doi:10.1016/j.jmb.2006.06.049

J. Mol. Biol. (2006) **361**, 1003–1034

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes

A. Maxwell Burroughs[1,2], Karen N. Allen[2,3]
Debra Dunaway-Mariano[4] and L. Aravind[1]*

[1]*National Center for Biotechnology Information National Library of Medicine National Institutes of Health Bethesda, MD 20894, USA*

[2]*Bioinformatics Program Boston University, Boston MA 02215, USA*

[3]*Department of Physiology and Biophysics, Boston University School of Medicine, 715 Albany Street, Boston MA 02118-2394, USA*

[4]*Department of Chemistry University of New Mexico Albuquerque, NM 87131, USA*

The HAD (haloacid dehalogenase) superfamily includes phosphoesterases, ATPases, phosphonatases, dehalogenases, and sugar phosphomutases acting on a remarkably diverse set of substrates. The availability of numerous crystal structures of representatives belonging to diverse branches of the HAD superfamily provides us with a unique opportunity to reconstruct their evolutionary history and uncover the principal determinants that led to their diversification of structure and function. To this end we present a comprehensive analysis of the HAD superfamily that identifies their unique structural features and provides a detailed classification of the entire superfamily. We show that at the highest level the HAD superfamily is unified with several other superfamilies, namely the DHH, receiver (CheY-like), von Willebrand A, TOPRIM, classical histone deacetylases and PIN/FLAP nuclease domains, all of which contain a specific form of the Rossmannoid fold. These Rossmannoid folds are distinguished from others by the presence of equivalently placed acidic catalytic residues, including one at the end of the first core β-strand of the central sheet. The HAD domain is distinguished from these related Rossmannoid folds by two key structural signatures, a "squiggle" (a single helical turn) and a "flap" (a beta hairpin motif) located immediately downstream of the first β-strand of their core Rossmanoid fold. The squiggle and the flap motifs are predicted to provide the necessary mobility to these enzymes for them to alternate between the "open" and "closed" conformations. In addition, most members of the HAD superfamily contains inserts, termed caps, occurring at either of two positions in the core Rossmannoid fold. We show that the cap modules have been independently inserted into these two stereotypic positions on multiple occasions in evolution and display extensive evolutionary diversification independent of the core catalytic domain. The first group of caps, the C1 caps, is directly inserted into the flap motif and regulates access of reactants to the active site. The second group, the C2 caps, forms a roof over the active site, and access to their internal cavities might be in part regulated by the movement of the flap. The diversification of the cap module was a major factor in the exploration of a vast substrate space in the course of the evolution of this superfamily. We show that the HAD superfamily contains 33 major families distributed across the three superkingdoms of life. Analysis of the phyletic patterns suggests that at least five distinct HAD proteins are traceable to the last universal common ancestor (LUCA) of all extant organisms. While these prototypes diverged prior to the emergence

Abbreviations used: HAD, haloacid dehalogenase; PNKP, polynucleotide kinase phosphatase; KDO 8-P, deoxy-D-mannose-octulosonate 8-phosphate; CTD, carboxyl-terminal domain; PNKP, polynucleotide kinase phosphatase; LUCA, last universal common ancestor.

E-mail address of the corresponding author: aravind@ncbi.nlm.nih.gov

0022-2836/$ - see front matter. Published by Elsevier Ltd.

of the LUCA, the major diversification in terms of both substrate specificity and reaction types occurred after the radiation of the three superkingdoms of life, primarily in bacteria. Most major diversification events appear to correlate with the acquisition of new metabolic capabilities, especially related to the elaboration of carbohydrate metabolism in the bacteria. The newly identified relationships and functional predictions provided here are likely to aid the future exploration of the numerous poorly understood members of this large superfamily of enzymes.

Published by Elsevier Ltd.

Keywords: Rossmann fold; catalytic diversity; lateral gene transfer; substrate specificity; domain mobility

*Corresponding author

Introduction

All cellular organisms depend extensively upon the biochemical reactions related to organo-phosphoesters and phosphoanhydrides. Hence, it is not surprising that an enormous diversity of phosphohydrolases have evolved on multiple occasions to catalyze the dephosphorylation of various compounds.[1,2] The majority of cellular phosphohydrolases belong to a relatively small set of evolutionarily distinct superfamilies, which are almost entirely dedicated to the catalysis of such reactions. These large superfamilies include the P-loop NTPases, which is the largest monophyletic assemblage of nucleotide triphosphatases encoded by cellular genomes,[3] the RNase H fold of ATPases, including actin, Hsp70 and their relatives,[4,5] the DHH,[6] HD,[7] PHP,[8] HAD,[9,10] calcineurin-like,[11] synaptojanin-like,[12] and the Receiver domain (CheY) superfamilies.[13,14] They span the entire range of structural basic classes with α-helical forms, such as the HD superfamily,[15] the beta-barrels such as the CYTH superfamily of phosphohydrolases,[16] three-layered α/β sandwiches, such as P-loop NTPases,[3] HAD[17] and DHH,[18,19] α/β barrels such as the PHP phosphoesterases,[20] and four- layered α/β-sandwiches such as the calcineurin-like[21] and synaptojanin-like phosphoesterases.[22]

The HAD superfamily, named after the archetypal enzyme haloacid dehalogenase,[9] includes enzymes catalyzing carbon or phosphoryl group transfer reactions on a diverse range of substrates, using an active site aspartate in nucleophilic catalysis (Figure 1(a)). The majority of the enzymes in this superfamily are involved in phosphoryl transfer, i.e. phosphate monoester hydrolases (phosphatases) or phosphoanhydride hydrolase P-type ATPases. These include variations such as a phosphonoacetaldehyde hydrolase (phosphonatase) and phosphotransferases, such as β-phosphoglucomutase and α-mannophosphomutase. Each of the phosphotransferase enzymes requires a Mg^{2+} cofactor for catalysis[9,10] (Figure 1(b)). The carbon group transfer reaction (Figure 1(a)) catalyzed by haloalkanoic acid dehalogenase (HAD)[23] is unique in that it does not utilize a metal ion cofactor, and that a water nucleophile attacks the Asp C═O in the hydrolysis partial reaction.

The HAD superfamily is represented in the proteomes of organisms from all three superkingdoms of life, and have colonized numerous very disparate biological functions, which vary in their degree of essentiality to the cell. We are primarily interested in understanding how the catalytic platform of the HAD superfamily has been adapted through evolution to act on a wide range of substrates, a process which has been termed the "evolutionary exploration of substrate space".[24] The accumulation of over 40 X-ray crystal structures and the enormous amount of sequence data available through genome sequencing projects have made the HAD superfamily amenable to understanding this process of evolution. Accordingly, here we present a comprehensive natural classification of the HAD superfamily using the information derived from relevant sequence and structural elements, phyletic distribution patterns, and phylogenetic tree analysis. This classification system offers a model for understanding the diversification of enzymes and allows us to predict important functional residues or regions in members of the superfamily having unknown function.

Results and Discussion

Structural and functional aspects of the HAD superfamily

Structural core of the HAD superfamily

To provide the basic context for a structure–function analysis of the HAD superfamily we first define its essential structural core, and compare it to other structurally related folds. The core catalytic domain of the HAD superfamily contains a three-layered α/β sandwich comprised of repeating β-α units which adopt the topology typical of the Rossmannoid class of α/β folds. The central sheet is parallel and is typically comprised of at least five strands in a 54123 strand order (Figures 2(a) and 3). These strands are hereinafter referred to as S1–S5. The HAD fold is distinguished from all other related Rossmannoid folds by two key structural motifs (Figure 3). First, immediately downstream of strand S1 is a unique, approximately six residue

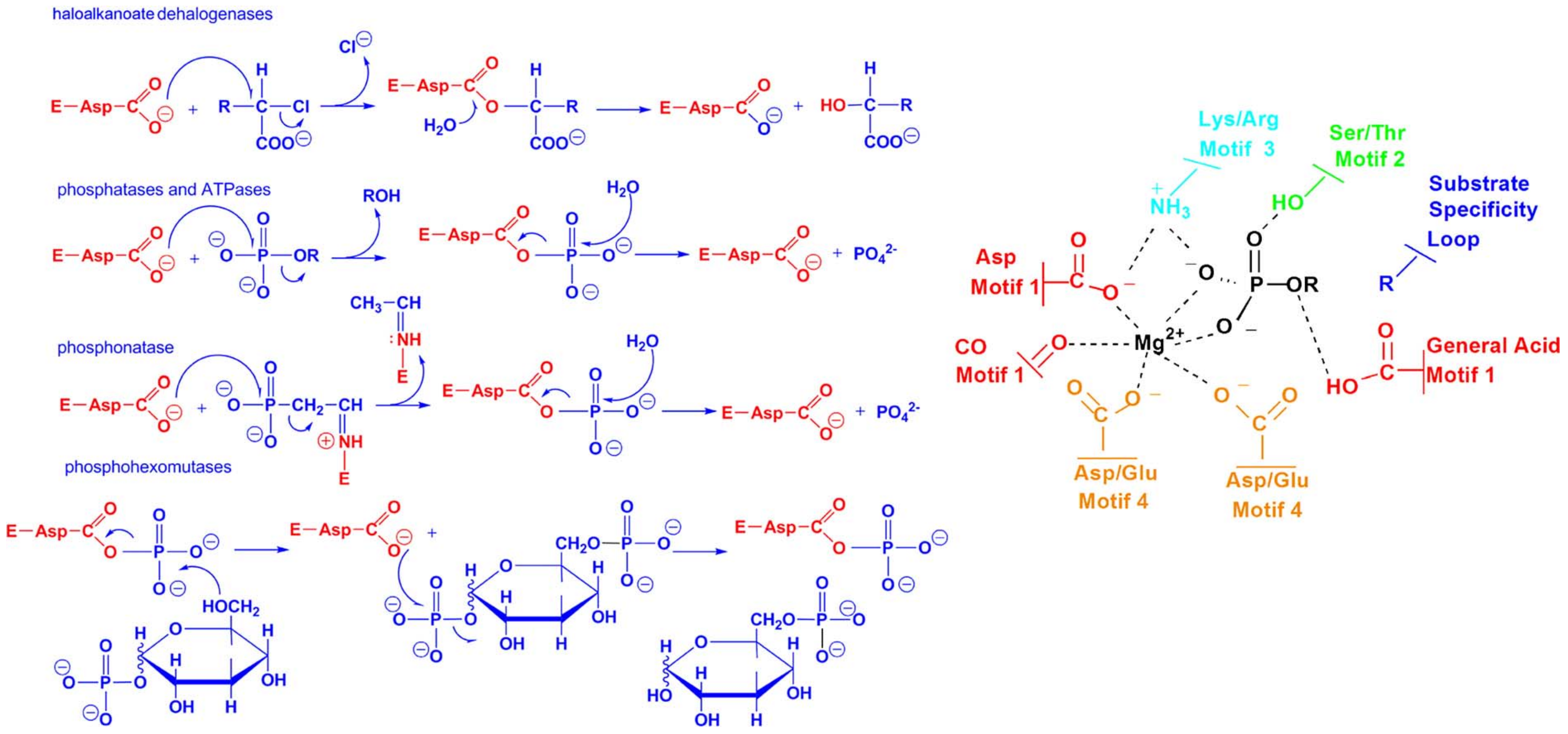


Figure 1. HAD reaction mechanisms. A schematic representation of the reaction pathway in carbon transfer and in phosphoryl transfer is depicted. The left panel shows the major types of reactions known to be catalyzed by the HAD superfamily can be distinguished by the identity of the leaving group of the substrate, the site of hydrolysis of the intermediate, and the identity of the phosphoryl acceptor group. Moieties originating from the substrate or solvent are colored blue and those originating from the enzyme are colored red. The right panel shows a schematic of the active-site template for phosphoryl transferases showing interactions of the substrate with the catalytic motifs (contributed by the core domain) and substrate specificity determinants (usually contributed by the cap domain). Residues contributed by each motif are coded: the substrate specificity component is colored in blue, while residues from each of the motifs are given a separate color.

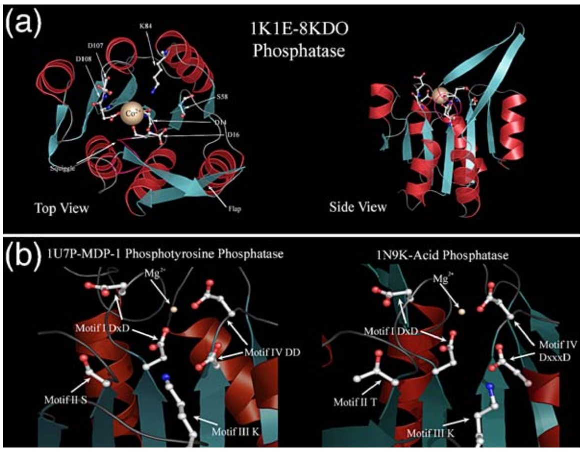


Figure 2. HAD catalytic domain. Cartoon representations of the structure of the HAD fold with close-ups of different active site configurations. Beta strands are colored blue while alpha-helices are colored red. (a) Top and side views are shown from deoxy-D-mannose-octulosonate 8-phosphate phosphatase (8KDO) from *Haemophilus influenzae* (PDB: 1K1E). The top view (upper left) reveals the typical spatial orientations of the conserved residues involved in catalysis, which are depicted as stick and ball figures. Conserved residues and two conserved structural motifs, the flap and the squiggle, are labeled. The side view (upper right) shows the Rossmann-like fold of the HAD superfamily, and the location of the cap domain relative to the core domain. The squiggle motif, central to the active site, is colored pink. Although HAD proteins typically employ a Mg^{2+} in catalysis, the tan sphere in this crystal structure represents a cobalt ion, a metal used for crystallization.[103] (b) Close-up active site views of two HAD representatives with distinct motif IV signatures. The left panel (MDP-1 from the MDP-1/FkbH family) has a motif IV DD signature (PDB: 1U7P) while the right panel (AphA from the Acid Phosphatase family) has a motif IV DxxxD signature (PDB: 1N9K).

structural motif which assumes a nearly complete single helical turn, not unlike those found in the catalytic domains of unrelated enzymes of the polymerase–β-fold.[25] We term this motif the "squiggle". In some members of the HAD superfamily the squiggle forms hydrogen bonds between the ith and *i*+5th position resulting in the rare pi-helix conformation.[26] Second, downstream of the squiggle there is a β-hairpin turn formed by two strands projecting from the core of the domain (Figure 2(a)). We term this structural motif the "flap". The squiggle and flap structural motifs play essential roles in HAD superfamily catalysis (see below for details).

Sequence comparisons have shown that practically all members of the HAD superfamily contain four highly conserved sequence motifs.[10] Sequence motif I corresponds to strand S1 and the DxD signature is present at the end of this strand (Figures 1(b) and 4). The carboxylate group of the first Asp and the backbone C=O of the second Asp coordinate the Mg^{2+} cofactor (Figure 1(b)). Additionally, the first Asp in motif I acts as a nucleophile that forms an aspartyl-intermediate during catalysis.[27–31] In phosphatase and phosphomutase members of the superfamily the second acidic residue acts as a general acid-base. It binds and, in many cases, protonates the substrate leaving group in the first step and deprotonates the nucleophile of the second step.[32] In the ATPases, the occurrence of a threonine at this position allows for a reduced rate of aspartyl phosphate hydrolysis, which may allow for the time lag necessary for the consequent conformational change. In the phosphonatases, there is an alanine instead of the second aspartate, which is consistent with the unique role played by the enamine intermediate (formed with the insert domain, see below) as a general acid-base catalyst in aspartyl phosphate hydrolysis by these proteins.

Motif II corresponds to the S2 strand, which is characterized by a highly conserved threonine or a serine at its end (Figures 1(b) and 4). Motif III is centered on a conserved lysine that occurs around the N terminus of the helix located upstream of S4 (Figures 2, 3 and 4). Motif II and motif III contribute to the stability of the reaction intermediates of the hydrolysis reaction. The lysine in motif III is reminiscent of the basic residues termed arginine fingers that stabilize the negative charge on reaction intermediates in many other phosphohydrolases, particularly those of the P-loop NTPase fold.[33] It is likely that they play a similar role even in the HAD hydrolases. An analysis of the available structures shows that the lysine in motif III may occur in either of two structural contexts in different HAD hydrolases. In the P-type ATPases, acid phosphatases, phosphoserine phosphatases and the Cof hydrolases the lysine is incorporated into the helix immediately preceding strand S4. However, in all other HAD hydrolases it emerges from the loop immediately prior to the helix. On account of this difference in the secondary structure context of the lysine, motif III is poorly conserved relative to the other motifs. The poor local conservation beyond the functionally critical basic residue is also comparable to the regions bearing the arginine finger in the AAA+ ATPases.[3] Motif IV maps to strand S4 and the conserved acidic residues located at its end. These terminal acidic residues of motif IV typically exhibit one of three basic signatures: DD, GDxxxD, or GDxxxxD (where x is any amino acid) ((Figures 1(b), 2 and 4)). These acidic residues along with those in motif I are required for coordinating the Mg ion in the active site.[27,32,34–39] Motifs I–IV are spatially arranged around a single "binding cleft" at the C-terminal end of the strands of the central sheet that forms the active site of the HAD superfamily (Figure 1 (b)). This binding cleft is partly covered by the β-hairpin flap occurring after S1 (Figures 2(a) and 3). Additional inserts occurring between the two strands of the flap or in the region immediately after S3 provide extensive shielding for the catalytic cavity. These inserts, termed caps, often contribute residues required for specificity or auxiliary

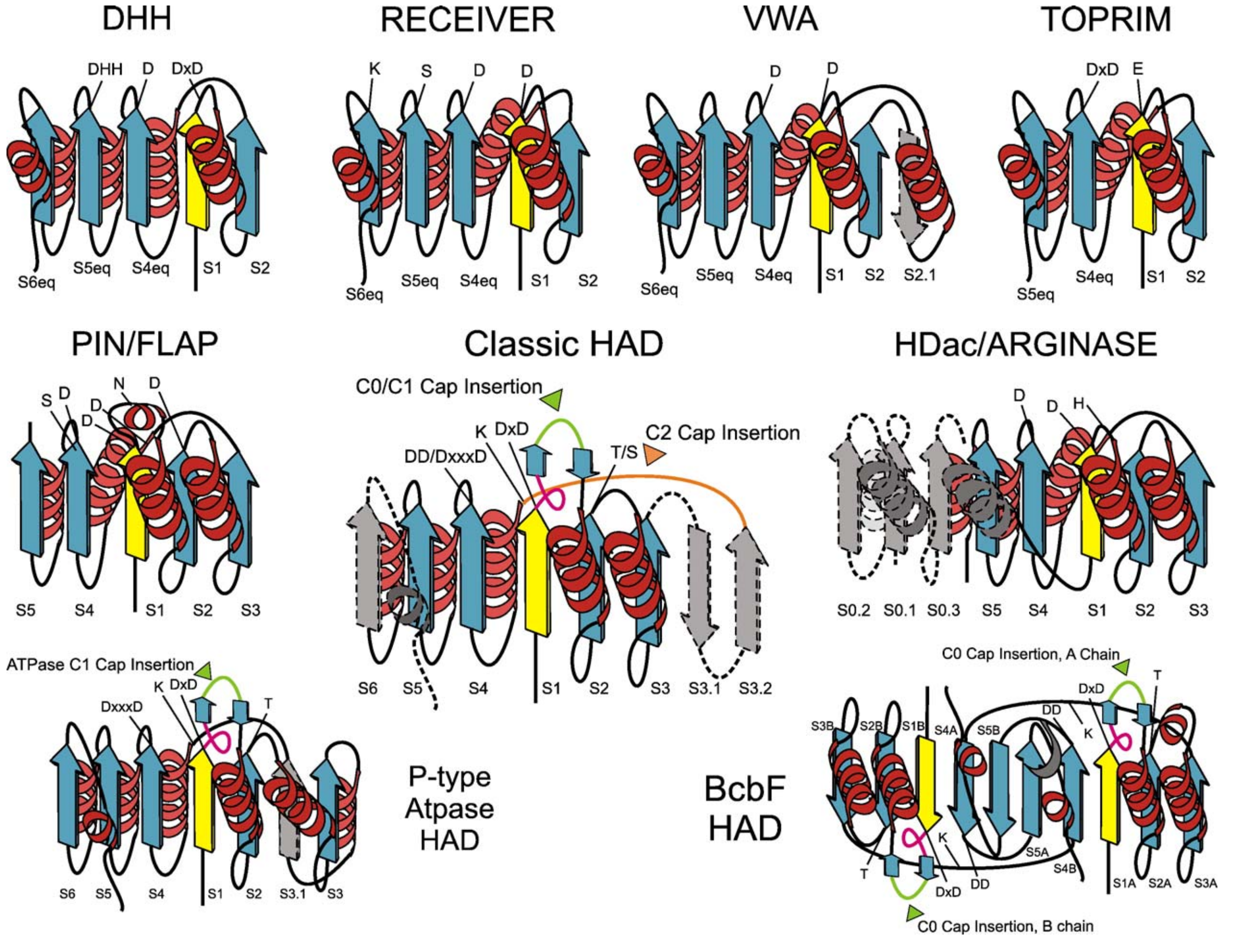


Figure 3. Rossmannoid domains. Topology diagrams of domains representative of the major divisions of Rossmann-like folds with catalytic acidic residues. Two versions of the HAD domain (P-type ATPase HAD and BcbF HAD) that show significant modifications to the classic HAD domain are also shown. Strands are shown as arrows with the arrowhead on the C-terminal end and are labeled from S1 to S6 in the classic HAD, with equivalent strands marked with eq in other Rossmannoid domains. The HAD domain of BcbF is an obligate dimer, and strands from the two dimers are differentiated as A and B (e.g. S1A and S1B). The initial strand containing the catalytic D residue is rendered in yellow; other core strands conserved across all members of the domain are in blue; non-conserved elements that may have been absent from the ancestral state of a domain are in gray. The HAD C1 cap insertion point is represented as a bright green line and the C2 cap insertion point is represented as an orange line. Broken lines indicate secondary structures elements not present in all members sharing the domain. The pink loop in the HAD domain represents the conserved squiggle. Residues conserved across all members of a particular domain, including the initial catalytic D residue, are shown.

catalytic functions, and play a central role in the reactions catalyzed by most HAD hydrolases[28,40,41] (see below for further discussion).

