliphatic residue) present after strand 2. Phylogenetic analysis and clustering also suggest the close relationship of two other families, namely the UBASH3 family of vertebrates and insects and the plant-specific At5g40190 family, to the above bacterial families. Together, all these families comprise the YjcG-like group (Table 1, Fig. 3). The sporadic distribution, sequence divergence and architectural diversity seen with this group suggest that its members have probably evolved to occupy a number of different niches.

Members of the YjcG family do not occur in conserved operons implicative of RNA metabolism, with the possible exception of the *Streptomyces* gene SC5G8.08 which is a gene-neighbor of the tryptophanyl tRNA synthetase. Furthermore, a spatial plot of the residues, uniquely conserved in the YjcG family, does not show any extensive interaction surface associated either with the face bearing the catalytic cavity or elsewhere (Fig. 4E). This suggests that the YjcG proteins are likely to function as stand-alone proteins on as yet unknown soluble small molecules with potential 2′,3′ cyclic

phosphoester linkages. The mll4975 family in contrast tends to occur in fairly conserved gene neighborhoods. At least five out of seven members of this family are encoded by genes co-occurring in predicted operons with genes that are implicated in phosphonate metabolism (Fig. 3). Phosphonates are organophosphorus molecules containing carbon-phosphate bond and are catabolized by a carbon-phosphate lyase pathway that is composed of a transport system and several proteins (75). This association suggests that in α-proteobacteria the assimilation of phosphonate could involve intermediates with cyclic phosphoesters that may be substrates of the 2H domains of the mll4975-like family.

The UBASH3 family, which is restricted to the coelomate animals, contains multidomain proteins where the 2H domain is fused to a UBA domain, an SH3 domain and a previously unrecognized phosphoglyceromutase domain (Fig. 3) (76). The combination with a phosphoglyceromutase, which could generate 2,3 phosphate-containing polyols, suggests that the 2H domain could perhaps cooperate with this activity in processing substrates such as RNA. The SH3 domain additionally suggests that it may participate in cytoskeletal interaction (77), and interact with the ubiquitin-mediated signaling system via the UBA domain (78). The plant-specific At5g40190 family, while belonging to the generic YjcG-like group, lacks the C-terminal Hx[TS]h motif suggesting that these proteins are either inactive or possess an alternative activity that depends on only one of the Hx[TS]h motifs.

The mlr3352-like group

This group of 2H-phosphodiesterases comprises of a single family typified by the protein mlr3352 from *M.loti*. Members are also present in various α-proteobacteria, *Synechocystis*, *Streptococcus* and CIV. Members of this family are strongly supported as a monophyletic clade in the maximum-likelihood tree (Fig. 3, RELL bootstrap: 100%) and have a synapomorphic insert after strand 2 (Fig. 2B) and a conserved N-terminal motif prior to strand 1. The presence of a member of this predominantly bacterial group in a large eukaryotic DNA virus represents a potential case of a horizontal transfer from a bacterial source into the virus. Several proteins of bacterial origin have been noticed in the insect viruses (L.M.Iyer, E.V.Koonin and L.Aravind, unpublished observations) and these appear to have been acquired from endosymbiotic or parasitic bacteria that share the same host cells with the viruses. Presence of 2H proteins in the proteomes of large DNA viruses (e.g. T4 57B protein and the Fowlpox virus FPV025) may point to some role for these proteins in regulating the viral tRNA metabolism.

