BMC Genomics

BioMed Central

Research article

Open Access

The catalytic domains of thiamine triphosphatase and CyaB-like adenylyl cyclase define a novel superfamily of domains that bind organic phosphates

Lakshminarayan M Iyer and L Aravind*

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

E-mail: Lakshminarayan M Iyer - lakshmin@ncbi.nlm.nih.gov; L Aravind* - aravind@ncbi.nlm.nih.gov

*Corresponding author

Published: 27 November 2002

BMC Genomics 2002, **3**:33

Received: 19 September 2002

Accepted: 27 November 2002

This article is available from: http://www.biomedcentral.com/1471-2164/3/33

Abstract

Background: The CyaB protein from *Aeromonas hydrophila* has been shown to possess adenylyl cyclase activity. While orthologs of this enzyme have been found in some bacteria and archaea, it shows no detectable relationship to the classical nucleotide cyclases. Furthermore, the actual biological functions of these proteins are not clearly understood because they are also present in organisms in which there is no evidence for cyclic nucleotide signaling.

Results: We show that the CyaB like adenylyl cyclase and the mammalian thiamine triphosphatases define a novel superfamily of catalytic domains called the CYTH domain that is present in all three superkingdoms of life. Using multiple alignments and secondary structure predictions, we define the catalytic core of these enzymes to contain a novel $\alpha+\beta$ scaffold with 6 conserved acidic residues and 4 basic residues. Using contextual information obtained from the analysis of gene neighborhoods and domain fusions, we predict that members of this superfamily may play a central role in the interface between nucleotide and polyphosphate metabolism. Additionally, based on contextual information, we identify a novel domain (called CHAD) that is predicted to functionally interact with the CYTH domain-containing enzymes in bacteria and archaea. The CHAD is predicted to be an alpha helical domain, and contains conserved histidines that may be critical for its function.

Conclusions: The phyletic distribution of the CYTH domain suggests that it is an ancient enzymatic domain that was present in the Last Universal Common Ancestor and was involved in nucleotide or organic phosphate metabolism. Based on the conservation of catalytic residues, we predict that CYTH domains are likely to chelate two divalent cations, and exhibit a reaction mechanism that is dependent on two metal ions, analogous to nucleotide cyclases, polymerases and certain phosphoesterases. Our analysis also suggests that the experimentally characterized members of this superfamily, namely adenylyl cyclase and thiamine triphosphatase, are secondary derivatives of proteins that performed an ancient role in polyphosphate and nucleotide metabolism.

Background

Organic phosphate compounds are the central metabolites of all biological systems [1,2]. Some are the basic building blocks of nucleic acids, some like ATP and GTP,

are additionally, cellular energy stores, others like cAMP or cGMP are messengers in signal transduction, and, yet others, such as FAD, NAD, thiamine phosphates and pyridoxal phosphate are cofactors for a range of enzymes [1,2]. Protein domains belonging to a relatively small set of structural folds are known to bind or catalyze reactions that utilize these organic phosphate compounds (see SCOP database: [http://scop.mrc-lmb.cam.ac.uk/scop/]) [3,4]. Several of these folds trace back to some of the earliest phases of the evolution of the protein world, and participate in a wide range of disparate biological functions in extant proteins [4,5]. Some folds, such as the P-loop fold, the Rossmann fold and the Hsp70-like fold, have been well studied, and comprise mainly of dedicated nucleotide binding or hydrolyzing proteins [6–9]. Others, such as the palm-domain, which is found in adenylyl cyclases and various nucleic acid polymerases, belong to more generalized protein folds that contain representatives with diverse biochemical activities [4,10,11]. Current availability of extensive genome sequence data, allows one to identify less numerous, nevertheless biological important organic phosphate-binding domains that may have previous eluded detection. The identification of such domain superfamilies, containing enzymes with several different activities, often throws considerable light on their evolution, structure and catalytic mechanisms [4,12].

The majority of previously known nucleotide cyclases belong to two major folds. The classical adenylyl cyclases, guanylyl cyclases and the GGDEF (diguanylate cyclase) domains share the catalytic palm domain with the family B DNA polymerases, reverse transcriptases, viral RNA dependent RNA polymerases and eukaryote-type primases [4,13,14]. The pathogenic adenylyl cyclases of several bacteria and the CyaA-like proteobacteria adenylyl cyclases are extremely divergent versions of the catalytic domain seen in the Pol-β family of nucleotidyl transferases [15] (also see SCOP database: [http://scop.mrc-lmb.cam.ac.uk/scop/]). While the catalytic domains of these two superfamilies have very different folds, they follow a similar reaction mechanism that is dependent on two Mg^{2+} ions coordinated by a cluster of acidic residues. However, the CyaB adenylyl cyclase, which was identified in the bacterium *Aeromonas hydrophila* is unrelated to any of these above superfamilies of enzymes [16]. Though close relatives of this enzyme exist in some bacteria, like *Yersinia pestis* and *Borrelia burgdorferi* and the archaea, its antecedents or catalytic mechanism have not been understood. Using sensitive sequence profile comparison methods, we show that the CyaB-like adenylyl cyclases are homologs of the soluble mammalian thiamine triphosphatases [17], and they define a novel superfamily of enzymes that utilize ATP and other organic phosphates. We present evidence that a representative of this domain was present in the last common ancestor of all extant life forms. The primary biological function of these proteins appears to be related to polyphosphate and nucleotide metabolism. Cyclic AMP generation and thiamine triphosphate hydrolysis appear to be secondarily acquired activities. We also identify the potential active site- and substrate interacting-residues and postulate that these enzymes are likely to catalyze a two-metal ion dependent reaction on structural scaffold that is completely different from that seen in the other two superfamilies of adenylyl cyclases.