Relationship of the HAD superfamily to other Rossmannoid folds

The topology of the central β-sheet of the HAD fold makes it a typical representative of the Rossmannoid class of three-layered α/β sandwich folds (Figure 3). It shares with other Rossmanoid folds the general location of the active site formed by residues at the C-terminal end of the central sheet. More specifically, the HAD fold shares with other Rossmannoid fold enzymes a critical substrate-binding site in the loop between S1 and the downstream α-helix, and a second active site residue positioned immediately downstream of the strand occurring after the crossover in the β-sheet, i.e. strand S4 (Figure 3).[42] Amongst the Rossmannoid folds two major divisions can be recognized: (1) the nucleotide binding domains with a nucleotide binding loop between strand 1 and the helix after it. This group includes many large monophyletic assemblages of proteins, namely the classic Rossmann NAD/FAD-dependent dehydrogenases,[43] Sir2-like deacetylases,[44] the *S*-AdoMet-binding methyltransferases,[45–47] the GTPase FtsZ,[48] the ISOCOT fold[49] and the HUP superclass (class I tRNA synthetases, HIGH nucleotidyltransferases, USPA, photolyase and electron transport flavoprotein).[42] Most members of this division are characterized by specific signatures, often glycine-rich, in their nucleotide-binding loops. (2) The second division comprises phosphohydrolases or divalent cation-chelating domains with a conserved acidic residue in the loop between the first strand and the helix that comes after it. This division includes the HAD superfamily, whose DxD motif is found in this loop, and several other enzymes superfamilies with similar active site configurations. These superfamilies are the DHH domain phosphoesterases (e.g. the DNAse involved in repair and recombination, RecJ),[6] the receiver or CheY domain of the two-component signaling system,[13,14] the TOPRIM domain, which is the shared catalytic domain of the topoisomerases and DnaG-type primases,[50] the PIN/5′-3′ nuclease domain,[51] the classical histone deacetylases/arginases[52] and the von Willebrandt factor A (vWA) domain[53] (second division only depicted in Figure 3). Most members of this division are also unified by a second acidic residue, which is

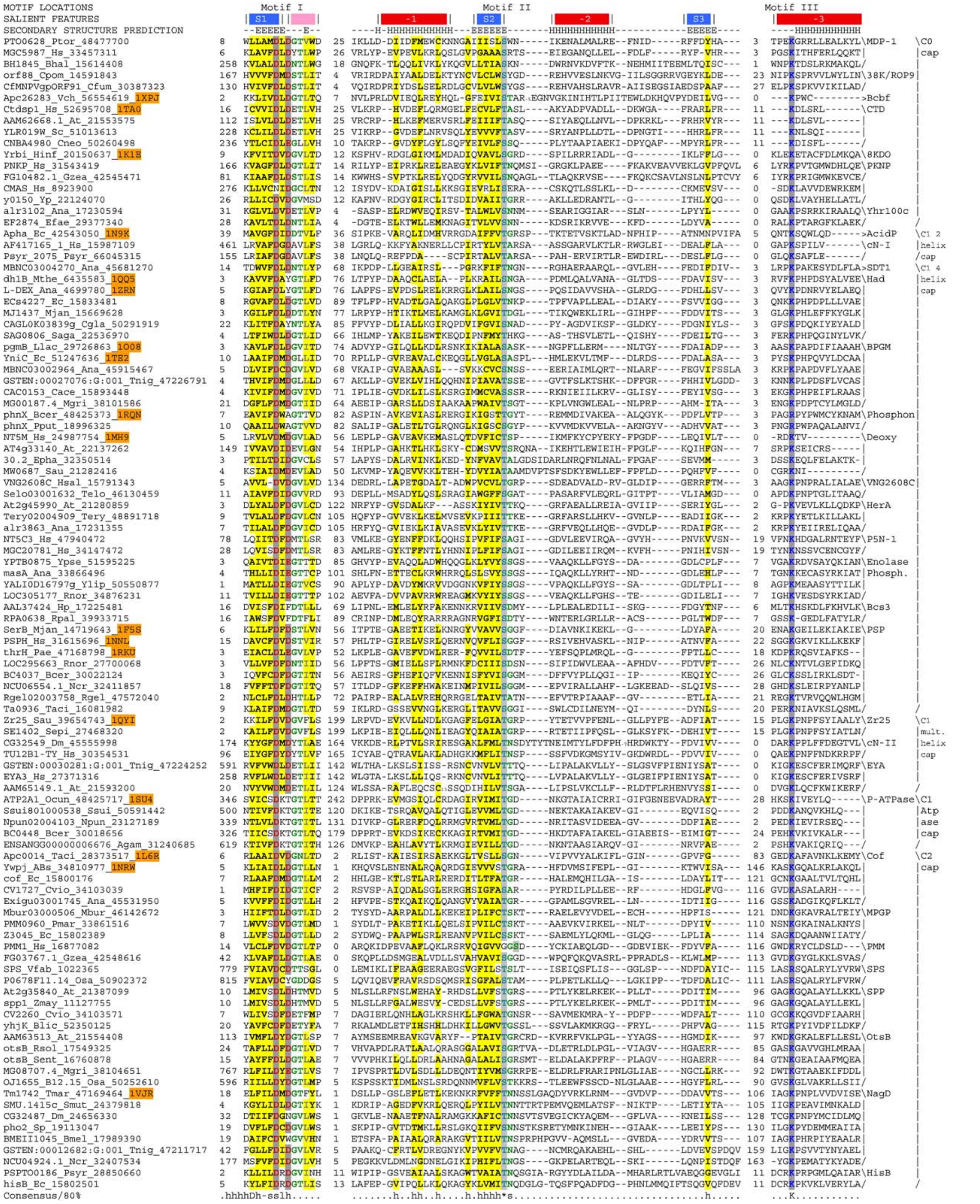
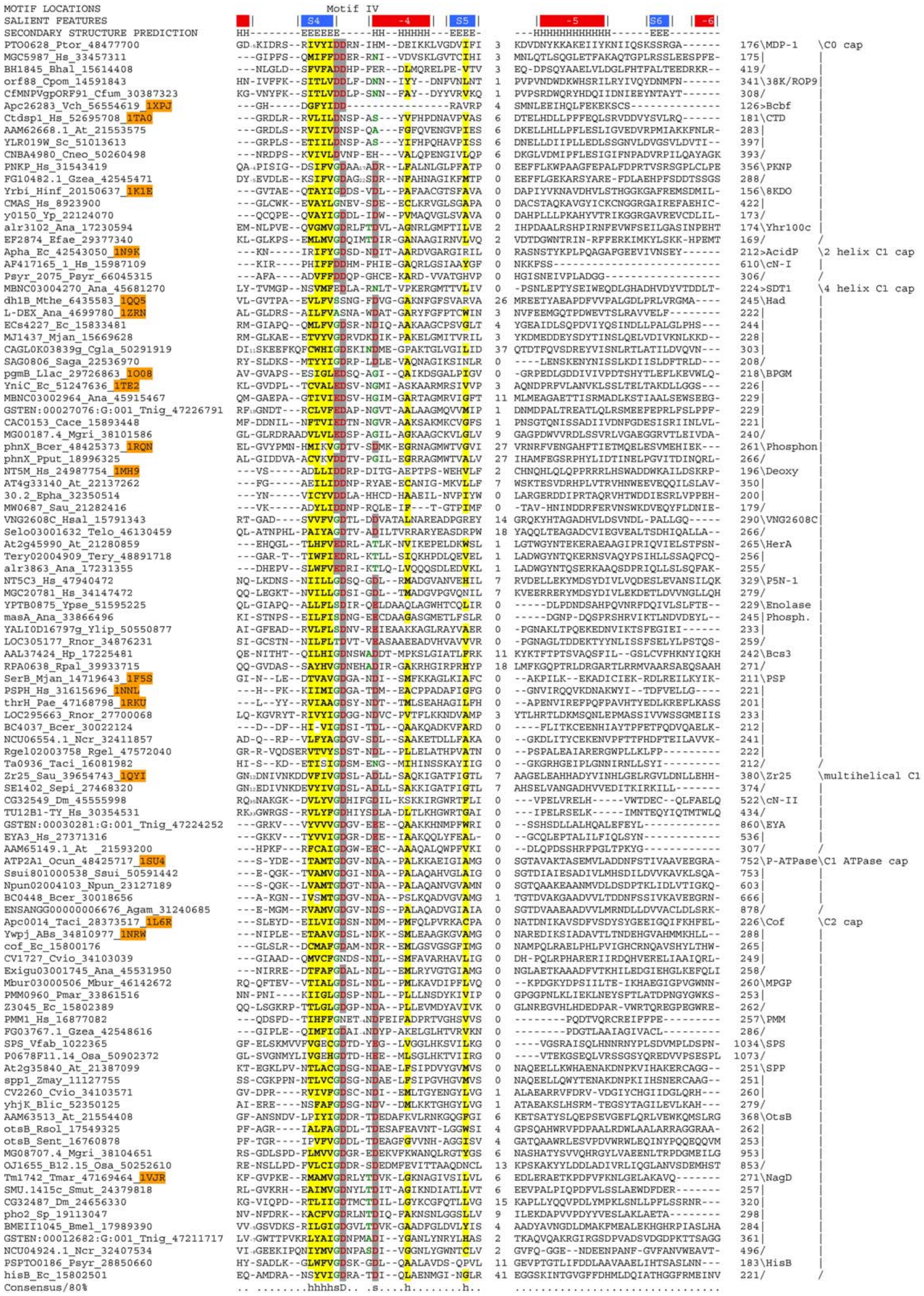


Figure 4 (*legend on page 1010*)

borne at the end of the strand adjacent to the first strand, which occurs after the crossover of the sheet to the opposite side (left of strand S1 in Figure 3). Like the HAD domains, the receiver domain forms an aspartyl phosphate intermediate,[54] which receives a phosphate from a histidinyl-phosphate on the histidine kinase.[54–56] Because of the mechanistic similarity, the receiver domain has previously been claimed to be a member of the HAD fold.[57,58] However, a careful examination of the active site organization and sheet topology of the receiver domains (Figure 3) shows that it does not share any



```
MOTIF LOCATIONS                                         Motif IV
SALIENT FEATURES                     ▇|      |S4 |   |  -4  | |S5|        |  -5   |     |S6|  |-6|
SECONDARY STRUCTURE PREDICTION       HH--------EEEEEE-----HH--HHHHH---EEEE-     ---HHHHHHHHHHHH------EEE--------
PT00628_Ptor_48477700                GD5KIDRS--RIVYIDDRN-IHM--DEIKKLVGDVIFI   3  KDVDNYKKAKEIIYKNIIQSKSSRGA------  176\MDP-1     \C0 cap
MGC5987_Hs_33457311                  ---GIPFS--QMIFFDDER-RNI--VDVSKLGVTCIHI   3  MNLQTLSQGLETFAKAQTGPLRSSLEESPFE-  175|          |
BH1845_Bhal_15614408                 ---NLGLD--SFVFADDHP-FER--DLMQRELPE-VTV   3  EQ-DPSQYAAELVLDGLFHTFALTEEDRKR--  419/          |
orf88_Cpom_14591843                  HN-IVFFK--SITLVDDLF-DNN--IYY--DNFVNLNT   1  PVPVNDWDKWHSRILRYIVQYDNMFN------  341\38K/ROP9|
CfMNPVgpORF91_Cfum_30387323          KG-VNYFK--SITLVDDLP-SNN--FAY--DYYVRVKQ   1  PVPSRDWQRYHDQIIDNIEEYNTAYT------  308/          |
Apc26283_Vch_56554619_1XPJ           ---GH-----DGFYIDD----------------RAVRP   4  SMNLEEIHQLFEKEKSCS--------------  126>Bcbf      |
Ctdspl_Hs_52695708_1TA0              ---GRDLR--RVLILDNSP-AS---YVFHPDNAVPVAS   6  DTELHDLLPFFEQLSRVDDVYSVLRQ------  181\CTD       |
AAM62668.1_At_21553575               ---GRDLS--RVIIVDNSP-QA---FGFQVENGVPIES   6  DKELLHLLPFLESLIGVEDVRPMIAKKFNLR-  283|          |
YLR019W_Sc_51013613                  ---GRPLS--ETIILDNSP-AS---YIFHPQHAVPISS   6  DNELLDIIPLLEDLSSGNVLDVGSVLDVTI--  397|          |
CNBA4980_Cneo_50260498               ---NRDPS--KVIVLDVNP-EH---VALQPENGIVLQP   6  DKGLVDMIPFLESIGIFNPADVRPILQAYAGK  393/          |
PNKP_Hs_31543419                     QA4PISIG--DSIFVGDAA17ADR--LFALNLGLPFATP  0  EEFFLKWPAAGFEPALFDPRTVSRSGPLCLPE  356\PKNP      |
FG10482.1_Gzea_42545471              DY5EVDLE--KSIFVGDAG22SDR--NFAHNAGIKFMTP  0  EEFFLGEKARSYARE-FDLAEHPFSDDTSSGS  288/          |
Yrbi_Hinf_20150637_1K1E              ---GVTAE--QTAYIGDDS-VDL--PAFAACGTSFAVA   0  DAPIYVKNAVDHVLSTHGGKGAFREMSDMIL-  156\8KDO      |
CMAS_Hs_8923900                      ---GLCWK--EVAYLGNEV-SDE--ECLKRVGLSGAPA   0  DACSTAQKAVGYICKCNGGRGAIREFAEHIC-  422|          |
y0150_Yp_22124070                    ---QCQPE--QVAYIGDDL-IDW--PVMAQVGLSVAVA   0  DAHPLLLPKAHYVTRIKGGRGAVREVCDLIL-  173/          |
alr3102_Ana_17230594                 EM-NLPVE--QVGMVGDRLFTDVL-AGNRLGMFTILVE   2  IHPDAALRSHPIRNFEVWFSEILGASINPEHT  174\Yhr100c   |
EF2874_Efae_29377340                 KL-GLKPS--EMLMVGDQIMTDIR-GANAAGIRNVLVQ   2  VDTDGWNTRIN-RFFERKIMKYLSKK-HPEMT  169/          /
Apha_Ec_42543050_1N9K                ---KN-----IRIFYGDSD-NDIT-AARDVGARGIRIL   0  RASNSTYKPLPQAGAFGEEVIVNSEY------  212>AcidP    \2 helix C1 cap
AF417165_1_Hs_15987109               ---KIR----PHIFFDDHM-FHIE-GAQRLGSIAAYGF   0  NKKFSS--------------------------  610\cN-I     |
Psyr_2075_Psyr_66045315              ---AFA----ADVFFDDQP-GHCE-KARD-VVATGHVP   0  HGISNEIVPLADGG------------------  306/          /
MBNC03004270_Ana_45681270            LY-TVMGP--NSVMFEDLA-RNLT-VPKERGMTTVLIV   0  -PSNLEPTYSEIWEQDLGHADHVDYVTDDLT-  224>SDT1      \4 helix C1 cap
dh1B_Mthe_6435583_1QQ5               VL-GVTPA--EVLFVSSNG-FDVG-GAKNFGFSVARVA  26  MREETYAEAPDFVVPALGDLPRLVRGMA----  245\Had       |
L-DEX_Ana_4699780_1ZRN               AL-GLDRS--AILFVASNA-WDAT-GARYFGFPTCWIN   3  NVFEEMGQTPDWEVTSLRAVVELF--------  222|          |
ECs4227_Ec_15833481                  RM-GIAPQ--QMLFVGDSR-NDIQ-AAKAAGCPSVGLT   4  YGEAIDLSQPDVIYQSINDLLPALGLPHS---  244|          |
MJ1437_Mjan_15669628                 RM-GLKAE--ETVYVGDRVDKDIK-PAKELGMITVRIL   3  YKDMEDDEYSDYTINSLQELVDIVKNLKKD--  228|          |
CAGL0K03839g_Cgla_50291919           DI11SKEEFKQFCWHIGDEKINDME-GPAKTGLVGILID 37  QTDTFQVSDREYVISNLRTLATILDVQVN---  303|          |
SAG0806_Saga_22536970                RY-SLDKS--MTYYIGDRP-LDLE-VAQNAGIKSINLR   0  ----LENSKENYNISSLKDIISLDFTRLD---  208/          |
pgmB_Llac_29726863_1O08              AV-GVAPS--ESIGLEDSQ-AGI--QAIKDSGALPIGV   0  -GRPEDLGDDIVIVPDTSHYTLEFLKEVWLQ-  218\BPGM      |
YniC_Ec_51247636_1TE2                KL-GVDPL--TCVALEDSV-NGMI-ASKAARMRSIVVP   3  AQNDPRFVLANVKLSSLTELTAKDLLGGS---  226|          |
MBNC03002964_Ana_45915467            QM-GAEPA--GTIVIEDSV-HGIM-GARTAGMRVIGFT  11  MLMEAGAETTISRMADLKSTIAALSEWSEEG-  229|          |
GSTEN:00027076:G:001_Tnig_47226791   RF10GNDT---RCLVFEDAP-NGVT-AALAAGMQVVMIP  1  DNMDPALTREATLQLRSMEEFEPRLFSLPPF-  229|          |
CAC0153_Cace_15893448                MF-DDNIL--NFTVIEDTN-NGVR-AAKSAKMKCVGFS   1  PNSGTQNISSADIIVDNFGDESISRIINLVL-  221|          |
MG00187.4_Mgri_38101586              GL-GLRDRAADVLVLEDSP-AGIL-AGKAAGCKVLGLV   9  GAGPDWVVRDLSSVRLVGAEGGRVTLEIVDA-  240/          |
phnX_Bcer_48425373_1RQN              EL-GVYPMN-HMIKVGDTV-SDMK-EGRNAGMWTVGVI  27  VRNRFVENGAHFTIETMQELESVMEHIEK---  261\Phosphon|
phnX_Pput_18996325                   AL-GIDDVA-ACVKVDDTV-PGIL-EGRRAGMWTVALV  27  IHAMFEGSRPHYLIDTINELPGVITDINQRL-  266/          |
NT5M_Hs_24987754_1MH9                ---VS-----ADLLIDDRP-DITG-AEPTPS-WEHVLF   2  CHNQHLQLQPPRRRLHSWADDWKAILDSKRP-  196\Deoxy     |
AT4g33140_At_22137262                ---FG-----AEILIDDNP-RYAE-ECANIG-MKVLLF   7  WSKTESVDRHPLVTRVHNWEEVEQQILSLAV-  350|          |
30.2_Epha_32350514                   ---YN-----VICYVDDLA-HHCD-HAAEIL-NVPIYW   0  LARGERDDIPRTAQRVHTWDDIESRLVPPEH-  200|          |
MW0687_Sau_21282416                  ---VK-----ADYLIDDNP-RQLE-IF---T-GTPIMF   0  -TAV-HNINDDRFERVNSWKDVEQYFLDNIE-  179/          |
VNG2608C_Hsal_15791343               RT-GAD----SVVFVGDTL-DDVATALNAREADPGREY  14  GRQKYHTAGADHVLDSVNDL-PALLGQ-----  290\VNG2608C|
Selo03001632_Telo_46130459           QL-ATNPHL-PAIYAGDTV-ADILTVRRARYEASDRPW  18  YAQQLTEAGADCVIEGVEALTSDHIQALLA--  266/          |
At2g45990_At_21280859                ---EHQGL--TLHFVEDRL-ATLK-NVIKEPELDKWSL   1  LGTWGYNTEKERAEAAGIPRIQVIELSTFSN-  265\HerA      |
Tery02004909_Tery_48891718           ---GAR-T--TIWFIEDRL-KTLL-SIQKHPDLQEVEL   1  LADWGYNTQKERNSVAQYPSIHLLSSAQFCQ-  256|          |
alr3863_Ana_17231355                 ---DHEPV--SLWFVEDRI-KTLQ-LVQQQSDLEDVKL   1  LADWGYNTQSERKAAQSDPRIQLLSLSQFAK-  255/          |
NT5C3_Hs_47940472                    NQ-LKDNS--NIILLGDSQ-GDL--RMADGVANVEHIL   7  RVDELLEKYMDSYDIVLVQDESLEVANSILQK  329\P5N-1     |
MGC20781_Hs_34147472                 QQ-LEGKT--NVILLGDSI-GDL--TMADGVPGVQNIL   7  KVEERRERYMDSYDIVLEKDETLDVVNGLLQH  279/          |
YPTB0875_Ypse_51595225               QL-GIAPQ--ALLFLSDIR-QELDAAQLAGWHTCQLIR   0  -----DLPDNDSAHPQVNRFDQIVLSLFTE--  229\Enolase   |
masA_Ana_33866496                    KI-STNPS--EILFISDNG-EECDAAGASGMETLFSLR   0  -----DGNP-DQSPRSHRVIKTLNDVDEYL--  245|Phosph.   |
YALI0D16797g_Ylip_50550877           AI-GFEAD--RVLFLSDNV-REIEAAKKAGLRAYVAER   0  -PGNAKLTPQEKEDNVIKTSFEGIEI------  233|          |
LOC305177_Rnor_34876231              SI-GCSTN--NILFLTDVT-VEASAAEEADVHVAVVVR   0  -PGNAGLTDDEKTYYNLISSFSELYLPSTQS-  259/          |
AAL37424_Hp_17225481                 QE-NITHT--QILHIGDNSWADDT-MPKSLGIATLFRK  11  KYKTFTPTSVAQSFIL--GSLCVFHKNYIQKH  242\Bcs3      |
RPA0638_Rpal_39933715                QQ-GVDAS--SAYHVGDNEHADIR-GAKRHGIRPRHYP  18  LMFKGQPTRLDRGARTLRRMVAARSAEQSAAH  271/          |
SerB_Mjan_14719643_1F5S              GI-N--LE--DTVAVGDGA-NDI--SMFKKAGLKIAFC   0  --AKPILK--EKADICIEK-RDLREILKYIK-  211\PSP       |
PSPH_Hs_31615696_1NNL                ---H--FK--KIIMIGDGA-TDM--EACPPADAFIGFG   0  --GNVIRQQVKDNAKWYI--TDFVELLG----  221|          |
thrH_Pae_47168798_1RKU               ---L--YY--RVIAAGDSY-NDT--TMLSEAHAGILFH   0  --APENVIREFPQFPAVHTYEDLKREFLKASS  201|          |
LOC295663_Rnor_27700068              LQ-KGVRYT-RIVYIGDGG-NDVC-PVTFLKKNDVAMP   3  YTLHRTLDKMSQNLEPMASSIVVWSSGMEIIS  233|          |
BC4037_Bcer_30022124                 ---D--DF--HI-VIGDSI-TDL--QAAKQADKVFARD   0  ----FLITKCEENHIAYTPFETFQDVQAELK-  212|          |
NCU06554.1_Ncr_32411857              AD-Q--RP--VLFYAGDGV-SDL--SAAKETDLLFAKA   0  --GKDLITYCEKENVPFTTFHDFTEILAVVK-  241|          |
Rgel02003758_Rgel_47572040           GR-R-VQDSERVTVYSDST-NDL--PLLELATHPVATN   0  --PSPALEAIARERGWPLLKLFP---------  222|          |
Ta0936_Taci_16081982                 HI-S--KD--ETISIGDSM-ENG--MIHINSSKAYIIG   0  --GKGRHGEIPLGNNIRNLLSYI---------  212/          /
Zr25_Sau_39654743_1QYI               GN12DNIVNKDDVFIVGDSL-ADLL-SAQKIGATFIGTL  7  AAGELEAHHADYVINHLGELRGVLDNLLEHH-  380\Zr25      \multihelical C1 cap
SE1402_Sepi_27468320                 GN12EDIVNKDEVYIVGDSL-ADLL-SAKKIGATFIGTL  7  AHSELVANGADHVVEDITKIRKILL-------  374/          |
CG32549_Dm_45555998                  RQ35NAKGK--DVLYVGDHIFGDIL-KSKKIRGWRTFLI  0  ---VPELVRELH-----VWTDEC--QLFAELQ  522\cN-II     |
TU12B1-TY_Hs_30354531                RK33GWRGS--RVLYFGDHIYSDLA-DLTLKHGWRTGAI  0  ---IPELRSELK-----IMNTEQYIQTMTWLQ  434/          |
GSTEN:00030281:G:001_Tnig_47224252   ---GRKV---VYVVVGDGV-EEE--QAAKKHNMPFWRI   0  --SSHSDLLALHQALEFEYL------------  860\EYA      |
EYA3_Hs_27371316                     ---GKKV---TYVVIGDGR-DEE--IAAKQQLYFEAL-   0  --GCQLEPTALILFIQLSYN------------  536|          |
AAM65149.1_At _21593200              ---HPKF---RFCAIGDGW-EEC--AAAQALQWPFVKI   0  --DLQPDSSHRFPGLTPKYG------------  307/          /
ATP2A1_Ocun_48425717_1SU4            ---S-YDE--ITAMTGDGV-NDA--PALKKAEIGIAMG   0  SGTAVAKTASEMVLADDNFSTIVAAVEEGRA-  752\P-ATPase\C1 ATPase cap
Ssui801000538_Ssui_50591442          ---E-QGK--TVAMVGDGI-NDA--PALAQAHVGLAIG   0  SGTDIAIESADIVLMHSDILDVVKAVKLSQA-  753|          |
Npun02004103_Npun_23127189           ---S-QGK--LVAMTGDGT-NDA--PALAQANVGVAMN   0  SGTQAAKEAANMVDLDSDPTKLIDLVTIGKQ-  603|          |
BC0448_Bcer_30018656                 ---A-KGN--IVSMTGDGV-NDA--PSLKQADVGVAMG   1  TGTDVAKGAADVVLTDDNFSSIVKAVEEGRN-  666|          |
ENSANGG00000006676_Agam_31240685     ---E-MGM--RVAMVGDGV-NDS--PALAQADVGIAIA   0  SGTDVAAEAADVVLMRNDLLDVVACLDLSRK-  878/          /
Apc0014_Taci_28373517_1L6R           ---SLEYD--EILVIGDSN-NDM--PMFQLPVRKACPA   0  NATDNIKAVSDFVSDYSYGEEIGQIFKHFEL-  226\Cof       \C2 cap
Ywpj_ABs_34810977_1NRW               ---NIPLE--ETAAVGDSL-NDK--SMLEAAGKGVAMG   0  NAREDIKSIADAVTLTNDEHGVAHMMKHLL--  288|          |
cof_Ec_15800176                      ---GLSLR--DCMAFGDAM-NDR--EMLGSVGSGFIMG   0  NAMPQLRAELPHLPVIGHCRNQAVSHYLTHW-  265|          |
CV1727_Cvio_34103039                 ---GIAAD--QMVCFGNDS-NDL--SMFAVARHAVLIG   0  DH-PQLRPHARERIIRDQHVERELIAAIQRL-  249|          |
Exigu03001745_Ana_45531950           ---NIRRE--DTFAFGDAL-NDL--EMLRYVGTGIAMG   0  NGLAETKAAADFVTKHILEDGIEHGLKEFQLI  258/          |
Mbur03000506_Mbur_46142672           RQ-QFTEV--VTIALGDSL-NDL--PMLKAVDIPFLVQ   0  --KPDGKYDPSIILTE-IKHAEGIGPVGWNN-  260\MPGP      |
PMM0960_Pmar_33861516                NN-PNI----KIIGLGDSP-NDL--PLLLNSDYKIVIP   0  GPGGPNLKLIEKLNEYSFTLATDPNGYGWKS-  253|          |
Z3045_Ec_15802389                    QQ-LSGKRP-TTLGLGDGP-NDA--PLLEVMDYAVIVK   0  GLNREGVHLHDEDPAR-VWRTQREGPEGWRE-  262/          |
PMM1_Hs_16877082                     ---QDSFD--TIHFFGNET4NDFEIFADPRTVGHSVVS   0  --------PQDTVQRCREIFFPE---------  257\PMM       |
FG03767.1_Gzea_42548616              ---GIPLE--QIMFIGDAI4NDYP-AKELGLHTVRVKN   0  --------PDGTLAAIAGIVACL---------  286/          |
SPS_Vfab_1022365                     GF-ELSKMVVFVGECGDTD-YEG--LVGGLHKSVILKG   0  ----VGSRAISQLHNNRNYPLSDVMPLDSPN-  1034\SPS      |
P0678F11.14_Osa_50902372             GL-SVGNMYLIVGEHGDTD-HEE--MLSGLHKTVIIRG   0  ----VTEKGSEQLVRSSGSYQREDVVPSESPL  1073/         |
At2g35840_At_21387099                KT-EGKLPV-NTLACGDSG-NDAE-LFSIPDVYGVMVS   0  NAQEELLKWHAENAKDNPKVIHAKERCAGG--  251\SPP       |
spp1_Zmay_11127755                   SS-CGKPPN-NTLVCGDSG-NDAE-LFSIPGVHGVMVS   0  NAQEELLQWYTENAKDNPKIIHSNERCAAG--  251|          |
CV2260_Cvio_34103571                 GV-DPR----RVIVFGDSC-NDI--EMLTGYENGYLVG   1  ALAEARRVFDRV-VDGIYCHGIIDGLQRH---  260|          |
yhjK_Blic_52350125                   AI-ERE----NSFAFGDSG-NDV--DMLKKTGHGYLVG   1  ATAEAKSLHSRM-TEGSYTAGILEVLKAH---  279/          |
AAM63513_At_21554408                 GF-ANSNDV-LPIYIGDDR-TDEDAFKVLRNKGQGFGI  6  KETSATYSLQEPSEVGEFLQRLVEWKQMSLRG  368\OtsB      |
otsB_Rsol_17549325                   PF-AGR----IALFAGDDL-TDESAFEAVNT-LGGWSI  4  GPSQAHWRVPDPAALRDWLAALARRAGGRAA-  262|          |
otsB_Sent_16760878                   PF-TGR----IPVFVGDDL-TDEAGFGVVNH-AGGISV  4  GATQAAWRLESVPDVWRWLEQINYPQQEQQVM  253|          |
MG08707.4_Mgri_38104651              RS-GDLSPD-FLMVVGDGR-EDEKVFKWANQLRGTYGS  6  NASHATYSVVQHRGYLVAEENLTRPDGMEILG  953|          |
OJ1655_B12.15_Osa_50252610           RE-NSLLPD-FVLCIGDDR-SDEDMFEVITTAAQDNCL  13  KPSKAKYYLDDLADIVRLIQGLANVSDEMHST  853/          |
Tm1742_Tmar_47169464_1VJR            KF-GVPKE--RMAMVGDRLYTDVK-LGKNAGIVSILVL  6  EDLERAETKPDFVFKNLGELAKAVQ-------  271\NagD      |
SMU.1415c_Smut_24379818              RL-GVKRH--EAIMVGDNYLTDIT-AGIKNDIATLLVT  6  EEVPALPIQPDFVLSSLAEWDFDER-------  257|          |
CG32487_Dm_24656330                  KG-VIQPD--RTLIIGDTMCTDIL-LGYKCGFQTLLVG  15  KAPLLYQQVPDLYMPKLSNLLPFLSSRNR---  320|          |
pho2_Sp_19113047                     NV-NFDRK--KACFVGDRLNTDIQ-FAKNSNLGGSLLV  9  ILEKDAPVVPDYYVESLAKLAETA--------  298|          |
BMEII1045_Bmel_17989390              VV5GSVDKS-RILGIGDGVLTDVK-GAADFGLDVLYIS  4  AADYAVNGDLDMAKFMEALEKHGHRPIASLHA  284|          |
GSTEN:00012682:G:001_Tnig_47211717   LV7GWTTPVKRLYAIGDNPMADI--YGANLYNRYLHAS  2  TKAQVQAKRGIRGSDPVADSVDGDPKTTSAGG  361|          |
NCU04924.1_Ncr_32407534              VI6GEEKIPQNIYMVGDNPASDI--VGGNLYGWNTCLV  2  GVFQ-GGE--NDEENPANF-GVFANVWEAVT-  496/          |
PSPTO0186_Psyr_28850660              HY-SADLK--GLWFVGDSK-GDL--QAALAVDS-QPVL  11  GEVPTGTLIFDDLAAVAAELIHTSASLNN---  183\HisB      |
hisB_Ec_15802501                     EQ-AMDRA--NSYVIGDRA-TDI--QLAENMGI-NGLR  41  EGGSKINTGVGFFDHMLDQIATHGGFRMEINV  221/          /
Consensus/80%                        .. ........hhhhsD.. .s....h........h..     ................................
```

