

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

**APPLICATION OF COMPUTATIONAL APPROACHES TO DECIPHER
FUNCTION AND EVOLUTION OF KEY PROTEIN FOLDS**

by

ALEXANDER MAXWELL BURROUGHS

B.A., Brigham Young University-Hawaii, 2003

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2008

Approved by

First Reader

Karen N. Allen, Ph.D.
Assistant Professor of Physiology, School of Medicine
Boston University

Second Reader

L. Aravind, Ph.D.
Senior Investigator
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

ACKNOWLEDGEMENTS

Research contained in this dissertation was supported by the intra-mural research program at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health. Portions of this research were also supported from the extramural NIH program grant GM1099. Data for one section of the dissertation were collected at beamline X12C of the National Synchrotron Light Source.

**APPLICATION OF COMPUTATIONAL APPROACHES TO DECIPHER
FUNCTION AND EVOLUTION OF KEY PROTEIN FOLDS**

(Order No.)

ALEXANDER MAXWELL BURROUGHS

Boston University Graduate School of Arts and Sciences, 2008

Major Professor: L. Aravind, Senior Investigator
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

ABSTRACT

The class of investigative methods collectively termed evolutionary or comparative genomics embraces the principle that features of a protein or nucleic acid are significantly influenced by ancestry and are the product of extensive natural selection. The application of different computational techniques exploiting this principle yields insight into evolution and function, and also provides a sound theoretical framework for further experimental investigation. This dissertation uncovers such insights by applying these techniques to selected protein folds, namely the PUA, P-loop NTPase, Rossman, and β -grasp folds.

Highlights of these investigations include the discovery of a conserved protein domain in the ASC-1 transcriptional coactivator complex with an ancestral linkage to the PUA fold and prediction of an RNA-binding role for the domain. Multiple evolutionary origins for viral DNA packaging ATPases of the P-loop NTPase fold were discovered, and novel components of viral DNA packaging systems were detected. Contrary to previous assumptions, portal proteins of diverse tailed phages, the primary interacting partners of packaging ATPases, were determined

to be of monophyletic origin. A third higher-order assemblage of domains belonging to the Rossmann-like fold was established, unified by conserved sequence features which contribute to metal-chelation. Determinants of catalytic activity in the Haloacid Dehalogenase (HAD) and E1-like protein superfamilies of the Rossmann fold were elucidated; these describe how structural and domain architecture variation have influenced exploration of substrate space by these superfamilies. These findings also led to the identification and crystallization of the first member of the HAD superfamily lacking catalytic activity, and functional predictions for these inactive HAD domains. The determinants of functional diversity in the β -grasp fold were also identified resulting in discovery of a novel superfamily of ligand-binding β -grasp domains, the first dedicated soluble-ligand binding activity identified in the fold. Functional predictions for domains in this superfamily were made for diverse systems such as polysaccharide export, DNA uptake, and intracellular redox reactions. Investigations into the β -grasp fold also led to the elucidation of the evolutionary origins of the eukaryotic ubiquitin signaling system and the prediction of several distinct, entire ubiquitin-like modification systems in prokaryotes featuring conjugation and deconjugation of ubiquitin-like proteins.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT.....	iv
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS.....	xiv
INTRODUCTION.....	1
Organization of the Dissertation	6
Overview of Methods Employed.....	7
TASS software package	7
Computing resources.....	9
Methods of Sequence Analysis.....	10
Methods of Structural Analysis.....	19
Phylogenetic tree-building.....	21
Contextual Inference Analysis	23
Differences and similarities in the application of above methods to other laboratories	24
INVESTIGATION RELATING TO THE PUA FOLD	27
The ASCH Superfamily: Novel Domains with a Fold Related to the PUA Domain and a Potential Role in RNA Metabolism.....	27
Introduction	27
Application of Methods.....	29
Results and Discussion.....	30
Evolutionary diversity of ASCH domains and general conclusions	37

INVESTIGATION RELATING TO THE P-LOOP NTPASE FOLD	41
Comparative Genomics And Evolutionary Trajectories Of Viral ATP Dependent DNA- Packaging Systems	42
Introduction	42
Application of Methods.....	44
Results and Discussion.....	45
Evolutionary considerations and general conclusions.....	59
Experimental validation of work presented above	61
INVESTIGATIONS RELATING TO THE ROSSMANN FOLD	62
Evolutionary Genomics Of The HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes.....	63
Introduction	63
Application of Methods.....	65
Results and Discussion.....	67
Evolutionary implications and general considerations	103
The X-Ray Crystallographic Structure and Activity Analysis of a <i>Pseudomonas</i> Specific Subfamily of the HAD Enzyme Superfamily Evidences a Novel Biochemical Function.....	110
Introduction	110
Materials and Methods.....	112
Results and Discussion.....	116
Structure Determination of Recombinant <i>P. syringae</i> PSPTO_2114.....	118
Conservation of Binding Surfaces.....	125
Functional Assessment Derived from Gene Context	127

Evolutionary History of the E1-Like Fold and Architectural Themes Contributing to the Catalytic Roles of E1-Like Domains.....	130
Introduction	130
Application of Methods.....	132
Identification and Classification of E1-like Protein Families	133
The Origins of the E1 Superfamily.....	136
Sequence and structural diversity within the E1 superfamily	141
Diverse Domain Architectures in the E1 Superfamily	148
Evolutionary Themes in the E1 Fold	152
Conclusions and General Observations.....	156
INVESTIGATIONS RELATING TO THE β -GRASP FOLD	159
Small but Versatile: the Extraordinary Functional and Structural Diversity of the β -Grasp Fold.....	159
Introduction	160
Application of Methods.....	164
Results and Discussion.....	168
The relative timeline of major adaptive radiations and functional transitions of the β -GF domains	182
Evolutionary trends in the domain architectures of β -GF domains.....	189
Structural correlates for functional diversity in the β -GF.....	193
General conclusions.....	202
A Novel Superfamily Containing the β -Grasp Fold Involved in Binding Diverse Soluble Ligands	206

Introduction	206
Application of Methods.....	207
Results and Discussion.....	208
Evolutionary History of the SLBB Domain and General Conclusions	219
Experimental validation of work presented above	221
The Prokaryotic Antecedents of the Ubiquitin Signaling System and the Early Evolution of Ubiquitin-like β -Grasp Domains.....	223
Introduction	223
Application of Methods.....	227
Results and Discussion.....	229
Functional implications of the prokaryotic systems with components related to the eukaryotic Ub-signaling network	251
Evolutionary implications of prokaryotic cognates of the Ub-signaling system	255
General conclusions	259
Experimental validation of work presented above	260
CONCLUSIONS AND GENERAL OBSERVATIONS.....	261
BIBLIOGRAPHY	267
CURRICULUM VITAE.....	318

LIST OF TABLES

Table 1. Natural classification of HAD superfamily.....	88
Table 2. Summary of Data Collection and Refinement Statistics.....	117
Table 3. Secondary structure features of major E1 fold structural categories.....	139
Table 4. Secondary structure features of major β -GF structural categories.....	167
Table 5. Phyletic distributions and conserved gene neighborhoods of prokaryotic Ub-like families.....	233

LIST OF FIGURES

Fig 1. Structures and domain architectures from the ASCH and PUA superfamilies	32
Fig 2. Multiple alignment of members of the ASCH superfamily	35
Fig 3. Molecular surfaces of observed binding cleft in ASCH superfamily	36
Fig. 4. Packaging ATPase presence/absence and position distribution in viral genomes.	46
Fig. 5. Topology diagrams depicting ASCE division of P-loop NTPases and accompanying cladogram depicting higher-order relationships.	51
Fig. 6. Phylogenetic tree of TLS depicting gene displacement among portal protein families.....	56
Fig 7. HAD reaction mechanisms.....	64
Fig 8. HAD catalytic domain.....	67
Fig 9. Rossmannoid Domains	68
Fig 10. Multiple sequence alignment of HAD-domain containing proteins	72
Fig 11. Topology diagrams of selected C0 and C1 cap HAD domains	77
Fig 12. Topology diagrams of selected HAD C2 cap domains.....	79
Fig 13. Interaction of cap modules with the active site in the HAD superfamily.....	84
Fig 14. Preliminary reconstructed evolutionary scenario for the HAD superfamily.....	87
Fig. 15. Domain architectures of selected multidomain members of the HAD superfamily.....	90
Fig. 16. Residues forming phosphonate catalytic site.....	111
Fig. 17. Sequence alignment of members of the PA2803 subfamily.	119
Fig. 18. Structures of PSPTO_2114 and phosphonate.	120
Fig. 19: Comparison of phosphonate and PSPTO_2114.....	123
Fig. 20. Surface representation of PSPTO_2114.....	126

Fig. 21: Gene context of PSPTO_2114 in <i>Pseudomonads</i>	128
Fig. 22. Reconstructed evolutionary scenario and major architecture and gene neighborhood domain associations of the E1 domain.	135
Fig. 23. Colored topology diagrams of representatives from the major divisions of Rossmannoid folds.	140
Fig. 24. Phylogenetic tree displaying interrelationships between lineages in bacterial ThiS/MoeB E1 domains and the conserved gene neighborhoods characteristic of lineages.	154
Fig. 25. Topology diagrams of selected β -GF members.	162
Fig. 26. Cartoon representations of distinct β -GF domains.	174
Fig. 27. Reconstructed evolutionary history of β -grasp fold.	175
Fig. 28. Reconstructed evolutionary history of eukaryotic ubiquitin superfamily.....	187
Fig. 29. Architectural complexity plot and novel domain architectures for β -grasp domains.	191
Fig. 30. Diagram of relative location of β -grasp interacting partners.	194
Fig. 31. Topology diagram of SLBB domain and multiple alignment of SLBB superfamily.....	210
Fig. 32. Domain architectures and conserved gene neighborhoods of the SLBB superfamily.	215
Fig. 33. ThiS/MoaD/Ubiquitin-based protein conjugation system.....	224
Fig 34. Multiple alignment of ThiS/MoaD-like ubiquitin domain containing proteins.....	233
Fig 35. Domain Architectures of the ThiS/MoaD-like ubiquitin domains and functionally associated proteins.	237
Fig 36. Gene neighborhoods of the prokaryotic ThiS/MoaD-like ubiquitin domains and functionally associated proteins.	240
Fig. 37. Multiple alignment of JAB domain-containing proteins.....	254

Fig. 38. Multiple alignment of E2 (UBC)-like proteins with a special emphasis on bacterial versions.....256

Fig 39. Network diagram of ThiS/MoaD-like beta-grasp domains.258

LIST OF ABBREVIATIONS

β -GF	β -grasp fold
BLAST	Basic Local Alignment Search Tool
DUBs	deubiquitinating peptidases.
HAD	Haloacid Dehalogenase
HMM	Hidden Markov Model
LECA	Last eukaryotic common ancestor
LUCA	Last Universal Common Ancestor
ML	Maximum-likelihood
MoCo/Wco	molybdenum/tungsten cofactor
NCBI	National Center for Biotechnology Information
NR	non-redundant database
PSSM	position-specific scoring matrix
SBHM	Sandwich β -hybrid Motif
SEALS	System of Easy Analysis of Lots of Sequences
SLBB	Soluble Ligand-binding β -grasp
TASS	Tools for the Analysis of Sequence and Structure
Ub	ubiquitin
Ubl	ubiquitin-like

INTRODUCTION

Traditionally, decoding interactions between biomolecules in the cell has relied on the techniques of biochemistry and molecular biology. While these investigations have been central to the current picture of fundamental life science, they tend to underplay an important aspect— all biological systems are a product of natural selection. This fact enforces an element of historical reasoning in biology which may be missed in the traditional methods of analysis on account of their focus on model systems which are studied in isolation. The key consequence of this evolutionary element in biology is that any given biological system must be viewed as a product which is downstream of a long line of stochastic variations that have been channeled through natural selection. This means that a given feature of a biomolecule such as a protein is heavily influenced by its ancestry rather than by its current functional properties alone. These basic assumptions underlying evolutionary reasoning have been applied by several workers even within the sphere of classical molecular biology and biochemistry. However, their relevance has come to the fore more recently due to the revolutionary developments in genomics over the past decade. These developments provide us with an unprecedented amount of data whose full significance can only be fully realized upon application of methods of evolutionary analysis.

Broadly, the methods of analysis which exploit evolutionary logic include a variety of techniques which identify homology between proteins or nucleic acid elements [1, 2], reconstruct evolutionary (phylogenetic) trees at the level of individual molecules, collection of molecules or whole genomes [3] and extract contextual information from genomic organization and high-throughput experimental data sets for gene and protein expression and protein-protein

interaction [4-9]. Collectively, this class of computational analyses may be termed evolutionary or comparative genomics [1].

On the technical side, comparative evolutionary genomic analysis of protein sequences and structures has mainly concentrated on the development of sensitive sequence and structure similarity search methods, development of statistics to assess significance of sequence and structure similarity, alignment techniques for protein sequences and structures, and construction of phylogenetic trees using multiple alignments (see overview of methods employed below). On the side of biologically relevant applications, the main focus of evolutionary studies on proteins has been prediction of organismal biology or biochemical function from sequence data. More specifically, the latter area has included: protein domain discovery which has vastly improved biochemical understanding of proteins, prediction of novel functional components of biological functional systems, understanding the biochemical mechanisms of enzymes, and reconstructing interaction networks for components of biological systems. The research comprising this dissertation chiefly focuses on the latter area, the application of computational approaches in comparative genomics to understand protein function, structure, and evolution.

The earliest efforts in comparative or evolutionary genomics typically relied on a limited sample of sequences from a limited number of organisms. The first studies recognizing similarity between biomolecular sequences and the possible evolutionary implications of these findings were performed by Drs. Russell Doolittle, Linus Pauling, and Emile Zuckerkandl [10], [11], [12], [13]. The first attempts to systematically classify large amounts of sequences manually collected and maintained within in-house databases were initiated in the lab of Dr. Margaret Dayhoff in the 1960s and 1970s [14, 15]. Improvements in sequencing techniques in molecular biology, notably the development of Edman degradation in determining amino acid sequences [16],

enabled researchers to more rapidly obtain sequences derived from genomic sources and contributed to an increase in sequence availability, leading to an increase in the 1980s in the number of public databases housing genomic sequence data [17-19]. However, the application of these computational techniques to investigation of biological phenomena truly came of age with the development and utilization of high-throughput genome sequencing technologies [20, 21] [22, 23] and the commencement of international initiatives to improve structure-solving technologies and increase the output of solved protein structures [24-26], resulting in a veritable flood of biological data. The rise of centralized and large, publicly-financed databases corresponded with this increase in data, and include, among others, the sequence storing GenBank (originally housed at Los Alamos [27]) and Protein databases at the National Center for Biotechnology Information (NCBI) [28] and the EMBL and SWISS-PROT databases at the European Bioinformatics Institute (EBI) [29, 30]. Databases housing other types of large-scale biological data have also emerged more or less concomitantly, structural data at the Protein Data Bank [31] hosted by Rutgers University and the University of California, San Diego, and metabolic and interaction network data at the Kyoto Encyclopedia for Genes and Genomes at Kyoto University [32], VisANT at Boston University [33], and BIOGRID at the Ontario Cancer Institute [34]. Algorithms capable of rapidly assessing similarity between sequences of interests and an entire database of biomolecule sequences were subsequently developed (see overview of methods employed below for detailed descriptions), most notably the omnipresent BLAST program [35], as well as its forerunners FASTP [36] and FASTA [37] and successor PSI-BLAST [38]. The increase in search space and efficiency of tools performing the searching has ultimately allowed researchers to comprehensively explore features on a genome-wide scale across numerous

organisms, whether these features are conserved across all organisms or restricted to a smaller set of organisms.