Divergent forms versions of the 2H domain

In addition to the major families, described above, we noticed several proteins that were clearly members of the 2H superfamily but did not group specifically with any of the other major groups. In most of these sequences the signals of their specific relationships appear to have been eroded by rapid sequence divergence. The vertebrate specific 2′,3′ phosphodiesterases, which are membrane-associated proteins found in the myelin sheaths of neurons, are one such set of divergent proteins. The mammalian forms are fused to a N-terminal P-loop nucleotide kinase domain (16) suggesting that they are bifunctional enzymes that could phosphorylate free –OH groups in addition to the phosphodiesterase activity. The identification of pumps that secrete nucleotides (79) suggests that there might be a neural signaling pathway that secretes 2′,3′ cyclic nucleotides, and is regulated by these phosphodiesterases. As these proteins form a tight vertebrate specific group, their provenance is unclear: they could have been derived through rapid divergence from any of the above-defined families that are represented in vertebrates. The Fowlpox virus FPV025 protein, yeast Cpd1p, *Dictyostelium* DD00921, *Corynebacterium* Cgl1020, *Ferroplasma* Faci_p1766 and the plant 2′,3′ CPDases represent other divergent members of this superfamily with no clear affinities. The maximum-likelihood phylogenetic tree moderately supports their grouping with the above-defined YjcG-like group and they may have rapidly diverged from it to occupy specific niches. The *M.mazei* MM1887 and Shrimp white spot syndrome virus WSV147, on the other hand, show no affinities to any group. Finally, another divergent set of 2H phosphoesterases is seen in the polyproteins of piscine retroviruses such as the Walleye epidermal hyperplasia virus, the Snakehead retrovirus, and the Zebrafish endogenous retrovirus N-terminal to the reverse transcriptase module. These 2H domains display no close affinity to the viral versions from either the other RNA viruses or the DNA viruses. However, it is possible that they have been derived from the same source as the RNA viral versions and have vastly diverged from them due to the rapid evolution typical of most of viral proteins. Additionally, in these retroviruses they might have acquired a different function related to the processing of the tRNA, which is typically used by retroviruses as a primer. Interestingly, these retroviral polyproteins also encode a phosphoesterase of the Histone 2A Macrodomain superfamily (Fig. 3), which in yeast acts downstream of the 2H enzyme, Cpd1p, to process the generated by it Appr-1″-p (18). The co-occurrence of these enzymes in these viruses may suggest the presence of a processing pathway similar to that seen in cellular tRNA splicing.

Overall evolutionary history of the 2H phosphoesterases

The overall phyletic pattern of the 2H phosphoesterases, with a nearly universal presence in the archaea and the eukaryotes, and a widespread presence in phylogenetically diverse bacteria, supports the presence of a 2H protein in the LUCA of all extant life forms. The above-presented data also suggest that in all likelihood, this ancestral version of the 2H phosphodiesterase resembled the LigT-family of proteins and had a generic role in processing 2′,3′ cyclic phosphates in the context of RNA metabolism. This also suggests that the internal duplication that gave rise to the two interacting subdomains of the extant form of the 2H domain had already been fixed prior to the LUCA itself. However, one major phylogenetic anomaly is observed in these proteins: the bacterial and the archaeal LigT-like proteins appear to form a monophyletic clade to the exclusion of the eukaryotic-LigT like proteins (Fig. 3). This picture is contrary to the standard tree topology that is observed for most RNA-metabolism proteins traceable to the LUCA; they normally show an archaeo-eukaryotic clade to the exclusion of the bacteria (8,80,81). One explanation for this anomaly is that the eukaryotic proteins have diverged extensively due to a functional shift, while the archaea and bacteria retain the

ancestral state and hence appear to be closer to each other. The differences in the predicted interaction surfaces between the archaeo-bacterial and the eukaryotic groups (Fig. 4) appear to favor this hypothesis. Further, we obtained some evidence for a close interaction between CCA addition and RNA processing in the archaea. In the eukaryotes this version of the CCA adding enzyme appears to have been functionally displaced by the version acquired from the pro-mitochondrial symbiont, suggesting a disruption of the interaction (82). This may have additionally contributed to the divergence of the eukaryotic enzyme.

In the eukaryotes, especially the multicellular forms, the core LigT-like groups appear to have diversified into various niches within the RNA metabolism function itself, perhaps as an adaptation related to the complex RNA processing events occurring in the eukaryotes (8). Further, several other lineage specific versions of the 2H phosphodiesterases appear to have emerged later in evolution and been utilized in new functions. Some of these such as UBASH3A and the myelin phosphodiesterase, especially in animals, may have been recruited to direct or indirect functions in diverse signal transduction pathways. Another interesting feature is the emergence of several divergent forms, especially in the eukaryotes. In part, this may reflect their function as stand-alone enzymes on soluble small molecule substrates like Appr>p that relieves most selective constraints on them beyond the maintenance of the active site. Additionally, some of the divergent forms, like the one associated with the fungal tRNA ligase, could have also been introduced from a viral source.