Figure 4 (*legend on next page*)

of the other specific features conserved throughout the HAD superfamily beyond the phosphorylated aspartate and other generic features of the acidic active-site-containing division of Rossmannoid folds (Figure 3).

Of the other Rossmannoid folds of this division, the DHH phosphoesterases contain a DxD signature, and the histone deacetylases/arginases a DxH signature at the end of strand 1, which chelate a metal ion, just as in the HAD superfamily. However, these enzymes also contain their own characteristic motifs further downstream (Figure 3) and there is no evidence for any aspartylphosphate intermediate being formed.[52,59,60] In the PIN/5′-3′ nuclease domains, a catalytic Mg^{2+} is chelated by the acidic residues including those occurring at the end of the S1 equivalent and the strand immediately to its left[51] activates a water for nucleophilic attack. In the TOPRIM domains of primases and topoisomerases the acidic residue at the end of the first strand is always a glutamate (Figure 3) that acts as a general acid or base in the hydrolysis of the phosphoester bond or polynucleotide transfer.[50,61] The DXD motif is instead borne at the end of the strand left of S1 (Figure 3) and coordinates a Mg^{2+}. In the vWA domain the first aspartate is part of the so called MIDAS metal-binding motif (DxSxS[53,62]), which is critical for metal chelation by these domains. Thus, different superfamilies of this division of the Rossmannoid folds, despite similarly positioned acidic catalytic residues and metal coordination sites, have acquired very distinct catalytic mechanisms. Large to moderate inserts within the core Rossmannoid domain are also seen in the TOPRIM, PIN/5′-3′ nuclease domains, histone deactylase/arginase and DHH superfamilies, suggesting that they might also form caps controlling access to the active site area, analogous to the HAD superfamily.

Structural variations in the core Rossmannoid domain of the HAD superfamily

The core Rossmannoid fold of the HAD superfamily is generally not prone to many modifications beyond the insertion of the cap modules. However, the central sheet often shows lateral modifications corresponding to the two ends of the sheet. The ancestral condition of the HAD appears to have been the five-stranded central sheet (Figure 3), to which a major division of the HAD superfamily appears to have added a C-terminal β-α unit after the fifth[f] strand-helix unit (S6), extending the sandwich further (at the left side of the sheet in Figure 3). The additional strand S6 was lost on rare occasions in members of this six-stranded division, especially in the context of C-terminal domain fusions. Likewise, on the opposite side (right end of the sheet in Figure 3) there are inserts of additional strands, which stack in the same plane as the core strands to extend the sheet. The simplest of these is a β-hairpin, which folds back and extends the central sheet, and is the defining feature of a large clade within the HAD superfamily that includes the sucrose phosphate phosphatases, the phosphomannomutases, the trehalose phosphate phosphatases, mannosyl-3-phosphoglycerate phosphatases and the cof-type phosphatases (Figure 3). A second independent insert in the “right side” of the sheet is seen in the P-type ATPases in the form of an additional α-β unit immediately after S3 (Figure 3, bottom left). This additional strand is accommodated in the sheet between the S2 and S3 and is a unique and defining feature of the P-type ATPases.

Figure 4. Multiple sequence alignment of HAD-domain containing proteins. The alignment shows only conserved structural regions. Structural regions not conserved, including cap regions, are replaced with numbers denoting the residues in these excised regions. The top line of the alignment indicates the approximate areas of the four conserved motifs considered essential for HAD domain catalytic activity. Conserved residues of these motifs are shaded in gray. Secondary structure motifs are colored and labeled in the second line of the alignment, blue representing β-sheets, red representing α-helices, and pink representing the squiggle motif. The third line of the alignment designates secondary structural elements; E for β-strand regions, H for α-helical regions, and – for coil regions. Widely conserved hydrophobic residues are shaded in yellow (A,C,F,I,L,M,V,W and Y) while conserved aliphatic residues appear in gray and are shaded in yellow (I,L and V). Conserved small residues appear in green (G,A and S), hydroxy residues appear in teal (S and T), positive residues appear in blue (K and R), and negative residues appear in red (E and D). Sequences are identified by the protein name, species name abbreviation, the GenBank GI number, and if applicable the PDB code; identifiers are demarcated by underscores. PDB codes are shaded in orange for added emphasis. Species names are abbreviated as follows: Agam, *Anopheles gambiae*; Ana, *Nostoc sp.*; At, *Arabidopsis thaliana*; BPRB69, enterobacteria phage RB69; Bcer, *Bacillus cereus*; Bhal, *Bacillus halodurans*; Blic, *Bacillus licheniformis*; Bmel, *Brucella melitensis*; Bs, *Bacillus subtilis*; Cace, *Clostridium acetobutylicum*; Cfum, *Choristoneura fumiferana*; Cgla, *Candida glabrata*; Cneo, *Cryptococcus neoformans*; CpGV, *Cydia pomonella granulovirus*; Cvio, *Chromobacterium violaceum*; Dm, *Drosophila melanogaster*; Ec, *Escherichia coli*; Efae, *Enterococcus faecalis*; Exsp, *Exiguobacterium sp.*; Gzea, *Gibberella zeae*; Hinf, *Haemophilus influenzae*; Hp, *Helicobacter pylori*; Hs, *Homo sapiens*; Hsp, *Halobacterium sp.*; Llac, *Lactococcus lactis*; Mbur, *Methanococcoides burtonii*; Mgri, *Magnaporthe grisea*; Mjan, *Methanocaldococcus jannaschii*; Msp., *Mesorhizobium sp.*; Ncr, *Neurospora crassa*; Npun, *Nostoc punctiforme*; Ocun, *Oryctolagus cuniculus*; Osa, *Oryza sativa*; Pae, *Pseudomonas aeruginosa*; Pmar, *Prochlorococcus marinus*; Pput, *Pseudomonas putida*; Psp., *Pseudomonas sp.*; Psyr, *Pseudomonas syringae*; Ptor, *Picrophilus torridus*; Rgel, *Rubrivivax gelatinosus*; Rnor, *Rattus norvegicus*; Rpal, *Rhodopseudomonas palustris*; Rsol, *Ralstonia solanacearum*; Saga, *Streptococcus agalactiae*; Sau, *Staphylococcus aureus*; Sc, *Saccharomyces cerevisiae*; Sent, *Salmonella enterica*; Sepi, *Staphylococcus epidermidis*; Smut, *Streptococcus mutans*; Sp, *Schizosaccharomyces pombe*; Ssui, *Streptococcus suis*; Syn, *Synechococcus sp.*; Taci, *Thermoplasma acidophilum*; Telo, *Thermosynechococcus elongatus*; Tery, *Trichodesmium erythraeum*; Tmar, *Thermotoga maritima*; Tnig, *Tetraodon nigroviridis*; Vfab, *Vicia faba*; Xaut, *Xanthobacter autotrophicus*; Ylip, *Yarrowia lipolytica*; Yp, *Yersinia pestis*; Ypse, *Yersinia pseudotuberculosis*; Zmay, *Zea mays*. Alignment was produced with the aid of the Chroma program.[196]

The most dramatic modification, however, is seen in the proteobacterial BcbF family of phosphatases, which exist as obligate dimers in the catalytic form (Figure 3, bottom right). In these proteins the helix immediately downstream of the conserved lysine in motif III is replaced by a loop, which displaces the strand S4 away from the core sheet and places it an anti-parallel configuration, where it stacks with the remaining three strands (S1–S3) of the second monomer in a parallel configuration (Figure 3). Thus, the S4 appears to be swapped between the two monomers and two identical active sites are formed by a combination of two monomers, one monomer supplying motifs I, II and III and the other monomer supplying motif IV associated with the swapped strand (Figure 3). Given that this configuration has a very limited phyletic spread, this dramatic modification appears to have evolved rather recently through a relatively simple process. We suspect that the ancestral version was a five-stranded version, which probably functioned as a tightly associated dimer with the active sites in each ancestral monomer facing in opposite directions (head–tail dimer). In such a head–tail dimer, accidental swapping of strand 4 between the monomeric subunits could have re-constituted a functional active enzyme, thereby allowing the emergence of the configuration seen in the BcbF family.

Cap modules of the HAD superfamily

The most notable inserts seen in the HAD superfamily are the caps, which, despite their diversity, can be classified in three generic categories: (1) C0 caps, the structurally simplest representatives of the HAD superfamily, have only small inserts in either of the two points of cap insertion. (2) The C1 are defined as inserts occurring in the middle of the β-hairpin of the flap motif, and fold into a structural unit distinct from the core domain. (3) The C2 caps are defined as inserts occurring in the linker immediately after strand S3 (Figure 3). Most representatives of the HAD superfamily have either a C1 cap or a C2, though in few cases proteins may simultaneously possess C1 and C2 caps.

The simplest C0 state with no elaboration of β-hairpin or additional inserts in the C2 position are rather infrequent in the HAD superfamily and are seen in proteins such as deoxy-D-mannose-octulosonate 8-phosphate (KDO 8-P) phosphatase (Figure 5). Slightly longer inserts are seen in the polynucleotide phosphatases, which have a long loop separating the two β-strands of the flap. In the case of the CTD phosphatases and MDP-1 like phosphatases, this basic condition is elaborated further, with the addition of a strand between the two sheets forming the β-hairpin; resulting in a cap in the form of three-stranded sheet. Some of these phosphatases have also acquired a rudimentary C2 cap in the form of a long loop that extends out of the core domain.

The classical C1 caps belong to two distinct structural classes, the α-helical C1 caps and the cap with the unique α+β fold seen in the P-type ATPases (Figure 5). The most basic α-helical cap in the form of bi-helical α-hairpin is observed in the acid phosphatase and the cN-I nucleotidase families (Figure 5). The next level of complexity is the tetra-helical bundle, which is the form of the C1 cap seen in the majority of HAD domains with a cap in this position. It includes three general subclasses that may be distinguished based on structural properties and conserved interactions. The first subclass, represented by β-phosphoglucomutases and deoxyribonucleotidases, has conserved contacts between the descending arm of the cap domain and the second helix of the Rossmannoid core. The second subclass seen in haloacid dehalogenases and their close relatives (see below) has conserved contacts involving the loop between the second and third helices of the cap and the linker between strand S3 and the core helix downstream of it. The third subclass, typified by the phosphoserine phosphatase family, shows contacts in the region between the third and fourth α-helices of the C1 cap and a smaller C2 cap that is unique to this family. Despite sharing the same topology, these three categories of tetra-helical C1 caps share little primary sequence conservation, and show notable differences in the packing of the helices. The largest helical caps are seen in the form of the globular multi-helical bundle found in the uncharacterized Zr25 family, with a core formed by eight prominent helices (Figure 5). Secondary structure prediction for the cN-II nucleotidase and Eyes absent (EYA) families reveals the presence of large caps, which are predicted to form multi-helical bundles similar to the Zr25 (the cap of cN-II has developed an additional beta meander).

The P-type ATPase C1 caps are unrelated to the helical caps and searches of the PDB database with the DALI program[63,64] do not recover any known fold. However, an analysis of the P-type ATPase caps showed that they contain an internal duplication of a simple α+β unit, with a core sheet formed by a three-stranded β-meander (Figure 5). This suggests they possibly arose from a single ancestral unit, which in turn could have itself emerged from a precursor resembling the C0 caps of the CTD phosphatases and MDP1 *via* the addition of a small α-helical hairpin to the three-stranded sheet. Subsequent duplication of this unit appears to have generated the C1 cap seen in extant P-type ATPases (Figure 5). However, the C1 cap of the extant P-type ATPases manifests considerable variability both in terms of sequence as well as in the form of some additional insertion and deletions. Thus, in the most parsimonious scenario, the classical C1 caps appear to have been independently invented at least twice. All the known α-helical caps can be conservatively pictured as an evolutionary series of α-helical bundles of increasing complexity emerging through serial duplication from a basic bihelical precursor, along with rapid sequence divergence and reorganization of the helical packing (Figure 5).

There are two major unrelated types of classical C2 caps, respectively, seen in the Cof-type phosphatases and the NagD-like phosphatases and its

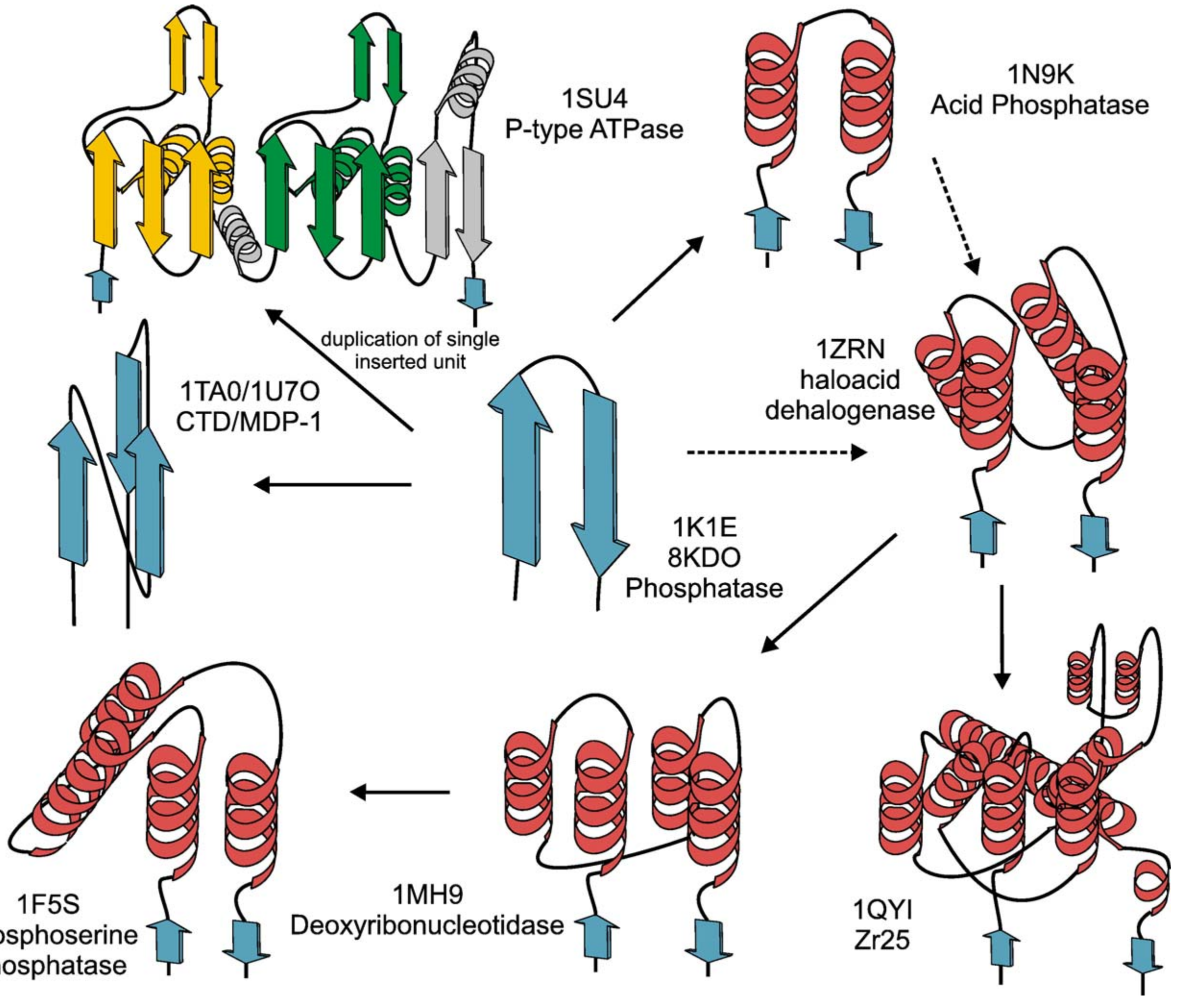


Figure 5. Topology diagrams of selected C0 and C1 cap HAD domains. Representatives are identified with the PDB code followed by one or more HAD family/clade name(s). With the exception of the P-type ATPase representative, strands are shown as blue arrows with the arrowhead on the C-terminal side and helices are represented by red coils. Central to the diagram is the ancestral strand–strand C0 cap. Remnants of the basal C0 cap can be seen in the small strands leading into and out of all other C1 caps. Arrows refer to the likely evolutionary progression leading to the diversification of C1 caps. Broken arrows pointing to the 1O08 tetra-helix cap domain reflect two possible progression scenarios. The 1SU4 P-type ATPase cap is colored to accentuate a possible duplication event. The first unit of the cap is colored in yellow and the second is colored in green. Other pieces of the cap that likely developed around the duplication event, including a strand–strand motif at the C terminus of the cap and a single helix linking the two units, are rendered in gray.

relatives (Figure 6). Both these types of C2 caps are distinctly α+β with a core β-sheet containing at least three strands. However, in structural similarity searches with the DALI program[63,64] and through manual examination of topologies, we were unable to detect any convincing similarity to other folds in the protein universe, or between themselves. In addition to these major classes of C2 caps there is a yet another small, unique C2 cap found in the histidinol phosphatase family. In the Cof-type phosphatases we observed a remarkable diversification of the C2 cap through accretion of secondary structure elements to a basic unit with a three-stranded anti-parallel β-sheet (Figure 6). The most basic version, seen in the protein Ta0175 (PDB: 1L6R) from *Thermoplasma acidophilum*,[65] contains a three-stranded antiparallel sheet. A slightly more complex form is seen in the trehalose-6-phosphatase ortholog (1U02) from the same organism, where a strand is added to the sheet at the N terminus. In some other forms (e.g. YedP from *Escherichia coli*, PDB: 1XVI) there is entire β-α unit, instead of single strand, added to the N terminus of the ancestral unit (Figure 6). In the uncharacterized phosphatase Tm0651 from *Thermotoga maritima* (PDB: 1NF2),[26] this trend is further exaggerated *via* the addition of three α-β units to the ancestral unit. In the related YwpJ (1NRW) from *Bacillus subtilis*, in contrast, we observe elaboration via duplication of a helix in one of the α-β units. Thus, as in the case of the helical C1 caps, it appears that the C2 caps of the Cof-type phosphatases evolved through a process of serial addition of simple secondary structure units, most probably through duplications limited to the N-terminal region of the cap.