Results and discussion

Identification of the CYTH domain

In order to understand the evolutionary affinities and provenance of the CyaB adenylyl cyclase from *Aeromonas hydrophila*, we carried out database searches using sensitive sequence profile analysis methods. As CyaB is a small protein with no detectable low complexity regions, we used it as a seed to initiate a PSI-BLAST search [18] (run to convergence, with expect-value for inclusion in profile = .01). The search resulted in the detection of its obvious orthologs from *Yersinia* and various archaea at significant expect (e) values ranging from from 3×10^{-43} to 8×10^{-5}. The second iteration recovered proteins from more archaea, eukaryotes (e-value: 3×10^{-7}), *Clostridium* (3×10^{-8}) and *Borrelia* (6×10^{-8}). Further iterations, run to convergence, recovered the soluble mammalian thiamine triphosphatases (3×10^{-4}), and the N-terminal region of *E. coli* YgiF. At convergence, several bacterial proteins, that showed a conserved EXEXK (where X is any amino acid) characteristic of this family, and a region C-terminal to a P-loop like uridine kinase domain in plants were also recovered at borderline e-values (e value ~0.01 – 0.05). Reciprocal searches initiated with the *E. coli* YgiF protein (residues 1–200), not only recovered its bacterial orthologs, archaeal CyaB homologs and eukaryotic proteins, but also several others such as *Bacillus subtilis* YjbK, *Methanosarcina* Ma2350, and *Mesorhizobium loti* Mll4592 with e-values in the range of 10^{-4}–10^{-6} upon first detection. Additionally, transitive searches initiated from the region C-terminal to the uridine kinase of the plants (49D11.13 from *Oryza sativa*, region: 250–410) recovered archaeal CyaB homologs confirming their relationship to with the other proteins detected in these searches. Regular expression searches with the conservation pattern found in these CyaB homologs also recovered the most of the members detected in the above-mentioned profile searches, but failed to recover any new candidates.

In all these searches, the alignments more or less spanned the entire length of the CyaB protein and typically contained the same set of conserved residues. The Gibbs sampling procedure revealed the presence of seven conserved motifs, with a probability of chance occurrence less than

10^{-14}, in the search space comprising of the 70 or so proteins that were identified in the above searches as having this domain. We clustered these proteins using the BLASTCLUST program in several smaller clusters and prepared multiple alignments for the individual clusters and predicted secondary structure for these set using the PHD program. A nearly complete congruence was seen in the comparison of the secondary structures of the individual clusters. In many cases, the region of similarity to CyaB comprised the entire length of the target protein detected in these searches. However, in some cases it only comprised a part of the protein, with rest of the protein being made of other globular domains. These observations, taken together, suggested that CyaB and soluble mammalian thiamine triphosphatases define a novel superfamily of conserved domains, which may either occur by itself or in combination with other domains. We named this domain the CYTH (CyaB, thiamine triphosphatase) domain after the two experimentally characterized proteins in which it is present.

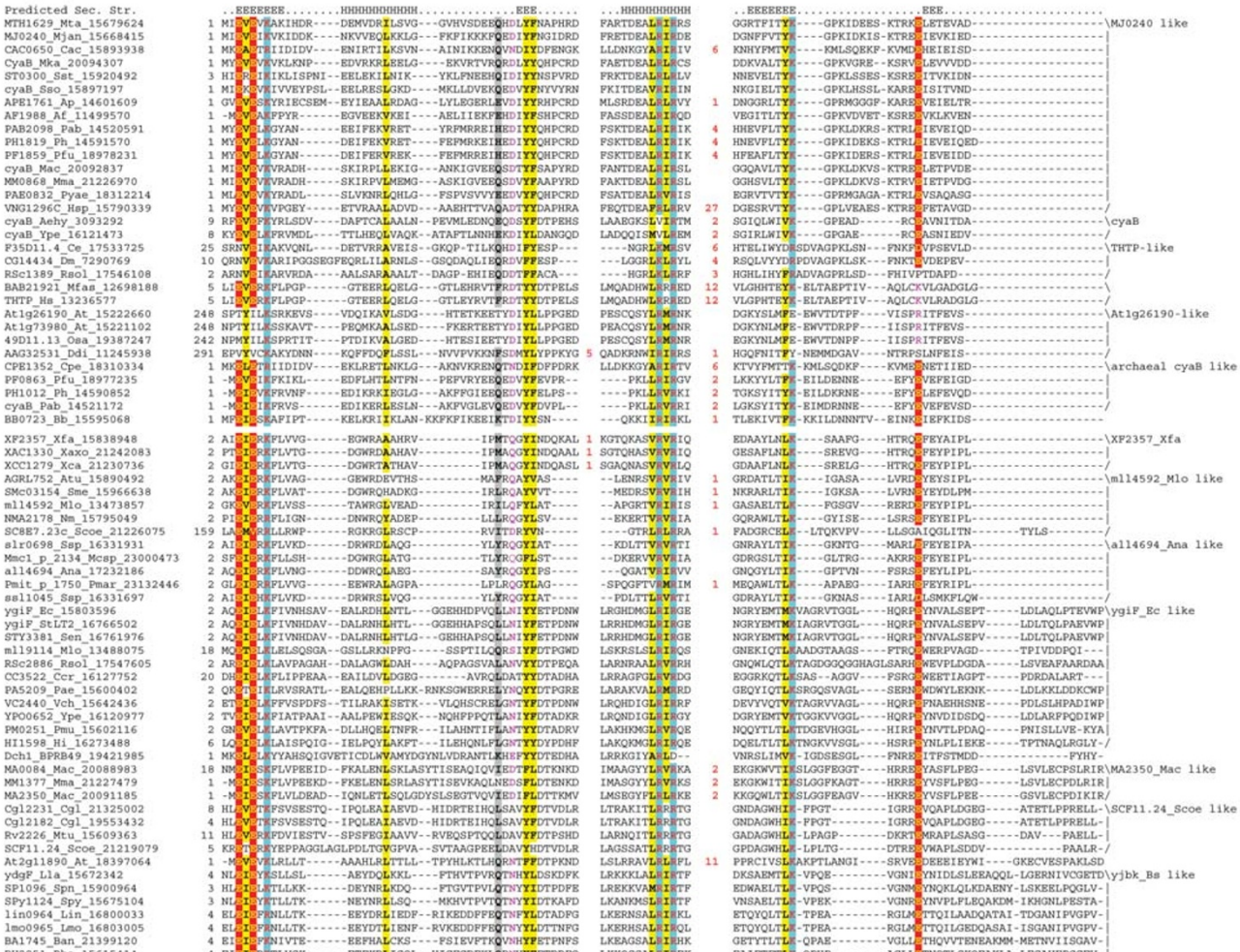
Sequence conservation, structure and biochemical activities of the CYTH domain