Several landmark research findings have been based on comparative evolutionary genomics, and while a comprehensive review of these findings is beyond the scope of this introduction, a sampling is instructive as to the role this field has played in shaping theoretical and experimental discussions in biology over the past 15 years. 1) Providing support for and describing the driving forces behind eukaryogenesis (the origin of eukaryotic cell) [39-41] 2) Novel genes contributing to human disease conditions have been analyzed and function, structure, and evolution have been predicted [42-44]. 3) Identification of genes and other conserved functional genomic elements [45-47]. 4) Elucidation of interacting networks of proteins and the evolutionary descent of characterized biological systems. 5) Providing evidence in support of the RNA world hypothesis and prediction of the functional importance of so-called “RNA genes” in extant organisms [48, 49]. 6) Providing novel insights into the function and evolution of various components of important metabolic and signaling pathways [50, 51]. 7) Comparison of prokaryotic genomes leading to the conclusion that horizontal gene transfer has played a fundamental role in prokaryotic evolution [52].

An additional accomplishment in the field of comparative evolutionary genomics that will frame the organization of this dissertation is the construction of a classification hierarchy for the protein domain universe. Most proteins, with the exception of certain transmembrane proteins and several proteins displaying primarily unstructured regions, can be decomposed into globular regions termed domains. The entire collection of protein domains forms the different components of the protein universe. Comparison of protein structures was again first investigated on a small scale by Pauling, Zuckerkandl, and Doolittle. The first systematic

approaches to classification were performed by Dayhoff and Dr. Michael Rossmann, with Dayhoff proposing in 1976 that related groups of protein domains form “superfamilies” based on shared sequence features [53], [54]. In the early 1980s, Jane Richardson developed the Ribbon diagram method of schematically representing protein structures in three dimensions; the underlying concepts of this method would later aid the development of structural comparison and prediction algorithms [55]. Because sequence diverges at a faster rate than structure, determining higher-order relationships exclusively from sequence data has proven impossible. The development of sensitive structural similarity-based search algorithms allowed researchers to begin probing more ancient homologies between domains. The first algorithm capable of searching against a database of structures was the Dali algorithm developed by Holm and Sander [56]; their research based on Dali-derived scores was the first attempt at a comprehensive classification of the protein domain universe [57]. Murzin and colleagues refined the classification and released the SCOP database [58] after a careful case-by-case analysis of individual domain structures. In brief, the top level of the hierarchy is formed by what are termed “classes”, which contain domains sharing very general sequence features; for example, domains comprised of all β -sheets form the all- β class, entirely α -helical domains form the α -class, the α/β class contains domains with mixed α/β elements, and the $\alpha+\beta$ class contains domains where the α -helices and β -strands are largely segregated. The next level down the hierarchy, the fold level, consists of domains containing the same topological arrangements of secondary structure elements. Within folds, domains with low sequence identity but containing shared structural or sequence features suggestive of a common evolutionary origin are grouped together into superfamilies. The family level at the bottom of the hierarchy typically contains orthologous sets of proteins performing the same function across different organisms [58].

Organization of the Dissertation

In my research, I will be utilizing state of the art tools (see below) for comparative evolutionary analysis to investigate specific problems relating to several key protein folds spanning several of the different classes of protein domains: the PUA fold of the all β class, the P-loop NTPase and Rossmann folds of the α/β class, and β -grasp fold of the $\alpha+\beta$ class. Each section of this dissertation will relate to one of these folds and contain one or more sub-sections representing specific studies in which I performed the bulk or all of the research involved, under the guidance of other researchers in the lab. The specific motivations for investigating each fold are introduced at the beginning of the relevant section along with a brief discussion of its known biological significance and the history surrounding its initial discovery.

This dissertation spans a range of different topics related to each fold listed above, tied together by a common set of methods used to analyze the folds. In order to simplify the reading of this dissertation and eliminate repetition, this introduction will proceed with a description of the methods used across all of the studies, discussing theory and specifically describing commonly-used algorithms. The specific implementation of these methods will then be detailed case-by-case within individual sections of the dissertation (see Table of Contents). Following the description of the methods, the introduction will conclude with an account of how these methods are similarly or differently applied in other labs.

Due to the broad range of topics contained in this dissertation, I have modularized each section to facilitate in the reading of a particular section, regardless of whether all sections have been read. Vital concepts and methods are thus described towards the beginning of individual sections; tailored to the specific content of that section. In this way the dissertation provides an easy reference for individuals interested in the findings of a single section.

Overview of Methods Employed

The methods employed in this research can be divided into four general categories: sequence analysis, structure analysis, phylogenetic tree-building, and genomic contextual comparison methods. Each category and its associated techniques are described in detail below, following a description of the software package used by the lab as a platform for performing computational procedures and the computing resources that were available for use.

TASS software package

All large scale procedures were directed by the TASS (Tools for the Analysis of Sequence and Structure) software package developed in our lab by Vivek Anantharaman, S. Balaji, Abhiman Saraswanti, and L Aravind. The TASS package shares many functionalities with the SEALS (System of Easy Analysis of Lots of Sequences) software package originally developed at the NCBI by Rolen Walker and Eugene Koonin [59]. TASS, currently being prepared for publication, is designed to run on open systems and be flexible enough to allow for easy modification of the different tools built into the system and the easy incorporation of other open-source tools. TASS is designed for use only on the Unix operating system, as it is dependent on the Unix pipe feature (see below for explanation and example). There are approximately 100 commands currently in the package; commands are constructed to be modular, permitting the construction of long “sentences” of several consecutive commands, as was the case in the SEALS package [59].

The primary “currency” of the TASS package mediating access and retrieval of specific sequences of interest is the gi number, a unique identifier for individual sequences housed in databases at the NCBI, including the Protein and Nucleotide databases. After retrieving a sequence or sequences through this gi number, TASS performs a range of processing functions

including conversion into an assortment of formats, retrieval of specific regions in sequences, and alignment construction through links to various alignment programs like T-COFFEE [60], MUSTANG [61], and MUSCLE [62]. TASS is also capable of evaluating sequences through scripted links to search tools like BLAST [35] and HMMer [63], transmembrane helix prediction programs like TMHMM2 [64] and SignalP, and phylogenetic tree-building programs like TREE-PUZZLE [65] and Weighbor [66]. Evaluation is also achieved through accessing gene neighborhoods of user-specified size from the NCBI PTT tables. TASS is also useful for constructing easy-to-view outputs, including alignment viewing in seqrows format (examples of which are scattered throughout the dissertation) and also through automatic generation of color-coded alignments in html format constructed according to various user-specified parameters.

TASS is therefore a means of performing complex analyses on sequence and structure through the use of simple and logical sentence structures. For example, an extended sentence containing multiple commands would be as follows:

```
{/home}> gi2fasta <gi number> | splishpgp -d nr -j 7 -h 0.01 | psi2normal | blast2gi -pcut 0.01 |
gi2fasta | muscle | fasta2seqrows
```

In this command, the sequence corresponding to gi number is retrieved in fasta format from the NCBI Protein database, and the PSI-BLAST program (see below) is called by the *splishpgp* command with several common parameters used in command-line BLAST searches. The *psi2normal* command removes all iterations from the completed PSI-BLAST search except for the final command, the *blast2gi* retrieves the gi numbers from a blast search above a certain specified cut-off (given in the *-pcut* option), and the *gi2fasta* command retrieves the sequences from the

PSI-BLAST output. The *muscle* command sends the sequences to the MUSCLE alignment program, and the final *fasta2seqrows* command converts the format provided by the MUSCLE output into an easier format to view.

Computing resources

For operations requiring a minimal amount of memory, one of five local general purpose Linux machines dedicated to research pursuits at the NCBI was used, four of these contain 32-bit memory and one is 64-bit machine. Operations that could be broken into multiple smaller jobs were performed on the Compute Farm at the NCBI, a Platform LSF cluster designed by the Platform Computing company (<http://www.platform.com/>). The Compute Farm contains roughly 50 interconnected processors each with 8 GB maximum available RAM, and is accessed through the set of commands standard to the LSF platform via the local Linux machines mentioned above. Several commands in TASS are designed to send input directly to the Compute Farm, although due to the need to interface with the LSF platform, these results can not be outputted to another TASS command.

For more computationally intense procedures the National Institutes of Health (NIH) Biowulf cluster was utilized, a Linux distributed memory parallel processing system available to the NIH community designed and constructed by NIH staff. The Biowulf cluster consists of 1560 compute nodes with over 3700 Opteron, Xeon, and XP/Athlon processors, combining for a floating-point 10 TFLOPS capacity (<http://biowulf.nih.gov/>). We generally did not have the need for this kind of computational power; however, it was on occasion used for rapid processing of batch searching of entire genomes.

Methods of Sequence Analysis

The primary search tool used for protein sequence profile analysis in this research is PSI-BLAST [38]. PSI-BLAST is based on the Basic Local Alignment Search Tool [35] (BLAST), a heuristic algorithm which searches for homology between a query sequence and a database of sequences. To describe simply, the BLAST algorithm breaks the query sequence down into composite “words”, small subsequences of a user-defined length (default for protein sequences $W=3$; nucleotide sequences $W=11$). These words are searched against a database of sequences, and when an exact match is made BLAST attempts to extend an ungapped alignment in both directions. If the score for the extended alignment falls above a certain threshold, BLAST constructs a complete gapped alignment between database and query sequence with a modified version of the Smith-Waterman algorithm.

Several methods for scoring alignments of proteins like those outputted by the BLAST algorithm have been developed; these revolve around the use of substitution matrices that assign scores for matched and mismatched proteins in an alignment; based on the frequency of changes observed across related sequences. In my research, two matrix types were commonly applied in conjunction with the BLAST algorithm: the PAM (Point Accepted Matrix) matrix developed by Dayhoff and the BLOSUM (BLOck SUBstitution Matrix) matrix developed by Henikoff and Henikoff [67]. In PAM matrices, scores are calculated through comparison of differences in proteins superimposed on an accurate or accepted phylogenetic model. The PAM1 matrix is normalized to reflect expected substitution rates if 1% of the amino acid residues across sequences had changed. Extrapolating from PAM1, matrices which reflect changes occurring between sequences over a longer period of time can be constructed, the largest commonly in use is the PAM250 matrix. In BLOSUM matrices, scores are calculated by clustering “blocks” of

highly conserved sequences at a specific sequence identity threshold in a multiple alignment, comparing substitutions within these blocks, and weighting each substitution according to the number of occurrences in a given block and across all blocks in a sequence. The scoring threshold for a given BLOSUM matrix is given in its name; the standard BLOSUM62 matrix implies that blocks of sequences within an alignment were clustered at 62% sequence identity [68].

These two matrix types have distinct differences, leading to relative advantages and disadvantages. PAM is based on explicit evolutionary models while BLOSUM is based on implicit models contained in multiple sequence alignments, giving PAM the advantage of more closely representing evolutionary substitution patterns. On the other hand, one common criticism of PAM is that a comparison of substitution events across a phylogenetic tree may fail to take into account multiple substitution events that may have occurred between two nodes in a phylogenetic tree [69]. Furthermore, PAM matrices are also based on global alignments across neighboring sequences in a phylogenetic tree; these alignments can include highly variable regions that could particularly exacerbate the multiple substitution problem. BLOSUM avoids these problems by focusing on blocks of highly-conserved regions in multiple alignments. One additional difference in matrix construction is observed in the method of calculating substitution events between closely-related sequences; PAM considers substitutions as equally important events while BLOSUM weights substitutions according to conservation across different clusters of sequences in a block of a multiple alignment. One other difference in terms of implementation is that higher-numbered PAM while lower-numbered BLOSUM matrices are designed to detect more distant evolutionary relationships between proteins. Since my research deals primarily with distantly-related proteins, these are the matrices that were generally employed [68].

In addition to detecting high-scoring matches and constructing pairwise alignments between query and database sequences, the BLAST algorithm also provides estimates of the statistical significance of a detection event, also referred to as a “hit.” Karlin and Altschul developed the BLAST statistics; determining that the significance of scores for a given hit found by the BLAST algorithm depends on the score of an alignment between two sequences identified by the search (S , which is calculated using match and mismatch scores from one of the scoring matrices described above against an expected frequency of observing a given amino acid pair), sequence lengths (m and n), and a pair of pre-calculated scaling parameters that gauge the search space size and the selected scoring model (K and λ , respectively). They additionally showed that the maximum of a large number of such scores roughly follows an extreme value distribution, with the following equation calculating the expected number (e-value) of hits with at least a score S :

$$E = Kmne^{-\lambda S}$$

The PSI-BLAST program was specifically designed to search for distant relationships between a protein and other proteins in a database beyond what might be identified in a simple BLAST search. PSI-BLAST builds a position-specific weight matrix (PSSM) from BLAST hits that score above the specified e-value. This constructed PSSM captures the sequence conservation patterns of a given protein domain and can be used to iteratively search the database in subsequent discrete steps, with each search being followed by reconstruction of the PSSM based on the results of the previous search. In my research, a PSSM was typically built using either a single sequence or a multiple alignment as starting queries; and was run against the non-redundant (nr) database for a set number of iterations or until convergence was achieved. I often employed the built-in statistical test for compositional bias (known as the SEG filter) associated

with PSI-BLAST might also be used to avoid the detection of spurious hits emerging from compositional bias in protein sequence [70].

Despite the solid statistical basis grounding BLAST and PSI-BLAST searches, it is important to remember that these statistical tests are not fail proof. All sequence analyses in this research were followed by extensive manual examination of the final or, if necessary, intermediate results falling below the scoring cut-off of the PSI-BLAST searches for the presence of conserved sequence and structural motifs diagnostic of a particular protein fold or superfamily [2, 51]. Likewise, sequences recovered above the cut-off were also examined, as genuine hits can also on occasion escape detection. In short, while the BLAST statistics can, and should, be used as a useful guide; ultimately it is a tool useful in answering the binary question (yes or no) of whether or not two sequences are related.