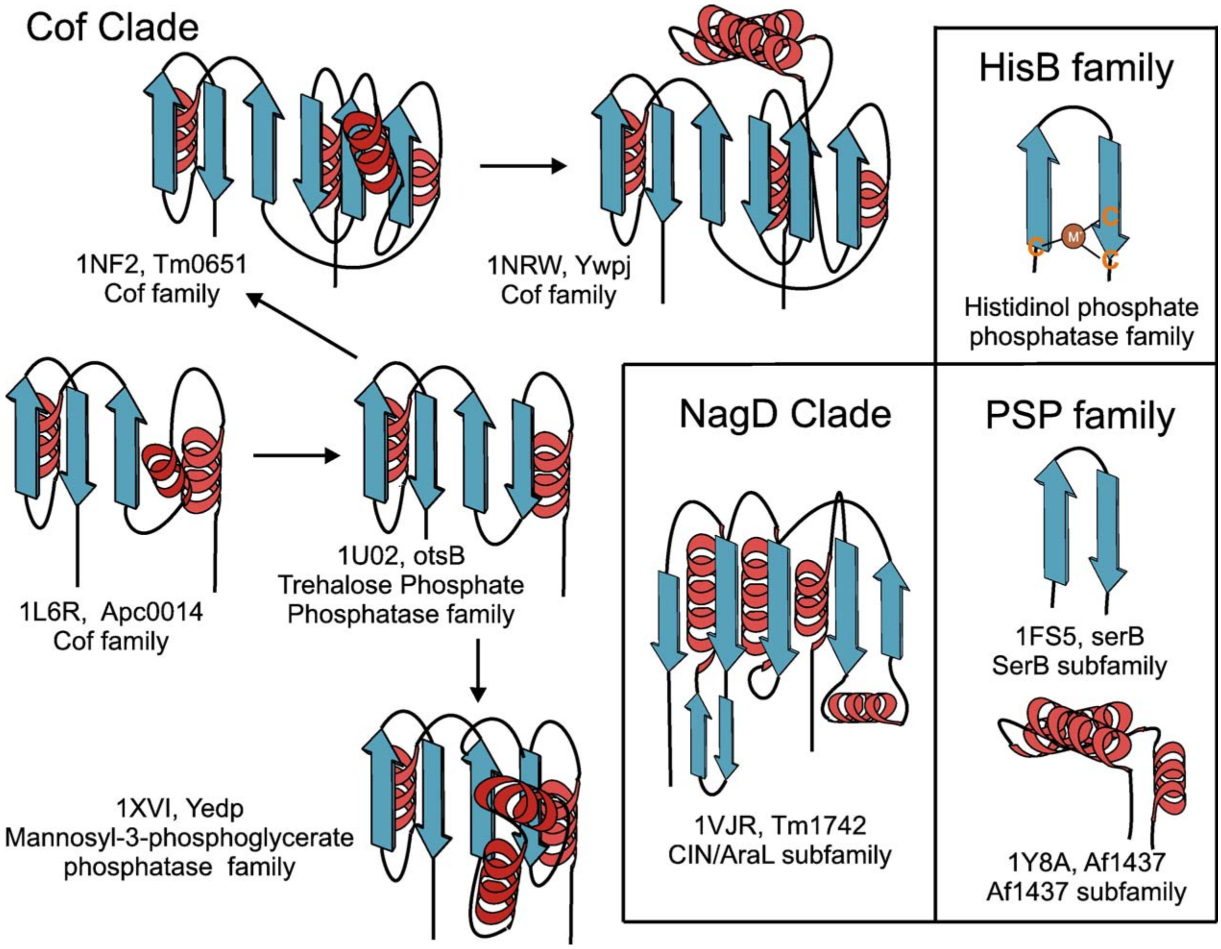


Figure 6. Topology diagrams of selected HAD C2 cap domains. Representatives are identified with PDB code and family/clade name. Strands are shown as arrows rendered in blue with the arrowhead on the C-terminal side. The main section of the diagram contains arrows that show the likely evolutionary progression of C2 caps in the Cof Hydrolase clade. The depiction of 1L6R represents the most likely ancestral state. The boxes to the right illustrate three additional independent innovations of the C2 caps in the NagD clade, the HisB clade, and the phosphoserine phosphatase (PSP) family. The depiction of the HisB C2 cap is predicted, as no structure is currently available. Orange Cs represent the likely locations of putative metal-chelating cysteine residues.

The C2 cap of the NagD-like phosphatases is an α/β domain with a core four-stranded parallel β-sheet, with an additional N-terminal antiparallel strand. The parallel configuration of the sheet, combined with the lack of specific similarities to any other known domain, suggests that it might have arisen *via* a duplication of the core domain which also has a parallel β-sheet. However, at the sequence level there is no significant similarity with the core domain. This group of C2 caps also contains a unique beta hairpin inserted after the third strand (Figure 6). An examination of the sequence of the C2 caps of the histidinol phosphatase family reveals a conserved CxHx(6-13) CxC signature (where x is any amino acid). This suggests that this C2 cap is stabilized through the chelation of a divalent metal ion, and is likely to assume a simple flap-like structure (Figure 6).

Several lineages of the HAD superfamily simultaneously possess both C1 and C2 caps, both of which may be similarly sized, or one of them may be the dominant cap. In the case of the enzymes with C0 caps such as the CTD phosphatase family and the related ROP9/38K family there is sometimes an additional C2 cap in the form of a small β-hairpin. Similarly, small β-hairpin C2 caps are also seen in the phosphoserine phosphatase and the pyrimidine 5- nucleotidase families, which also contain helical C1 caps (Figure 6). In an archaeal sub-family of the phosphoserine phosphatases, typified by the protein AF1437 (PDB: 1Y8A), a small C2 cap assuming the form of a tri-helical bundle is seen, suggesting that there have been multiple independent innovations of such smaller C2 caps. In all these families the C1 cap is clearly the dominant cap with the C2 cap packing against it and probably providing an additional solvent exclusion module (see below).

Role of the cap modules in the catalytic mechanism of the HAD superfamily

Several studies have revealed that HAD enzymes with C1 caps are likely to follow a similar catalytic cycle comprised of the steps outlined below.[17,27,32,37,66–68] The enzyme in the "open" configuration allows the substrate (typically a phosphoester) to enter the active site. Once the substrate is bound the enzyme assumes the "closed" configuration and the Mg^{2+} in the active site interacts with the negatively charged phosphate, preparing it for nucleophilic attack by the first conserved aspartate at the end of strand one (Figure 1(a)). As a result an acyl phosphate intermediate is formed with the carboxyl group of this aspartate.[27–31] Subsequently, the enzyme enters the open configuration again and allows the leaving group to escape (Figure 1(a)). In the open state bulk solvent enters the active site and a water is deprotonated by the second aspartate of strand one; hydrolyzing the acyl phosphate intermediate and returning the enzyme to the native state.[32] A variation on this theme is seen in the haloacid dehalogenases which release a halide ion along with the formation of a regular ester linkage.[69] In the phosphonatases and sugar phosphate mutases there are differences in the initial and the terminal stages of the reaction, respectively[27,32,38,70–72] (Figure 1(a)) but the core phosphoryl transfer mechanism remains the same.

Key aspects of the HAD catalytic mechanism that emerged from these studies are: (1) the alternation between open and close states and (2) a preliminary reaction favored by solvent exclusion and a subsequent step favored by extensive solvent contact. The principal features of the core domain responsible for this process are the squiggle and the flap. The squiggle, being close to a helical conformation, appears to be a structure that can be alternatively tightly or loosely wound (Figures 2(a) and 3). This differential winding in turn induces a movement in the flap immediately juxtaposed to the active site (Figures 2(a) and 3) and alternatively results in the closed and open states. Given the strict conservation of the squiggle and the flap across the HAD superfamily found herein, they are likely to be part of a universal essential functional feature of this superfamily. The conformational changes in the squiggle and flap are likely to comprise the minimal apparatus for solvent exclusion and access at the active site of these enzymes. Given this ground state, natural selection appears to have favored the emergence of cap modules as they made the process of solvent exclusion and acyl phosphate formation more efficient. In addition to aiding the basic catalytic mechanism the emergence of diverse caps also provided a means of substrate recognition by supplying new surfaces for interaction with substrates, which was not afforded by the ancestral active site alone.[28,40,41]

The simplest structures add the cap to the flap motif itself, so as to completely seal the active site in the closed state (Figure 7). Thus, the flap region was a hotspot for the insertion of the various C1 caps, which appears to suggest intense natural selection for efficient solvent exclusion.[27,32,38,69,70,73] In the case of the HAD enzymes with C2 caps there is no evidence from either biochemical or structural studies, thus far, for extensive movement of the cap itself to result in open and closed states. However, an examination of the internal cavities of the available structures of the HAD enzymes with C2 caps shows that the C2 cap forms a cavernous structure over the active site with the flap sealing off the aperture to this cavity (Figure 7). This implies that although the C2 caps likely lack mobility comparable to the C1 caps, even in these cases the squiggle–flap elements likely exhibit drastic movements similar to that observed in C1 caps. As a result there would be an open state in which the substrate, solvent and leaving group can be exchanged with the active site cavity and a closed state where the flap occludes the cavity formed by the C2 cap completely and excludes the solvent. In most cases where both C1 and C2 caps are present such as the phosphoserine phosphatase family, the C1 cap is the principal functional moiety that closes the active site. The subsidiary C2 cap packs against the C1 cap and completes the occlusion by sealing off potential channels to the active site that exist in these C1 caps. In most of the C0 Caps the rudimentary caps forms a crater-like structure associated with the active site (e.g. MDP-1 and the CTD phosphatase families) (Figure 7). In the case of the polynucleotide kinase phosphatases (PNKP) this crater-like structure is also walled by a unique insert occurring immediately after strand S4 with motif IV. These crater-like accesses to the active site of the C0 cap enzymes are unlikely to completely occlude the solvent, but their substrates are large molecules (proteins and polynucleotides), which may block the rest of the active site from solvent while being bound to it. Another C0 cap enzyme, the 8KDO phosphatase, adopts an unusual strategy for solvent exclusion by using the particularly elongated strands of its flap to form a tetramer interface. As a result, each monomer in the tetrameric unit forms a cap over the active site in the adjacent monomer, effectively performing the same function of solvent exclusion (Figure 7). A similar strategy of occlusion *via* cooperation between two subunits is also seen in the aberrant BcbF family, which shows strand swapping between adjacent subunits of the obligate dimer.

Natural classification of the HAD superfamily

Identification and clustering of the HAD superfamily enzymes

We identified all available structures of the HAD superfamily by using the DALI program[63] to search the PDB database with the coordinates of previously well-known HAD domains. HAD structures were typically recovered with *Z*-scores >9.0 regardless of the type of cap present in the structure initiating the search, suggesting strong, detectable relationships between all members of the superfamily (see Materials and Methods and Supplementary Data). We then defined the conserved sequence features (along with their structural cognates) of all HAD superfamily enzymes by means of a structure-based sequence alignment of all available structures (Figure 4). Individual

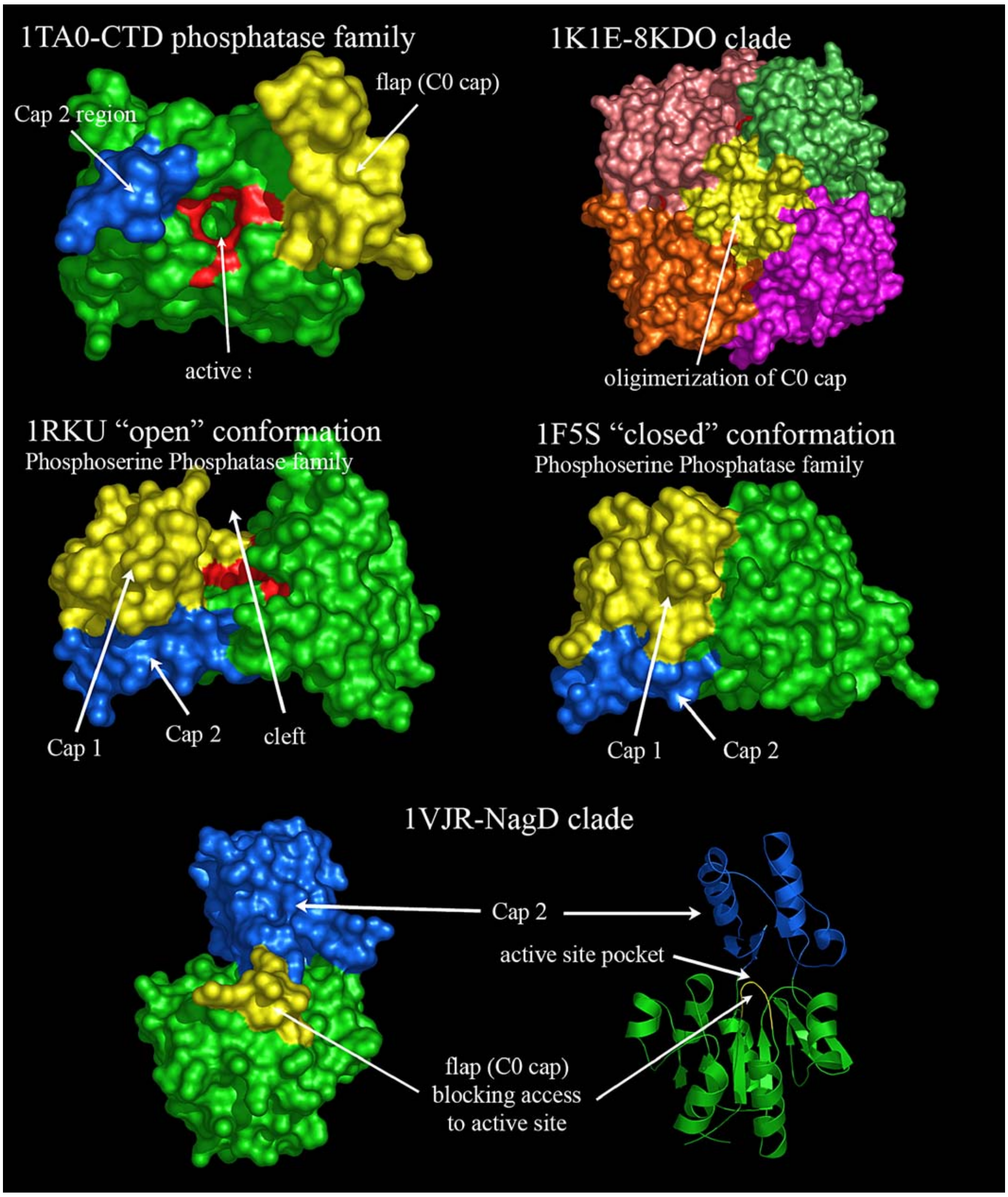


Figure 7. Interaction of cap modules with the active site in the HAD superfamily. Molecular surface diagrams illustrating possible different roles played by the cap domain in substrate recognition and solvent exclusion in different HAD families. The top left shows proximity of primitive C0 and C2 caps to the active site crater found in CTD phosphatases. The top right depicts the role of the C0 cap in tetramerization in 8KDO phosphatases. The middle diagrams show the open and closed states associated with C1 caps, as well as the presence of the β-hairpin C2 cap in the phosphoserine phosphatase family, likely adding an extra layer of solvent exclusion. The bottom diagrams depict the putative role of the C0 cap as a gate to the active site in immobile C2 cap-dominant HAD lineages. With the exception of the top right depiction, core domains are colored green; C0/C1 cap inserts are colored yellow while C2 cap regions are colored blue. In the top right 8KDO phosphatase tetramer rendering, the cap domains are colored yellow while the core domain from each monomer is distinctly colored. Crystal structures are denoted by PDB identifiers followed by family names.

sequences from this alignment were used to initiate iterative PSI-BLAST searches[74] to identify all possible members of the HAD superfamily in the NR database (Materials and Methods and Supplementary Data). Searches were carried out until exhaustion, recovering sequence representatives from known families of HAD domain-containing proteins. For example, a search initiated with the sequence of the crystal structure of 8KDO phosphatase from *H. influenzae*(gi: 20150626, PDB: 1K1E) returns other members of the 8KDO phosphatase family in the first PSI-BLAST iteration. In subsequent iterations, sequences from the Cof hydrolase assemblage (gi: 28373517, iteration 2, *E*-value: 4e-07), P-type ATPase family (gi: 82407772, iteration 3, *E*-value: 4e-11), and phosphoserine phosphatase family (gi: 18160539, iteration 6, *E*-value: 0.002) were recovered. A search initiated with the sequence of a crystal structure from the NagD family (gi: 47169464, PDB: 1VJR) recovered

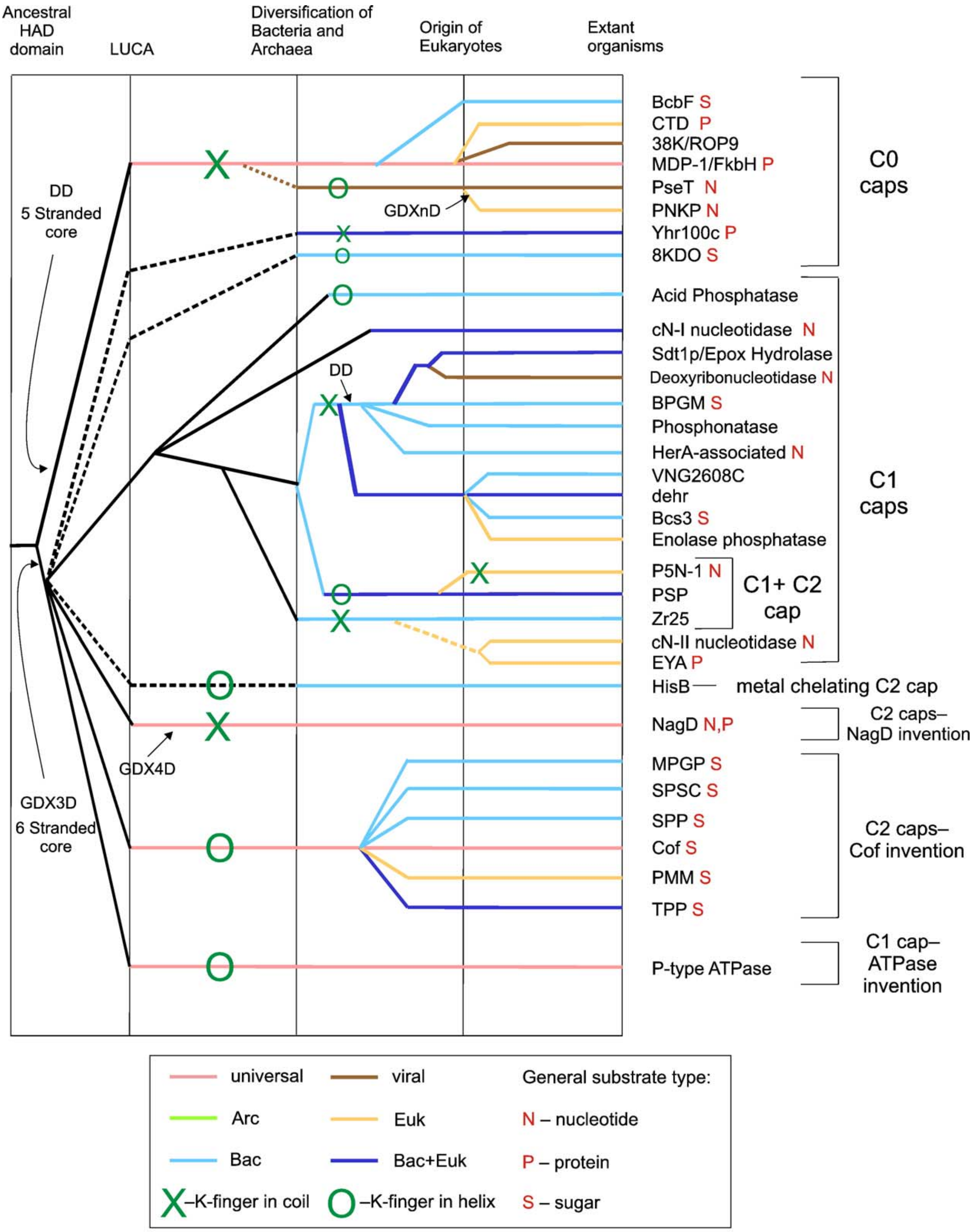


Figure 8. Reconstructed evolutionary scenario for the HAD superfamily. Inferred evolutionary history of HAD domain families. The chart shows relative temporal eras that are demarcated by vertical black lines representing major evolutionary transitions. Individual HAD lineages are listed in a column on the right side of the chart. Horizontal colored lines illustrate the maximum depth to which a HAD domain family can be traced relative to the temporal periods corresponding to the major transitions. Broken horizontal lines indicate that the lineage cannot be traced to a definite starting point. Green Xs and Os lying over horizontal lines indicate position of the lysine finger in a particular HAD lineage or set of HAD lineages. An X indicates the lysine finger is found in the coil region immediately preceding the α-helix N-terminal to S4. An O indicates the lysine finger is incorporated within the aforementioned helix. Red letters to the right of the HAD domain family names represent generalized substrate type(s) known to be processed by a family: N, nucleotide; P, protein; S, sugar. Color key: pink, universal; dark blue, bacteria and eukaryota; brown, virus; green, archaea; light blue, bacteria; orange, eukaryota.