CONCLUSIONS

Previous studies had indicated that a variety of 2′,3′ phosphodiesterases share a common catalytic site comprising of two copies of the HX[TS]h motif (2H motif) and that these proteins may share a common catalytic core that was typified by the structure of the plant 2′,3′ phosphodiesterase. Here we used sensitive sequence comparison methods to analyze these proteins and show that they define a sizeable superfamily of proteins with the characteristic 2H signature, widely represented in all the three superkingdoms of life. We identified several new families within the 2H superfamily, as well as sporadically distributed divergent versions. Based on phylogenetic analysis and detection of group-specific conserved motif we present an evolutionary classification of these proteins. The evolutionary analysis suggests that the LUCA had at least one 2H phosphoesterase that probably functioned in RNA metabolism.

Through the analysis of the spatial distribution of group-specific motifs, we show that both the prokaryotic and eukaryotic LigT-like proteins are likely to interact with a large substrate along the face that exposes the catalytic pocket. Yet, there are significant differences between the prokaryotic and eukaryotic versions in the mode of the interaction with their potential substrates. Two of the new families of 2H proteins, which we identified in the bacteria, may be involved in metabolizing small molecules. Further, we present evidence that some of the eukaryotic 2H proteins may be points of regulation of RNA processing and post-transcriptional gene silencing, while others may have been recruited for different roles in signal transduction. Interestingly we were able to identify 2H proteins to be encoded by the genomes of some RNA viruses as well as large DNA viruses. These may have been recruited for diverse virus specific functions such as tRNA repair or processing, capping or priming. The identification of several new versions of this superfamily could help in the experimental analysis of as yet unexplored biochemical activities in a range of organisms.

REFERENCES

1. Aravind,L. and Koonin,E.V. (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.*, **23**, 469–472.
2. Aravind,L. (1999) An evolutionary classification of the metallo-beta-lactamase fold proteins. *In Silico Biol.*, **1**, 69–91.
3. Aravind,L. and Koonin,E.V. (1998) Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucleic Acids Res.*, **26**, 3746–3752.
4. Aravind,L. and Koonin,E.V. (1998) A novel family of predicted phosphoesterases includes *Drosophila* prune protein and bacterial RecJ exonuclease. *Trends Biochem. Sci.*, **23**, 17–19.
5. Koonin,E.V. and Tatusov,R.L. (1994) Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.*, **244**, 125–132.
6. Koonin,E.V. (1996) A duplicated catalytic motif in a new superfamily of phosphohydrolases and phospholipid synthases that includes poxvirus envelope proteins. *Trends Biochem. Sci.*, **21**, 242–243.
7. Ponting,C.P. and Kerr,I.D. (1996) A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: identification of duplicated repeats and potential active site residues. *Protein Sci.*, **5**, 914–922.
8. Anantharaman,V., Koonin,E.V. and Aravind,L. (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.*, **30**, 1427–1464.
9. Ponting,C.P., Aravind,L., Schultz,J., Bork,P. and Koonin,E.V. (1999) Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.*, **289**, 729–745.
10. Phizicky,E.M. and Greer,C.L. (1993) Pre-tRNA splicing: variation on a theme or exception to the rule? *Trends Biochem. Sci.*, **18**, 31–34.
11. Culver,G.M., McCraith,S.M., Zillmann,M., Kierzek,R., Michaud,N., LaReau,R.D., Turner,D.H. and Phizicky,E.M. (1993) An NAD derivative produced during transfer RNA splicing: ADP-ribose 1"-2" cyclic phosphate. *Science*, **261**, 206–208.
12. Culver,G.M., Consaul,S.A., Tycowski,K.T., Filipowicz,W. and Phizicky,E.M. (1994) tRNA splicing in yeast and wheat germ. A cyclic phosphodiesterase implicated in the metabolism of ADP-ribose 1",2"-cyclic phosphate. *J. Biol. Chem.*, **269**, 24928–24934.
13. Westaway,S.K. and Abelson,J. (1995) Splicing of tRNA precursors. In Soel,D. and RajBhandary,U.L. (eds), *tRNA: Structure, Biosynthesis and Function*. American Society for Microbiology, Washington DC, pp. 79–92.
14. Xu,Q., Teplow,D., Lee,T.D. and Abelson,J. (1990) Domain structure in yeast tRNA ligase. *Biochemistry*, **29**, 6132–6138.
15. Amitsur,M., Levitz,R. and Kaufmann,G. (1987) Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA. *EMBO J.*, **6**, 2499–2503.
16. Koonin,E.V. and Gorbalenya,A.E. (1990) Related domains in yeast tRNA ligase, bacteriophage T4 polynucleotide kinase and RNA ligase and mammalian myelin 2′,3′-cyclic nucleotide phosphohydrolase revealed by amino acid sequence comparison. *FEBS Lett.*, **268**, 231–234.
17. Nasr,F. and Filipowicz,W. (2000) Characterization of the *Saccharomyces cerevisiae* cyclic nucleotide phosphodiesterase involved in the metabolism of ADP-ribose 1",2"-cyclic phosphate. *Nucleic Acids Res.*, **28**, 1676–1683.
18. Martzen,M.R., McCraith,S.M., Spinelli,S.L., Torres,F.M., Fields,S., Grayhack,E.J. and Phizicky,E.M. (1999) A biochemical genomics approach for identifying genes by the activity of their products. *Science*, **286**, 1153–1155.