All complete CYTH domain sequences were aligned using the T_Coffee program, and this alignment was further refined based on the PSI-BLAST HSPs, conserved motifs detected in the Gibbs sampling procedure and predicted secondary structure for the individual groups (Fig. 1). A text copy of the alignment is being provided as an additional file (see additional file1 and additional file 2). The predicted secondary structure for this domain indicated an α+β fold, with 6 conserved β-strands and 6 conserved α-helices. Neither the predicted secondary structure, nor the pattern of the conserved residues revealed an obvious relationship with any previously recognized fold. Given that the CYTH domain is the only globular domain in the enzymes, thiamine triphosphatase and CyaB, it is predicted to be an enzymatic domain. While the reactions catalyzed by these two enzymes are distinct, their substrates, respectively thiamine triphosphate and ATP, are both organic triphosphates. The CYTH domain is also present C-terminal to the catalytic P-loop containing domain in the plant and slime mold uridine kinase homologs. In these proteins it is likely to interact with nucleoside diphosphates or triphosphates, which are substrates for these kinases. These observations suggest that the CYTH domains are likely to be domains specialized to bind nucleotides and other organic phosphates. A multiple alignment of this superfamily reveals the presence of several nearly universally conserved charged residues that are likely to form the active site of these enzymes (Fig. 1). The most prominent of these are an EXEXK motif associated with strand-1 of the domain, two basic residues in helix-2, a K at the end of strand 3, an E in strand 4, a basic residue in helix-4, a D at the end of strand 5 and two acidic residues (typically glutamates) in strand 6 (Fig. 1). The presence of around 6 conserved acidic positions in the majority of the CYTH domains suggests that it coordinates two divalent metal ions. Analogous active sites, that coordinate two metal ions, are observed in domains with similar activities, such as the classical adenylyl/guanylyl cyclases, family B DNA polymerases, pol-β fold nucleotidyl transferases, and triphosphatases or phosphoesterases of the HD and DHH superfamilies [11,15,19,20]. Consistent with these observations, both CyaB and ThTPase have been shown to require Mg^{2+} ions for their nucleotide cyclase and phosphatase activities [16,17].

The four conserved basic residues in the CYTH domain are most probably involved in the binding of acidic phosphate moieties of their substrates (Fig. 1). The conservation of these two sets of residues in the majority of CYTH domains suggests that most members of this group are likely to possess an activity dependent on two metal ions, with a preference for nucleotides or related phosphate-moiety-bearing substrates. The proposed biochemical activity, and the arrangement of predicted strands in the primary structure of the CYTH domain imply that the may adopt a barrel or sandwich-like configuration, with metal ions and the substrate bound in the central cavity. The only prominent exceptions to the basic conservation pattern of the CYTH domains are the versions found in the plant and *Dictyostelium* pyrimidine kinase homologs (Fig. 1, At1g26190-like). These versions lack 5 of the 6 conserved acidic residues, but retain 3 of the 4 conserved basic residues (Fig. 1). This leads to the prediction that these CYTH domains are catalytically inactive. However, as they retain the basic residues, they probably still bind the organic phosphate substrates, and function as regulatory domains that are linked to P-loop kinase domains.

Phyletic patterns, evolutionary history and potential biological functions of the CYTH domains

CYTH domains are present in most of the major lineages, for which sequence information is currently available, from the three principal superkingdomains of life (Fig. 2). We used the multiple alignment of the CYTH domain to construct phylogenetic trees using the least squares, neighbor joining and maximum likelihood methods (see Materials and Methods). The monophyletic clusters that emerged in this analysis were essentially the same as those that were derived through similarity based clustering using BLASTCLUST. The majority of archaeal and eukaryotic proteins formed a monophyletic cluster to the exclusion of most of the bacterial proteins (RELL Bootstrap support 77%) (Fig. 2). This cluster was also supported by a unique synapomorphy (a shared derived character) in the form of a conserved motif (Dh; where h is a hydrophobic residue) associated with the second strand (Fig. 1). This phylogenetic tree topology resembles that of several proteins in-



(Figure 1 - Continued on next page)