As an independent confirmation of sequence homology, the HMMer software package that also searches for homology between a query sequence and a database of sequences was also employed. Instead of constructing a PSSM, HMMer utilizes Hidden Markov Models (HMMs) constructed from a multiple alignment of a protein domain. Briefly, HMMs are models containing a set of states with emission probabilities determining what will be emitted in a state and a set of transition probabilities determining the likelihood of moving from one state to another state. For example, when modeling a multiple alignment, there would be three states for each column in an alignment, an amino acid state, a gap state, and an insertion state. The amino acid state would contain emission probabilities representing the likelihood of observing a given amino acid in that position, based on the observed count of all twenty amino acids in that position gauged against an amino acid background frequency, often calculated from the total count of a particular amino acid observed throughout a multiple alignment. Likewise, the gap

and insertion states model the likelihood of observing the introduction of a gap or insert region into the alignment. If a column in an alignment corresponds to a common hotspot for insertions in a protein scaffold, the transition state probability for entering the insert state will be higher at this point in the HMM model than at other points.

Using an HMM constructed from a protein domain multiple alignment to test other sequences for possible homology is an example of a classic HMM evaluation problem where a model is provided and the probability that a given sequence of events was generated by the model is in question. Evaluation problems like these are solved through application of the forward-backward algorithm, which considers all possible paths through a model and scores the match between an HMM and an observed sequence. This algorithm is rooted in the Markov assumption, which states that the probability of transitioning to a state is dependent only on the current state and not any previous states. The algorithm is thus able to compute probabilities of a sequence of states given a model while expending a minimal amount of computational time as each state depends only on the previous calculation as well as the transition and emission probabilities characteristic of the current state. The HMMer software package builds an HMM from a multiple sequence alignment using the *hmmbuild* command, and then uses the forward-backward algorithm in the *hmmsearch* command to search a database of sequences for homology [68]. Studies have shown that the probabilities generated by HMMer and PSI-BLAST are comparable over the same sets of data, the major difference being that PSI-BLAST runs considerably faster [71].

Often after retrieving large sets of protein sequences from sequence homology-based searches it is often of interest to finely cluster these proteins based on shared sequence features, as it has been shown that differences in sequence feature conservation often correlate with

functional divergence [72, 73]. In this research the BLASTCLUST program (<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>) has primarily been used to accomplish this type of clustering. BLASTCLUST employs a single-linkage clustering method grouping sequences in a cluster if it is defined as being a neighbor to at least one sequence in that cluster. Because this method of clustering is inclusive relative to other methods like complete linkage clustering in which a sequence is added to a cluster if it was found to be a neighbor with all other sequences in a cluster and average linking clustering in which a sequence is added if it is a neighbor with at least half of the sequences in a cluster; care must be taken to manually examine the results for false positives. In BLASTCLUST, sequences are defined as neighbors if they meet two user-defined parameters, the coverage and score density. Coverage is calculated by maximum (or minimum, if desired) length of the alignment relative to the total sequence length while score density is calculated from the score of a BLAST hit alignment divided by the minimum alignment length of the two sequences being aligned. BLASTCLUST was run at different parameter values until stable clusters of related sequences fell out of the operation; implying that the clusters represented cores of legitimate families or subfamilies of the set of sequences in question. BLASTCLUST was also used to remove redundancies from a set of related sequences by clustering at very high score density and coverage values.

Multiple alignment building can be used to infer key functional residues by gauging conservation across multiple species and can form the basis for several other kinds of computational analysis, including the construction of HMMs discussed above and construction of phylogenetic trees, discussed in detail below. I chiefly employed three multiple alignment programs: T-Coffee [60], MUSCLE [62], and PCMA [74]. The forerunner for all of these programs was the CLUSTAL W program [75] which introduced the progressive alignment method, which

builds a distance matrix based on pair-wise alignment scores of all sequences to be included in an alignment and then uses these scores to construct a rough guide phylogenetic tree utilizing the neighbor-joining algorithm (see below). The multiple alignment is then built by aligning the closest-related sequences on the tree, progressively moving up from the terminal branches of the tree until reaching the root.

T-Coffee (tree-based consistency objective function for alignment evaluation) is a weighted progressive multiple alignment program that also constructs a library of pairwise alignments of all sequences that will comprise the alignment. T-Coffee, however, in contrast to CLUSTAL W, aligns each pairwise alignment with all other single sequences in the multiple alignment for the purpose of calculating a value for each position in the alignment that depends on observed residue matches and mismatches. These values are then used as guides during the alignment building stage, which proceeds by progressively aligning the closest sequences as determined by the guide tree [60].

The PCMA (profile consistency multiple sequence alignment) program employs a strategy that combines attractive elements of the algorithms underlying both the T-Coffee and CLUSTAL W programs, primarily for the purpose of reducing computational time. PCMA identifies groups of sequences with a sequence identity of over 42%, and aligns these groups using CLUSTAL W. Similar to T-Coffee, PCMA then calculates scores for each position in an alignment for all sequence triplets (or groups of sequence as in the CLUSTAL W aligned groups), but it calculates PSSMs for these, similar to those used in PSI-BLAST (see above). Using the PSSMs, PCMA can then quickly calculate rough E-values for groups of sequences, and then progressively build an alignment by combining pairs or groups of sequences beginning with the best scoring sequences. PCMA typically performs around 20 times faster than T-Coffee [74].

The MUSCLE alignment algorithm is considerably more complex, featuring two rounds of progressive alignment followed by an iterative refinement process. In the first step, a distance matrix based on *k*mer distances (calculated from the fraction of common *k*mers across sequences being aligned [76]) and guide tree is constructed using UPGMA instead of neighbor-joining (see below) and an initial progressive alignment is formed. In step 2, this initial tree is refined through recalculation of distances, this time finding the Kimura distance (which is more accurate than *k*mer distance but requires a pre-existing alignment) and again constructing a UPGMA-assembled guide tree and progressive alignment based on that tree. The final step consists of systematically breaking the guide tree formed in the second step by removal of a single edge, and then constructing and aligning PSSMs for the two sets of sequences. Alignments between PSSMs can be scored using the SP (sum of pairs) method which essentially adds normalized match and mismatch penalties across a column using a substitution matrix and gap penalties [77]. If the resulting new alignment scores higher than the previous, the previous is rejected and vice versa. The final step is repeated until a (local) maximum is achieved. Remarkably, the MUSCLE algorithm is twice the speed of CLUSTAL W and capable of aligning many more sequences in a single run [78].

Each of these aforementioned alignment programs occupied a distinctive niche in my research. MUSCLE, which is extremely fast but generally not as accurate as other methods, was often used to generate a quick overview of the general shared features in an aligned set of sequences. T-Coffee was used to construct more reliable alignments at the expense of greater time taken to construct the alignment. PCMA, meanwhile, was adept at aligning sequences that contain blocks of conserved regions that were otherwise extremely divergent.

Another method for extracting information stored in a multiple alignment is through application of programs predicting secondary structure from sequence. Several reviews discuss the strides made in accurately predicting secondary structure over the last 40 years [79-81], initial methods were based on the propensities and physicochemical properties of individual amino acid residues observed in individual sequences, these methods, however, barely crested the 60% accuracy mark. The application of evolutionary information to prediction in the 1990s through incorporation of distant homologues into multiple sequence alignments, nearest neighbor approaches, and training of neural networks on an annotated data set of multiple alignments led to breakthroughs pushing accuracy beyond the 70% mark. Combinatory algorithms incorporating several or all of the above methods developed at the turn of the century brought accuracy near the 80% mark; my research utilized one of these algorithms, the JPRED algorithm [82]. Recent developments in the field involve the incorporation of known small peptide structural information into prediction; researchers claim to have surpassed the 80% accuracy mark [83].

JPRED combines the following prediction methods: PHD, a 3-level neural network algorithm [84]; DSC, an algorithm that relies on amino acid propensity and attribute [85]; NNSSP, a nearest-neighbor algorithm that locates related sequences known to fold in a particular manner and applies this to the existing alignment [86]; PREDATOR, which creates pair-wise alignments of all sequences and uses a modified nearest-neighbor approach to predict secondary structure elements which are ultimately combined for a final prediction [87]; ZPRED, an prediction algorithm based on amino acid properties which are in this case weighted according to the differing conservation observed across secondary structure elements (i.e. β -strand, α -helix, loop) [88]; and MULPRED, which is an algorithm devised by the JPRED builders combining a series of

single sequence prediction methods. The final prediction of secondary structure in JPRED depends on a simple majority vote of the results from the different techniques, with the PHD algorithm casting the tie-breaking vote. Several other methods of determining the final prediction were attempted by the authors including modified voting schemes where greater weight is given to techniques showing better predictive ability on specific element types and neural network training, but no technique improved on the simple majority vote [89].

Prediction of bulk properties was accomplished through several programs. Transmembrane regions of proteins were predicted primarily with the TMHMM2.0 [90] program, which has been shown to predict accurately 97-98% of all transmembrane helices. This detection algorithm is based on an HMM (see above) trained against a known set of experimentally-characterized transmembrane helices and contained several states capturing the architecture of a transmembrane helix and the regions surrounding it, including helix core, helix cap, and loop states. The program discriminates well between signal peptides and transmembrane helices, and also accurately predicts topology of transmembrane helices based on several well-characterized features typical of cytoplasmic vs. non-cytoplasmic regions of these helices [90]. The presence of signal peptide protein regions were predicted using the SignalP program [91], which similarly uses an HMM to predict signal peptide length and the most likely site of peptide cleavage. The SEG program detects globular regions in proteins which typically correspond to the presence of a domain, as opposed to a disordered region with low sequence complexity [92], [93].

Methods of Structural Analysis

Several programs designed for structural examination and manipulation are currently available, I favored two programs in this research: SWISS-PDB viewer [94] and PyMol (<http://www.pymol.org>). SWISS-PDB viewer is particularly useful due to its intuitive interface;

atom-atom distances can be quickly calculated and simulated mutations easily performed using this program. PyMol is considerably more complex, but produces superior images, and is also useful for qualitatively assessing molecular surface diagrams of protein domains.

Comparisons between solved crystal structures were performed using the DALI program [56]. The DALI program converts three-dimensional protein structures into two-dimensional distance matrices, also called distance plots or distance maps. The algorithm then breaks these distance matrices into overlapping subsets of hexapeptide patterns and searches for matches between the patterns from the two query proteins. DALI stores all matches in a pair list, and then uses a Monte Carlo optimization search to perform a random walk designed to extend the matches until a score an optimal score is reached. Conceptually, Monte Carlo methods begin with a match and then randomly select another match, ultimately rejecting or accepting the match depending on whether the final score of the newly-aligned match decreases or improves. The scoring function chosen by Holm and Sander adds the set of all scores from matched residues in a given alignment based on the $C\alpha$ - $C\alpha$ distance of the matched residues. The basic scoring function ϕ is given below:

$$\phi(i,j)=\theta-|d(a)_{ij}-d(b)_{ij}|$$

where d_{ij} is the distance between equivalenced elements as determined by the distance matrix and θ is equal to 1.5 angstroms, or the presumed level of absolute similarity [56].

When searching against a database, the DALI program provides a Z-score as an indicator of the level of significance for a given match. Z-scores calculate the standard deviation from the expected score across a normally-distributed set of scores. When gauging relationships between different proteins in this research, I employed the scoring scheme established by Koonin and colleagues suggesting that Z-scores >10 are characteristic of obvious relationships, such as those

between two closely related proteins of the same family. Z-scores between 10 and 6 typify more distant relationships that might be recovered through sequence profile analysis and searches using HMMs. Z-scores <3 fall in the realm of remote structural relationships and require additional analysis, such as comparisons of topologies to make further inference regarding these relationships [95, 96].

Phylogenetic tree-building

Several problems related to my investigations required the application of phylogenetics, including the need to distinguish between very closely related orthologs in a protein family and the need to discern overarching trends in the evolution of a particular protein domain family or groups of closely related protein domain families. I did not favor any particular analysis over another; instead seeking and reported findings where a consensus across several different phylogenetic techniques was reached. To this end, several different tree-building methods were often applied to the same problem, including distance-based, maximum likelihood, and parsimony methods. Due to the relative ease of construction and algorithm speed, trees were first built using distance-based methods, most commonly with the neighbor-joining and UPGMA algorithms (as implemented by the MEGA [97] and WEIGHBOR [66] software packages). At the core, distance-based methods all entail the construction of a pairwise distance matrix which is calculated from similarity-based scores. Despite the speed of these algorithms, several disadvantages are inherent; first, a degree of evolutionary information is inherently lost through the use of pair-wise scoring systems [98] and second, these algorithms are based on evolutionary assumptions that are, charitably-put, tenuous at best in many situations (for example, the molecular clock hypothesis [99] in the UPGMA method assumes a constant rate of evolution across diverging lineages).

Maximum-likelihood methods, which select the tree from the set of all possible tree topologies that maximizes the likelihood given the data and a standard model of evolution (like a substitution matrix, see above), are among the most statistically robust methods of tree-building. However, they are also the most time-consuming and for a large number of sequences, like the amounts this research typically traffics in, ML-methods border on intractability. To circumvent this, I often built distance-based trees as initial guide trees using the FITCH program of the PHYLIP package [100], and then performed local rearrangements using the PROTML program of the MOLPHY package to obtain a maximum likelihood tree [101]. While this tree is not guaranteed to be the global optimum, it can be reasonably assumed to be a close approximation [102]. In cases when the total number of sequences was tractable, full ML trees were constructed using the Proml program of the PHYLIP package [100]. ML distance analyses used TreePuzzle 4.02 [65] to calculate ML distance matrices along with Puzzleboot for 1000 replicates (<http://www.tree-puzzle.de>); resampled matrices were then analyzed using Fitch from the PHYLIP package [100] and the WEIGHBOR program [66]. We additionally employed Bayesian inference, a method rooted in likelihood analysis, with Markov chain Monte Carlo simulations implemented in the MRBAYES package [103] (see above for a brief discussion of Monte Carlo simulations).

Maximum parsimony-based methods, which select a tree topology minimizing the number of differences observed between sequences, were also implemented. The disadvantages of using parsimony tend to surface when a large number of differences are present across the sequences being examined, i.e. when the degree of homoplasy increases. When this occurs, parsimony methods are less likely to determine the true tree topology because of their inability to discern the different intermediate steps minimizing differences. On the other hand, a powerful

advantage to this approach is that it can be done in the absence of an evolutionary model (like a particular substitution matrix) [104, 105], and is therefore not reliant on the assumptions of that model. It is possible this approach will become more reliable as the number of completely sequenced genomes increases, thus minimizing the level of divergence between the two closest-related sequences in a tree. Another kind of parsimony commonly employed in this research is cladistic parsimony, which asserts that shared derived, preferably irreversible characters (also known as synapomorphies) are useful determinants of phylogenetic relationships [106]. This analysis has historically been associated with the use of SINE and LINE elements as shared characters [106] and the use of intron presence and intron position has been gaining popularity [107, 108]. Rare amino acid replacements have also been recently proposed as a novel type of character [109]. Identify of shared structural characters is often at the core of higher-order evolutionary classifications constructed in this research.