19. Tyc,K., Kellenberger,C. and Filipowicz,W. (1987) Purification and characterization of wheat germ 2′,3′-cyclic nucleotide 3′-phosphodiesterase. *J. Biol. Chem.*, **262**, 12994–30000.
20. Genschik,P., Hall,J. and Filipowicz,W. (1997) Cloning and characterization of the *Arabidopsis* cyclic phosphodiesterase which hydrolyzes ADP-ribose 1",2"-cyclic phosphate and nucleoside 2′,3′-cyclic phosphates. *J. Biol. Chem.*, **272**, 13211–13219.
21. Edgell,D.R., Belfort,M. and Shub,D.A. (2000) Barriers to intron promiscuity in bacteria. *J. Bacteriol.*, **182**, 5281–5289.
22. Reinhold-Hurek,B. and Shub,D.A. (1992) Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature*, **357**, 173–176.
23. Greer,C.L., Javor,B. and Abelson,J. (1983) RNA ligase in bacteria: formation of a 2′,5′ linkage by an *E.coli* extract. *Cell*, **33**, 899–906.
24. Arn,E.A. and Abelson,J.N. (1996) The 2′-5′ RNA ligase of *Escherichia coli*. Purification, cloning and genomic disruption. *J. Biol. Chem.*, **271**, 31145–31153.
25. Sprinkle,T.J. (1989) 2′,3′-cyclic nucleotide 3′-phosphodiesterase, an oligodendrocyte-Schwann cell and myelin-associated enzyme of the nervous system. *Crit. Rev. Neurobiol.*, **4**, 235–301.
26. Olafson,R.W., Drummond,G.I. and Lee,J.F. (1969) Studies on 2′,3′-cyclic nucleotide-3′-phosphohydrolase from brain. *Can. J. Biochem.*, **47**, 961–966.
27. Hofmann,A., Tarasov,S., Grella,M., Ruvinov,S., Nasr,F., Filipowicz,W. and Wlodawer,A. (2002) Biophysical characterization of cyclic nucleotide phosphodiesterases. *Biochem. Biophys. Res. Commun.*, **291**, 875–883.
28. Hofmann,A., Zdanov,A., Genschik,P., Ruvinov,S., Filipowicz,W. and Wlodawer,A. (2000) Structure and mechanism of activity of the cyclic phosphodiesterase of Appr>p, a product of the tRNA splicing reaction. *EMBO J.*, **19**, 6207–6217.
29. Ballestero,R.P., Dybowski,J.A., Levy,G., Agranoff,B.W. and Uhler,M.D. (1999) Cloning and characterization of zRICH, a 2′,3′-cyclic-nucleotide 3′-phosphodiesterase induced during zebrafish optic nerve regeneration. *J. Neurochem.*, **72**, 1362–1371.
30. Hofmann,A., Grella,M., Botos,I., Filipowicz,W. and Wlodawer,A. (2002) Crystal structures of the semireduced and inhibitor-bound forms of cyclic nucleotide phosphodiesterase from *Arabidopsis thaliana*. *J. Biol. Chem.*, **277**, 1419–1425.
31. Francis,S.H., Colbran,J.L., McAllister-Lucas,L.M. and Corbin,J.D. (1994) Zinc interactions and conserved motifs of the cGMP-binding cGMP-specific phosphodiesterase suggest that it is a zinc hydrolase. *J. Biol. Chem.*, **269**, 22477–22480.
32. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
33. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
34. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
35. Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.
36. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
37. Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pacific Symposium on Biocomputing*, *Hawaii*. World Scientific, Singapore, pp. 119–130.
38. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
39. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.
40. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
41. Wolf,Y.I., Rogozin,I.B., Grishin,N.V., Tatusov,R.L. and Koonin,E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, **1**, 8.