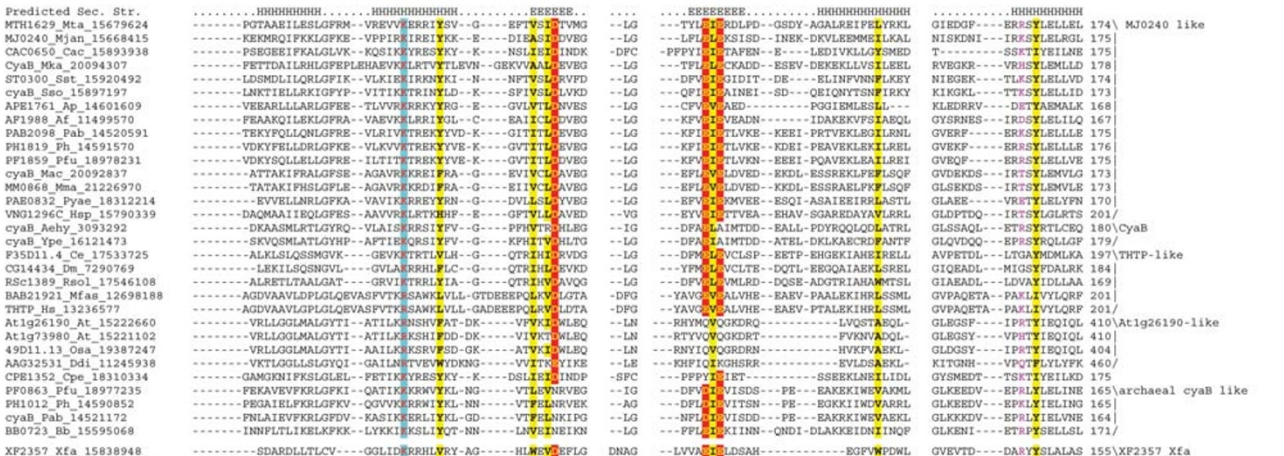


Figure 1
Multiple alignment of CYTH domains Proteins are represented by their corresponding gene names, followed by a species abbreviation, followed by the Genbank gi number. The coloring reflects the amino acid conservation at 90% consensus. The consensus abbreviations and coloring scheme are as follows: h: hydrophobic residues (L,I,Y,F,M,W,A,C,V), l: aliphatic (L,I,A,V) and a: aromatic (F,Y,W,H) residues shaded yellow, c: charged (K,E,R,D,H) residues, and p: polar (S,T,E,C,D,R,K,H,N,Q) residues colored purple; +: basic (K,R,H) residues shaded blue, -: acidic (D,E) residues shaded red, s: small (S,A,C,G,D,N,P,V,T) and u:tiny (G,A,S) residues, colored green; b: big (L,I,F,M,W,Y,E,R,K,Q) residues shaded gray. Secondary structure assignments: H: Helix, E: Extended (Strand).

volved in core cellular functions such as the DNA recombinase RecA, aminoacyl tRNA synthetases, RNA polymerase and other RNA metabolism proteins [21–23]. This suggests that a CYTH domain was present in the last universal common ancestor of all extant life forms and the extant forms are in part vertically inherited from this ancestral form.

However, there are certain anomalies to this pattern. The CYTH domains are entirely absent from the small genomes of pathogenic bacteria such as *Rickettsia* and *Chlamydia* as well as some of the large genomes such as *Deinococcus radiodurans*. At least, a single copy of the CYTH domain is seen in most archaeal and eukaryotic genomes sampled to date, with the exception of *Thermoplasma* and the yeasts, where it is absent. This implies that the CYTH domain has been lost independently on a number of occasions in evolution. CyaB homologs from *Aeromonas*, *Clostridium*, *Borrelia*, and *Ralstonia* lie firmly (RELL Bootstrap >= 70%) within the archaeal and eukaryotic clusters, rather than with their bacterial counterparts (Fig. 2). These bacterial forms also share the unique sequence signature in the sec-

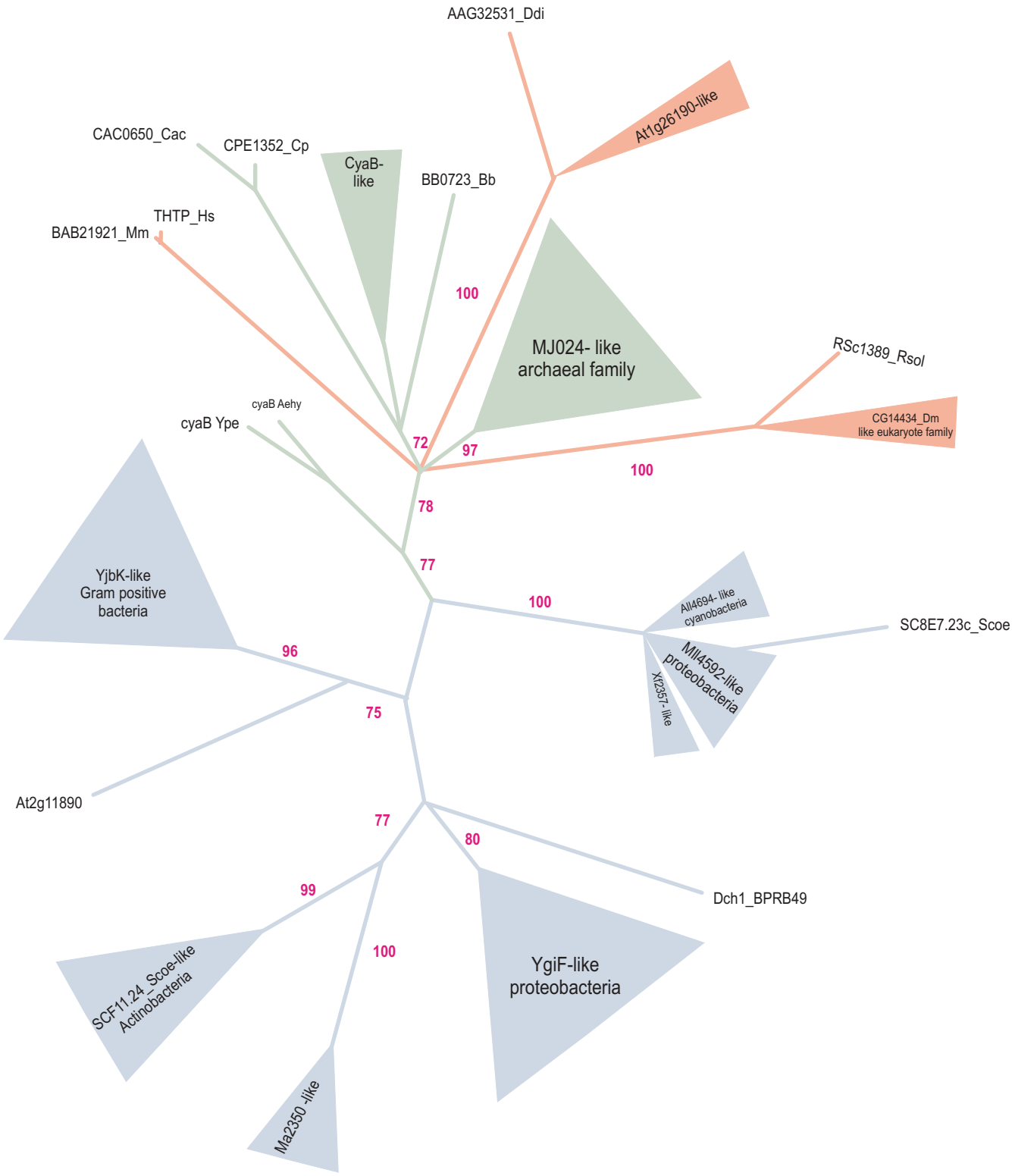


Figure 2
Maximum-likelihood phylogenetic tree of CYTH domains. RELL bootstrap values are shown below the branches and branches with bootstrap values <50% are collapsed. The protein nomenclature follows the convention given in the legend to Fig. 1.