Contextual Inference Analysis

The clustering of functionally related genes in prokaryotic genomes into co-transcribed and co-regulated units, gene neighborhoods, often allows functional assignments through the principle of 'guilt by association' [4-8]. Generally, genes whose products physically interact to form a complex or are involved in successive steps in a biochemical pathway form gene neighborhoods that are conserved over large evolutionary distances. Well-established examples of this phenomenon include the clustering of ribosomal proteins and several functionally associated metabolic proteins in prokaryotes. Assessment of the statistical significance of a prediction based on a conserved gene neighborhood has been an ongoing issue in the comparative genomics community, the currently accepted standard requires conservation across three or more genomes with 87% or less pairwise SSU rRNA identity, below which gene order is

randomized in most genomes [7]. However, it is important to note that even in closely-related genomes; gene order can serve as predictors of functional association between gene products, an example being the relatively poorly-conserved lac operon studied in introductory biochemistry textbooks. Notable examples of functional assignment using conserved gene neighborhoods include identification of novel proteins involved in DNA repair complexes and their components [110], novel ribonucleotide reduction genes [7], and identification of several novel components of the translation machinery [4]. The gene neighborhood for a given gene is collected from the NTT tables at the NCBI which contain the gene order for all complete genome sequences and whole shot-gun sequences via the TASS command `gi2operons`.

Other forms of contextual analysis are also useful in elucidating function where it is not well-understood. Phyletic profiles, or the pattern of co-occurrence of orthologs of a particular gene in a set of genomes under comparison [111, 112] can be used in implicating the function of an uncharacterized gene that shares a profile with a gene of known function. Domain architectures may also be used as contextual information. Certain kinds of protein domain fusions are useful in predicting function, particularly those occurring in the more terminal domains branches of the tree of life and are observed separately in several genomes. These fusions often indicate direct physical interaction between the proteins encoded by solo versions of these genes [113].

Differences and similarities in the application of above methods to other laboratories

Several labs focusing on comparative evolutionary genomics have emerged in the past fifteen years. A few of these labs include those headed by Eugene Koonin, Peer Bork, Alex Bateman, Erik Sonnhammer, and Alexei Murzin. In general, differences between the labs manifest themselves not in techniques used, but in the kinds of biological problems that are

investigated. For example in this lab the focus is mainly on protein sequence and structure, with genome analysis secondary while the labs of Drs. Koonin and Bork spend significant time directly analyzing DNA sequence and genome organization at the level of nucleic acids.

That being said, minor differences in the application of the techniques outlined above do exist across the different labs. Although all labs tend to incorporate both PSSM and HMM-based sequence database methods into experimental design, the primary method of searching varies. Drs. Bateman and Sonnhammer tend to prefer HMM-based searching while the others (this lab included) tend to favor PSSM-based searches. There are compelling arguments for both sides, PSSM-based searching is generally considered better at capturing higher-order relationships. This is due to the implementation of the PSSM-based algorithms which use all positions simultaneously to identify matches, implying that sequences retaining ancestral position similarities will be detected [114] more easily than HMM-based techniques, which assume position independence during searches. Additionally, PSSM-based methods are consistently much faster at identifying and scoring matches relative to HMM-based techniques [71]. HMM-based methods, on the other hand, are advantageous due to a more robust probabilistic basis; HMMs use probability theory to guide scoring while PSSMs rely on additional parameters like database size and the assumptions underlying scoring-matrices.

Another difference is seen in Dr. Bork's lab, which has pioneered novel statistical tests designed to robustly assess the significance of contextual associations compiled from a variety of sources including gene neighborhoods, gene fusions, text mining, and co-occurrence or co-exclusion of a gene in phylogenetic profiles. A web-based implementation of the algorithm is available [115]. Our lab has kept using the standards described above, as questions regarding the

efficacy of the annotations used in text-mining [116] and problems associated with over-representation of certain bacterial genomes [117] may undermine the efficacy of the algorithm.

Our lab is unique from most other labs in the emphasis of cladistics in reconstruction of evolutionary history, although another lab is also known to apply similar techniques at times [118]. As described above, many researchers in the molecular phylogenetics employ cladistics-based approaches in examining irreversible character states; we feel that unique, conserved structural features constitute irreversible states, as they are under selective pressure to retain that state. Some molecular phylogeneticists do not share this view, claiming that sequence features should be the sole source in assessing evolutionary affinities [119].

INVESTIGATION RELATING TO THE PUA FOLD

The PUA fold belongs to the all- β class of protein domains and was initially characterized by Koonin and Aravind in 1999 [120] as a novel class of RNA-binding domains. The ensuing chapter describes the discovery of the ASCH superfamily of proteins, their evolutionary connection to the PUA fold, and the different lines of evidence predicting an RNA-binding role in the cell for ASCH domains. A recent review by Pérez-Arellano and colleagues gives a good overview of the structural diversity in the fold, the recruitment of the domain to RNA-binding in a variety of functional contexts, and touches on the discovery of the ASCH domains described in the investigation below [121].

The ASCH Superfamily: Novel Domains with a Fold Related to the PUA Domain and a Potential Role in RNA Metabolism

(based on reference [122])

Introduction

Systematic analyses of the proteins involved in RNA metabolism have suggested that despite the complexity of this system the majority of proteins are constructed from a relatively small set of conserved globular domains (for summary see [123]). The phyletic profiles of these conserved domains derived from large-scale comparative analyses of genomes from the three super-kingdoms of life show certain interesting features [1, 123]. Many of the RNA-binding domains, typically those present in ribosomal proteins, translation factors and tRNA and rRNA-modifying enzymes, are widely represented across the three superkingdoms of life. These appear to be ancient innovations, which were originally utilized in core RNA metabolism processes that

are likely to have been already present in the last universal common ancestor (LUCA) of all cellular life forms. In some cases, a subset of these ancient domains also appear to have been secondarily recruited to many of the unique eukaryotic innovations such as splicing, post-transcriptional gene silencing, mRNA capping and polyadenylation, and nonsense mediated RNA decay [123, 124]. Identification of these ancient RNA-binding domains have helped considerably in uncovering aspects of RNA-protein interactions that hold good across a wide range of biological functional contexts, and in clarifying the roles of uncharacterized conserved proteins from phylogenetically distant organisms (For example see: [125-129]).

Given these antecedents, my lab was interested in the identification of any potentially novel ancient conserved domains that might throw light on poorly understood ribonucleoprotein complexes that have been identified in the cellular transcription apparatus. The activating signal cointegrator 1 or the thyroid hormone receptor interactor protein 4 (ASC-1/TRIP4) is a transcriptional coactivator that is widely conserved in eukaryotes and is part of a potential RNA interacting protein complex [130, 131]. ASC-1 directly interacts with a wide range of unrelated transcription factors such as the serum response factor, NF κ B, AP-1 and nuclear hormone receptors, and has been shown to be part of a protein complex that bridges these specific transcription factors to the basal transcriptional apparatus [131]. One of the proteins of this coactivator complex is an RNA helicase, while the other one has an RNA-binding KH domain fused to a 2H RNA phosphoesterase [131, 132]. ASC-1 itself contains a conserved cysteine-rich Zn-chelating domain, which binds transcription factors [131] and a conserved C-terminal domain which has thus far not been characterized.

Using sensitive sequence profile searches and structural comparisons we show that C-terminal domain of ASC-1 domain defines a superfamily of domains that is widely distributed

across the 3 superkingdoms of the life. We show that this superfamily assumes a protein fold, which was originally observed in the RNA-binding PUA domain. Our findings suggest that this unique β -barrel fold, which is encountered both in the new superfamily of domains typified by the C-terminal domain of ASC-1 and the PUA superfamily, defines an ancient structural theme in RNA-protein interactions.

Application of Methods

Sequence analysis for this project was performed by Lakshminarayan Iyer, a staff scientist in the lab. The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTPGP program [38]. Iterative sequence profile searches were done using the PSI-BLAST program either with a single sequence or an alignment used as the query, with a profile inclusion expectation (e) value threshold of 0.01, and were iterated until convergence [38]. For all searches with compositionally biased proteins, the statistical correction for this bias was employed [70]. Multiple alignments were constructed using the T_Coffee program, followed by manual correction based on the PSI-BLAST results [60]. Hidden Markov Models (HMMs) were built from alignments using the *hmmbuild* program and searches carried out using the *hmmsearch* program from the HMMer package [63]. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED and PHD programs [82, 89, 133]. Preliminary clustering of proteins was done using the BLASTCLUST program with empirically determined length and score threshold cut-off values (For documentation see <ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>). I performed structural analysis tests and contributed to genome contextual analysis along with the aid of Dr. Iyer, including structure similarity searches that were conducted using the DALI program [96]. Structure manipulations and the construction of ribbon and surface diagrams were

performed using the Pymol program [134]. 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 (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>) with an in-house PERL script incorporated into the TASS package.

Results and Discussion

Identification of the ASCH domain

The ASC-1 proteins from animals are relatively large proteins (around 580-650 amino acids), and the only characterized globular domain in them is a unique Zn-chelating domain with 7 cysteines and 1 histidine. This domain was shown to be critical for the interaction of ASC-1 with specific transcription factors and is likely to form a binuclear metal cluster chelating two Zn atoms [131]. Given that other polypeptides of the ASC-1-containing complex have characteristic RNA-interaction domains, we further investigated the ASC-1 proteins to identify potential links to RNA interaction. Analysis of the human ASC-1 protein with the SEG program revealed that it contains additional globular segments, including a C-terminal globular segment (gi: 6013191, 434-581), which in searches of the NR database with the BLASTPGP gave significant hits to the proteins SAP1p60 from the bacterium *Streptomyces avermitilis* and Mbur03000455 from the archaeon *Methanococcoides burtonii* ($e=10^{-5}$ and 10^{-3} respectively). This region of similarity did not map to any previously published protein domain and more or less encompassed the entire length of the prokaryotic proteins, suggesting that it might define a novel protein domain. Further iterations of the search retrieved a large number of uncharacterized proteins from vertebrates, prokaryotes and bacteriophages such as LOC541578 from *Homo sapiens* (iteration 3; $e=10^{-3}$), gp69

from the Mycobacteriophage Che9c (iteration 2; $e=10^{-6}$), PF0238 from *Pyrococcus furiosus* (iteration 3; $e=10^{-4}$) and the TTC18981 protein from *Thermus thermophilus* (iteration 3; $e=10^{-3}$) whose crystal structure has been determined (pdb id: 1wk2). All the sequences showed a highly conserved GxKxxxxR motif that they shared with the ASC-1 protein. At convergence, the search also retrieved several proteins with the GxKxxxxR motif with e-values of border-line significance ($e > .01$). In order to retrieve all possible homologs for a comprehensive analysis, transitive sequence profile searches were performed by seeding with several homologs of the ASC-1 protein which were recovered in the above search. As a result, several additional significant hits from diverse species from all three superkingdoms were recovered, and proteins whose structures have been determined as part of various structural genomics project ($e < 10^{-2}$) such as the uncharacterized proteins YqfB (pdb:1TE7) from *Escherichia coli*, PF0455 (pdb: 1S04) from *P.furiosus*, and EF3133 (pdb: 1T62) from *Enterococcus faecalis* (Fig. 1). Some these proteins had been classified into separate families of domains of unknown function, DUF437, DUF984 and DUF1530, in the PFAM database [135].

The sequence affinities between the proteins recovered in the above searches were also independently corroborated by searches with HMMs derived using a seed alignment of the originally detected set of ASC-1 homologs. Furthermore, comparisons of the predicted secondary structures for different sub-groups of these homologous domains with the above-mentioned

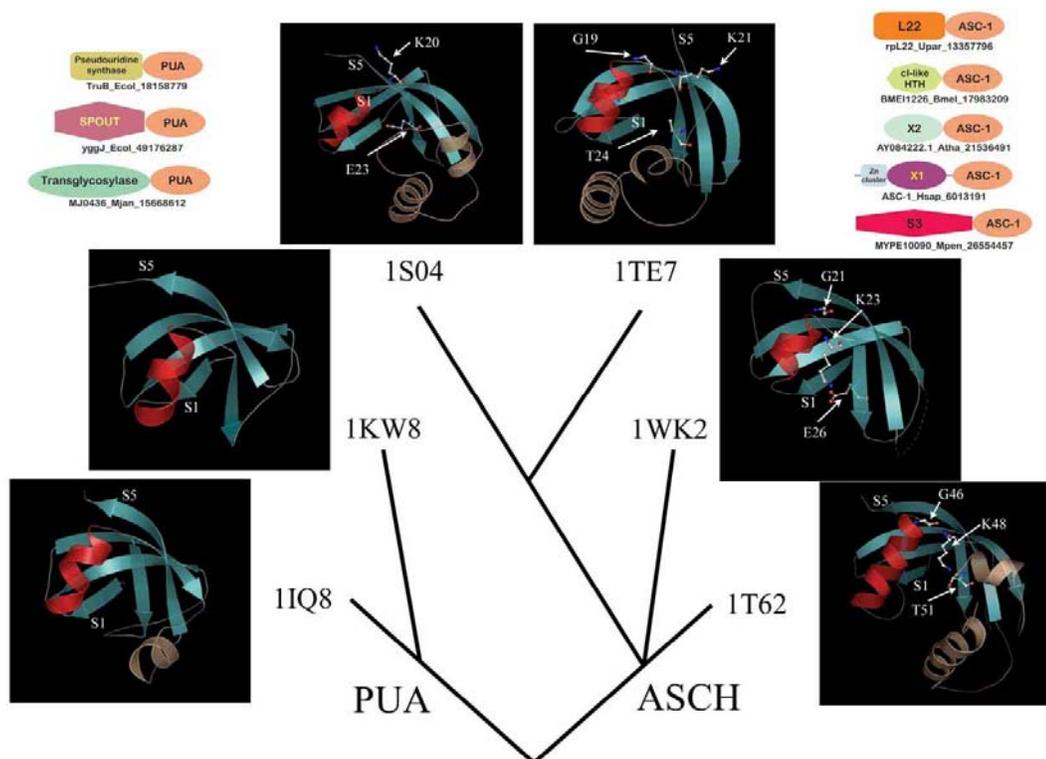


Fig. 1. Structures and domain architectures from the ASCH and PUA superfamilies

Cartoon representations of structures from the ASCH and the PUA superfamilies are mapped on a tree showing the inferred higher order relationships between the two superfamilies. The clustering was derived using distances from pairwise DALI Z-scores. Each structure is labeled with its Protein Data Bank (PDB) identifier. Conserved β -strands are shown in light blue while the characteristic conserved α -helix is shown in red. Variable helical inserts located between strand-4 and strand-5 are colored tan. The structures are shown with strand-2 vertical and approximately central to the depiction. S1 and S5, which are the first and the last strands of the PUA-ASCH fold, are labeled. Key conserved residues lining this cleft in the ASCH superfamily are rendered as ball and stick. Domain architectures of the ASCH superfamily and those of a subset of the PUA superfamily are shown in the top right and top left panels, respectively.