42. Felsenstein,J. (1996) Inferring phylogenies from protein sequences by parsimony, distance and likelihood methods. *Methods Enzymol.*, **266**, 418–427.
43. Hasegawa,M., Kishino,H. and Saitou,N. (1991) On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.*, **32**, 443–445.
44. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
45. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. App. Crystallogr.*, **24**, 946–950.
46. Murzin,A.G. and Bateman,A. (2001) CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. *Proteins*, **45**, 76–85.
47. Aravind,L. and Koonin,E.V. (2001) A natural classification of ribonucleases. *Methods Enzymol.*, **341**, 3–28.
48. Aravind,L., Tatusov,R.L., Wolf,Y.I., Walker,D.R. and Koonin,E.V. (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.*, **14**, 442–444.
49. Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A. *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*, **399**, 323–329.
50. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
51. Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.
52. Huynen,M., Snel,B., Lathe,W.,III and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
53. Iyer,L.M., Koonin,E.V. and Aravind,L. (2002) Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52. *BMC Genomics*, **3**, 8.
54. Lykke-Andersen,J., Aagaard,C., Semionenkov,M. and Garrett,R.A. (1997) Archaeal introns: splicing, intercellular mobility and evolution. *Trends Biochem. Sci.*, **22**, 326–331.
55. Thompson,L.D., Brandon,L.D., Nieuwlandt,D.T. and Daniels,C.J. (1989) Transfer RNA intron processing in the halophilic archaebacteria. *Can. J. Microbiol.*, **35**, 36–42.
56. Broida,J. and Abelson,J. (1985) Sequence organization and control of transcription in the bacteriophage T4 tRNA region. *J. Mol. Biol.*, **185**, 545–563.
57. Aravind,L. and Koonin,E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J. Mol. Biol.*, **287**, 1023–1040.
58. Lodish,H., Baltimore,D., Berk,A., Zipursky,S.L., Matsudaira,P. and Darnell,J. (1995) *Molecular Cell Biology*. W.H. Freeman and Co., New York, NY.
59. Filipowicz,W., Billy,E., Drabikowski,K. and Genschik,P. (1998) Cyclases of the 3′-terminal phosphate in RNA: a new family of RNA processing enzymes conserved in eucarya, bacteria and archaea. *Acta Biochim. Pol.*, **45**, 895–906.
60. Spinelli,S.L., Malik,H.S., Consaul,S.A. and Phizicky,E.M. (1998) A functional homolog of a yeast tRNA splicing enzyme is conserved in higher eukaryotes and in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **95**, 14136–14141.
61. Lichtarge,O., Sowa,M.E. and Philippi,A. (2002) Evolutionary traces of functional surfaces along G protein signaling pathway. *Methods Enzymol.*, **344**, 536–556.
62. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
63. Iyer,L.M., Koonin,E.V. and Aravind,L. (2002) Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. *Genome Biol.*, **3**, research 0012.1–0012.11.
64. Garcia,A.D., Aravind,L., Koonin,E.V. and Moss,B. (2000) Bacterial-type DNA holliday junction resolvases in eukaryotic viruses. *Proc. Natl Acad. Sci. USA*, **97**, 8926–8931.
65. Jung,D.J., Sung,H.S., Goo,Y.W., Lee,H.M., Park,O.K., Jung,S.Y., Lim,J., Kim,H.J., Lee,S.K., Kim,T.S. *et al.* (2002) Novel transcription coactivator complex containing activating signal cointegrator 1. *Mol. Cell. Biol.*, **22**, 5203–5211.