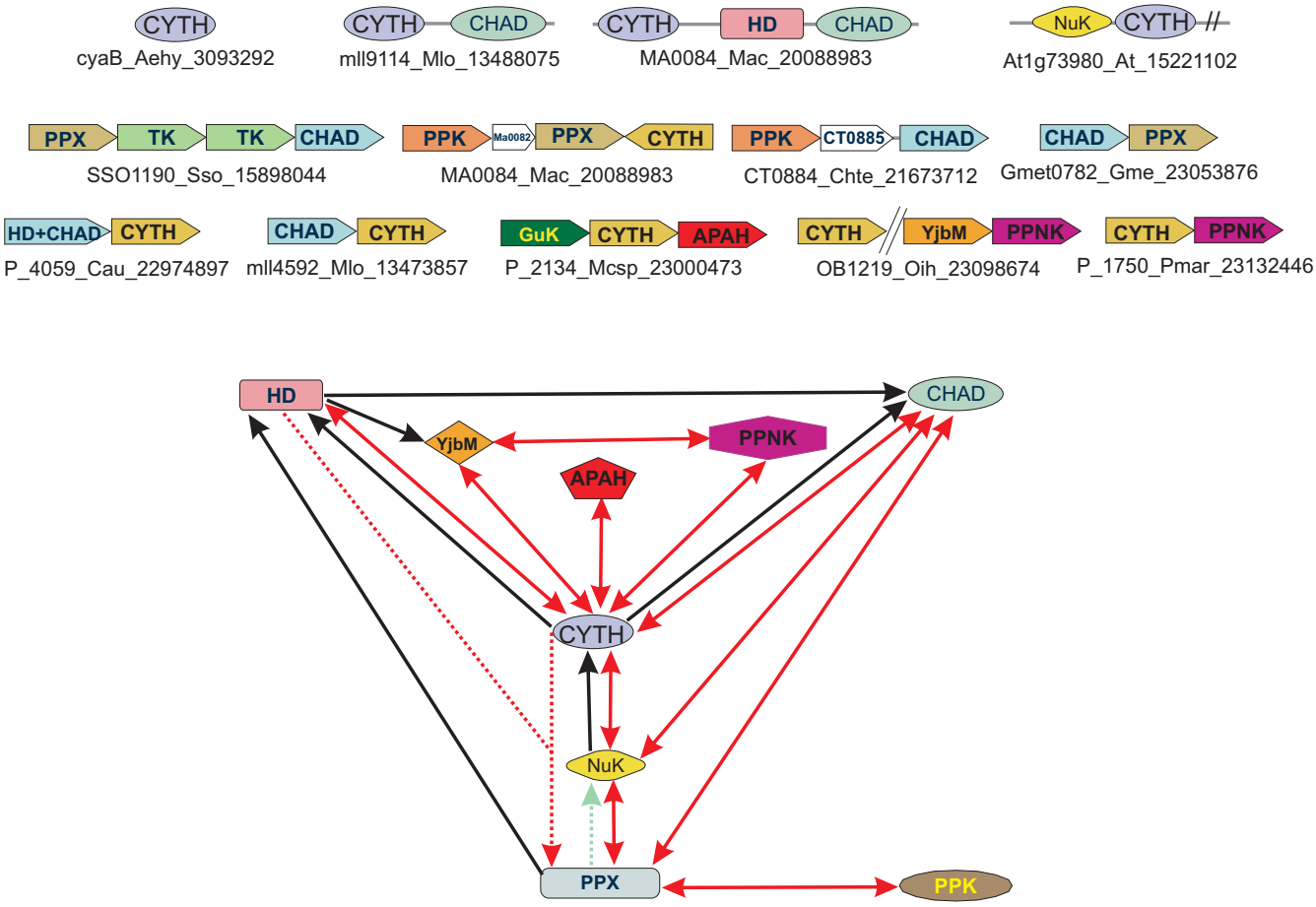


Figure 3
Domain architecture, predicted operons and contextual information map for CYTH domains Proteins are represented by their gene names, species abbreviations and gi as in Fig. 1. Operons are shown with genes represented as box-arrows. The contextual map shows different types of associations between the domains. Unidirectional black arrows represent domains co-occuring in the same protein. Bidirectional red arrows represent domains co-occuring in operons, the dotted red arrow represents adjacent gene transcribed in opposite directions, and the green arrow represents an experimentally derived functional association. Domain abbreviations: GuK, Guanylate kinase, NuK, Nucleotide kinase, TK, Thymidylate kinase.

ond strand with this group suggesting that they have been derived through horizontal transfer from different archaeal and eukaryotic sources (Fig. 1). In particular the CyaB homolog from *Ralstonia* groups very strongly with the animal versions and appears to be a recent horizontal transfer in this bacterium from the latter clade. The possibility of lateral transfer of *Aeromonas* CyaB from an archaeal source has been previously suggested, and is consistent with the enzyme being optimally functional under high temperature [16]. There are 3 distinct CYTH domains in the euryarchaeon *Methanosarcina acetivorans*, in addition to the version which groups with the CYTH domains that are found, in single or duplicate copies, in other archaea and eukaryotes. These former versions, strongly group with CYTH domains from actinobacteria (Rell Boostrap 77%) to the exclusion of other lineage (Fig. 2). Furthermore, they share a fusion to a novel conserved domain with characteristic histidines (see below) with the actinobacterial versions. Thus they appear to have been transferred laterally from the actinobacteria into the *Methanosarcina* lineage followed by a small lineage specific expansion in the latter. Bacteriophages, like RB49, that contain a solo CYTH domain, which is closer to the version seen in its proteobacterial hosts, could have served as conduits for the lateral distribution of this domain.