120 amino acid long domains define a novel monophyletic superfamily (Fig. 1). We refer to this superfamily, containing over 180 distinct representatives in the NR database from viruses and cellular organisms belonging to all three superkingdoms of life, as the ASC-1-homology (ASCH) superfamily. Structure similarity searches with members of the ASCH superfamily showed that it contains a fold, which was previously noted in the PUA domain (Fig. 2) (DALI Z-scores 4.5-6).

For example, DALI searches with the *Thermus* TTC18981 protein (pdb id: 1wk2) retrieved the PUA domains from pseudouridine synthase (pdb id: 1k8w, Z score: 4.8), ATP sulfurylase (1g8f, Z score: 4.8) and Archaeosine tRNA-guanine transglycosylase (1k8w Z score: 3.6) in addition to the bona fide ASCH proteins derived from structural genomics projects (pdb ids: 1t62, 1xne, 1zce, 1t5y, 1nxz; Z scores: 4.6-5.8). The PUA domain is an ancient RNA-binding domain, which is fused to the catalytic domains of a variety of RNA-modifying enzymes such as pseudouridine synthetases of the TruB family, the archaeosine transglycosylase, Rossmann fold methylases, YggJ-type SPOUT domain RNA methylases and thiouridine synthases, and also occurs as stand-alone forms [120, 136, 137]. However, PUA domains were not recovered in any of the sequence profile searches seeded with the ASCH domain or vice-versa, suggesting that these two classes of domains form distinct sequence superfamilies, despite them sharing a common fold. We propose that the fold be renamed the PUA-ASCH fold to reflect the two distinct superfamilies of the fold. The ASCH domains contain a conserved core of 5 strands that form a β -barrel, and a characteristic helix between strand-1 and strand-2 (Fig. 2). Additionally, most versions of the ASCH domain, unlike the majority PUA domains, contain a long insert between strand 4 and 5 that usually forms two or more helical segments (Fig. 2). In terms of sequence conservation, the most characteristic feature of the ASCH superfamily is a GxK motif (where x is any amino acid) that is found in the distinctive turn between the core helix and strand-2 (Fig. 1, 2). Members of the ASCH superfamily also contain a highly conserved polar position, two residues downstream of this GXK motif, which is typically occupied by either glutamate or threonine (Fig.1, 2).

Fig. 2. Multiple alignment of members of the ASCH superfamily

Proteins are shown with their gene name, species abbreviations and genbank ID (gi) numbers separated by underscores. The pdb codes of proteins with X-ray crystal or NMR structures are shown in brackets after the gi number. Columns in the alignment are colored based on the residue conservation profile at 90 and 70% consensus. Sample operons and domain architectures of interest are shown to the right of the alignment. The domains in the architectures are separated by a '+' symbol, whereas genes in operons are separated by '->' symbol with the '>' pointing from the 5' to the 3' directions of the coding sequence. X1 and X2 refer to uncharacterized domains, which were found fused with certain ASCH domains. The Pfam domains of unknown function, DUF437, DUF1530 and DUF984, include some of the representatives, respectively, from families 1, 3 and 4 defined by us. The consensus for residue conservation and the coloring scheme are as follows: h, hydrophobic residues (ACFILMVWY), shaded yellow; b, big residues (LIYERFQKMW), shaded gray; s, small residues (AGSVCDN) colored green; p, polar residues (STEDKRNQHC) colored magenta. The lysine residue that is characteristic of the ASCH superfamily is shaded red. Species abbreviations are given at the end of the paper.

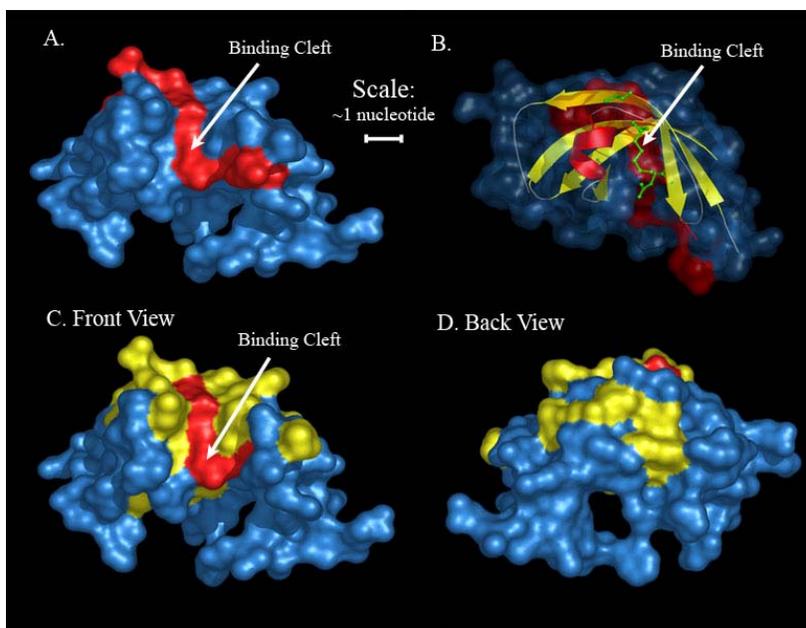
Predicted functions of members of ASCH superfamily

In order to obtain functional insights regarding members of the ASCH superfamily, we used the combined evidence gleaned from different forms of contextual connections, namely physical interactions, gene fusions and conserved operons. In different Gram-positive bacteria such as *Mycoplasma*, *Ureaplasma* and *Lactococcus lactis*, members of the ASCH superfamily are embedded or associated with the ribosomal protein operon (Fig. 1). Specifically, in *M. penetrans* the ASCH domain is fused to the ribosomal protein S3, whereas in *U. parvum* it is fused to ribosomal protein L22 (Fig. 2). Other members of the ASCH family are also found tightly linked with genes encoding RNA-binding proteins with RRM (e.g. in *Acinetobacter*, gene ACIAD0497) or R3H (e.g. *Listeria*, gene lmo2852) domains, implying that they are co-transcribed and probably functionally cooperate. These associations with ribosomal and RNA-metabolism proteins are consistent with the physical interactions of the vertebrate ASC-1 with proteins involved RNA processing and the potential requirement for RNA-protein interactions for transcriptional co-activation by the ASC-1 containing complex [131]. A study of the available structures of four distinct members of the

ASCH superfamily indicates that they contain a prominent cleft, whose scaffold is formed by the conserved helix and the downstream strand-2 (Fig. 2, 3). The above-described conserved residues of the ASCH superfamily, like the lysine from the GXK motif, and other polar residues associated with strand-2, line this cleft forming a positively charged surface (Fig. 3). A similarly positioned cleft has been observed in the structures of the PUA domain found in the Archaeosine tRNA-guanine transglycosylase, Pseudouridine synthase II TruB and the predicted RNA methylase [127, 138, 139], and is likely to form its RNA-binding surface. Taken together the above observations suggest that the ASCH domains are likely to possess RNA-binding activity.

Fig 3. Molecular surfaces of observed binding cleft in ASCH superfamily

Structure of ASCH superfamily (PDB: 1WK2) depicted in four different ways. In A, B and C the protein is oriented to expose the potential binding cleft, located between the helix and strand 2. In the top left (A), the predicted three-dimensional surface of the protein is shown with the conserved residues lining the binding cleft of family 1 colored in red while other surfaces are colored in blue. On the top right (B), cartoons indicating secondary structure features are shown against the transparent outline of the predicted molecular surface of the protein colored in dark blue. The most highly conserved residues found along the cleft are rendered as ball and sticks and are colored in green (G21, K23, E26). In the bottom left (C) and right (D) the front and back views of the predicted molecular surface are shown. Surfaces of residues are colored according to consensus conservation across the entire ASCH superfamily; red denotes positions with at least 90% conservation, while yellow denotes positions with at least 70% conservation.



Structure of ASCH superfamily (PDB: 1WK2) depicted in four different ways. In A, B and C the protein is oriented to expose the potential binding cleft, located between the helix and strand 2. In the top left (A), the predicted three-dimensional surface of the protein is shown with the conserved residues lining the binding cleft of family 1 colored in red while other surfaces are colored in blue. On the top right (B), cartoons indicating secondary structure features are shown against the transparent outline of the predicted molecular surface of the protein colored in dark blue. The most highly conserved residues found along the cleft are rendered as ball and sticks and are colored in green (G21, K23, E26). In the bottom left (C) and right (D) the front and back views of the predicted molecular surface are shown. Surfaces of residues are colored according to consensus conservation across the entire ASCH superfamily; red denotes positions with at least 90% conservation, while yellow denotes positions with at least 70% conservation.

Over the past few years a number of studies have shown that coactivator complexes are often bi-functional proteins that not only co-activate transcription mediated by specific transcription factors, like nuclear hormone receptors, but also participate in pre-mRNA processing [140-142] and regulation of splicing. Furthermore, a regulatory pseudouridylated RNA termed the steroid receptor coactivator RNA (SRA), together with specific RNA-binding proteins with which it interacts, have been shown to be a part of coactivator complexes that couple nuclear hormone receptors to the basal transcription machinery [143-145]. Given these observations, it is likely that the ASCH domain mediates some of the interactions between RNA and the ASC-1 coactivator complex. Its RNA partner could either be the pre-mRNA generated from the transcription of its target genes or a regulatory RNA like SRA. The association with the ribosomal proteins might indicate that some of the prokaryotic versions might be involved in translational regulation.

The prokaryotic and phage ASCH domains, with a few exceptions, occur as standalone versions (Fig. 1), which are encoded by genes in predicted co-transcribed arrays containing a wide variety of other genes. In several of these cases they are found adjacent to a gene encoding a helix-turn-helix protein, which is the transcriptional regulator of the predicted operon (Fig. 1). In *Brucella* an ASCH domain is fused to a CI-like HTH domain within the same polypeptide (Fig. 2). These associations suggest that solo ASCH proteins of prokaryotes functionally cooperate with transcription regulators, probably by binding the transcripts generated from particular operons, and thereby regulate their expression.

Evolutionary diversity of ASCH domains and general conclusions

The ASCH superfamily encompasses considerable diversity and can be sub-divided into several families that are unified by specific sequence signatures. The ASC-1 proper family is

typified by a unique insert between strand-3 and strand-4. It is present in animals (two paralogous versions, with and without a fusion to the Zn-chelating domain are seen in vertebrates, respectively typified by human ASC-1 and LOC541578; Fig. 1), plants, and trypanosomes amongst the eukaryotes and in certain cyanobacteria, actinobacteria and their phages, *Burkholderia* and the archaeon *Methanococoides*. The two copies in the vertebrates appear to have emerged from a relatively recent duplication in the common ancestor of the extant vertebrates with sequenced genomes. Related to the ASC-1 family is family 1 typified by the *Thermus* protein TTC1891 (termed DUF437 in PFAM) that is present in *Thermus*, *Pyrococcus* and *Archaeoglobus*. Family 2 (typified by the standalone ASCH domain protein *Zymomonas* protein ZM00922) is predominantly found in bacteria and archaea, with isolated eukaryotic representatives from the filamentous fungi such as *Neurospora* and *Magnaporthe* (Fig. 1). Likewise sporadic eukaryotic representatives from plants are seen in the otherwise prokaryotic family typified by the *Pyrococcus* protein PH0447 protein (family 3). All the other families of ASCH domains, such as families 4 (DUF984, e.g. EF3133), 5, 6, 7, 8 and 9 are restricted to prokaryotes and their phages. This phyletic pattern of the ASCH superfamily suggests that it diversified in the prokaryotes followed by multiple lateral transfers to the eukaryotes. The Zn-chelating domain and a predicted globular segment immediately downstream of it (Fig. 2) in ASC-1 are conserved in all eukaryotes, and occur as a standalone unit independent of the ASCH domain in basal eukaryotes like *Giardia* (Supplementary information). Hence, the transfer of the ASCH domain from prokaryotes that gave rise to eukaryotic ASC-1 appears to have happened after the divergence of the basal eukaryotic lineages like *Giardia*, followed by a fusion to the above-mentioned standalone unit. This was followed by losses of the ASCH domain in crown group eukaryotes, such as in the fungi. In addition to the emergence of ASC-1, there appear to have

been independent sporadic transfers of other prokaryotic ASCH family members to specific lineages of crown group eukaryotes (Fig. 1).

In terms of phyletic patterns, the PUA domains can be confidently traced back to the LUCA of all cellular life forms. The ancient versions of the PUA domain include those fused to key RNA metabolism enzymes such as the pseudouridine synthetase which are conserved in all the three superkingdoms of life [123, 138]. In the case of the ASCH domain no single family is conserved across the 3 superkingdoms of life, making it unclear whether it was present in LUCA. However, its broad phyletic range in the prokaryotes suggests that the ASCH domain emerged very early in the evolution of the prokaryotic superkingdoms. It is however not universally represented in all prokaryotic genomes and has been lost in some eukaryotes such as the fungi. This suggests that they are likely to belong to the more easily dispensable regulatory apparatus rather than the core aspects of RNA metabolism. No ASCH domain occur as multiple repeats in the same polypeptide unlike many other RNA binding domains such as the KH or the RRM domains. This suggests that it is likely to form single isolated contacts with specific features on RNA rather than extended multi-site contact with long RNA molecules. Furthermore, unlike the structurally similar PUA domains, which typically occur in multi-domain proteins fused to other RNA modifying or interacting domains [120, 123, 137], the ASCH domains typically occur as the sole globular domain in the polypeptide (Fig. 1, 2). The conserved residues on the surface of the predicted cleft are also distinct in the PUA and ASCH superfamilies, suggesting that they bind very different types of target RNAs. The PUA domain appears to have mainly colonized core functional niches related to rRNA and tRNA modification, while the ASCH domains appear to have to been recruited to a distinct set of functional niches, including transcription co-activation and regulation of translation. Thus, the ASCH and PUA domains appear to have emerged from a

common RNA-binding precursor and subsequently diversified to perform distinct functional roles, probably as result of the diversification of their binding clefts.