Table 1. Natural classification of HAD superfamily

I. HADS WITH C0 CAPS
A. *Basal 5-stranded core assemblage*
Three-stranded C0 cap, 5-stranded core, rudimentary C2 cap in the form of simple extended hairpin if present, DD motif IV
MDP-1/FkbH family DxDxTxW motif I and DD motif IV
FkbH/BryA subfamily (several Bacteria)
MDP-1 tyrosine phosphatase subfamily (plants, animals, fungi, kinetoplastids)
FFDDE motif IV and H near motif III
PDB: 1U7O
SSO0580 subfamily (Several archaea) TWN motif II and DDR motif IV
RNA polymerase carboxyl terminal domain (CTD) phosphatase family
Psr1p subfamily (animals, fungi, slime molds, plants, kinetoplastids, *Giardia, apicomplexa*, ciliates)
PDB: 1TA0
Nem1p-dullard subfamily (animals, fungi)
Tim50p subfamily (animals, fungi, plants)
Fcp1p-CPL subfamily (animals, fungi, plants, slime molds, *Cryptosporidium*)
355R subfamily (iridoviruses)
Ublcp1 subfamily (animals, fungi, plants, slime molds)
HSPC129 subfamily (animals, plants, slime molds)
38K/ROP9 phosphatase family (DDxxxN motif IV)
38K subfamily (Baculoviruses). Conserved R in core helix 1, DW in core helix 5;
ROP9 subfamily (Apicomplexa) HSGG motif in C0 cap
BcbF family (proteobacteria, bacteriophages). Unusual dimer-formed *via* strand swapping in core domain
PDB: 1XPJ
Polynucleotide kinase phosphatase (PNKP) family D in core helix 1, Dx3K in core helix 3, SGR motif II
Bacteriophage (PseT) subfamily (bacteriophages)
PDB: 1LTQ
Eukaryotic (PNKP) subfamily (eukaryotes) variable inserts between D residues of motif IV

B. *8KDO (3-Deoxy-D-manno-octulosonate-8-phosphate) phosphatase family (bacteria, Methanobacterium, vertebrates) (GDxxxD motif IV, GGxGAxRE motif at phosphatase C-term), tetramerizes through flap strands*
PDB: 1K1E

C. *Yhr100c family (Firmicutes, Cyanobacteria, Deinococcus-Thermus, Thermotoga, plants, fungi, slime molds) conserved W in core helix 2, R core helix 3*

II. C1 CAP-CONTAINING HAD PROTEINS
Simple Bi-helical Cap Families
A. *Acid Phosphatase family*
Bi-helical cap
non-specific acid phosphatases (NSAPs)
AphA subfamily (*Streptomyces,* enterobacteria)
PDB: 1N9K
P4 subfamily (several bacteria) NPxYGxWE motif at phosphatase C-term;
VSP subfamily (plants, *Streptomyces, Coxiella, Legionella*) GYR preceding motif IV

B. *cN-I nucleotidase family bi-helical cap (vertebrates, proteobacteria, Thermosynechococcus, Arthrobacter)*

Tetra-helical C1 cap assemblage
C. *Motif IV DD assemblage*
Phosphonotase family DxG motif I
Classic phosphonotase subfamily (proteobacteria) DFG motif in squiggle, small helical segment downstream of S5
PDB: 1RQN, 1FEZ
PA2803 subfamily (*Pseudomonas*) degenerate subfamily with loss of cap
Sdt1p-Epoxide Hydrolase C-terminal domain family
sEHCT/Acad10 subfamily (animals, several α-proteobacteria)
PDB: 1S8O, 1EK1
PHM8-SDT1 subfamily (fungi, plants, microsporidians, proteobacteria)
YrfG subfamily (proteobacteria)
YihX subfamily (most bacteria, some fungi, plants)
Deoxyribonucleotidase family (vertebrates, fungi (*Cryptococcus* only), plants, *Giardia*, several bacteria, bacteriophages (caudoviruses), mimivirus) W at the end of strand 6
PDB: 1MH9
HerA-associated family (cyanobacteria, plants) WGY motif at end of strand 5, TxK motif II
β-phosphoglucomutase (BPGM) family KPxP motif III
β-PGM proper subfamily (mainly firmicutes, some actinobacteria, *E.coli*, *Thermotoga*) conserved H, GxxR in cap domain
PDB: 1O08
CbbY subfamily (plants, cyanobacteria, *Chlamydia, Legionella, Yersinia,* several α-proteobacteria) conserved H in cap domain
DOG (2-deoxyglucose-6-phosphate phosphatase) subfamily (fungi, several bacteria, methanogenic euryarchaea) conserved HG in cap domain
YniC subfamily (most bacteria, fungi, animals, plants, *Giardia*)
PDB: 1TE2

D. *dehalogenase-Enolase-phosphatase assemblage*
dehr (dehalogenase-related) family
dehr subfamily I (most bacteria, many archaea, fungi, animals, plants)
Isr subfamily (plants) EWE motif I, SNxxxE motif IV

(continued on next page)

Table 1 (*continued*)

II. C1 CAP-CONTAINING HAD PROTEINS
D. *dehalogenase-Enolase-phosphatase assemblage*
dehr (dehalogenase-related) family
dehr subfamily II (most bacteria, with sporadic transfers to various eukaryotes and archaea)
PDB: 1ZRN, 1QQ5
Enolase-phosphatase family (animals, fungi, γ-proteobacteria, cyanobacteria, *Aquifex*, *Streptomyces*) conserved FVxxxLFPY and DxKxxxLKxLQGxxW regions in cap domain
Bcs3 family (some proteobacteria and *Streptococcus*)
VNG2608C family (cyanobacteria and some euryarchaea)

E. *PSP-P5N-1 assemblage*
P5N-1 (Pyrimidine 5-nucleotidase) family (animals) Additional small C2 cap present
Phosphoserine Phosphatase (PSP) family
SerB subfamily (bacteria, some euryarchaea, fungi (recent bacterial transfers), animals, plants)
PDB: 1F5S, 1NNL
ThrH subfamily (few proteobacteria)
PDB: 1RKU
phosphoserine:homoserine phosphotransferase
PHOSPHO1 subfamily (fungi, animals, plants, bacteria (mainly firmicutes), several Archaea; generates inorganic phosphate for skeletal matrix mineralization; small C2 cap contains 3 conserved cysteine residues that may be involved in metal chelation
CicA subfamily (proteobacteria, actinobacteria, *Parachlamydia, Porphyromonas*)
NapD subfamily (several bacteria, some filamentous ascomycetes, *Methanosarcina*)
AF1437 subfamily (few archaea) C2 cap has three helices stacking above C1 cap
PDB: 1Y8A
Multihelical C1 cap assemblage

G. *cN-II nucleotidase family β-hairpin insertion in core domain after motif III*
cN-II subfamily 1 (animals, slime molds)
cN-II subfamily 2 (animals, plants, slime molds)
cN-II subfamily 3 (animals, plants, slime molds, *Legionella*, *Bdellovibrio*) cystolic 5′-nucleotidases

H. *EYA (Eyes Absent) family (animals)*

I. *Zr25 family (Staphylococcus) Insertion in core domain after motif III*
PDB: 1QYI

P-type ATPase family
Strand 3.1 present between strands S2 and S3, DKTGT motif I, GDGXND motif IV, unique $\alpha+\beta$ cap with conserved K
Type I subfamily (bacteria, archaea, eukaryotes) heavy metal, K^+ transporting pumps
Type II subfamily (bacteria, eukaryotes) Ca^{2+}, Na^+/K^+, H^+/K^+ transporters
PDB: 1SU4
Type III subfamily (eukaryotes, bacteria, archaea) eukaryotic, archaeal proton pumps; bacterial Mg^{2+} transporters
Type IV subfamily (eukaryotes) aminophospholipid transporters
Type V subfamily (eukaryotes) Ca^{2+} transporters
Type VI subfamily (Euryarchaea) soluble phosphatases

III. C2 CAP-CONTAINING HAD PROTEINS
C. *HisB (Histidinol phosphatase) family (bacteria, only Thermoplasmales amongst archaea) metal-chelating C2 cap with conserved CxHxnCxC region; histidine biosynthesis/ADP-D-β-D-heptose synthesis/ADP-D-α-D-heptose synthesis*

A. *NagD family*
GDxxxxD motif IV, distinct α/β C2 cap
AraL subfamily (archaea, firmicutes, actinobacteria) conserved D in cap domain
PDB: 1VJR, 1YV9, 1WVI, 1YDF, 1YS9
chronophin (CIN) subfamily (fungi, animals, slime molds, plants, kinteoplastids) conserved D in cap domain and glycine patch downstream of motif II; putative cofilin-activating phosphatase
Cut-1/CECR5 subfamily (proteobacteria, eukaryotes) conserved D in cap domain
Phosphohistidine/phospholysine phosphatase subfamily (animals, several diverse bacteria)

B. *Cof hydrolase assemblage and constituent families*
β-Sandwich domain cap structure showing considerable diversity, core strands 3.1, 3.2 present
Cof family (archaea, bacteria)
PDB: 1L6R, 1NRW, 1NF2, 1YMQ, 1RLO, 1RKQ, 1WR8
Trehalose phosphate phosphatase (TPP) family
TPP 1 (animals, plants, proteobacteria, few archaea)
TPS2 (plants, fungi, slime molds, microsporidians, very few bacteria and archaea)
Mannosyl-3-phosphoglycerate phosphatase (MPGP) family (some bacteria and Euryarchaea)
PDB: 1XVI
Phosphomannomutase (PMM) family (eukaryotes, *Propionibacterium, Bifidobacterium, Lactococcus, Sphingomonas*)
Sucrose phosphate synthase C-terminal domain (SPSC) family (plants, cyanobacteria, few proteobacteria)
Sucrose phosphate phosphatase (SPP) family (plants, cyanobacteria, firmicutes)
PDB: 1U2T

sequences from the dehr family (gi: 691747, iteration 2, *E*-value: 9e-08), β-phosphoglucomutase family (gi: 1495997, iteration 2, *E*-value: 4e-04), phosphonatase family (gi: 48425373, iteration 2, *E*-value: 0.006), HisB family (gi: 29541277, iteration 3, *E*-value: 0.001), Zr25 family (gi: 39654743, iteration 4, *E*-value: 0.002), and Cof hydrolase assemblage (gi: 28373517, iteration 6, *E*-value: 0.007). Another search with a member of the deoxyribonucleotidase family recovers sequences from the P-type ATPase family (gi: 45359204, iteration 4, *E*-value: 3e-04), Enolase-phosphatase family (gi: 2984225, iteration 6, *E*-value: 0.004), acid phosphatase family (gi: 58176631, iteration 9, *E*-value: 0.001), and NagD family (gi: 10197682, iteration 10, *E*-value 9e-04). Preliminary classification was carried out by means of similarity-based clustering using the BLASTCLUST program (Supplementary Data). Distinct clusters which fell out of this operation were aligned throughout their length and unique signatures beyond the four basic HAD motifs were noted. These extended regions of conservation helped in identifying specific families and objectively distinguishing them from other families with signatures of their own. Within such families the internal relationships, where relevant, were determined using conventional phylogenetic analysis methods, namely neighbor-joining and maximum likelihood, and the phyletic profiles of the members. All major conclusions based on phylogenetic results discussed here were supported by bootstrap support 80% or greater in all the above-stated phylogenetic methods. Higher-order relationships between families were determined by comparing shared structural features, and determining synapomorphies (shared derived characters). Lastly, phyletic patterns, domain architectures, and predicted operon organization of representatives were used to infer likely function if it was not known and also to reconstruct a coherent evolutionary scenario for all branches of the HAD superfamily.

The higher-order relationships within the HAD superfamily are presented graphically in Figure 8 and the resultant natural classification is shown in Table 1 along with phyletic patterns, representatives in the PDB, and functional annotation while Figure 9 depicts domain architectures observed within each family. The most basic split appears to separate a group of C0 cap proteins with a core five-stranded sheet from the rest of the superfamily, which is unified by a six-stranded core sheet. Within this six-stranded assemblage the most basal members retain C0 caps, while the rest of the division is characterized either by dominant C1 or C2 caps. The distinct cap morphologies suggest five major radiations, namely the α-helical C1 cap assemblage, the P-type ATPases with their own C1 cap, and three distinct groups of dominant C2 cap proteins (Figure 8). We describe the details of the classification below, using the cap morphology as a handle.

The C0 cap assemblages and their constituent families

The basal-most clade of the HAD superfamily is comprised of an assemblage of C0 proteins with a five-stranded core sheet and currently includes five distinct families, which are briefly described below. Two additional families showing the C0 cap condition, whose precise evolutionary affinities are not clear, are also discussed in this section (Figure 8; Table 1).

The MDP-1/FkbH family

This family is prototyped by the eukaryotic MDP-1 type Mg(II)-dependent protein tyrosine phosphatases,[34,75,76] which appears to be widely distributed in eukaryotes suggesting a basic cellular function. We also recovered a number of bacterial MDP-1-like proteins typified by FkbH and BryA, and archaeal representatives typified by SSO0580 from *Sulfolobus*. FkbH and BryA are in the biosynthetic pathways for ascomycin and bryostatin in *Streptomyces*[77] and the bacterial symbiont *Candidatus* Endobugula sertula,[78] respectively. The FkbH protein combines an N-terminal HAD domain with a C-terminal acetyltransferase domain (FkbH_Shy in Figure 9) containing a highly conserved cysteine residue. Given its role in synthesis of methoxymalonyl-ACP and gene context, it is quite likely that the incoming substrate is an acyl phosphate, which is cleaved by the HAD domain and the acyl group may then be transferred to the internal cysteine in the acetyl transferase domain and then to the ACP. The presence of one distinct lineage of the MDP-1/FkbH family in each of the three superkingdoms of life is indicative of their possible presence in the last universal common ancestor (LUCA) of cellular life (Figure 8). Related to this ancient family are three other families detailed below with restricted phyletic patterns, and could have been potentially derived from the former family in a lineage-specific fashion (Figure 8; Table 1).

Notes to Table 1:
Clades/families are generally grouped according to dominant cap domain type, i.e. C0, C1, or C2. Indents indicate the inferred hierarchy of evolutionary relationships within each of these major groups of HAD domain containing proteins. Phyletic distribution of families and subfamilies are given in parentheses. Distinct sequence/structural features are listed underneath clade names or next to phyletic distributions of families. The defining characteristic(s) for each family and/or other distinct features are shown underneath clade names or next to phyletic distributions. PDB identifiers of solved crystal structures are indented and listed underneath their respective family/subfamily. Any known enzymatic function not intuitively associated with a family/subfamily name is also listed beneath said family/subfamily. A + refers to a positive conserved residue (lysine or arginine) and a – refers to a negative conserved residue (aspartate, glutamate, or histidine).

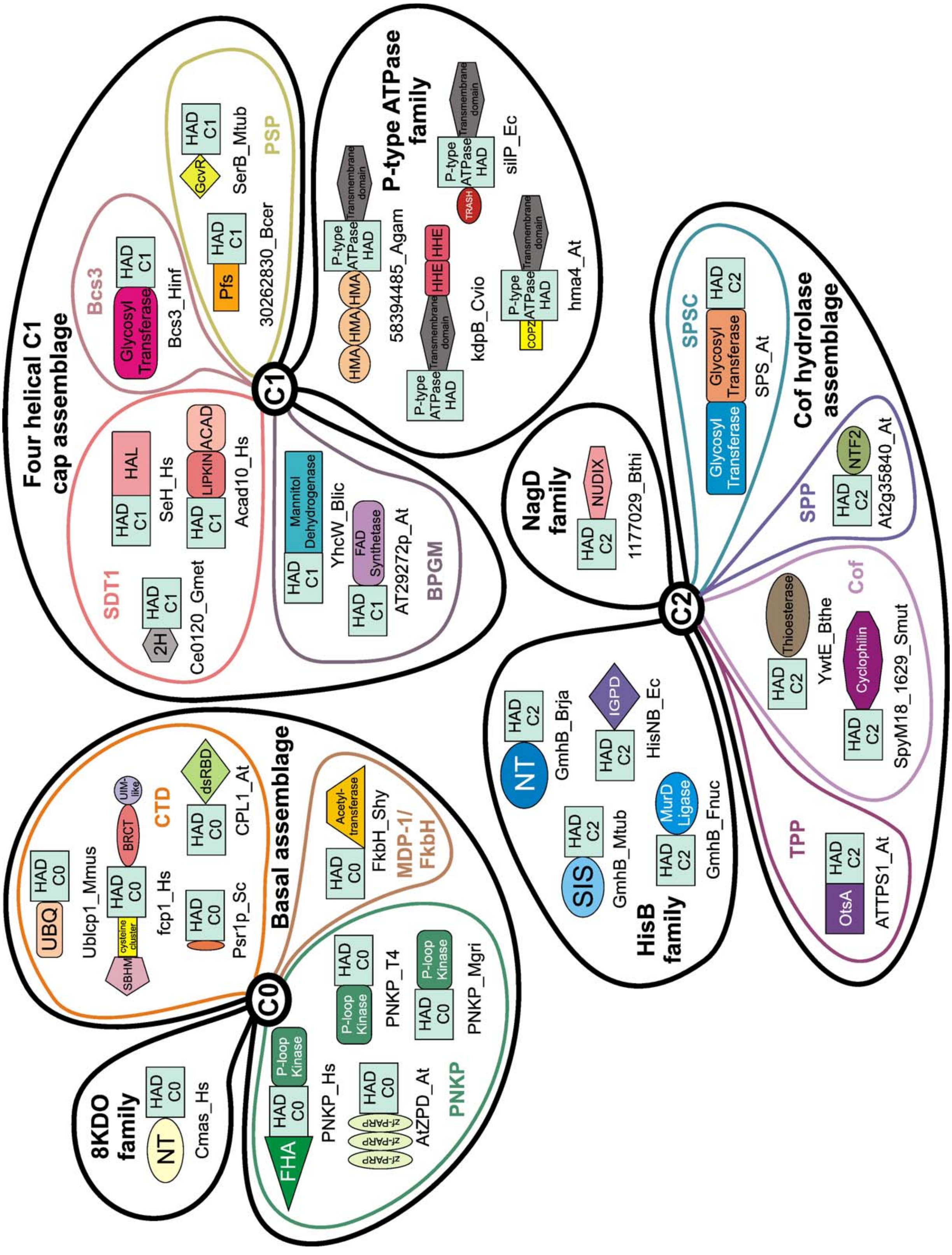


Figure 9 (*legend on opposite page*)

The RNA polymerase carboxyl-terminal domain (CTD) phosphatase family

This family is unique to the eukaryotes and shows an extensive radiation in them (Table 1). The prototypical version of this family, typified by yeast Fcp1p, is required for the dephosphorylation of specific serine residues in the carboxyl-terminal tail of the RNA polymerase catalytic subunit,[79–85] a feature essential for the reinitiation of transcription by the RNA polymerase.[86] This family has diversified into seven subfamilies in the eukaryotes and their viruses (Table 1). The most widespread of these is the Psr1p subfamily which is conserved throughout the eukaryotes and is typified by a conserved N-terminal module required for membrane localization,[87,88] and a conserved cysteine (Psr1p_Sc in Figure 9). Members of this subfamily are slow evolving and are likely to be the principal CTD phosphatases of eukaryotes, and ancient components of the nuclear membrane. The Nem1p/dullard subfamily is seen in animals and fungi, localizes to the nuclear membrane, and might act on nuclear pore complex proteins such as Nup84p.[89] The Tim50 subfamily is also a membrane protein with a peculiar N-terminal membrane-spanning segment.[90] It associates with the mitochondrial inner membrane and regulates the translocation of internal mitochondrial proteins. Recently, the Tim50a isoform has also been shown to localize to the nuclear membrane. Hence, it is likely that this entire group of membrane associated CTD phosphatases diversified as nuclear membrane proteins and Tim50 was subsequently recruited for a mitochondrial function.[91]

The remaining CTD phosphatase subfamilies are soluble proteins and include the Fcp1p-CPL subfamily, typified by the eponymous protein from *S. cerevisiae*.[80] Several versions of this subfamily are characterized by an N-terminal sandwich-barrel hybrid motif (SBHM) domain, followed by a downstream metal-chelating cysteine cluster, and a C-terminal BRCT domain (fcp1_Hs in Figure 9). The BRCT domain in these proteins has been implicated in recognizing the phosphorylated RNA-polymerase II substrate.[92] It is possible the SBHM of the Fcp1p subfamily interacts with the SBHM domains in the catalytic subunits of the RNA polymerase. The plant representatives of this subfamily, the CPL proteins, are implicated in regulating osmotic stress-responsive and abscisic acid-responsive transcription[93,94] and contain one or two double-stranded RNA-binding domains (dsRBD) at the C terminus, suggesting that they might be downstream of the RNA-mediated silencing pathway seen in plants (CPL1_At in Figure 9). Ublcp1 subfamily is also found throughout the crown group of eukaryotes and is typified by an N-terminal ubiquitin domain fused to the phosphatase domain (Ublcp1_Mmus in Figure 9)[95] and might regulate RNA polymerase stability through the ubiquitin pathway.[96]

The 38K/ ROP9 and BcbF families

The remaining two families, which might have arisen from the MDP-1/FkbH family, are much smaller and show even more restricted phyletic patterns. The 38K/ROP9 family shows a very unusual phyletic pattern, with one of the subfamilies being limited to the baculoviruses (the 38K subfamily) and the other to the apicomplexa (ROP9). This family is defined by a characteristic insert that is likely to form a rudimentary C2 cap. The ROP9 family has been experimentally determined to be a secreted protein localizing to the rhoptry, an apicomplexan organelle,[97] suggesting that it might act as a phosphatase in the assembly of the rhoptry or the parasitophorous vacuole. The BcbF family, despite its dramatic structural modifications, is largely limited to the proteobacteria and their viruses, suggesting that it might have arisen relatively recently in evolution. The predicted neighborhoods for these genes suggests that it is often embedded in operons for capsular polysaccharide biosynthesis,[98] suggesting that it might act as a phosphatase on one of the building blocks of the polysaccharide.

Figure 9. Domain architectures of selected multidomain members of the HAD superfamily. The architectures are grouped around a central circle indicating principal cap type (as outlined in Table 1). Domain architectures are further grouped according to the higher-order classification (Table 1), with clades encircled by thick black lines and designated in black lettering and families encircled by lighter colored, thinner lines and designated in the same colored lettering. However, clades and families without any currently known multi-domain architecture are not included in this figure. Each rectangle or other geometric shape represents a single conserved domain. The HAD domain is in light blue and is labeled with the dominant type of cap found in that protein: C0, C1, C2, or P-type ATPase. Proteins are identified with a protein name or abbreviation and an organism name abbreviation. Domain designations: BRCT, breast cancer susceptibility protein carboxy-terminal domain; SBHM, sandwich barrel hybrid motif domain; dsRBD, double-stranded RNA binding motif domain; UBQ, ubiquitin domain; NT, nucleotidyltransferase domain; MurD ligase, glutamate ligase domain; SIS, a sugar isomerase domain; IGPD, imidazole glycerol-phosphate dehydratase domain; zf-PARP, poly(ADP-ribose) polymerase and DNA-ligase Zn-finger domain; FHA, forkhead-associated domain involved in phosphopeptide binding; LIPKIN, antibiotic kinase type small molecule kinase; ACAD, acyl-CoA dehydrogenase; 2H, 2H phosphoesterase domain; Gcvr, repressor of glycine cleavage enzyme system domain; Pfs, nucleoside phosphorylase domain; HHE, possible metal binding domain; TRASH, metallochaperone-like domain; copz, copper chaperone domain; HMA, heavy-metal-associated domain; NTF2, small molecule binding domain of the nuclear transport factor twofold; OtsA, trehalose-6-phosphate synthase domain. The orange ellipse associated with the CTD phosphatase is a specialized membrane-targeting signal with a conserved cysteine. Organism abbreviations are the same as in the alignment, Figure 4. Additional abbreviations: Shy, *Streptomyces hygroscopicus*; Fnuc, *Fusobacterium nucleatum*; Mtub, *Mycobacterium tuberculosis*; Brja, *Bradyrhizobium japonicum*; T4, bacteriophage T4; Bthi, *Bacillus thiaminolyticus*; Gmet, *Geobacter metallireducens*; Bthe, *Bacteroides thetaiotaomicron*; Mmus, *Mus musculus*.

Polynucleotide kinase phosphatase (PNKP) family

The next major lineage of basal C0 cap HADs is the PNKP family which plays a role in both RNA and DNA repair[99] by removing 3′-terminal phosphate groups.[100] There are two subfamilies of these proteins (Table 1) with distinct motif IV sequence signatures (Figure 8), the first being the bacteriophage subfamily (PseT in Figure 8) with an N-terminal P-loop polynucleotide kinase domain (PNKP_T4 in Figure 9). The second is the eukaryotic subfamily (PNKP in Figure 8), which is seen in most major eukaryotic lineages and often contains a C-terminal polynucleotide kinase domain (PNKP_Mgri in Figure 9). In plants, the phosphatase is fused to a Zn-finger found in poly(ADP-ribose) polymerases and DNA-ligases (AtZPD_At in Figure 9).[101,102] Another previously uncharacterized eukaryotic subfamily is found in animals and is fused to the phosphopeptide-binding forkhead-associated domain (FHA) (PNKP_Hs in Figure 9). Both the Zn-finger and FHA domain are likely to be independent means of recruiting these phosphatases to regions of DNA damage.

8KDO family

While this family has a C0 configuration, its core sheet is six-stranded like the rest of the HAD superfamily suggesting that it is closer to the remaining groups of the HAD fold (Figure 8). These enzyme remove a phosphate group from 3-deoxy-D-manno-octulosonate 8-phosphate (Kdo-8) in the course of the biosynthesis of the polysaccharide chain in the bacterial lipid A pathway[103,104] and bacterial capsular polysaccharides.[105] The 8KDO family shows a conserved K residue in the cap domain that points in the direction of the active site and might participate in recognition of negatively charged substrates. Several bacteria and all vertebrate members of this family are fused to a nucleotidyltransferase that potentially catalyzes the subsequent step in the biosynthesis pathways (Cmas_Hs in Figure 9).