The phyletic pattern of the CYTH domain is not very typical of signaling enzymes like nucleotide cyclases. Classic adenylyl and guanylyl cyclases show a far more sporadic distribution, and are often present in multiple copies fused to a variety of signaling domains such as the cyclic nucleotide binding domains [24]. Cyclic nucleotide gen-

erating activity is not known to exist in a subset of the archaea [25], though most of them contain a well-conserved copy of the CYTH domain. Experimental analysis on the *Methanococcus* CyaB homolog revealed no adenylyl cyclase activity comparable to that seen in *Aeromonas* CyaB [16]. Likewise, there is no evidence for a widespread presence of the ThTPase activity in the organisms that contain CYTH domain proteins [17]. Hence, the principal biological function of the CYTH domains may be different from those of the experimentally characterized members of this family. General conservation of this domain across a range of organisms, and its presence in the LUCA suggests the possibility of an important general role in cellular process of free-living organisms. In order to decipher the potential biological roles of this domain we used different forms of contextual information regarding these domains, in the form of domain architectures, gene neighborhoods (predicted operons) and experimental evidence for interactions between proteins. Both domain architecture and operon analysis have been used extensively to make functional predictions of poorly characterized domains or genes [26–29]. Moreover, the presence of evolutionarily conserved operons often correlates with the involvement of the component genes in a sequential pathway or physically interacting complex [30,31]. We summarize the different forms of contextual information that we extracted from the CYTH domains in the form a network diagram (Fig. 3).

The CYTH domain shows a small array set of fusions to other conserved domains (Fig. 3). One of the most prevalent fusions is to an uncharacterized domain, with a characteristic pattern of conserved histidines and other charged residues. This domain is predicted to adopt an α-helical fold and is according referred to as CHAD (CHAD: conserved histidine α-helical domain; Fig. 4). The sequence conservation pattern (Fig. 4) suggests that this domain is likely to contain two repeat units, with at least 4 helices each, at its core. While no clear functional prediction can be made for the CHAD, the conserved charged residues could form a strongly polar surface that could participate, either in metal chelation, or act as phosphoacceptors. Another notable fusion is with a specific version of the HD hydrolase domain [19], which is also found fused to the C-terminus of the HSP70-fold domain in the exopolyphosphatase (PPX) (Fig. 3). HD domains typically possess phosphoesterase activity, and are fused to catalytic domains that possess nucleotide kinase, nucleotidyltransferase, nucleotide cyclase or diguanylate cyclase activity [19,29]. This fits well with the observed cyclase activity seen in CyaB, but is also consistent with phosphotransferase or nucleotidyl transferase for the CYTH domain. Finally, catalytically inactive versions of the CYTH domain are found fused to the nucleotide kinase domain in uridine kinase homologs and may serve as an allosteric nucleotide-binding site in these enzymes (Fig. 3).

In terms of conserved gene neighborhoods, CYTH-domain-encoding genes, like mll4592 from α-proteobacteria, are frequently found in the neighborhood of genes encoding solo CHADs (Fig. 3). Additionally, in *Methanosarcina* the CYTH-encoding genes are found in predicted operons or in the neighborhood of genes encoding exopolyphosphatase (PPX) and polyphosphate kinase (PPK). Furthermore, in *Sulfolobus*, a gene for a CHAD protein is in a predicted operon along with genes for thymidylate kinase and PPX. CHAD- and CTYH-encoding genes are also found in the neighborhood of PPK and PPX in *Chlorobium tepidum* and *Geobacter metallireducens*, respectively (Fig. 3). Genes for CYTH domains also co-occur in predicted operons with genes for another polyphosphate utilizing enzyme, the polyphosphate-dependent NAD kinase (PPNK), in certain cyanobacteria (eg. *Prochlorococcus marinus*) and Gram positive bacteria like *Oceanobacillus iheyensis*. Other potential connections are furnished by the co-occurrence of genes for CYTH-domain proteins with genes involved in with nucleoside polyphosphate metabolism. These include co-occurrence with the gene for adenosine tetraphosphatase (APAH; eg. in *Magnetococcus sp.*) and with genes encoding the YjbM-domain in Gram-positive bacteria. The YjbM domain, most often, occurs fused to pentaphosphate guanosine-3'-pyrophosphohydrolase (SpoT) and GTP pyrophosphokinase (RelA), suggesting a role for it in the metabolism of the stringent-response nucleoside polyphosphate.

These contextual connections are consistent with the participation of CYTH domains in organo-phosphate biochemistry, and circumstantially associate it with the metabolic network related to polyphosphates and nucleoside polyphospates (Fig. 3) [32]. Specifically, PPK and PPX have been shown to, respectively, lengthen or shorten polyphosphate polymers [32]. These two enzymes also appear to interact with the nucleotide metabolism of the cell. In particular PPK and PPNK can utilize Poly(P) to synthesize nucleoside polyphosphates, while PPK along with adenylate kinase can carry out polyphosphate-dependent phosphorylation AMP [33–35]. Hence, it is likely that the CYTH domain proteins participate directly in this biochemical network along with these proteins. One possibility is that the CYTH domains utilize polyphosphates to synthesize different organo-phosphate derivatives including nucleotides. Alternatively, they could also function as phosphoesterases that hydrolyze particular nucleoside polyphosphates. These leads could aid further experimental investigations on the CYTH domain that might help in uncovering ancient, as-yet-unexplored links between polyphosphate and nucleotide metabolism.

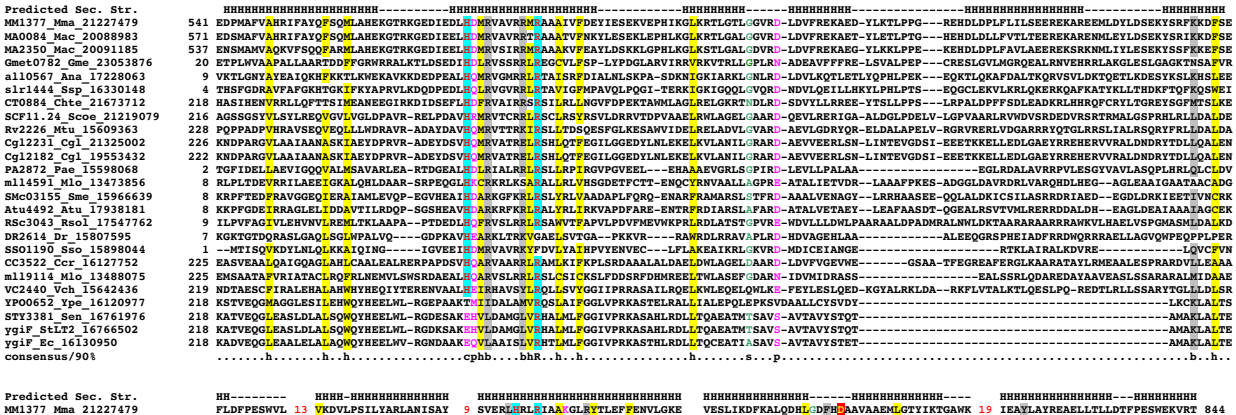


Figure 4
Multiple alignment of the CHAD domain The coloring scheme, secondary structure abbreviations and species abbreviations are as in Fig. 1. The coloring reflects the consensus at 90% conservation.