66. Kim,H.J., Yi,J.Y., Sung,H.S., Moore,D.D., Jhun,B.H., Lee,Y.C. and Lee,J.W. (1999) Activating signal cointegrator 1, a novel transcription coactivator of nuclear receptors and its cytosolic localization under conditions of serum deprivation. *Mol. Cell. Biol.*, **19**, 6323–6332.
67. Ponting,C.P. (2000) Proteins of the endoplasmic-reticulum-associated degradation pathway: domain detection and function prediction. *Biochem. J.*, **351**, 527–535.
68. van Nues,R.W. and Beggs,J.D. (2001) Functional contacts with a range of splicing proteins suggest a central role for Brr2p in the dynamic control of the order of events in spliceosomes of *Saccharomyces cerevisiae*. *Genetics*, **157**, 1451–1467.
69. Fraser,I.D., Tavalin,S.J., Lester,L.B., Langeberg,L.K., Westphal,A.M., Dean,R.A., Marrion,N.V. and Scott,J.D. (1998) A novel lipid-anchored A-kinase anchoring protein facilitates cAMP-responsive membrane events. *EMBO J.*, **17**, 2261–2272.
70. Trendelenburg,G., Hummel,M., Riecken,E.O. and Hanski,C. (1996) Molecular characterization of AKAP149, a novel A kinase anchor protein with a KH domain. *Biochem. Biophys. Res. Commun.*, **225**, 313–319.
71. Aravind,L., Watanabe,H., Lipman,D.J. and Koonin,E.V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl Acad. Sci. USA*, **97**, 11319–11324.
72. Shuman,S. and Schwer,B. (1995) RNA capping enzyme and DNA ligase: a superfamily of covalent nucleotidyl transferases. *Mol. Microbiol.*, **17**, 405–410.
73. Chen,D., Luongo,C.L., Nibert,M.L. and Patton,J.T. (1999) Rotavirus open cores catalyze 5′-capping and methylation of exogenous RNA: evidence that VP3 is a methyltransferase. *Virology*, **265**, 120–130.
74. Sawicki,S.G. and Sawicki,D.L. (1998) A new model for coronavirus transcription. *Adv. Exp. Med. Biol.*, **440**, 215–219.
75. Metcalf,W.W. and Wanner,B.L. (1993) Evidence for a fourteen-gene, phnC to phnP locus for phosphonate metabolism in *Escherichia coli*. *Gene*, **129**, 27–32.
76. Wattenhofer,M., Shibuya,K., Kudoh,J., Lyle,R., Michaud,J., Rossier,C., Kawasaki,K., Asakawa,S., Minoshima,S., Berry,A. *et al.* (2001) Isolation and characterization of the UBASH3A gene on 21q22.3 encoding a potential nuclear protein with a novel combination of domains. *Hum. Genet.*, **108**, 140–147.
77. Pawson,T. and Scott,J.D. (1997) Signaling through scaffold, anchoring and adaptor proteins. *Science*, **278**, 2075–2080.
78. Hofmann,K. and Bucher,P. (1996) The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. *Trends Biochem. Sci.*, **21**, 172–173.
79. Kruh,G.D., Zeng,H., Rea,P.A., Liu,G., Chen,Z.S., Lee,K. and Belinsky,M.G. (2001) MRP subfamily transporters and resistance to anticancer agents. *J. Bioenerg. Biomembr.*, **33**, 493–501.
80. Doolittle,R.F. and Handy,J. (1998) Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr. Opin. Genet. Dev.*, **8**, 630–636.
81. Leipe,D.D., Wolf,Y.I., Koonin,E.V. and Aravind,L. (2002) Classification and evolution of P-loop GTPases and related ATPases. *J. Mol. Biol.*, **317**, 41–72.
82. Aravind,L. and Koonin,E.V. (1999) DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucleic Acids Res.*, **27**, 1609–1618.