Yhr100c family

This family specifically recovers the NagD family of C2 cap proteins (see below), and *vice versa* in sequence searches; however, beyond general core sequence similarity there are no particular features that link these families. Gene neighborhood analysis suggests linkages with genes in the chorismate metabolism pathway, such as AroE (Shikimate 5-dehydrogenase) and chorismate synthase suggesting a possible regulatory role by acting on phosphorylated intermediates in the pathway. The results from the yeast protein–protein interaction map suggest that the eukaryotic members may be part of the Gip1p–Glc7p phosphatase complex required for organization of septins; implying that these proteins possibly function as protein phosphatases during cell division.

The helical C1 cap assemblage

The categories of α-helical caps are discussed in terms of their basic cap morphologies, namely the bihelical, tetrahelical and multi-helical caps. Of these the tetra-helical cap families form the bulk of the assemblage and include several large families (Table 1).

Simple bi-helical cap families

The simplest of the α-helical caps are the bi-helical caps seen in the acid phosphatase and cN-I nucleotidase families. However, there are no other features supporting a specific relationship between these families suggesting that they are basal lineages retaining the ancestral condition of the α-helical clade (Figure 8). The acid phosphatase family is characterized by an N-terminal signal peptide, which suggests that they are secreted proteins that function in periplasmic or extracellular environments. Plants show a lineage-specific expansion of members of this family, which are believed to function as vegetative storage proteins.[106–108] The cN-I family is a family of cytostolic 5′-nucleotidases found in vertebrates and several proteobacteria which regulate pyrimidine pools in the cytosol.[109,110]

Tetra-helical caps: the motif IV DD assemblage

This assemblage is distinguished by the presence of a DD signature in motif IV and contains the phosphonatase, SDT1-epoxide hydrolase C-terminal domain, deoxyribonucleotidase, HerA-associated (HA) and β-phosphoglucomutase (BPGM) families.

The phosphonatase family includes the phosphonoacetaldehyde phosphatases, which hydrolyze phosphonoacetaldehyde to orthophosphate and acetylaldehyde.[111–116] Experimental results suggest a role for cap residues in the catalytic activity of the classic phosphonatases of this family.[27] The family contains a group of degenerate versions from the bacterium *Pseudomonas* (PA2803 subfamily), which have rather partly lost their cap and show disruptions of motifs II and III and IV suggesting that they are catalytically inactive proteins which have take up a secondary binding function. The Sdt1p-epoxide hydrolase C-terminal domain family is widely represented in both bacteria and eukaryotes and appears to have diversified into four major subfamilies (Table 1). Several members of the sEHCT/Acad10 subfamily are fused to a C-terminal α/β hydrolase domain related to the haloalkane dehalogenase domain (HAL) (SeH_Hs in Figure 9). The animal enzyme has been shown to have hydroxyl lipid phosphate phosphatase activity in lipid degradation.[117–119] Some animal members of this subfamily, like Acad10, are fused to two C-terminal domains (Acad10_Hs in Figure 9); a lipid kinase domain related to the protein kinases and an Acyl-CoA dehydrogenase (ACAD) domain, which also suggests a role for them as phospholipid phosphatases. Phm8p of the eponymous subfamily is induced under low phosphate conditions and is likely to release soluble phosphate by hydrolysis of intracellular organo-phosphate compounds[120] while its paralog Sdt1p has been shown to be a pyrimidine 5′-nucleotidase.[121]

The deoxyribonucleotidase family includes one of the major types of 5′ (3′)-deoxyribonucleotidases responsible for dephosphorylating uracil and thymine deoxyribonucleotides.[122–124] The eukaryotic forms do not group together in phylogenetic analysis, suggesting that they might have been acquired from bacterial or phage sources on multiple occasions. The presence in large DNA viruses and mitochondria is consistent with other similarities between their DNA replication processes[125,126] and is indicative of the similar selective pressures faced by these replicons from excess uracil and thymine dNTs. The HerA-associated family is typified by its operonic association with the HerA-type ATPases and the NurA nuclease which are predicted to form a system for chromosome segregation and pumping in prokaryotes.[127] These contextual associations predict that this family might have a role in processing terminal phosphates on DNA, which might emerge due to nuclease action during the pumping process.[127] The β-phosphoglucomutase (BPGM) family is a large group that contains multiple subfamilies with different catalytic activities (Table 1). The archetypal subfamily of this group is the β-PGMs proper, which catalyze the inter-conversion of β-D-glucose 1-phosphate and D-glucose 6-phosphate.[128] This family contains a conserved histidine and GxxR motif in the cap, which are critical for substrate recognition by contacting the phosphate and sugar moieties, respectively.[129] In the related CbbY subfamily (typified by *Rhodobacter* CbbY;[130] Table 1), the histidine is likewise universally conserved, but the arginine is present only in a subset of proteins. The DOG subfamily is typified by the 2-deoxyglucose-6-phosphate phosphatase from fungi[131,132] and other fungal members of this subfamily have been characterized as glycerol 3-phosphatases.[133] The remaining members of this family constitute the large YniC family, which is widely represented throughout the bacteria and the eukaryotes, but not archaea. In plants the HAD domain is fused to the FAD synthetase (AT29272p_At in Figure 9), which adenylates FMN to form FAD.[134,135] The HAD domain might dephosphorylate a precursor in the pathway such as FMN and probably regulates FAD synthesis. Several proteobacterial members are fused to a predicted mannitol dehydrogenase domain (YhcW_Blic in Figure 9), suggesting they might dephosphorylate substrates in sugar metabolism.

Tetra-helical caps: dehalogenase-enolase-phosphatase assemblage

This assemblage contains two major families; the dehalogenase related family (dehr) and the enolase phosphatase family, as well as two other relatively small families; all of which are unified by their sequence similarities in motif IV (Table 1). The dehr family, despite being widespread, remains largely enigmatic with the only well characterized member being the type II D-L-haloalkanoic acid dehalogenase subfamily,[41,136] which is also the archetype of the entire HAD superfamily. The dehr family shows two clear subfamilies (dehr subfamily I and subfamily II). One distinct orthologous group in subfamily I found only in plants, Isr (inhibitor of striate) proteins, is characterized by an unusual EWE signature in motif I and a SNxxxE signature in motif IV. The dehr subfamily II shows even greater diversity in motif IV (e.g SSNxxD, SSxxxD and AAxxxD) with wide differences in the conservation of the acidic residues and is consistent with the acquisition of non-phosphate substrates such as haloalkanoic acids.[137,138] The enolase-phosphatase family of enzymes catalyzes the oxidative dephosphorylation (in combination with the enolase) of 2,3-diketo-1-phosphohexane to 2-keto-pentanoate in the latter steps of the methionine salvage pathway.[139,140] Members of the restricted bacterial Bcs3 family are fused to an N-terminal glycosyltransferase domain (Bcs3_Hinf in Figure 9) and might function as sugar phosphatases in the biosynthesis of capsular polysaccharides in certain pathogenic bacteria[141] (Table 1).

Tetra-helical caps: PSP-P5N-1 assemblage

This assemblage of tetra-helical cap proteins (Table 1) is unified by the presence of an additional insert, which forms a small secondary C2 cap that stacks against the tetra-helical cap. Within this assemblage the P5N-1 family is restricted to animals and catalyzes the dephosphorylation of the pyrimidine 5′ monophosphates UMP and CMP to the corresponding nucleosides.[142,143] The cap region contains highly conserved charged residues likely to be the substrate specificity determinants of this family. Its highly restricted phyletic pattern suggests that the P5N-1 family was possibly derived from the much larger phosphoserine phosphatase family in the animal lineage (Figure 8).

The large phosphoserine phosphatase (PSP) family includes a number of subfamilies, of which the classical phosphoserine phosphatases (SerB) constitute the most widespread subfamily (Table 1). The SerB proteins catalyze the dephosphorylation of L-3-phosphoserine or an exchange reaction between L-serine and L-phosphoserine in the biosynthetic pathway of serine.[144,145] We found a fusion of several prokaryotic SerBs (e.g. *Mycobacterium* and proteobacteria) with GcvR, the repressor of glycine cleavage (GCV) enzyme system (SerB_Mtub in Figure 9). Given the connection between serine catabolism and glycine metabolism,[146–148] this fusion might allow SerB to feedback regulate the glycine cleavage pathway. The related ThrH subfamily, which is restricted to the proteobacteria, participates in the threonine biosynthesis pathway by catalyzing a phosphoserine–homoserine phosphotransfer reaction, similar to the phosphate exchange reaction of SerB.[149,150] The PHOSPHO1 subfamily contains a peculiar C2 cap, which has three conserved cysteine residues, suggesting that it is stabilized by metal chelation. The vertebrate versions of this subfamily are believed to mobilize inorganic phosphate for skeletal matrix mineralization mineralization through their action on phosphocholine and phosphoethanolamine.[151,152] The fusion of this subfamily in some Gram positive bacteria to a nucleoside phosphorylase involved in methionine metabolism[153] might implicate it in this pathway.

The multi-helical cap assemblage

The multi-helical cap assemblage includes three families with strikingly sporadic distributions (Table 1). Among these the cN-II family is another family of cystolic 5′-nucleotidases[154,155] that appear to have convergently evolved this activity, similar to other families in the HAD superfamily (Table 1, see above). This family is unified by a unique β hairpin immediately downstream of motif III, which is unlikely to interact with the cap and might have a distinct function in multimerization or interactions with other proteins. The EYA family (Table 1), defined by the *Drosophila* Eyes Absent protein, functions as a protein tyrosine phosphatase and a transcription factor[156] with EYA itself and RNA polymerase II CTD repeats as its targets.[157,158] This family is characterized by large clusters of conserved charged and polar residues in the cap domain.

The P-type ATPase family

The P-type ATPases contain a cap with a conserved lysine residue at the end of a conserved three-strand stretch in the cap which contributes to the active site of the enzyme and appears to be required for activity.[36] All except one subfamily of these proteins are fused to membrane spanning regions and additional potential metal-ion binding domains (Figure 9). As the P-type ATPase clade has previously been subjected to extensive phylogenetic analysis,[159–161] we only briefly summarize the relationships within this family (Table 1). The type I P-type ATPase subfamily are heavy metal and K^+ transporting pumps and are found in all three superkingdoms of life,[159,160] but their evolutionary history appears to include many lateral transfer events between distantly related organisms. The type II subfamily predominantly consists of Ca^{2+} transporters, but also includes Na^+/ K^+ and H^+/K^+.[159,160] The type III subfamily includes eukaryotic and archaeal proton pumps and bacterial Mg^{2+} transporters.[160] Type IV ATPases are aminophospholipid transporters[160] and type V ATPases were recently characterized as eukaryotic Ca^{2+} transporters.[162] A small subfamily related to the P-type ATPases found only in euryarchaeota and lacking transmembrane regions and the conserved lysine and threonine residues of this family was recently studied experimentally[163] and proposed to be a phosphatase.[164] We propose naming this group of proteins the type VI P-type ATPase subfamily as their structure and sequence features suggest that they are the only surviving form close to the precursor of all other P-type ATPases.

C2 caps: the HisB family

There are several distinct lineages wherein a C2 cap emerged as the principal cap (Figure 6; Table 1) and of these the HisB family shows the simplest version of a C2 cap. These caps contain a CxHxnCxC motif, which is likely to chelate a metal ion that stabilizes the cap. Some of the enzymes in this family are a part of the histidine biosynthesis pathway in prokaryotes (Table 1) and catalyze the hydrolysis of histidinol phosphate[165] (HisNB_Ec in Figure 9). Other bacterial members of the HisB family, the GmhB proteins, catalyze the formation of D-α-D-heptose 1-P from an initial D-alpha-D-heptose 1,7-PP substrate or ADP-D-β-D-heptose 1-P from an initial ADP-D-β-D-heptose 1,7-PP substrate.[166] These members of the HisB family often show operonic association or fusions with sugar metabolism and cell surface glycolipid metabolism enzymes (GmhB_Fnuc, GmhB_Brja, and GmhB_Mtub; Figure 9).

C2 caps: the NagD family

The NagD family is unified by a distinct α/β C2 cap, which is unrelated to all other cap domains seen in the HAD superfamily. While the family is large and widely distributed (Figure 8; Table 1) with several subfamilies, few members have been experimentally characterized. The name of the family is derived from its initial characterization in the *N*-acetylglucosamine (NAG) operon in *E. coli*,[167] although it is not required for the production of NAG.[168] The AraL subfamily (Table 1) has potentially diversified to accommodate a range of substrates. In *Paenibacillus* the HAD domain of this subfamily is fused to a NUDIX domain (1177029_Bthi in Figure 9) which hydrolyzes a variety of substrates with a nucleoside diphosphate linked to another moiety[169,170] implying that its most likely substrate is a nucleotide. A related subfamily is the cronophin phosphatase (CIN) subfamily, which has recently been identified as a cofilin-activating protein phosphatase.[171] The Cut-1 subfamily (after the Cut-1 protein from *Neurospora*) is encoded in a predicted operon in α-proteobacteria with the bifunctional riboflavin kinase/FAD synthetase protein (RibF) and an adenyltransferase that catalyzes the formation of FAD, and might function in cofactor biosynthesis. Except for the phosphohistidine/phospholysine phosphatase[172,173] subfamily (Table 1), all the other members of the NagD family contain a highly conserved aspartate in the C2 Cap (D149 in 1VJR), which points towards the active site and likely acts as a substrate recognition feature.

C2 caps: the Cof phosphatase assemblage and its constituent families

The largest group of C2 cap proteins is the Cof assemblage, which includes several families unified by a C2 cap sharing a common sheet topology (Figure 6). Six distinct families with diverse phyletic patterns can be clearly identified within this assemblage (Table 1) and are briefly summarized below.

The fundamental split in the Cof family is between the archaeal and bacterial subgroups, suggesting that there was probably at least one member of the Cof phosphatase assemblage in the LUCA. A member of the archaeal subgroup, Apc014 from *Thermoplasma acidophilum*, has been shown to exhibit phosphoglycolate phosphatase activity *in vitro*,[65] but there is no evidence that this is its endogenous substrate. An examination of the caps of the Cof family reveals the presence of several conserved residues specific to particular subgroups suggesting that there might be considerable substrate diversity within this family. Members of the trehalose phosphate phosphatase (TPP) family function in conjunction with the trehalose-6-phosphate synthase synthesizing trehalose from glucose-6-phosphate and UDP-glucose[174,175] (Table 1; Figure 9) The broad phyletic pattern suggests that TPP-dependent trehalose biosynthesis or assimilation is one of the most prevalent of the three known catalytic pathways for trehalose biosynthesis.[176,177] The mannosyl-3-phosphoglycerate phosphatase (MPGP) family is a small family comprised of proteins catalyzing the dephosphorylation of mannosyl-3-phosphoglycerate to mannosylglycerate[178] as part of a two-step pathway to synthesize the latter compound from GDP-mannose and D-glycerate. It is found in several hyperthermophilic archaea and some thermophilic bacteria like *Thermus*, where it generates mannosylglycerate, a solute with a protective role against osmotic and thermal stress.[178–180]

The phosphomannomutase (PMM) family catalyzes the isomerization of mannose 6-phosphate and mannose 1-phosphate, which is required in the synthesis of GDP-mannose, a precursor for the dolichol-linked oligosaccharide and GPI anchors, which is unique to eukaryotes[181] (Table 1). The sucrose phosphate synthase C-terminal domain (SPSC) family is comprised of the C-terminal domains of a key enzyme in the sucrose synthesis pathway, which contains an N-terminal two domain glycosyltransferase module (related in structure to glycogen synthase) fused to a C-terminal HAD domain (SPS_At in Figure 9).[182–184] It is likely to regulate the accumulation of sucrose by hydrolyzing the sucrose phosphate formed by the N-terminal domains. The sucrose phosphate phosphatase (SPP) family is closely related to the previous family and catalyzes the dephosphorylation of sucrose phosphate to form sucrose.[185,186] The SPP plant versions additionally have a highly conserved C-terminal domain (At2g35840_At in Figure 9), which we show belongs to the NTF2 class of $\alpha+\beta$ domains.[187] These domains have been previously found in a variety of enzymes, such as the steroid delta-isomerase and scytalone dehydratase, as well as small molecule-binding proteins such as the orange carotenoid protein. This domain been suggested to be involved in increasing catalytic efficiency,[188] and probably binds a small molecule effector to function as an allosteric regulatory site. We note the presence of highly conserved acidic and cysteine residues in this C-terminal domain which might play a role in ligand interactions. The previous two families have been transferred to plants from the cyanobacterial chloroplast precursor.[182]

Evolutionary implications and general considerations

The origin and early evolution of the HAD fold

The higher-order structural relationships of the HAD fold suggest that it first emerged as a part of the radiation phosphoesterase or Mg^{2+} chelating class of Rossmannoid folds. The ancestral version of this division of Rossmanoids folds was characterized by a conserved acidic residue in the first β-α unit of the Rossmannoid fold and another at the end of the strand immediately after the "cross-over" in the sheet (Figure 3). This division of Rossmannoid folds had already expanded to include several distinct representatives in the LUCA of extant cellular life forms, suggesting that the divergence of the HAD fold from related Rossmannoid folds occurred prior to the LUCA. The emergence of the squiggle and flap motifs might have allowed for a rudimentary solvent exclusion mechanism that allowed the HAD superfamily to acquire a catalytic mechanism based on the concomitant formation of an acyl phosphate intermediate. As hardly any HAD enzymes are core components of biological systems such as the RNA metabolism or translation apparatus, they do not show comparable conservation to these proteins. Thus, their phyletic patterns are more drastically affected by gene loss and lateral gene transfer. An examination of the phyletic patterns and phylogenetic relationships of the extant families of the HAD superfamily (Table 1) allows us to potentially extrapolate up to five distinct lineages to the LUCA. The proteins extrapolated to LUCA include (1) the common precursor of the MDP-1/FkbH and CTD phosphatases; (2) a representative of the NagD family; (3) a representative of the Cof clade; (4) a representative of the P-type ATPases; (5) a possible representative of the helical C1 cap assemblage. This suggests that the HAD superfamily had already diversified into the major subtypes, with distinct versions of C0, C1 or C2 caps with duplications and divergence prior to the emergence of the LUCA. We suggest that the ancestral HAD phosphatase, like the ancestral version of the Rossmannoid folds, might have used nucleotides as substrates. Consistent with this, nucleotide substrates are encountered in all the major branches of the HAD superfamily including members of the earliest branching C0 assemblage, specifically the polynucleotide kinase phosphatases (Figure 8; Table 1). Given the role of the PNKP in RNA repair, it is possible that they retain the primitive functional features of the ancestral C0 clade in early biological systems when RNA was the dominant genetic material.

This early branching C0 clade also appears to have specialized in large substrates such as proteins and nucleic acids, which precluded the need for large solvent-excluding caps. The emergence of various caps appears to have provided an additional structurally variable interaction module that allowed different representatives of HAD superfamily to accept a diverse range of substrates, typically small molecules. This process was accompanied by the extensive radiation of the various C1 and C2 cap-containing enzymes and capture of numerous functional niches in the cell. Of these the P-type ATPases represent an early adaptation, wherein the conformational change associated with the catalytic mechanism of the HAD phosphatases was used to drive ion transport. Most of the other members of the superfamily evolved specific catalytic functions in various metabolic pathways. In some cases, such as the Cof assemblage, most enzymes appear to have acquired sugar phosphate substrates early on in their evolution. In other cases, such as the tetra-helical C1 cap assemblage, there is no evidence that any of the early versions had already acquired preferences for a particular category of substrates. Irrespective of the emergence of early substrate preferences, almost none of the HAD enzymes catalyze any of the core reactions in ancient cellular metabolic pathways. Thus, while the prototypes of most major HAD lineages had emerged prior to LUCA, the expansion and diversification of most families occurred well after the separation of the three major superkingdoms of life.

Post-LUCA evolution of HAD superfamily

Phyletic patterns suggest that an explosive radiation of subfamilies occurred in the bacteria and to a smaller extent in the eukaryotes. There are several predominantly bacterial families, but few families that are purely archaeal in their distribution (Figure 8; Table 1). Furthermore, there are at least 26 monophyletic lineages within the HAD superfamily that contain multiple bacteria and eukaryotic representatives, but no or very rare archaeal representatives. The rare archaeal representatives, if any, in these lineages do not preferentially group with the eukaryotic representatives. Given that the eukaryotes have vertically inherited most of their core biological systems from archaeal sources, it is most likely that the lineages of the HAD superfamily shared by eukaryotes and bacteria were acquired laterally by the former. At least four distinct lineages of the HAD superfamily (e.g. the YniC subfamily, the Yhr100c subfamily and the phosphomannomutase family; see Table 1) are present throughout the eukaryotic tree, suggesting that they were acquired early in eukaryotic evolution, most possibly from the mitochondrial precursor. However, about 22 lineages of the HAD superfamily are restricted to only a small section of the eukaryotic superkingdom. Several of these might represent secondary independent transfers from other bacterial sources. In the case of the families shared by the plants and cyanobacteria, such as the SPSC and SPP families and the VSP subfamily of acid phosphatases it is most likely that the plants acquired their versions from chloroplast precursors. More interestingly, we observe that at least four lineages (e.g. 8KDO phosphatase family; see Table 1) are shared by bacteria and animals, but are absent in other eukaryotes. While in principle some of these instances might arise due to losses in earlier eukaryotes, they are likely to represent occurrences of late transfers to the animal line. These are of particular interest because of the potential role of genes of bacterial origin in the emergence of particular metabolic abilities of animals, such as the ability to synthesize or metabolize certain carbohydrates and lipids.

An examination of the bacterial diversification of the HAD superfamily shows that some of the early lineages within bacteria appear to have specialized in particular aspects of amino-acid metabolism, such as the phosphoserine phosphatase and histidinol phosphate phosphatase. Specific roles in amino acid metabolism continued to be acquired in specific lineages of the bacterial tree; for example, the enolase phosphatase and the phosphoserine:homoserine phosphotransferase respectively in methionine and threonine metabolism.[139,140,150] The other major bacterial innovations were related to sugar metabolism and appear to have occurred somewhat later in bacterial evolution. These sugar metabolism enzymes arose throughout the HAD superfamily, though the cof assemblage appears to be the most dominant amongst them. The ancestral ability to use a nucleotide substrate probably served as a pre-adaptation that allowed the emergence of several phosphosugar related activities on multiple occasions. Most of these functions appear to have coincided with the extensive development of storage oligosaccharides and polysaccharide secondary metabolites including components of the cell wall, capsule, and extracellular matrix in bacteria. The other major class of activities colonized by the HAD superfamily in bacteria concerned nucleotide interconversion and salvage, in the form of the various nucleotidases. Interestingly, similar catalytic activities were “invented” within the HAD family on multiple occasions. For example, nucleotidase activity appears to have emerged on at least five different occasions in versions with both C1 and C2 caps (cN-I, Sdt1p, deoxyribonucleotidase, pyrimidine 5-nucleotidase and cN-II). Likewise, phosphosugar mutase activity appears to have arisen on at least two different occasions, once each in lineages with C1 and C2 caps (respectively β-phosphoglucomutase and α-phosphomannomutase). The HAD enzymes with larger caps also appear to have acquired protein phosphatase activity independently on at least three different occasions in evolution, mainly in eukaryotes. Finally, members of the HAD superfamily with the ability to tackle substrates containing non-phosphate ester linkages, such as carbon-phosphorus and carbon-halogen bonds, emerged in bacteria, particularly in the tetrahelical C1 cap assemblage.