Finally, at least in some lineages, the CYTH domain proteins may have been secondarily recruited for other functions. The CyaB protein may represent one such case where after the original transfer from an archaeon into the proteobacterial lineage it may have acquire the novel function as an adenylyl cyclase. However, it is entirely possible that even in this case the adenylyl cyclase activity is secondary to some other uncharacterized metabolic activity. The vertebrate soluble thiamine triphosphatase has undergone accelerated divergence, as it is present on a long branch in the phylogenetic tree (Fig. 2). ThTPase is also unusual in lacking the 3rd conserved acidic domain of the CYTH domain. Hence, it may represent a case of relatively recent acquisition of a new catalytic activity.

Conclusions

We show that *Aeromonas* adenylyl cyclase CyaB and thiamine triphosphatase define a novel superfamily of catalytic domains that act on nucleotides and organo-phosphate substrates. These domains are widely distributed in all the 3 superkingdoms of life and can be traced back to the last ancestor of all life forms. We identify 6 conserved acidic residues, that are likely to form the active site of these enzymes, and 4 conserved basic residues, that may participate in interactions with phosphate-moiety-containing substrates. We postulate that these enzymes are likely to chelate 2 divalent cations and are likely follow a bimetal reaction mechanism similar to what has been proposed for nucleotide cyclases, nucleic acid polymerases, or certain phosphoesterases such as those of the HD and DHH superfamilies. A version of the CYTH domain, which is fused to the catalytic domain of nucleotide kinases, lacks the predicted catalytic residues, and probably function as an allosteric regulatory domain. Additionally, we detected a novel domain, termed CHAD, which occurs fused to the CYTH domain or is encoded by genes occurring in the same operon as those encoding CYTH domains. CHAD contains conserved histidines that are predicted to either chelate metals or serve as phosphoacceptors. Based on phyletic distribution and contextual information, we conclude that these enzymes may play a critical role in the interface between nucleotide and polyphosphate metabolism.

Methods

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTP and PSI-BLAST programs [18]. Profile searches using the PSI-BLAST program were conducted either with a single sequence or an alignment used as the query, with a profile inclusion expectation (E) value threshold of 0.01 and were iterated until convergence. Additionally, hidden Markov model based searches using a multiple alignment of known members were run using the HMMER2 package [36]. The Gibbs sampling procedure, as implemented in the MACAW program was used to detect and evaluate statistically significant conserved motifs [37]. Multiple alignments were constructed using the T_Coffee program [38], followed by manual correction based on the PSI-BLAST results. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED and PHD programs [39,40]. Preliminary clustering of proteins was done using the BLASTCLUST program with empirically determined length and score threshold cut off values (For documentation see [ftp://ftp.ncbi.nih.gov/blast/documents/README.bcl]). Phylogenetic analysis was performed using the neighbor joining or least square method followed by local rearrangements using the maximum likelihood algorithm to predict the most likely tree. 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 phylogenetic analyses [41,42].

The species abbreviations used in the alignments are: Aehy: *Aeromonas hydrophila*, Af: *Archaeoglobus fulgidus*, Aga: *Anopheles gambiae*, Ana: *Anabaena* sp. PCC 7120, Ap: *Aeropyrum pernix*, At: *Arabidopsis thaliana*, Atu: *Agrobacterium tumefaciens*, Ban: *Bacillus anthracis*, Bb: *Borrelia burgdorferi*, Bha: *Bacillus halodurans*, BPRB49: Bacteriophage RB49, Bs: *Bacillus subtilis*, Cac: *Clostridium acetobutylicum*, Cau: *Chloroflexus aurantiacus*, Ccr: *Caulobacter crescentus*, Ce: *Caenorhabditis elegans*, Chte: *Chlorobium tepidum*, Cgl: *Corynebacterium glutamicum*, Cpe: *Clostridium perfringens*, Ddi: *Dictyostelium discoideum*, Dm: *Drosophila melanogaster*, Ec: *Escherichia coli*, Gme: *Geobacter metallireducens*, Hi: *Haemophilus influenzae*, Hs: *Homo sapiens*, Hsp: *Halobacterium* sp., Lin: *Listeria innocua*, Lla: *Lactococcus lactis*, Lmo: *Listeria monocytogenes*, Mac: *Methanosarcina acetivorans*, Mcsp: *Magnetococcus sp.* Mfas: *Macaca fascicularis*, Mjan: *Methanococcus jannaschii*, Mka: *Methanopyrus kandleri*, Mlo: *Mesorhizobium loti*, Mma: *Methanosarcina mazei*, Mta: *Methanothermobacter thermautotrophicus*, Mtu: *Mycobacterium tuberculosis*, Nm: *Neisseria meningitidis*, Oih: *Oceanobacillus iheyensis*, Osa: *Oryza sativa*, Pa: *Pyrococcus abyssi*, Pae: *Pseudomonas aeruginosa*, Pfu: *Pyrococcus furiosus*, Ph: *Pyrococcus horikoshii*, Pmar: *Prochlorococcus marinus*, Pmu: *Pasteurella multocida*, Pyae: *Pyrobaculum aerophilum*, Rsol: *Ralstonia solanacearum*, Sa: *Staphylococcus aureus*, Scoe: *Streptomyces coelicolor*, Sen: *Salmonella enterica*, Sme: *Sinorhizobium meliloti*, Spn: *Streptococcus pneumoniae*, Spy: *Streptococcus pyogenes*, Sso: *Sulfolobus solfataricus*, Ssp: *Synechocystis* sp. PCC 6803, Sst: *Sulfolobus tokodaii*, StLT2: *Salmonella typhimurium* LT2, Vch: *Vibrio cholerae*, Xaxo: *Xanthomonas axonopodis*, Xca: *Xanthomonas campestris*, Xfa: *Xylella fastidiosa*, Ype: *Yersinia pestis*.