Supplementary material

A complete of alignment of all ASCH domains in the NR-database and other domains found fused to the ASCH can be retrieved from the following website:

<http://www.ncbi.nlm.nih.gov/CBBresearch/Lakshmin/aschsupplementary.html>.

INVESTIGATION RELATING TO THE P-LOOP NTPASE FOLD

The phosphate-binding loop (P-loop) protein domain fold belongs to the α/β protein domain class and was initially identified as an important assemblage of ATP- and GTP-binding proteins [146] with several conserved motifs involved in the binding and transfer of phosphate groups to substrates [147, 148]. It has also been recognized as the largest assemblages of paralogous globular protein domains of the cellular proteomes, likely due to its fundamental role in driving catalytic reactions depending on ATP, the currency of energy in the cell [149].

This section consists of a comparative analysis of viral DNA packaging systems. Recent landmark studies identified a common domain (the β -jelly roll domain) in viral capsid proteins across several virus types including dsDNA, ssDNA, and ssRNA viruses suggesting a common evolutionary origin for capsid proteins from diverse viral lineages [150, 151]. At the same time, it has been recognized for some time that the packaging ATPases of diverse DNA viruses all belong to the P-loop (phosphate-binding loop) NTPase fold [152] [153], suggesting possible evolutionary relationships existing between these packaging systems. This lab has performed a numerous studies on the higher-order relationships of different members of the P-loop ATPase fold [154-158] and as such was poised to study these packaging proteins as well as other components of the system in depth to establish their precise origins, affinities, and evolutionary trajectories. Results from this study provide novel insight into the DNA viral packaging mechanism and also into the general evolution of DNA viruses.

Comparative Genomics And Evolutionary Trajectories Of Viral ATP Dependent DNA-Packaging Systems

(based on the following reference [159])

Introduction

Proper segregation of chromosomes and their partitioning into daughter cells or capsids is a common problem faced by cellular and viral replicons. While diverse solutions to this problem have evolved in different viruses, they may all be categorized under two broad mechanistic themes [160, 161]. Most RNA viruses and several small DNA viruses do not appear to require an active energy-dependent process for packaging their genomes, and the process simply proceeds via coating of the nucleic acids by capsid subunits. Coating is usually initiated by packaging signals in the form of sequence or structural features in the nucleic acid, resulting in condensation of the capsid proteins on the nucleic acid scaffold [162, 163]. In the second theme, an active ATP-dependent process drives the genome of larger double stranded (ds) DNA viruses and single (ss) stranded DNA viruses of the Inovirus family into empty capsids [158, 161, 164].

Extreme sequence divergence of viral proteins has hampered understanding of relationships between components of chromosome segregation and packaging systems of different viruses. However, recent availability of a wealth of crystal structures and complete sequences of numerous viral genomes allows us to address this problem using a variety of sequence and structure analysis techniques and comparative genomics. Some recent developments in this regard include structural studies on viral coat proteins revealing that barring a few exceptions, the principal capsid or coat protein of the majority of characterized viruses contains a distinctive β -strand fold with a β -jelly-roll topology [150, 151]. Remarkably, this structural conservation of capsid proteins transcends the diversity of viruses, which might be

otherwise unrelated in terms of their genomic nucleic acid or replication and packaging mechanisms. This raised the intriguing possibility that principal capsid proteins of a majority of viruses might have descended from a common ancestor [151].

Similarly, sensitive sequence comparisons showed that packaging ATPase motors of diverse large eukaryotic and prokaryotic DNA viruses belong to the HerA-FtsK superfamily which includes the DNA pumps involved in prokaryotic cellular chromosome segregation and related DNA pumps of several conjugative plasmids and transposons [158]. Packaging ATPases of the HerA-FtsK superfamily are encountered in the recently unified Nucleo-cytoplasmic Large DNA virus (NCLDV) assemblage and in several dsDNA phages like PRD1 and the Inovirus family [158, 164]. The other major functionally characterized ATP-dependent DNA-packaging system is the terminase-portal protein system first noticed in caudoviruses (tailed prokaryotic viruses) and herpesviruses [152, 153]. In its most basic form the system consists of the two-subunit terminase complex and a multimeric portal protein (PP) providing a conduit for nucleic acid entry into capsids. The terminase large subunit (TLS) has both ATPase activity that powers DNA translocation and nuclease activity, which cleaves the replicating DNA into genome-sized fragments [165-168]. In addition to these systems, there are smaller families of packaging ATPases in ϕ 29-like bacteriophages and the adenoviruses whose evolutionary affinities were previously unclear [169, 170].

This study builds upon these previous studies to provide a synthetic overview of the protein components of various ATP-dependent phage DNA-packaging systems. This study establishes the evolutionary affinities of several poorly understood components and also describe new potential components. Relationships and structural features of packaging proteins presented

here also throw light on various functional aspects of DNA packaging, with general implications for the origins of chromosome segregation.

Application of Methods

Due to the enormous amount of profile searches ran, often using as queries all of the proteins in a given viral genome, Dr. Iyer and I worked together to conduct profile searches using the PSI-BLAST program with either a single sequence or alignment as query [38]. I also performed all of the contextual analyses outlined in the paragraph below, as well as the statistical analyses on gene distribution in the virus genomes (see description in main text below). Drs. Iyer and Aravind both aided in guiding the investigation through its different phases.

PSSM searches were typically run with a PSSM inclusion expectation (E) value threshold of 0.01, and were iterated until convergence. All other sequence analyses were performed by myself; multiple alignments of protein sequence were constructed using the T_Coffee [171], PCMA [172] and MUSCLE [62] software packages, followed by manual correction based on the PSI-BLAST results. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED program, with information extracted from a PSSM, HMM and the seed alignment itself [173]. Similarity based clustering of proteins was carried out using the BLASTCLUST program (<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>). I also performed all structure and phylogenetic-related analyses; searches of the PDB database with query structures were conducted using the DALI program [96]. Protein structures were visualized and manipulated using the Swiss-PDB viewer program [174]. Phylogenetic analysis was carried out using the maximum-likelihood, neighbor-joining and least squares methods. This process involved the construction of a least squares tree using the FITCH program or a neighbor joining tree using the NEIGHBOR program (both from the Phylip package) [175], followed by local

rearrangement using the Protml program of the Molphy package [176] to arrive at the maximum likelihood (ML) tree. The statistical significance of various nodes of this ML tree was assessed using the relative estimate of logarithmic likelihood bootstrap (Protml RELL-BP), with 10,000 replicates. Gene neighborhoods were collected and analyzed using the custom script of the TASS package (V.Anantharaman, S.Balaji and L.A unpublished) that derives tables of gene neighbors centered on a query gene.

Results and Discussion

The demography of packaging ATPases in large DNA viruses

Amongst large eukaryotic DNA viruses all the NCLDVs, which includes poxviruses, iridoviruses, African Swine Fever Virus, phycodnaviruses and the mimivirus, share a packaging ATPase of the HerA-FtsK superfamily [158, 177]. The herpesviruses contain a terminase-portal packaging system similar to the bacteriophages [152]. Additionally, a HerA/FtsK-type ATPase is also encoded by a novel DNA transposon that is widespread in *Trichomonas*, ciliate and nematode genomes. It also has some relationship with adenoviruses in terms to its DNA polymerase and processing protease, suggesting that it might assemble into virus like particles aided by this ATPase [158, 178]. The unique packaging ATPase of the adenoviruses has thus far not been seen in any other viral lineage [170, 179]. Packaging ATPases, if any, of certain large DNA viruses like baculoviruses and the shrimp white spot syndrome virus are unknown, but they are unlikely to define large new lineages of packaging enzymes.

A systematic survey of phage packaging system components in completed genomes of 239 prokaryotic DNA viruses showed that they encompass a comparable diversity in terms of their ATPases. Up to a genome size of about 20 kb there is a steady increase in the fraction of phages encoding packaging ATPases (Fig. 4A). Beyond this size, 95% of phages encode a

packaging ATPase. The majority of small phages lacking packaging ATPases are microviruses, which initiate their packaging through a passive interaction with a small genomically encoded polypeptide [163]. This suggests that 20 kb is the approximate size threshold above which packaging appears to require an active energy-dependent process. The most common packaging ATPase in currently available phages is the terminase-type ATPase (seen in ~70% of the phages), whereas ~20% of phages utilize a version of the HerA-FtsK ATPase superfamily (see supplementary material: SM). The presence of a terminase-type ATPase is strongly correlated with the tailed capsid morphology typical of caudoviruses, the most common type of bacteriophage. The HerA-FtsK family appears to exclusively occur in phages with internal lipid

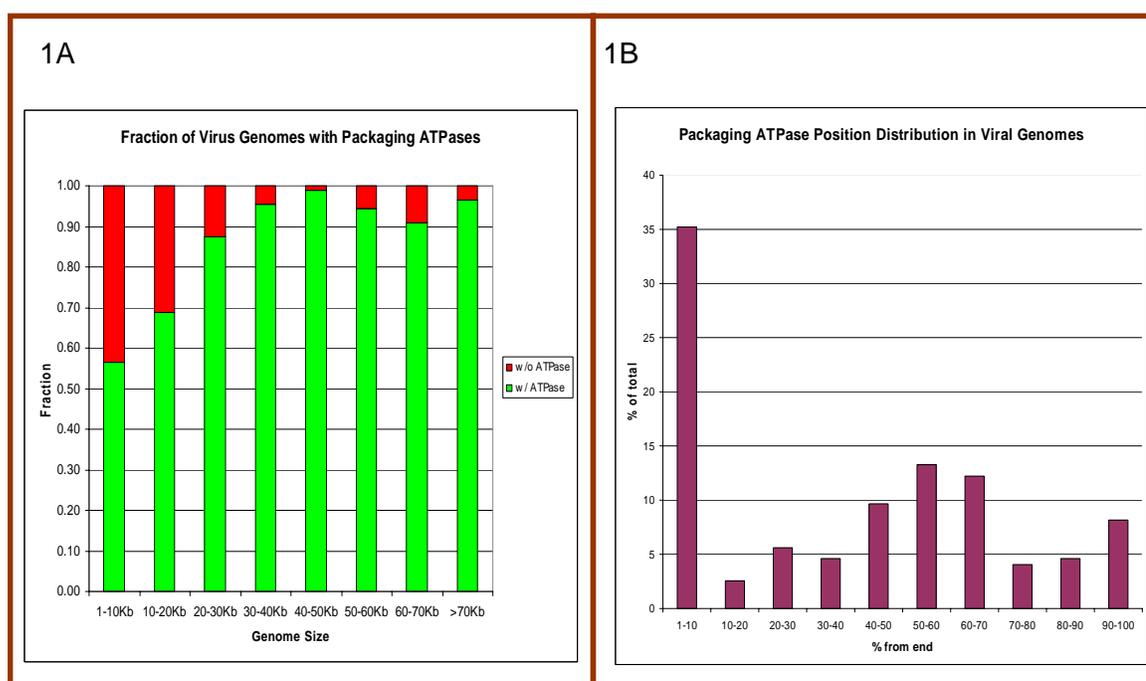


Fig. 4. Packaging ATPase presence/absence and position distribution in viral genomes.

(A) Presence/absence of a packaging ATPase in completely-sequenced viral genomes is depicted as a stacked column graph. Percentages of genomes containing a packaging ATPase within a certain genome size range are green columns while percentages lacking an ATPase are red. (B) Genome position frequency distribution of packaging ATPases from completely-sequenced viral genomes with linear chromosomes is shown as bars graph. Statistically significant preference for placement in the middle or termini of viral genomes is observed (χ^2 : $p < 10^{-5}$).

membranes, such as tectiviruses, corticoviruses and *Sulfolobus* turreted virus, irrespective of their outer protein coat morphology (SM) [158, 164, 180]. These phages also often contain terminal inverted repeats in their genomes. Furthermore, 82% of viruses with terminase-type ATPases have linear chromosomes, while 63% of those with HerA-FtsK type ATPases have circular chromosomes (SM). This suggests that while each system can handle either chromosome type, there might be a preferred type for each of them.

A study of the positional distribution of genes for packaging ATPases in phages with linear genomes revealed that in 80% of the cases they are either located at an end or close to the center of the genome (Fig. 4B). This unusual distribution is highly significant ($p < 10^{-5}$ by Chi-test) and appears to be related to the time of transcription of the packaging ATPase in the virus life cycle. Placement of these genes towards the chromosome termini or middle may allow late transcription, thereby making the packaging apparatus available only at the last phase of the viral cycle. This bias in chromosomal position of the gene for packaging ATPases provides a contextual means of predicting potential packaging ATPases of uncharacterized viruses. Two archaeal globuloviruses are observed (*Thermoproteus tenax* spherical virus 1: TTSV and *Pyrobaculum* spherical virus: PSV) with genome sizes greater than 20 kb lacking any known packaging ATPases. However, both viruses encode an uncharacterized P-loop NTPase at the termini of their genomes (TTSV: ORF1 and PSV: ORF582). Sequence searches with these proteins showed no close relation to any other ATPases involved in replication such as helicases or clamp loaders; supporting its possible role as a packaging ATPase.

Multiple origins for different packaging ATPases within the P-loop NTPase fold

All known and predicted packaging ATPases of DNA viruses belong to the P-loop NTPase fold, one of the most prevalent protein folds in both cellular and viral genomes [153, 158, 170]. Members of the P-loop NTPase fold are unified by the conserved nucleotide binding (Walker A) and Mg²⁺ binding motifs (Walker B) and belong to one of two major divisions; the KG division which includes P-loop kinases and GTPases, and the ASCE (additional strand conserved E (glutamate)) division [154, 181]. The latter division is characterized by an additional conserved acidic residue (typically a glutamate occurring immediately after the conserved Walker B aspartate) and a conserved polar residue (Sensor 1) occurring at the end of the 4th core strand of the domain [181, 182]. Examination of all characterized packaging ATPase domains, namely the TLS N-terminal domain, the HerA-FtsK ATPase domain, the ϕ 29-like phage ATPase domain, and the adenoviral packaging ATPase domain revealed hallmark features of the ASCE division, indicating derivation from within this radiation of the P-loop fold [158, 170]. This observation is consistent with the fact that majority of highly active ATPases mediating energy dependent processes in biological systems belong to the ASCE division [154, 181, 182].