These trends suggest that the HAD fold was one of the players in the diversification of the metabolic potential of organisms by providing the raw evolutionary material for the innovation of enzymes that could catalyze new reactions. The five major types of reactions that are known to date to be catalyzed by the superfamily are: (1) phosphatase, (2) ATPase, (3) dehalogenase, (4) phosphosugar mutase and (5) phosphonatase (Figure 1(a)). These reactions show mechanistic similarity[129] and can be accommodated by means of relatively small changes to the active site. Consistent with this, the superfamily is remarkably conservative with respect to the active-site residues, with only small deviations either in the core motifs (e.g. P-type ATPases and the dehalogenases[41,136]) or additions from the cap (phosphonatase). These observations suggest that the intricate active site of the HAD superfamily, with contribution from four distinct core elements and sometimes the cap, taken together with the general asymmetry in the position of the active site in the Rossmannoid fold, precluded them from a extensive evolutionary exploration of "reaction space". However, the location of the active site between a catalytic core and cap allowed the exploration of a vast range of "substrate space". The phyletic patterns of the various lineages of this superfamily suggest that a major component of this evolutionary exploration of substrate space occurred in the Post-LUCA period in the bacteria. Some of these innovations were transmitted *via* lateral gene transfers to the eukaryotes at various points in their evolution, and used as is (e.g. sucrose phosphate phosphatase[185,186]) or recruited for new functions (e.g. the chronophin subfamily[171]). However, there also appear to be a few genuine innovations in the eukaryotes such as PMM and EYA protein phosphatases.[189,156] The apparent lower diversity of these proteins in available archaeal genomes is a potential puzzle. It has also been noticed that another enzyme family forming phospho-aspartyl intermediates, the receiver domains of the two-component systems, are rare in hyperthermophilic archaea.[14] Hence, it is possible that the inherent instability of these aspartyl phosphates in high temperatures might have limited the enzyme's spread in the archaeal superkingdom, particularly in thermophilic and hyperthermophilic members.

More generally the predictions provided here regarding catalytic mechanisms and potential substrate interaction residues can serve as a guide for future biochemical investigations of these enzymes.

Materials and Methods

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) was searched using the BLASTP program.[74] Iterative database searches were performed using the PSI-BLAST program with an alignment or a single sequence serving as the query and a typical expectation value (*E*-value) of 0.01 for inclusion in the position-specific scoring matrix (PSSM); searches were iterated until convergence.[74] For all searches containing computationally biased proteins, the statistical correction option built into the BLAST program was employed. Multiple alignments were constructed using the MUSCLE[190] and/or the T-COFFEE[191] programs, followed by a manual refinement based on PSI-blast results and structural information. All large-scale sequence-analysis procedures were carried out using the TASS package (S.Balaji, V.Anantharaman and L.A., unpublished results). Transmembrane regions were predicted in individual proteins using the default parameters in the TMPRED† and the TMMH2.0[192] programs. Signal peptides in individual proteins were predicted using the SignalP program.[193] Protein structures were visualized and manipulated with the Swiss-PDB viewer[194] and PyMOL‡ programs. Predicted molecular surfaces diagrams and ribbon diagrams were created using the PyMOL program. Protein secondary structures were predicted by feeding multiple alignments into the JPRED2[195] program. The DALI program was used for structural comparisons [63] (see Supplementary Data for details). Similarity-based clustering of proteins was accomplished using BLASTCLUST§.

Gene neighborhoods were obtained by isolating all conserved genes in the neighborhood of the gene under consideration that showed a separation of less than 70 nucleotides between their termini. Genes fulfilling this criterion were considered likely to form operons. Gene neighborhoods were determined by searching the NCBI PTT tables‖ with an in-house PERL script. Phylogenetic analysis was carried out using maximum-likelihood, neighbor-joining, and minimum evolution (least squares) methods (see Supplementary Data for details).

Supplementary information

A collection of the tree files in the Newick format of all the HAD families discussed in the text, along with the corresponding alignments will be made available for download at the ftp-site¶. A table providing a list of all families with potential lateral transfers between bacteria and eukaryotes is also made available at the same site.

Acknowledgements

M.B. and L.A.'s research is supported by the intramural research program of the National Center for Biotechnology Information, NIH. K.N.A. and D.D.-M acknowledge support from the extramural NIH program, grant GM61099.

† http://www.ch.embnet.org/software/TMPRED_form.html

‡ http://www.pymol.org

§ ftp://ftp.ncbi.nih.gov/blast/documents/xml/README.blxml

‖ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome

¶ ftp://ftp.ncbi.nih.gov

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.06.049

References

1. Vincent, J. B., Crowder, M. W. & Averill, B. A. (1992). Hydrolysis of phosphate monoesters: a biological problem with multiple chemical solutions. *Trends Biochem. Sci.* **17**, 105–110.
2. Vetter, I. R. & Wittinghofer, A. (1999). Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer. *Quart. Rev. Biophys.* **32**, 1–56.
3. Iyer, L. M., Leipe, D. D., Koonin, E. V. & Aravind, L. (2004). Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* **146**, 11–31.
4. Bork, P., Sander, C. & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc. Natl Acad. Sci. USA*, **89**, 7290–7294.
5. Haren, L., Ton-Hoang, B. & Chandler, M. (1999). Integrating DNA: transposases and retroviral integrases. *Annu. Rev. Microbiol.* **53**, 245–281.
6. Aravind, L. & Koonin, E. V. (1998). A novel family of predicted phosphoesterases includes *Drosophila* prune protein and bacterial RecJ exonuclease. *Trends Biochem. Sci.* **23**, 17–19.
7. Aravind, L. & Koonin, E. V. (1998). The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* **23**, 469–472.
8. Aravind, L. & Koonin, E. V. (1998). Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucl. Acids Res.* **26**, 3746–3752.
9. Koonin, E. V. & 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.
10. Aravind, L., Galperin, M. Y. & Koonin, E. V. (1998). The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23**, 127–129.
11. Goldberg, J., Huang, H. B., Kwon, Y. G., Greengard, P., Nairn, A. C. & Kuriyan, J. (1995). Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature*, **376**, 745–753.
12. Whisstock, J. C., Romero, S., Gurung, R., Nandurkar, H., Ooms, L. M., Bottomley, S. P. & Mitchell, C. A. (2000). The inositol polyphosphate 5-phosphatases and the apurinic/apyrimidinic base excision repair endonucleases share a common mechanism for catalysis. *J. Biol. Chem.* **275**, 37055–37061.
13. Grebe, T. W. & Stock, J. B. (1999). The histidine protein kinase superfamily. *Adv. Microb. Physiol.* **41**, 139–227.
14. Koretke, K. K., Lupas, A. N., Warren, P. V., Rosenberg, M. & Brown, J. R. (2000). Evolution of two-component signal transduction. *Mol. Biol. Evol.* **17**, 1956–1970.
15. Hogg, T., Mechold, U., Malke, H., Cashel, M. & Hilgenfeld, R. (2004). Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response (corrected). *Cell*, **117**, 57–68.
16. Iyer, L. M. & Aravind, L. (2002). The catalytic domains of thiamine triphosphatase and CyaB-like adenylyl cyclase define a novel superfamily of domains that bind organic phosphates. *BMC Genomics*, **3**, 33.
17. Allen, K. N. & Dunaway-Mariano, D. (2004). Phosphoryl group transfer: evolution of a catalytic scaffold. *Trends Biochem. Sci.* **29**, 495–503.
18. Yamagata, A., Kakuta, Y., Masui, R. & Fukuyama, K. (2002). The crystal structure of exonuclease RecJ bound to Mn^{2+} ion suggests how its characteristic motifs are involved in exonuclease activity. *Proc. Natl Acad. Sci. USA*, **99**, 5908–5912.
19. Ahn, S., Milner, A. J., Futterer, K., Konopka, M., Ilias, M., Young, T. W. & White, S. A. (2001). The "open" and "closed" structures of the type-C inorganic pyrophosphatases from Bacillus subtilis and *Streptococcus gordonii*. *J. Mol. Biol.* **313**, 797–811.
20. Teplyakov, A., Obmolova, G., Khil, P. P., Howard, A. J., Camerini-Otero, R. D. & Gilliland, G. L. (2003). Crystal structure of the *Escherichia coli* YcdX protein reveals a trinuclear zinc active site. *Proteins: Struct. Funct. Genet.* **51**, 315–318.
21. Knofel, T. & Strater, N. (1999). X-ray structure of the *Escherichia coli* periplasmic 5′-nucleotidase containing a dimetal catalytic site. *Nature Struct. Biol.* **6**, 448–453.
22. Mol, C. D., Kuo, C. F., Thayer, M. M., Cunningham, R. P. & Tainer, J. A. (1995). Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature*, **374**, 381–386.
23. Collet, J. F., Stroobant, V., Pirard, M., Delpierre, G. & Van Schaftingen, E. (1998). A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif. *J. Biol. Chem.* **273**, 14107–14112.
24. Anantharaman, V., Aravind, L. & Koonin, E. V. (2003). Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**, 12–20.
25. Aravind, L. & Koonin, E. V. (1999). DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. *Nucl. Acids Res.* **27**, 1609–1618.
26. Shin, D. H., Roberts, A., Jancarik, J., Yokota, H., Kim, R., Wemmer, D. E. & Kim, S. H. (2003). Crystal structure of a phosphatase with a unique substrate binding domain from *Thermotoga maritima*. *Protein Sci.* **12**, 1464–1472.
27. Morais, M. C., Zhang, W., Baker, A. S., Zhang, G., Dunaway-Mariano, D. & Allen, K. N. (2000). The crystal structure of *Bacillus cereus* phosphonoacetaldehyde hydrolase: insight into catalysis of phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily. *Biochemistry*, **39**, 10385–10396.
28. Baker, A. S., Ciocci, M. J., Metcalf, W. W., Kim, J., Babbitt, P. C., Wanner, B. L. *et al.* (1998). Insights into the mechanism of catalysis by the P-C bond-cleaving enzyme phosphonoacetaldehyde hydrolase derived from gene sequence analysis and mutagenesis. *Biochemistry*, **37**, 9305–9315.
29. Qian, N., Stanley, G. A., Hahn-Hagerdal, B. & Radstrom, P. (1994). Purification and characterization of two phosphoglucomutases from *Lactococcus lactis* subsp. lactis and their regulation in maltose- and glucose-utilizing cells. *J. Bacteriol.* **176**, 5304–5311.

30. Collet, J. F., Gerin, I., Rider, M. H., Veiga-da-Cunha, M. & Van Schaftingen, E. (1997). Human L-3-phosphoserine phosphatase: sequence, expression and evidence for a phosphoenzyme intermediate. *FEBS Letters*, **408**, 281–284.
31. Seal, S. N. & Rose, Z. B. (1987). Characterization of a phosphoenzyme intermediate in the reaction of phosphoglycolate phosphatase. *J. Biol. Chem.* **262**, 13496–13500.
32. Lahiri, S. D., Zhang, G., Dunaway-Mariano, D. & Allen, K. N. (2002). Caught in the act: the structure of phosphorylated beta-phosphoglucomutase from *Lactococcus lactis*. *Biochemistry*, **41**, 8351–8359.
33. Ahmadian, M. R., Stege, P., Scheffzek, K. & Wittinghofer, A. (1997). Confirmation of the arginine-finger hypothesis for the GAP-stimulated GTP-hydrolysis reaction of Ras. *Nature Struct. Biol.* **4**, 686–689.
34. Peisach, E., Selengut, J. D., Dunaway-Mariano, D. & Allen, K. N. (2004). X-ray crystal structure of the hypothetical phosphotyrosine phosphatase MDP-1 of the haloacid dehalogenase superfamily. *Biochemistry*, **43**, 12770–12779.
35. Hisano, T., Hata, Y., Fujii, T., Liu, J. Q., Kurihara, T., Esaki, N. & Soda, K. (1996). Crystal structure of L-2-haloacid dehalogenase from *Pseudomonas* sp. YL. An alpha/beta hydrolase structure that is different from the alpha/beta hydrolase fold. *J. Biol. Chem.* **271**, 20322–20330.
36. Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. (2000). Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature*, **405**, 647–655.
37. Wang, W., Kim, R., Jancarik, J., Yokota, H. & Kim, S. H. (2001). Crystal structure of phosphoserine phosphatase from Methanococcus jannaschii, a hyperthermophile, at 1.8 Å resolution. *Structure (Camb)*, **9**, 65–71.
38. Lahiri, S. D., Zhang, G., Dunaway-Mariano, D. & Allen, K. N. (2003). The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction. *Science*, **299**, 2067–2071.
39. Rinaldo-Matthis, A., Rampazzo, C., Reichard, P., Bianchi, V. & Nordlund, P. (2002). Crystal structure of a human mitochondrial deoxyribonucleotidase. *Nature Struct. Biol.* **9**, 779–787.
40. Olsen, D. B., Hepburn, T. W., Moos, M., Mariano, P. S. & Dunaway-Mariano, D. (1988). Investigation of the *Bacillus cereus* phosphonoacetaldehyde hydrolase. Evidence for a Schiff base mechanism and sequence analysis of an active-site peptide containing the catalytic lysine residue. *Biochemistry*, **27**, 2229–2234.
41. Kurihara, T., Liu, J. Q., Nardi-Dei, V., Koshikawa, H., Esaki, N. & Soda, K. (1995). Comprehensive site-directed mutagenesis of L-2-halo acid dehalogenase to probe catalytic amino acid residues. *J. Biochem. (Tokyo)*, **117**, 1317–1322.
42. Aravind, L., Anantharaman, V. & Koonin, E. V. (2002). Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. *Proteins: Struct. Funct. Genet.* **48**, 1–14.
43. Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). Chemical and biological evolution of nucleotide-binding protein. *Nature*, **250**, 194–199.
44. Zhao, K., Chai, X. & Marmorstein, R. (2003). Structure of the yeast Hst2 protein deacetylase in ternary complex with 2′-*O*-acetyl ADP ribose and histone peptide. *Structure*, **11**, 1403–1411.
45. Martin, J. L. & McMillan, F. M. (2002). SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. *Curr. Opin. Struct. Biol.* **12**, 783–793.
46. Schubert, H. L., Blumenthal, R. M. & Cheng, X. (2003). Many paths to methyltransfer: a chronicle of convergence. *Trends Biochem. Sci.* **28**, 329–335.
47. Sistla, S. & Rao, D. N. (2004). *S*-Adenosyl-L-methionine-dependent restriction enzymes. *Crit. Rev. Biochem. Mol. Biol.* **39**, 1–19.
48. Lowe, J. & Amos, L. A. (1998). Crystal structure of the bacterial cell-division protein FtsZ. *Nature*, **391**, 203–206.
49. Anantharaman, V. & Aravind, L. (2006). Diversification of catalytic activities and ligand interactions in the protein fold shared by the sugar isomerases, eIF2B, DeoR transcription factors, acyl-CoA transferases and methenyltetrahydrofolate synthetase. *J. Mol. Biol.* **356**, 823–842.
50. Aravind, L., Leipe, D. D. & Koonin, E. V. (1998). Toprim–a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucl. Acids Res.* **26**, 4205–4213.
51. Clissold, P. M. & Ponting, C. P. (2000). PIN domains in nonsense-mediated mRNA decay and RNAi. *Curr. Biol.* **10**, R888–R890.
52. Finnin, M. S., Donigian, J. R., Cohen, A., Richon, V. M., Rifkind, R. A., Marks, P. A. *et al.* (1999). Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature*, **401**, 188–193.
53. Whittaker, C. A. & Hynes, R. O. (2002). Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere. *Mol. Biol. Cell.* **13**, 3369–3387.
54. Robinson, V. L., Buckler, D. R. & Stock, A. M. (2000). A tale of two components: a novel kinase and a regulatory switch. *Nauret Struct. Biol.* **7**, 626–633.
55. Wolanin, P. M., Thomason, P. A. & Stock, J. B. (2002). Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol.* **3**, REVIEWS3013.1–3013.8.
56. West, A. H. & Stock, A. M. (2001). Histidine kinases and response regulator proteins in two-component signaling systems. *Trends Biochem. Sci.* **26**, 369–376.
57. Ridder, I. S. & Dijkstra, B. W. (1999). Identification of the Mg^{2+}-binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY. *Biochem. J.* **339**, 223–226.
58. Meng, E. C., Polacco, B. J. & Babbitt, P. C. (2004). Superfamily active site templates. *Proteins*, **55**, 962–976.
59. Merckel, M. C., Fabrichniy, I. P., Salminen, A., Kalkkinen, N., Baykov, A. A., Lahti, R. & Goldman, A. (2001). Crystal structure of *Streptococcus mutans* pyrophosphatase: a new fold for an old mechanism. *Structure (Camb)*, **9**, 289–297.
60. Fabrichniy, I. P., Lehtio, L., Salminen, A., Zyryanov, A. B., Baykov, A. A., Lahti, R. & Goldman, A. (2004). Structural studies of metal ions in family II pyrophosphatases: the requirement for a Janus ion. *Biochemistry*, **43**, 14403–14411.
61. Chen, S. J. & Wang, J. C. (1998). Identification of

active site residues in *Escherichia coli* DNA topoisomerase I. *J. Biol. Chem.* **273**, 6050–6056.

62. Lee, J. O., Rieu, P., Arnaout, M. A. & Liddington, R. (1995). Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18). *Cell*, **80**, 631–638.
63. Holm, L. & Sander, C. (1996). The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucl. Acids Res.* **24**, 206–209.
64. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480.
65. Kim, Y., Yakunin, A. F., Kuznetsova, E., Xu, X., Pennycooke, M., Gu, J. *et al.* (2004). Structure- and function-based characterization of a new phosphoglycolate phosphatase from *Thermoplasma acidophilum*. *J. Biol. Chem.* **279**, 517–526.
66. Li, Y. F., Hata, Y., Fujii, T., Hisano, T., Nishihara, M., Kurihara, T. & Esaki, N. (1998). Crystal structures of reaction intermediates of L-2-haloacid dehalogenase and implications for the reaction mechanism. *J. Biol. Chem.* **273**, 15035–15044.
67. Ridder, I. S., Rozeboom, H. J., Kalk, K. H., Janssen, D. B. & Dijkstra, B. W. (1997). Three-dimensional structure of L-2-haloacid dehalogenase from *Xanthobacter autotrophicus* GJ10 complexed with the substrate-analogue formate. *J. Biol. Chem.* **272**, 33015–33022.
68. Calderone, V., Forleo, C., Benvenuti, M., Cristina Thaller, M., Maria Rossolini, G. & Mangani, S. (2004). The first structure of a bacterial class B Acid phosphatase reveals further structural heterogeneity among phosphatases of the haloacid dehalogenase fold. *J. Mol. Biol.* **335**, 761–773.
69. Ridder, I. S., Rozeboom, H. J., Kalk, K. H. & Dijkstra, B. W. (1999). Crystal structures of intermediates in the dehalogenation of haloalkanoates by L-2-haloacid dehalogenase. *J. Biol. Chem.* **274**, 30672–30678.
70. Zhang, G., Mazurkie, A. S., Dunaway-Mariano, D. & Allen, K. N. (2002). Kinetic evidence for a substrate-induced fit in phosphonoacetaldehyde hydrolase catalysis. *Biochemistry*, **41**, 13370–13377.
71. Morais, M. C., Zhang, G., Zhang, W., Olsen, D. B., Dunaway-Mariano, D. & Allen, K. N. (2004). X-ray crystallographic and site-directed mutagenesis analysis of the mechanism of Schiff-base formation in phosphonoacetaldehyde hydrolase catalysis. *J. Biol. Chem.* **279**, 9353–9361.
72. Zhang, G., Dai, J., Wang, L., Dunaway-Mariano, D., Tremblay, L. W. & Allen, K. N. (2005). Catalytic cycling in beta-phosphoglucomutase: a kinetic and structural analysis. *Biochemistry*, **44**, 9404–9416.
73. Wang, W., Cho, H. S., Kim, R., Jancarik, J., Yokota, H., Nguyen, H. H. *et al.* (2002). Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic "snapshots" of intermediate states. *J. Mol. Biol.* **319**, 421–431.
74. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
75. Selengut, J. D. & Levine, R. L. (2000). MDP-1: a novel eukaryotic magnesium-dependent phosphatase. *Biochemistry*, **39**, 8315–8324.
76. Selengut, J. D. (2001). MDP-1 is a new and distinct member of the haloacid dehalogenase family of aspartate-dependent phosphohydrolases. *Biochemistry*, **40**, 12704–12711.
77. Wu, K., Chung, L., Revill, W. P., Katz, L. & Reeves, C. D. (2000). The FK520 gene cluster of *Streptomyces hygroscopicus* var. ascomyceticus (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units. *Gene*, **251**, 81–90.
78. Hildebrand, M., Waggoner, L. E., Liu, H., Sudek, S., Allen, S., Anderson, C. *et al.* (2004). bryA: an unusual modular polyketide synthase gene from the uncultivated bacterial symbiont of the marine bryozoan *Bugula neritina*. *Chem. Biol.* **11**, 1543–1552.
79. Archambault, J., Chambers, R. S., Kobor, M. S., Ho, Y., Cartier, M., Bolotin, D. *et al.* (1997). An essential component of a C-terminal domain phosphatase that interacts with transcription factor IIF in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **94**, 14300–14305.
80. Archambault, J., Pan, G., Dahmus, G. K., Cartier, M., Marshall, N., Zhang, S. *et al.* (1998). FCP1, the RAP74-interacting subunit of a human protein phosphatase that dephosphorylates the carboxyl-terminal domain of RNA polymerase IIO. *J. Biol. Chem.* **273**, 27593–27601.
81. Chambers, R. S. & Dahmus, M. E. (1994). Purification and characterization of a phosphatase from HeLa cells which dephosphorylates the C-terminal domain of RNA polymerase II. *J. Biol. Chem.* **269**, 26243–26248.
82. Chambers, R. S. & Kane, C. M. (1996). Purification and characterization of an RNA polymerase II phosphatase from yeast. *J. Biol. Chem.* **271**, 24498–24504.
83. Cho, H., Kim, T. K., Mancebo, H., Lane, W. S., Flores, O. & Reinberg, D. (1999). A protein phosphatase functions to recycle RNA polymerase II. *Genes Dev.* **13**, 1540–1552.
84. Kobor, M. S., Archambault, J., Lester, W., Holstege, F. C., Gileadi, O., Jansma, D. B. *et al.* (1999). An unusual eukaryotic protein phosphatase required for transcription by RNA polymerase II and CTD dephosphorylation in S. cerevisiae. *Mol. Cell*, **4**, 55–62.
85. Lin, P. S., Marshall, N. F. & Dahmus, M. E. (2002). CTD phosphatase: role in RNA polymerase II cycling and the regulation of transcript elongation. *Prog Nucl. Acid Res. Mol. Biol.* **72**, 333–365.
86. Orphanides, G. & Reinberg, D. (2002). A unified theory of gene expression. *Cell*, **108**, 439–451.
87. Siniossoglou, S., Hurt, E. C. & Pelham, H. R. (2000). Psr1p/Psr2p, two plasma membrane phosphatases with an essential DXDX(T/V) motif required for sodium stress response in yeast. *J. Biol. Chem.* **275**, 19352–19360.
88. Yeo, M., Lin, P. S., Dahmus, M. E. & Gill, G. N. (2003). A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5. *J. Biol. Chem.* **278**, 26078–26085.
89. Siniossoglou, S., Santos-Rosa, H., Rappsilber, J., Mann, M. & Hurt, E. (1998). A novel complex of membrane proteins required for formation of a spherical nucleus. *EMBO J.* **17**, 6449–6464.
90. Guo, Y., Cheong, N., Zhang, Z., De Rose, R., Deng, Y., Farber, S. A. *et al.* (2004). Tim50, a component of the mitochondrial translocator, regulates mitochondrial integrity and cell death. *J. Biol. Chem.* **279**, 24813–24825.
91. Xu, H., Somers, Z. B., Robinson, M. L., 2nd & M.D. (2005). Tim50a, a nuclear isoform of the mitochondrial Tim50, interacts with proteins involved in snRNP biogenesis. *BMC Cell. Biol.* **6**, 29.
92. Yu, X., Chini, C. C., He, M., Mer, G. & Chen, J. (2003).

The BRCT domain is a phospho-protein binding domain. *Science*, **302**, 639–642.