Authors' contributions

Author 1 (LMI) contributed to the discovery process, preparation of the multiple sequence alignments and figures, Author 2 (LA) contributed to the discovery process, preparation of the figures and manuscript and conceived the study. All authors read and approved the final manuscript.

Additional material

Additional file 1

A text copy of the alignments of the CYTH domain and the CHAD are provided in simple alignment and Clustal aln formats
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-3-33-S1.txt]

Additional file 2

Alignments presented in the article in PDF format.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-3-33-S2.pdf]

References

1. Stryer L: *Biochemistry Newyork, NY: W H Freeman and Co* 1995
2. Nelson DL, Cox MM: *Lehninger Principles of Biochemistry Worth Publishers Inc* 2000
3. 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
4. Aravind L, Mazumder R, Vasudevan S, Koonin EV: **Trends in protein evolution inferred from sequence and structure analysis.** *Curr Opin Struct Biol* 2002, **12:**392-399
5. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci U S A* 1996, **93:**10268-10273
6. Saraste M, Sibbald PR, Wittinghofer A: **The P-loop – a common motif in ATP- and GTP-binding proteins.** *Trends Biochem Sci* 1990, **15:**430-434
7. Gorbalenya AE, Koonin EV: **Superfamily of UvrA-related NTP-binding proteins. Implications for rational classification of recombination/repair systems.** *J Mol Biol* 1990, **213:**583-591
8. Vetter IR, Wittinghofer A: **Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer.** *Q Rev Biophys* 1999, **32:**1-56
9. Bork P, Sander C, Valencia A: **An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins.** *Proc Natl Acad Sci U S A* 1992, **89:**7290-7294
10. Artymiuk PJ, Poirrette AR, Rice DW, Willett P: **A polymerase I palm in adenylyl cyclase?** *Nature* 1997, **388:**33-34
11. Murzin AG: **How far divergent evolution goes in proteins.** *Curr Opin Struct Biol* 1998, **8:**380-387

12. Leipe DD, Wolf YI, Koonin EV, Aravind L: **Classification and evolution of P-loop GTPases and related ATPases.** *J Mol Biol* 2002, **317:**41-72
13. Pei J, Grishin NV: **GGDEF domain is homologous to adenylyl cyclase.** *Proteins* 2001, **42:**210-216
14. Koonin EV, Wolf YI, Kondrashov AS, Aravind L: **Bacterial homologs of the small subunit of eukaryotic DNA primase.** *J Mol Microbiol Biotechnol* 2000, **2:**509-512
15. Aravind L, Koonin EV: **DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27:**1609-1618
16. Sismeiro O, Trotot P, Biville F, Vivares C, Danchin A: **Aeromonas hydrophila adenylyl cyclase 2: a new class of adenylyl cyclases with thermophilic properties and sequence similarities to proteins from hyperthermophilic archaebacteria.** *J Bacteriol* 1998, **180:**3339-3344
17. Lakaye B, Makarchikov AF, Antunes AF, Zorzi W, Coumans B, De Pauw E, Wins P, Grisar T, Bettendorff L: **Molecular characterization of a specific thiamine triphosphatase widely expressed in mammalian tissues.** *J Biol Chem* 2002, **277:**13771-13777
18. 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:**3389-402
19. Aravind L, Koonin EV: **The HD domain defines a new superfamily of metal-dependent phosphohydrolases.** *Trends Biochem Sci* 1998, **23:**469-472
20. Aravind L, Koonin EV: **A novel family of predicted phosphoesterases includes Drosophila prune protein and bacterial RecJ exonuclease.** *Trends Biochem Sci* 1998, **23:**17-19
21. Leipe DD, Aravind L, Grishin NV, Koonin EV: **The bacterial replicative helicase DnaB evolved from a RecA duplication.** *Genome Res* 2000, **10:**5-16
22. 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
23. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002
24. Ponting CP, Aravind L, Schultz J, Bork P, Koonin EV: **Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer.** *J Mol Biol* 1999, **289:**729-745
25. Schultz JE, Klumpp S: **Cyclic GMP in lower forms.** *Adv Pharmacol* 1994, **26:**285-303
26. Huynen M, Snel B, Lathe W 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210
27. Iyer LM, Koonin EV, Aravind L: **Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52.** *BMC Genomics* 2002, **3:**8
28. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10:**1074-1077
29. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30:**482-496
30. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328
31. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11:**356-372
32. Kornberg A, Rao NN, Ault-Riche D: **Inorganic polyphosphate: a molecule of many functions.** *Annu Rev Biochem* 1999, **68:**89-125
33. Shiba T, Tsutsumi K, Ishige K, Noguchi T: **Inorganic polyphosphate and polyphosphate kinase: their novel biological functions and applications.** *Biochemistry (Mosc)* 2000, **65:**315-323
34. Ishige K, Noguchi T: **Polyphosphate:AMP phosphotransferase and polyphosphate:ADP phosphotransferase activities of Pseudomonas aeruginosa.** *Biochem Biophys Res Commun* 2001, **281:**821-826
35. Ishige K, Noguchi T: **Inorganic polyphosphate kinase and adenylate kinase participate in the polyphosphate:AMP phosphotransferase activity of Escherichia coli.** *Proc Natl Acad Sci U S A* 2000, **97:**14168-71
36. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-63
37. Schuler GD, Altschul SF, Lipman DJ: **A workbench for multiple alignment construction and analysis.** *Proteins* 1991, **9:**180-190
38. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-17
39. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584-99
40. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server.** *Bioinformatics* 1998, **14:**892-893
41. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266:**418-27
42. Hasegawa M, Kishino H, Saitou N: **On the maximum likelihood method in molecular phylogenetics.** *J Mol Evol* 1991, **32:**443-5