However, relationships between different viral packaging ATPases and affinities to other major classes of ATPases of the ASCE group have remained largely unclear. Previous systematic analysis of the HerA-FtsK superfamily revealed that viral packaging ATPases of this superfamily do not form an exclusive virus-specific clade but are successive out-groups of the crown-group formed by cellular HerA and FtsK families. The basal-most clade was comprised of packaging ATPases of filamentous inoviruses with ssDNA genomes while those from remaining diverse groups of lipid membrane-containing dsDNA viruses of prokaryotes and eukaryotes formed a large assemblage, an immediate sister group of the cellular and plasmid members of this

superfamily [158]. Preliminary sequence searches with ϕ 29-like ATPases recovered only cognate ATPases of other related viruses and secondary structure prediction revealed that ϕ 29-like ATPases contained an α - β unit C-terminal to strand-2 as seen in the FtsK, RecA, helicase and PilT assemblage within the ASCE division. Furthermore, the ϕ 29-like ATPases bore a conserved arginine at the base of strand-4, equivalent to identically positioned arginine fingers in HerA/FtsK ATPases. ϕ 29-like ATPases also possessed a conserved asparagine at the end of the sensor-1 strand, equivalent to the glutamine seen in the HerA/FtsK superfamily (SM). These observations together with the statistically significant recovery of ϕ 29-ATPases by sensitive profiles of the HerA/FtsK superfamily indicate that the former are a distinct branch of the latter superfamily. However, there were no specific features that unified ϕ 29-ATPases with HerA/FtsK-type packaging ATPases of other dsDNA viruses, suggesting that they are likely a rapidly diverging independent lineage within the HerA/FtsK superfamily.

TLSs are almost always two domain proteins with an N-terminal ASCE-type P-loop ATPase domain and a C-terminal nuclease domain with a RuvC-like version of the RNaseH fold [183, 184]. The secondary structure of the terminase ATPase domain revealed the presence of at least one additional strand after strand-2 (SM, Fig.5), placing them in a monophyletic assemblage of the ASCE division along with HerA/FtsK, PilT, RecA and helicase superfamilies [158]. However, they lack the C-terminal β -hairpin or any other specific features characteristic of most members of the above assemblage [158] (SM). The TLS ATPase domain is distinguished from other related ATPases by the presence of a poorly conserved but universally present insert after the second β - α unit which includes strand-2. They also contain a characteristic arginine at the third position in the Walker A motif. While it could potentially act as an arginine finger in the terminase multimer, such a function remains uncertain as the arginine is absent in a few active

terminases, like that of phage T1 (SM). Thus, it appears that TLS ATPase domains comprise a separate lineage within the above monophyletic assembly of the ASCE division.

Adenoviral packaging ATPases (IVA2) consistently retrieved ABC ATPases as best hits. They specifically share with the ABC ATPases two polar residues at the end of the sensor-1 strand, one of which is a highly conserved histidine. Secondary structure predictions also suggest that they contain an insert with β -strands after helix-1 which might be equivalent to the corresponding insert found in all ABC ATPases [185]. Hence, adenoviral ATPases were probably derived from the ABC superfamily. Adenoviral ATPases additionally contain a distinct C-terminal extension predicted to form an $\alpha+\beta$ domain with two conserved aromatic positions and several polar residues (SM). It might play a role in recognizing packaging-initiation signals in genomic regions. Thus, it appears that DNA packaging has been derived on at least three independent occasions within the ASCE division of P-loop NTPases, with two of them being exclusively comprised of packaging ATPases (HerA/FtsK and terminase) and the third from a superfamily of ATPases that were ancestrally associated with DNA (ABC ATPases). The predicted packaging ATPases of archaeal globuloviruses can currently only be identified as members of the ASCE division with no close relationships to any of the other three classes, and might represent a fourth independent innovation. Interestingly, the only characterized packaging ATPases of dsRNA viruses, those of cystoviruses (e.g. $\phi 12$), represent another independent recruitment for packaging function from within the RecA superfamily [186, 187] (Fig. 5).

Ancillary components of DNA packaging systems

Functional studies to date have not uncovered any conserved system of interacting proteins that function along with viral members of the HerA/FtsK superfamily. Previous studies have shown their cooperation with diverse nucleases in resolving target DNA during active pumping by

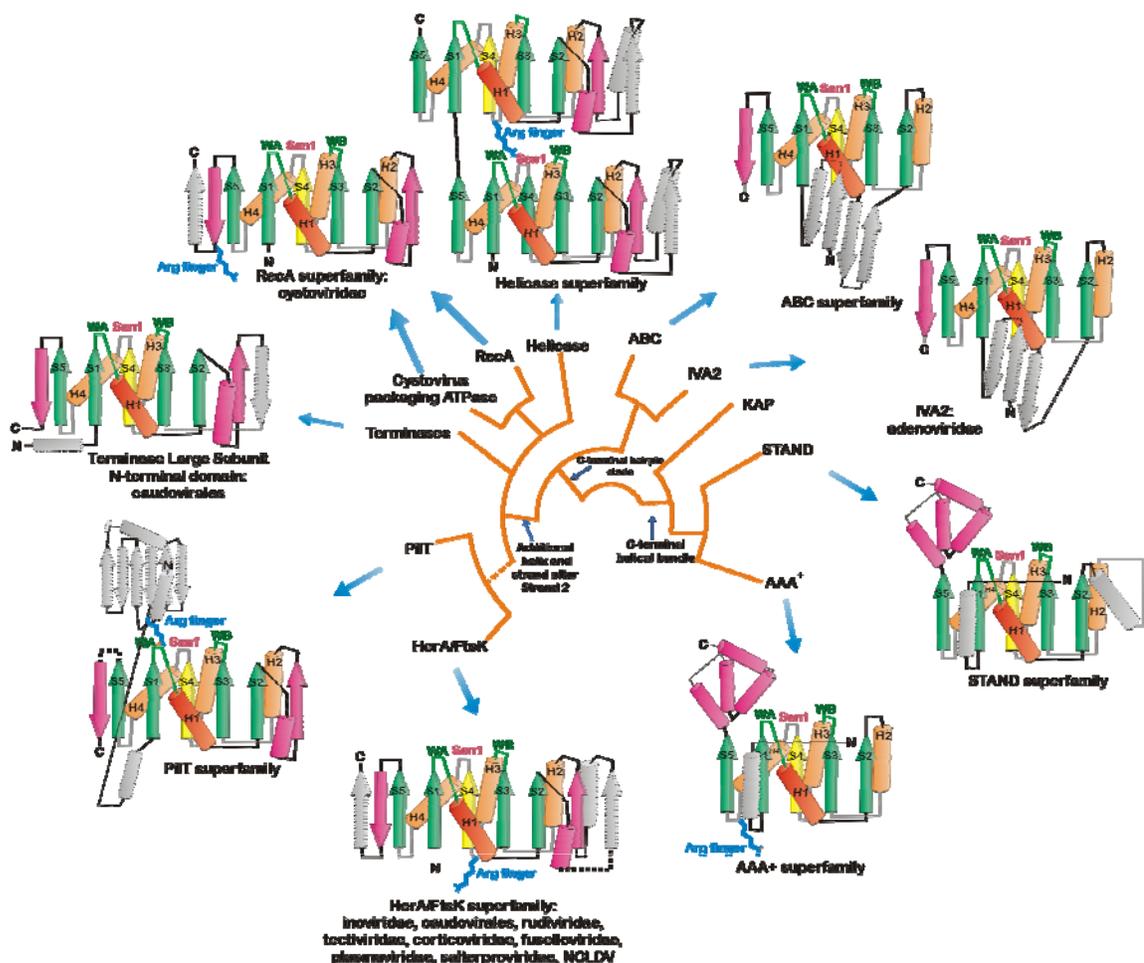


Fig. 5. Topology diagrams depicting ASCE division of P-loop NTPases and accompanying cladogram depicting higher-order relationships.

Strands and helices forming the core of the ASCE P-loop NTPase domain are numbered and colored. Strands are in green with the central strand S4 in yellow and helices in orange. Synapomorphies shared across different lineages are colored pink, elements not conserved across lineages are colored gray and outlined in broken lines. Lines connecting different lineages represent higher-order relationships constructed by comparison of shared structural and/or sequence similarities. Broken lines represent relationships with more uncertainty. Abbreviations: WA, Walker A; WB, Walker B and Sen1, sensor-1.

these ATPases [158]. Hence, it is likely that these ATPases cooperate during packaging with different resolvases, including the frequently present RuvC-like resolvase, in NCLDVs and prokaryotic dsDNA viruses [180, 188, 189]. The ϕ -29 lineage of the HerA/FtsK superfamily

appears to utilize a distinct portal protein (PP) containing a globular domain of the SH3 fold [190]. This domain forms a multimeric ring similar to those formed by other nucleic acid binding members of this fold such as the RNA binding Sm domain [190-192]. Adenoviruses possess a unique ancillary protein, not observed elsewhere in the viral universe, which functions in conjunction with the IVA2 ATPase to interact with genomic packaging sequences [193, 194]. Secondary structure predictions of this protein indicate a lineage-specific α -helical fold. Terminase systems show considerable diversity with different types of PPs and ancillary components like terminase small subunits (TSS). Given this diversity, their origins were investigated and new interacting components were identified using genomic context information.

Diversity of the terminase-dependent packaging systems: a common origin for portal proteins of all tailed bacteriophages

In addition to the TLS whose two domains supply ATP-dependent motor and nuclease activity, packaging systems of all characterized caudoviruses also require a PP. PPs form homo-multimers providing a conduit for DNA into the viral prohead Guasch, 1998 #176; Simpson, 2000 #178; Bazinet, 1988 #183}. In contrast to the common origin of the TLSs, the PPs of these viruses were believed to belong to distinct families, typified by versions found in phage T4, T5, λ , A118 and Mu [195]. To investigate evolutionary affinities of PPs, systematic transitive sequence profile searches were initiated from all known versions of PPs. As a result of these searches, PPs were recovered from a variety of phages or their equivalents such as the head-tail connector protein (gp8) of phage T7; consequently unifying all known PPs of tailed bacteriophages. These searches showed that every TLS-containing phage also encoded one predicted PP suggesting a strict functional association (SM). The above unification of PPs of diverse phage families also suggested descent from a common ancestor like their terminase counterpart. However, they have

subsequently undergone rather drastic sequence divergence. Secondary structure prediction of the conserved core shared by PPs indicates a six-stranded region embedded between two predominantly α -helical elements. The most prominent sequence conservation is in the β -strand rich region and includes a Gxs (where 'x' is any amino acid and 's' a small residue) prior to the first conserved strand (SM). Sequence similarity-based clustering and examination of conserved shared motifs in the alignment helped us to discern eight distinct families (Supplementary table 1), which further grouped together into four higher order clades (T1/T5/⊙-like clade, the T4/SPP1/ ϕ g1e-like clade, the phage μ -like clade and the phage T3/T7-like clade).

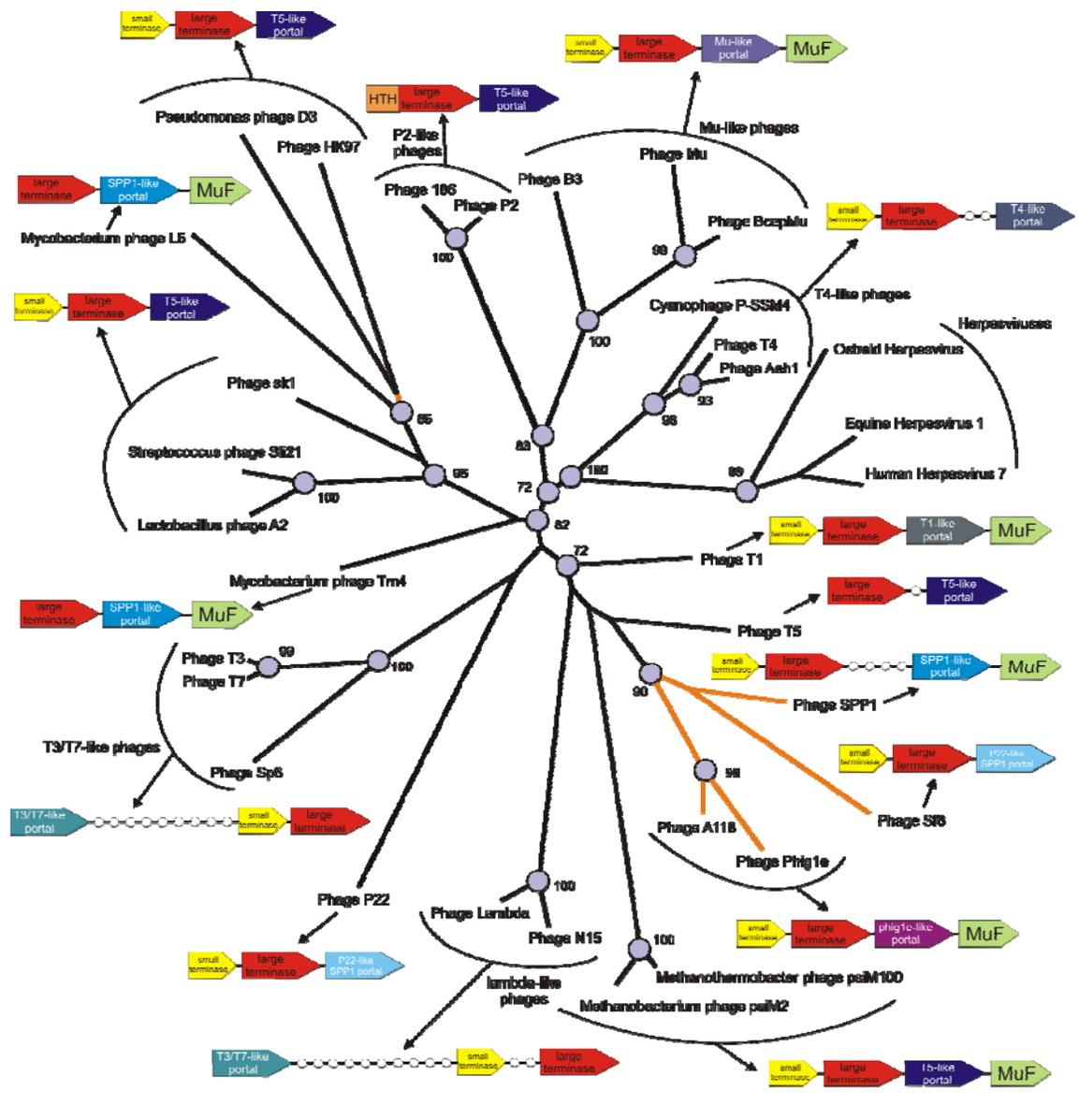
The presence of a conserved β -strand-rich region in the PPs is reminiscent of the SH3 fold β -barrel in the ϕ 29-type PP. The conserved Gxs motif in the former superfamily is also reminiscent of a similar motif seen in the corresponding position of the SH3-like barrels [196]. Hence, despite the lack of significant sequence similarity, it is not impossible that a similar β -barrel might be present in the PPs of the terminase-dependent systems. Likewise, the herpesviral PPs, while displaying no significant sequence similarity to those of the phages, also contain a core β -strand rich region suggesting the presence of a similar structure (data not shown). I propose that this β -strand rich region might form a comparable β -barrel domain, which multimerizes to give rise to the funnel shaped portal.