93. Hugouvieux, V., Kwak, J. M. & Schroeder, J. I. (2001). An mRNA cap binding protein, ABH1, modulates early abscisic acid signal transduction in *Arabidopsis*. *Cell*, **106**, 477–487.
94. Xiong, L., Lee, H., Ishitani, M., Tanaka, Y., Stevenson, B., Koiwa, H. *et al.* (2002). Repression of stress-responsive genes by FIERY2, a novel transcriptional regulator in *Arabidopsis*. *Proc. Natl Acad. Sci. USA*, **99**, 10899–10904.
95. Zheng, H., Ji, C., Gu, S., Shi, B., Wang, J., Xie, Y. & Mao, Y. (2005). Cloning and characterization of a novel RNA polymerase II C-terminal domain phosphatase. *Biochem. Biophys. Res. Commun.* **331**, 1401–1407.
96. Wu, X., Chang, A., Sudol, M. & Hanes, S. D. (2001). Genetic interactions between the ESS1 prolyl-isomerase and the RSP5 ubiquitin ligase reveal opposing effects on RNA polymerase II function. *Curr. Genet.* **40**, 234–242.
97. Reichmann, G., Dlugonska, H. & Fischer, H. G. (2002). Characterization of TgROP9 (p36), a novel rhoptry protein of *Toxoplasma gondii* tachyzoites identified by T cell clone. *Mol. Biochem. Parasitol.* **119**, 43–54.
98. Boyce, J. D., Chung, J. Y. & Adler, B. (2000). Genetic organisation of the capsule biosynthetic locus of *Pasteurella multocida* M1404 (B:2). *Vet. Microbiol.* **72**, 121–134.
99. Jilani, A., Ramotar, D., Slack, C., Ong, C., Yang, X. M., Scherer, S. W. & Lasko, D. D. (1999). Molecular cloning of the human gene, PNKP, encoding a polynucleotide kinase 3′-phosphatase and evidence for its role in repair of DNA strand breaks caused by oxidative damage. *J. Biol. Chem.* **274**, 24176–24186.
100. Soltis, D. A. & Uhlenbeck, O. C. (1982). Isolation and characterization of two mutant forms of T4 polynucleotide kinase. *J. Biol. Chem.* **257**, 11332–11339.
101. Petrucco, S., Volpi, G., Bolchi, A., Rivetti, C. & Ottonello, S. (2002). A nick-sensing DNA 3′-repair enzyme from *Arabidopsis*. *J. Biol. Chem.* **277**, 23675–23683.
102. Betti, M., Petrucco, S., Bolchi, A., Dieci, G. & Ottonello, S. (2001). A plant 3′-phosphoesterase involved in the repair of DNA strand breaks generated by oxidative damage. *J .Biol. Chem.* **276**, 18038–18045.
103. Parsons, J. F., Lim, K., Tempczyk, A., Krajewski, W., Eisenstein, E. & Herzberg, O. (2002). From structure to function: YrbI from *Haemophilus influenzae* (HI1679) is a phosphatase. *Proteins: Struct. Funct. Genet.* **46**, 393–404.
104. Wu, J. & Woodard, R. W. (2003). *Escherichia coli* YrbI is 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase. *J. Biol. Chem.* **278**, 18117–18123.
105. Tzeng, Y. L., Datta, A., Strole, C., Kolli, V. S., Birck, M. R., Taylor, W. P. *et al.* (2002). KpsF is the arabinose-5-phosphate isomerase required for 3-deoxy-D-manno-octulosonic acid biosynthesis and for both lipooligosaccharide assembly and capsular polysaccharide expression in *Neisseria meningitidis*. *J. Biol. Chem.* **277**, 24103–24113.
106. Leelapon, O., Sarath, G. & Staswick, P. E. (2004). A single amino acid substitution in soybean VSPalpha increases its acid phosphatase activity nearly 20-fold. *Planta*, **219**, 1071–1079.
107. Utsugi, S., Sakamoto, W., Murata, M. & Motoyoshi, F. (1998). *Arabidopsis thaliana* vegetative storage protein (VSP) genes: gene organization and tissue-specific expression. *Plant Mol. Biol.* **38**, 565–576.
108. Gomez, L. & Faurobert, M. (2002). Contribution of vegetative storage proteins to seasonal nitrogen variations in the young shoots of peach trees (Prunus persica L. Batsch). *J Exp Bot*, **53**, 2431–2439.
109. Hunsucker, S. A., Spychala, J. & Mitchell, B. S. (2001). Human cytosolic 5′-nucleotidase I: characterization and role in nucleoside analog resistance. *J. Biol. Chem.* **276**, 10498–10504.
110. Sala-Newby, G. B., Skladanowski, A. C. & Newby, A. C. (1999). The mechanism of adenosine formation in cells. Cloning of cytosolic 5′-nucleotidase-I. *J. Biol. Chem.* **274**, 17789–17793.
111. La Nauze, J. M. & Rosenberg, H. (1968). The identification of 2-phosphonoacetaldehyde as an intermediate in the degradation of 2-aminoethylphosphonate by *Bacillus cereus*. *Biochim. Biophys. Acta*, **165**, 438–447.
112. Dumora, C., Lacoste, A. M. & Cassaigne, A. (1989). Phosphonoacetaldehyde hydrolase from Pseudomonas aeruginosa: purification properties and comparison with *Bacillus cereus* enzyme. *Biochim. Biophys. Acta*, **997**, 193–198.
113. Lee, K. S., Metcalf, W. W. & Wanner, B. L. (1992). Evidence for two phosphonate degradative pathways in *Enterobacter aerogenes*. *J. Bacteriol.* **174**, 2501–2510.
114. Jiang, W., Metcalf, W. W., Lee, K. S. & Wanner, B. L. (1995). Molecular cloning, mapping, and regulation of Pho regulon genes for phosphonate breakdown by the phosphonatase pathway of Salmonella typhimurium LT2. *J Bacteriol*, **177**, 6411–6421.
115. Ternan, N. G. & Quinn, J. P. (1998). In vitro cleavage of the carbon-phosphorus bond of phosphonopyruvate by cell extracts of an environmental *Burkholderia cepacia isolate*. *Biochem. Biophys. Res. Commun.* **248**, 378–381.
116. Parker, G. F., Higgins, T. P., Hawkes, T. & Robson, R. L. (1999). *Rhizobium* (*Sinorhizobium*) *meliloti* phn genes: characterization and identification of their protein products. *J. Bacteriol.* **181**, 389–395.
117. Imig, J. D., Zhao, X., Capdevila, J. H., Morisseau, C. & Hammock, B. D. (2002). Soluble epoxide hydrolase inhibition lowers arterial blood pressure in angiotensin II hypertension. *Hypertension*, **39**, 690–694.
118. Fang, X., Kaduce, T. L., Weintraub, N. L., Harmon, S., Teesch, L. M., Morisseau, C. *et al.* (2001). Pathways of epoxyeicosatrienoic acid metabolism in endothelial cells. Implications for the vascular effects of soluble epoxide hydrolase inhibition. *J. Biol. Chem.* **276**, 14867–14874.
119. Newman, J. W., Morisseau, C., Harris, T. R. & Hammock, B. D. (2003). The soluble epoxide hydrolase encoded by EPXH2 is a bifunctional enzyme with novel lipid phosphate phosphatase activity. *Proc. Natl Acad. Sci. USA*, **100**, 1558–1563.
120. Ogawa, N., DeRisi, J. & Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell*, **11**, 4309–4321.
121. Nakanishi, T. & Sekimizu, K. (2002). SDT1/SSM1, a multicopy suppressor of S-II null mutant, encodes a novel pyrimidine 5′-nucleotidase. *J. Biol. Chem.* **277**, 22103–22106.
122. Fritzson, P. & Smith, I. (1971). A new nucleotidase of rat liver with activity toward 3′-and 5′-nucleotides. *Biochim. Biophys. Acta*, **235**, 128–141.

123. Rampazzo, C., Johansson, M., Gallinaro, L., Ferraro, P., Hellman, U., Karlsson, A. *et al.* (2000). Mammalian 5′(3′)-deoxyribonucleotidase, cDNA cloning, and overexpression of the enzyme in *Escherichia coli* and mammalian cells. *J. Biol. Chem.* **275**, 5409–5415.
124. Rampazzo, C., Gallinaro, L., Milanesi, E., Frigimelica, E., Reichard, P. & Bianchi, V. (2000). A deoxyribonucleotidase in mitochondria: involvement in regulation of dNTP pools and possible link to genetic disease. *Proc. Natl Acad. Sci. USA*, **97**, 8239–8244.
125. Leipe, D. D., Aravind, L., Grishin, N. V. & Koonin, E. V. (2000). The bacterial replicative helicase DnaB evolved from a RecA duplication. *Genome Res.* **10**, 5–16.
126. Iyer, L. M., Aravind, L. & Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* **75**, 11720–11734.
127. Iyer, L. M., Makarova, K. S., Koonin, E. V. & Aravind, L. (2004). Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucl. Acids Res.* **32**, 5260–5279.
128. Levy, H. R. (1979). Glucose-6-phosphate dehydrogenases. *Adv. Enzymol. Relat. Areas Mol. Biol.* **48**, 97–192.
129. Lahiri, S. D., Zhang, G., Dai, J., Dunaway-Mariano, D. & Allen, K. N. (2004). Analysis of the substrate specificity loop of the HAD superfamily cap domain. *Biochemistry*, **43**, 2812–2820.
130. Gibson, J. L. & Tabita, F. R. (1997). Analysis of the cbbXYZ operon in *Rhodobacter sphaeroides*. *J. Bacteriol.* **179**, 663–669.
131. Sanz, P., Randez-Gil, F. & Prieto, J. A. (1994). Molecular characterization of a gene that confers 2-deoxyglucose resistance in yeast. *Yeast*, **10**, 1195–1202.
132. Randez-Gil, F., Blasco, A., Prieto, J. A. & Sanz, P. (1995). DOGR1 and DOGR2: two genes from *Saccharomyces cerevisiae* that confer 2-deoxyglucose resistance when overexpressed. *Yeast*, **11**, 1233–1240.
133. Norbeck, J., Pahlman, A. K., Akhtar, N., Blomberg, A. & Adler, L. (1996). Purification and characterization of two isoenzymes of DL-glycerol-3-phosphatase from *Saccharomyces cerevisiae*. Identification of the corresponding GPP1 and GPP2 genes and evidence for osmotic regulation of Gpp2p expression by the osmosensing mitogen-activated protein kinase signal transduction pathway. *J. Biol. Chem.* **271**, 13875–13881.
134. Coquard, D., Huecas, M., Ott, M., van Dijl, J. M., van Loon, A. P. & Hohmann, H. P. (1997). Molecular cloning and characterisation of the ribC gene from *Bacillus subtilis*: a point mutation in ribC results in riboflavin overproduction. *Mol. Gen. Genet.* **254**, 81–84.
135. Mack, M., van Loon, A. P. & Hohmann, H. P. (1998). Regulation of riboflavin biosynthesis in *Bacillus subtilis* is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC. *J. Bacteriol.* **180**, 950–955.
136. Hill, K. E., Marchesi, J. R. & Weightman, A. J. (1999). Investigation of two evolutionarily unrelated halocarboxylic acid dehalogenase gene families. *J. Bacteriol.* **181**, 2535–2547.
137. Murdiyatmo, U., Asmara, W., Tsang, J. S., Baines, A. J., Bull, A. T. & Hardman, D. J. (1992). Molecular biology of the 2-haloacid halidohydrolase IVa from *Pseudomonas cepacia* MBA4. *Biochem. J.* **284**, 87–93.
138. Tsang, J. S. & Pang, B. C. (2000). Identification of the dimerization domain of dehalogenase IVa of *Burkholderia cepacia* MBA4. *Appl. Environ. Microbiol.* **66**, 3180–3186.
139. Myers, R. W., Wray, J. W., Fish, S. & Abeles, R. H. (1993). Purification and characterization of an enzyme involved in oxidative carbon-carbon bond cleavage reactions in the methionine salvage pathway of *Klebsiella pneumoniae*. *J. Biol. Chem.* **268**, 24785–24791.
140. Balakrishnan, R., Frohlich, M., Rahaim, P. T., Backman, K. & Yocum, R. R. (1993). Appendix. Cloning and sequence of the gene encoding enzyme E-1 from the methionine salvage pathway of *Klebsiella oxytoca*. *J. Biol. Chem.* **268**, 24792–24795.
141. Satola, S. W., Schirmer, P. L. & Farley, M. M. (2003). Complete sequence of the cap locus of *Haemophilus influenzae* serotype b and non-encapsulated b capsule-negative variants. *Infect. Immun.* **71**, 3639–3644.
142. Valentine, W. N., Fink, K., Paglia, D. E., Harris, S. R. & Adams, W. S. (1974). Hereditary hemolytic anemia with human erythrocyte pyrimidine 5′-nucleotidase deficiency. *J. Clin. Invest.* **54**, 866–879.
143. Paglia, D. E. & Valentine, W. N. (1975). Characteristics of a pyrimidine-specific 5(-nucleotidase in human erythrocytes. *J. Biol. Chem.* **250**, 7973–7979.
144. Borkenhagen, L. F. & Kennedy, E. P. (1958). The enzymic equilibration of L-serine with *O*-phospho-L-serine. *Biochim. Biophys. Acta*, **28**, 222–223.
145. Neuhaus, F. C. & Byrne, W. L. (1958). O-Phosphoserine phosphatase. *Biochim. Biophys. Acta*, **28**, 223–224.
146. Schirch, L. & Gross, T. (1968). Serine transhydroxymethylase. Identification as the threonine and allothreonine aldolases. *J. Biol. Chem.* **243**, 565–5651.
147. Ulevitch, R. J. & Kallen, R. G. (1977). Purification and characterization of pyridoxal 5′-phosphate dependent serine hydroxymethylase from lamb liver and its action upon beta-phenylserines. *Biochemistry*, **16**, 5342–5350.
148. Szebenyi, D. M., Musayev, F. N., di Salvo, M. L., Safo, M. K. & Schirch, V. (2004). Serine hydroxymethyltransferase: role of glu75 and evidence that serine is cleaved by a retroaldol mechanism. *Biochemistry*, **43**, 6865–6876.
149. Patte, J. C., Clepet, C., Bally, M., Borne, F., Mejean, V. & Foglino, M. (1999). ThrH, a homoserine kinase isozyme with *in vivo* phosphoserine phosphatase activity in *Pseudomonas aeruginosa*. *Microbiology*, **145**, 845–853.
150. Singh, S. K., Yang, K., Karthikeyan, S., Huynh, T., Zhang, X., Phillips, M. A. & Zhang, H. (2004). The thrH gene product of *Pseudomonas aeruginosa* is a dual activity enzyme with a novel phosphoserine: homoserine phosphotransferase activity. *J. Biol. Chem.* **279**, 13166–13173.
151. Houston, B., Seawright, E., Jefferies, D., Hoogland, E., Lester, D., Whitehead, C. & Farquharson, C. (1999). Identification and cloning of a novel phosphatase expressed at high levels in differentiating growth plate chondrocytes. *Biochim. Biophys. Acta*, **1448**, 500–506.
152. Houston, B., Paton, I. R., Burt, D. W. & Farquharson, C. (2002). Chromosomal localization of the chicken and mammalian orthologues of the orphan phosphatase PHOSPHO1 gene. *Anim. Genet.* **33**, 451–454.
153. Beeston, A. L. & Surette, M. G. (2002). pfs-dependent regulation of autoinducer 2 production in Salmonella enterica serovar *Typhimurium*. *J. Bacteriol.* **184**, 3450–3456.
154. Allegrini, S., Scaloni, A., Ferrara, L., Pesi, R., Pinna,

P., Sgarrella, F. *et al.* (2001). Bovine cytosolic 5′-nucleotidase acts through the formation of an aspartate 52-phosphoenzyme intermediate. *J. Biol. Chem.* **276**, 33526–33532.

155. Oka, J., Matsumoto, A., Hosokawa, Y. & Inoue, S. (1994). Molecular cloning of human cytosolic purine 5′-nucleotidase. *Biochem. Biophys. Res. Commun.* **205**, 917–922.
156. Rebay, I., Silver, S. J. & Tootle, T. L. (2005). New vision from Eyes absent: transcription factors as enzymes. *Trends Genet.* **21**, 163–171.
157. Li, X., Oghi, K. A., Zhang, J., Krones, A., Bush, K. T., Glass, C. K. *et al.* (2003). Eya protein phosphatase activity regulates Six1-Dach-Eya transcriptional effects in mammalian organogenesis. *Nature*, **426**, 247–254.
158. Tootle, T. L., Silver, S. J., Davies, E. L., Newman, V., Latek, R. R., Mills, I. A. *et al.* (2003). The transcription factor Eyes absent is a protein tyrosine phosphatase. *Nature*, **426**, 299–302.
159. Moller, J. V., Juul, B. & le Maire, M. (1996). Structural organization, ion transport, and energy transduction of P-type ATPases. *Biochim. Biophys. Acta*, **1286**, 1–51.
160. Axelsen, K. B. & Palmgren, M. G. (1998). Evolution of substrate specificities in the P-type ATPase superfamily. *J. Mol. Evol.* **46**, 84–101.
161. Fagan, M. J. & Saier, M. H., Jr (1994). P-type ATPases of eukaryotes and bacteria: sequence analyses and construction of phylogenetic trees. *J. Mol. Evol.* **38**, 57–99.
162. Cronin, S. R., Rao, R. & Hampton, R. Y. (2002). Cod1p/Spf1p is a P-type ATPase involved in ER function and Ca^{2+} homeostasis. *J. Cell Biol.* **157**, 1017–1028.
163. Ogawa, H., Haga, T. & Toyoshima, C. (2000). Soluble P-type ATPase from an archaeon, *Methanococcus jannaschii*. *FEBS Letters*, **471**, 99–102.
164. Bramkamp, M., Gassel, M., Herkenhoff-Hesselmann, B., Bertrand, J. & Altendorf, K. (2003). The *Methanocaldococcus jannaschii* protein Mj0968 is not a P-type ATPase. *FEBS Letters*, **543**, 31–36.
165. le Coq, D., Fillinger, S. & Aymerich, S. (1999). Histidinol phosphate phosphatase, catalyzing the penultimate step of the histidine biosynthesis pathway, is encoded by ytvP (hisJ) in *Bacillus subtilis*. *J. Bacteriol.* **181**, 3277–3280.
166. Kneidinger, B., Marolda, C., Graninger, M., Zamyatina, A., McArthur, F., Kosma, P. *et al.* (2002). Biosynthesis pathway of ADP-L-glycero-beta-D-manno-heptose in *Escherichia coli*. *J. Bacteriol.* **184**, 363–369.
167. Plumbridge, J. A. (1989). Sequence of the nagBACD operon in *Escherichia coli* K12 and pattern of transcription within the nag regulon. *Mol. Microbiol.* **3**, 505–515.
168. Peri, K. G., Goldie, H. & Waygood, E. B. (1990). Cloning and characterization of the N-acetylglucosamine operon of *Escherichia coli*. *Biochem. Cell Biol.* **68**, 123–137.
169. Perraud, A. L., Fleig, A., Dunn, C. A., Bagley, L. A., Launay, P., Schmitz, C. *et al.* (2001). ADP-ribose gating of the calcium-permeable LTRPC2 channel revealed by Nudix motif homology. *Nature*, **411**, 595–599.
170. Bessman, M. J., Frick, D. N. & O'Handley, S. F. (1996). The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J. Biol. Chem.* **271**, 25059–25062.
171. Gohla, A., Birkenfeld, J. & Bokoch, G. M. (2005). Chronophin, a novel HAD-type serine protein phosphatase, regulates cofilin-dependent actin dynamics. *Nature Cell Biol.* **7**, 21–29.
172. Hiraishi, H., Ohmagari, T., Otsuka, Y., Yokoi, F. & Kumon, A. (1997). Purification and characterization of hepatic inorganic pyrophosphatase hydrolyzing imidodiphosphate. *Arch. Biochem. Biophys.* **341**, 153–159.
173. Yokoi, F., Hiraishi, H. & Izuhara, K. (2003). Molecular cloning of a cDNA for the human phospholysine phosphohistidine inorganic pyrophosphate phosphatase. *J. Biochem. (Tokyo)*, **133**, 607–614.
174. Vandercammen, A., Francois, J. & Hers, H. G. (1989). Characterization of trehalose-6-phosphate synthase and trehalose-6-phosphate phosphatase of *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **182**, 613–620.
175. Kaasen, I., Falkenberg, P., Styrvold, O. B. & Strom, A. R. (1992). Molecular cloning and physical mapping of the otsBA genes, which encode the osmoregulatory trehalose pathway of *Escherichia coli*: evidence that transcription is activated by katF (AppR). *J. Bacteriol.* **174**, 889–898.
176. De Smet, K. A., Weston, A., Brown, I. N., Young, D. B. & Robertson, B. D. (2000). Three pathways for trehalose biosynthesis in mycobacteria. *Microbiology*, **146**, 199–208.
177. Wolf, A., Kramer, R. & Morbach, S. (2003). Three pathways for trehalose metabolism in *Corynebacterium glutamicum* ATCC13032 and their significance in response to osmotic stress. *Mol. Microbiol.* **49**, 1119–11134.
178. Empadinhas, N., Marugg, J. D., Borges, N., Santos, H. & da Costa, M. S. (2001). Pathway for the synthesis of mannosylglycerate in the hyperthermophilic archaeon *Pyrococcus horikoshii*. Biochemical and genetic characterization of key enzymes. *J. Biol. Chem.* **276**, 43580–43588.
179. Borges, N., Marugg, J. D., Empadinhas, N., da Costa, M. S. & Santos, H. (2004). Specialized roles of the two pathways for the synthesis of mannosylglycerate in osmoadaptation and thermoadaptation of *Rhodothermus marinus*. *J. Biol. Chem.* **279**, 9892–9898.
180. Empadinhas, N., Albuquerque, L., Henne, A., Santos, H. & da Costa, M. S. (2003). The bacterium *Thermus thermophilus*, like hyperthermophilic archaea, uses a two-step pathway for the synthesis of mannosylglycerate. *Appl. Environ. Microbiol.* **69**, 3272–3279.
181. Tomavo, S., Dubremetz, J. F. & Schwarz, R. T. (1992). Biosynthesis of glycolipid precursors for glycosylphosphatidylinositol membrane anchors in a *Toxoplasma gondii* cell-free system. *J. Biol. Chem.* **267**, 21446–21458.
182. Lunn, J. E. (2002). Evolution of sucrose synthesis. *Plant Physiol.* **128**, 1490–1500.
183. Langenkamper, G., Fung, R. W., Newcomb, R. D., Atkinson, R. G., Gardner, R. C. & MacRae, E. A. (2002). Sucrose phosphate synthase genes in plants belong to three different families. *J. Mol. Evol.* **54**, 322–332.
184. Castleden, C. K., Aoki, N., Gillespie, V. J., MacRae, E. A., Quick, W. P., Buchner, P. *et al.* (2004). Evolution and function of the sucrose-phosphate synthase gene families in wheat and other grasses. *Plant Physiol.* **135**, 1753–1764.
185. Hawker, J. S. & Hatch, M. D. (1966). A specific sucrose phosphatase from plant tissues. *Biochem. J.* **99**, 102–107.
186. Lunn, J. E. & ap Rees, T. (1990). Apparent equilibrium constant and mass-action ratio for sucrose-phosphate

synthase in seeds of *Pisum sativum*. *Biochem. J.* **267**, 739–743.
187. Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394.
188. Lunn, J. E. (2003). Sucrose-phosphatase gene families in plants. *Gene*, **303**, 187–196.
189. Bonini, N. M., Leiserson, W. M. & Benzer, S. (1993). The eyes absent gene: genetic control of cell survival and differentiation in the developing *Drosophila* eye. *Cell*, **72**, 379–395.
190. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797.
191. Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
192. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.
193. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795.
194. Guex, N. & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
195. Cuff, J. A. & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **40**, 502–511.
196. Goodstadt, L. & Ponting, C. P. (2001). CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics*, **17**, 845–846.

Edited by B. Honig

(Received 1 December 2005; received in revised form 16 June 2006; accepted 20 June 2006)
Available online 7 July 2006