Contextual information and inference of novel components of the terminase portal systems

Conserved gene neighborhoods (operons) and gene fusions have proven to be a powerful method for predicting previously unknown functional associations and protein-protein interactions in prokaryotes and their viruses [4, 197]. In order to identify other functional links to the terminase-portal system, all gene neighborhoods of terminase-portal pairs were systematically explored in bacteriophages. In terms of gene neighborhood, the most commonly

found association is between the TLS and the PP with the two typically occurring as neighboring genes in several viral genomes (some exceptions include λ , T3/T7 and T5) (SM, Fig. 6). PP genes are rarely fused to other genes suggesting that multimerization and strict interactions with TLS are likely to select against fusion proteins. One notable fusion of the PP is with a lysozyme (e.g. *Burkholderia* prophage, gi: 78061894; Fig. 6) which might correlate with the incorporation of lysozymes in viral capsids for their role in host entry.

Terminase small subunits (TSS) have been characterized in phages such as T4, T7, λ and SPP1, but corresponding small subunits have not been found in many of the other tailed bacteriophages [198-200]. Examination of gene neighborhoods suggested a strong association between the genes for the TSS and the TLS (Fig. 6). The crystal structure of the λ small subunit shows a specialized derivative of the winged HTH—the MerR-like HTH, which lacks the first of three conserved helices in the HTH domain [201]. This suggests that the primary role of the TSS is binding DNA. Accordingly, the contextual information of gene neighborhood and sequence profile searches were combined to characterize the other TSSs and identify previously undetected versions. Our searches identified TSSs in 151 of the 206 phages containing terminase-portal systems. While all these small subunits contain the HTH fold, they included versions distinct from the MerR-type HTH seen in λ -like TSSs. In total, seven distinct families of TSS were identified and also few sporadic unclassified HTH domains. Of these, the largest families were SPP1-type TSS and D3-like TSS. The SPP1-like family was shown to contain a simple trihelical HTH module of the FIS type, while the remaining families did not belong to any previously characterized type of HTH domain and likely represent phage-specific divergent versions of the fold (SM). In a subset of phages, including P2, the SPP1-like TSS is fused to the TLS, supporting



Domain Architectures

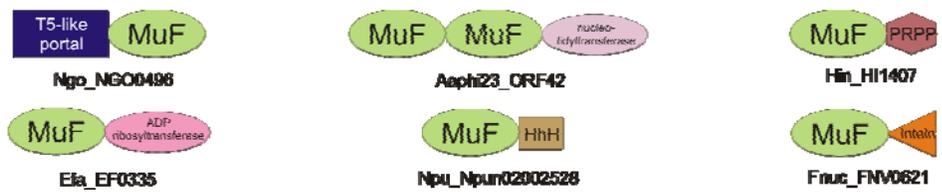


Fig. 6. Phylogenetic tree of TLS depicting gene displacement among portal protein families.

Phylogenetic trees were built using the least-square method with subsequent local rearrangement to obtain the maximum likelihood tree (see SM for details). Reliability of the tree topology was assessed using the RELL bootstrap method of MOLPHY, with 10,000 replications (SM). Branches where gene displacement has occurred as discussed in the text are colored orange for emphasis. Gene neighborhoods corresponding to TLSs are adjacent to branch ends, genes are shown as boxed arrows. TLS genes are colored in red, TSS colored in yellow, MuF colored in green, and PPs are colored according to family type. Nodes with bootstrap support >70% are linked by circles and labeled by bootstrap value. Domain architectures are also given below the tree, with organism abbreviations and gene names (separated by an underscore) written below. Abbreviations: HhH, helix-hairpin-helix; PRPP, PRPP amidotransferase. Please see SM for phage abbreviations.

the strong functional association between the two subunits through physical interaction (SM).

The above observations suggest that unlike the TLS, the TSS has been derived from the HTH fold on multiple occasions, and convergently evolved similar functional associations with TLS.

The next major family of proteins, often encoded in the same conserved gene neighborhoods as other components of the terminase system, is the so-called MuF family. This family, typified by phage SPP1 gp7 protein, is a component of the phage prohead. In Gram-positive bacteria, the MuF protein is known to associate with the PP and is believed to be led into the prohead by the latter [202, 203]. In our sequence profile searches, MuF proteins were detected in representatives of all major tailed prokaryotic virus families and their prophage derivatives (including one in archaeon *Methanococcus*: MJ0329). Nevertheless, several phages in each of these families lacked MuF, suggesting that it might not be an essential component of terminase-portal systems (Fig. 6). The MuF gene is almost always immediately downstream of the PP gene and is associated with genes for several distinct families of portal proteins in different phages like T1, T5, Mu, λ and SPP1 (Fig. 6). In one instance it is fused to a T5-like portal gene (*Neisseria* prophage, gi: 59800934), reinforcing the strong functional association between these two components.

MuF contains a characteristic C-terminal region with conserved cysteines, histidines and acidic residues suggesting it might form a distinct metal-chelating domain which might be involved in MuF-mediated DNA binding activity. MuF proteins show a number of fusions to other domains in several (pro)phages. These include fusions to the DNA-binding HhH domain (*Nostoc* prophage, gi: 23130420) and several catalytic domains such as ADP ribosyltransferase (*Enterococcus* prophage, gi: 29374974), pol- β -fold nucleotidyltransferase (phage Aa ϕ 23, gi: 31408074), PRPP amidotransferase (*Haemophilus* prophage, gi:16273315) and multiple intein-type HINT peptidase domains (*Fusobacterium*, gi: 34763916; *Bifidobacterium* gi: 23335596). ADP ribosyltransferases have been observed in a variety of phages, including T4 and eukaryotic NCLDV, like PBCV and mimivirus [180]. T4 ADP ribosyltransferases ModA, ModB and Alt are packaged into phage heads, and are involved in modifying a range of host proteins [204]. Hence, the MuF might help in loading ADP-ribosyltransferase and other catalytic activities in the phage head for modification of host or viral proteins. HINT peptidases fused to MuF are related to the BUBL1 peptidase of ciliates which is involved in cleaving tandemly-fused ubiquitin repeats and ADP ribosyltransferase domains [205]. Consequently, MuF associated HINT peptidases might be similarly involved in phage head maturation. In this context, it should be noted that the portal-terminase system genes including MuF are often combined with another conserved gene neighborhood, which contains proteases involved in capsid maturation belonging to ClpP or herpesvirus assemblin-like folds.

In situ gene displacement in terminase portal gene neighborhoods

Diversification of PPs into several distinct subgroups and recruitment of several distinct types of HTH domains as TSS raised the question of whether there was a correlation between the distinct families of these proteins and the phylogeny of the TLS. Only TLSs show sufficient

sequence conservation to reconstruct a suitably resolved phylogenetic tree through conventional methods (Fig. 6). Hence, this tree is used as a reference to study the distribution of other components of the terminase-portal system and structures of their gene neighborhoods. This distribution showed the following features. 1) MuF proteins show a sporadic distribution with related phages often differing in its presence or absence. 2) Phages with related TLS might often differ in the type of PP or TSS they are associated with. For example, phage SPP1 has an SPP1-like PP (PP2 family) while the related phage Sf6 contains a P22-like version. Likewise, related TLSs of phages P2 and B3 differ in terms of associated TSS and PP and presence or absence of MuF (Fig. 6).

These observations suggest that terminase-portal gene neighborhoods are prone to: 1) frequent gene loss and acquisition, evidenced by sporadic distribution of MuF and 2) *in situ* displacement of functionally equivalent proteins by evolutionarily unrelated or distantly related counterparts. This situation is parallel to previously observed gene neighborhoods of phage single strand annealing proteins and capsid maturation proteases [206, 207]. The presence of relatively strict gene orders (TSS followed by TLS, PP and MuF) suggest strong constraints with respect to their synthesis and interactions. General rarity or absence of gene fusions among TSS, TLS and PP suggest that their interactions are strongly coupled without much scope for additional associations. Based on gene order and nature of domain fusions, I speculate that TSS is synthesized first and associates with viral DNA. It subsequently recruits the TLS which processes DNA and recruits the PP through which DNA is loaded into the prohead. The PP in turn appears to recruit MuF, which might help position DNA into proheads and recruit other catalytic activities for capsid maturation.

Evolutionary considerations and general conclusions

The systematic survey of diverse active viral DNA-packaging systems suggests that their motors have been derived from two major superfamilies of ASCE ATPases (HerA/FtsK and TLS N-terminal domain). The remaining packaging motors are also derived from the ASCE division, but are very limited in their spread and appear to lack an extended evolutionary history. Taken together with monophyly of the capsid proteins of several DNA and RNA viruses, this suggests an early origin for the two major ATP-dependent packaging systems in the context of ancient pre-existing capsid-like envelopes [180].

Interestingly, both ancient superfamilies of packaging ATPases function in conjunction with DNAses that process or manipulate the products of genome replication. While TLSs contain the C-terminal RNaseH fold nuclease domain, the HerA/FtsK superfamily functions with a range of distinct nucleases in cellular and viral systems, such as XerC/XerD, NurA, RCR, pT181/Rep, Sir2 and possibly RuvC-like resolvases (in several NCLDV) [158, 188, 189]. The RNaseH-fold domain in TLS is most closely related in terms of its conserved active site to the RuvC resolvases and nuclease domains of several transposases (such as TnpA, mariner, hermes, Rag1/Transib and retroviral integrases) [184, 208, 209]. Thus, ATPases of both packaging systems probably associated with an ancestral DNA manipulating nuclease of the RNaseH fold, which appears to have diversified into nuclease, integrase or resolvase families of viral and cellular replicons. HerA/FtsK ATPases form ring-structures and lack domain fusions with their nuclease partners. This appears to have allowed more frequent evolutionary displacements of their nuclease partners by functionally equivalent nucleases [158]. In contrast, there is no evidence for TLSs forming comparable arginine finger-stabilized rings, and fusion with their nuclease partner appears to have been retained throughout their evolution. In general, functional associations

between nucleases and packaging ATPases suggest that from inception packaging systems were closely associated with post-replication genome segregation. Increasing size of DNA-based genomes probably provided the selection pressure for emergence of such systems [180].

Interestingly, diversification of several other superfamilies in the ASCE division of P-loop NTPases might be linked to emergence and expansion of DNA-based replication systems. These include DNA helicases of AAA+, recombinases of RecA, and higher order chromosome condensation proteins of ABC superfamilies. Hence, the two major DNA packaging systems probably arose as part of this diversification of ASCE NTPases concomitant with diversification of DNA-based replicons that occurred well before the emergence of the Last Universal Common Ancestor (LUCA) of cellular life. The nature of the envelope of early replicons, lipid membranes or purely protein capsids, appears to have played a principal role in emergence of the two independent packaging motors. In this context, it is notable that cellular systems (bacteria and archaea) use related packaging ATPases as viruses with lipid inner membranes. Thus, precursors of cellular compartments could have emerged from systems similar to lipid containing viral capsids [180]. While both major packaging systems remained largely mutually exclusive, on rare occasions potential hybrid systems were found. The $\phi 29$ system uses a HerA/FtsK ATPase but depends on a PP analogous to caudoviruses. Like the latter, it lacks an inner membrane, and has a unique hexameric RNA component [210]. It remains unclear if this pRNA is a remnant of a more ancient system or merely a lineage-specific innovation of $\phi 29$ -like phages. Similarly, evolution of adenoviruses, might have involved displacement of the HerA/FtsK ATPase of the above-mentioned DNA elements by a neomorphic packaging system. Our unification of PPs suggests that the terminase-dependent system emerged with a PP partner from earliest stages of their existence. The observation that most of these viruses also contain a version of the HTH

domain (TSS) suggests that there might have been a third DNA-interacting component recruiting nucleic acids to the motor in ancestral versions of this system.

I hope this investigation might help in further experimental investigations on functional interactions in these systems.

Supplementary material

Supplementary material is provided in a single file that can be accessed at http://www.ncbi.nlm.nih.gov/CBBresearch/Lakshmin/supplementary_material.html.

Experimental validation of work presented above

While our manuscript was in proof, Rao and colleagues published the crystal structure of the bacteriophage T4 packaging NTPase [211]. This structure confirms the location of the arginine finger and the oligomerization strategy predicted in the above pages. Other predictions in this research await further experimental confirmation.

INVESTIGATIONS RELATING TO THE ROSSMANN FOLD

Rossmann fold is an ancient fold of the mixed α/β class consisting of a three-layered alpha/beta sandwich composed of repeating beta-alpha units with a characteristic cross-over occurring in the central sheet, initially identified over thirty years ago through some of the first experiments involving comparison of conserved structural elements in proteins [54]. Members of this fold have been recruited to a wide range of functional roles, and at least two higher-order assemblages of the fold have been previously recognized. The first, referred to as classic Rossmann folds, are unified primarily by the presence of a glycine-rich nucleotide-binding loop found after the first β -strand in the fold and includes the FAD/NAD(P)-binding superfamily which contains the C-terminal domain of alcohol dehydrogenase-like domains and the S-adenosyl-L-methionine-dependent methyltransferase superfamily which includes the FtsJ-type RNA methyltransferase domains [54]. The second, referred to as the HUP assemblage, is unified by a conserved set of structural and sequence features and includes among others the phosphoryl-group transferring and AMP-generating superfamilies like the class I aminoacyl-tRNA synthetase HIGH Nucleotidyltransferase superfamily, and the adenosine phosphate-binding redox reaction-catalyzing domains functioning in electron transfer reactions in the ETFP (electron transport flavoprotein) superfamily [212].

The preceding section consists of three studies, each containing a comparative analysis of different structural and evolutionary facets of the Rossmann fold. In the first study, I establish a novel higher-order assemblage of Rossmann fold domains and investigate in detail the higher-order relationships and conserved features of one of the members of this assemblage, the

Haloacid Dehalogenase (HAD) superfamily. In the second study, I selected a previously uncharacterized target protein from the HAD superfamily and in conjunction with Drs. Ezra Peisach and Karen Allen we solved the crystal structure for the protein and predicted a possible functional role for the protein using genome contextual information. The final study describes the E1-like superfamily of the Rossmann fold, establishing its higher-order evolutionary relationships and a paradigm that describes an aspect of substrate interaction in members of the superfamily.

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

(based on reference [213])

Introduction

All cellular organisms extensively depend 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[214],[147]. 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 [181], the RNaseH fold of ATPases, including actin, Hsp70 and their relatives [161, 215], the DHH [216], HD [217], PHP [218], HAD [219, 220], calcineurin-like [221], synaptojanin-like [222], and the Receiver domain (CheY) superfamilies [223], [224]. They span the entire range of structural basic classes with α -helical forms, such as the HD superfamily [225], the beta-barrels such as the CYTH

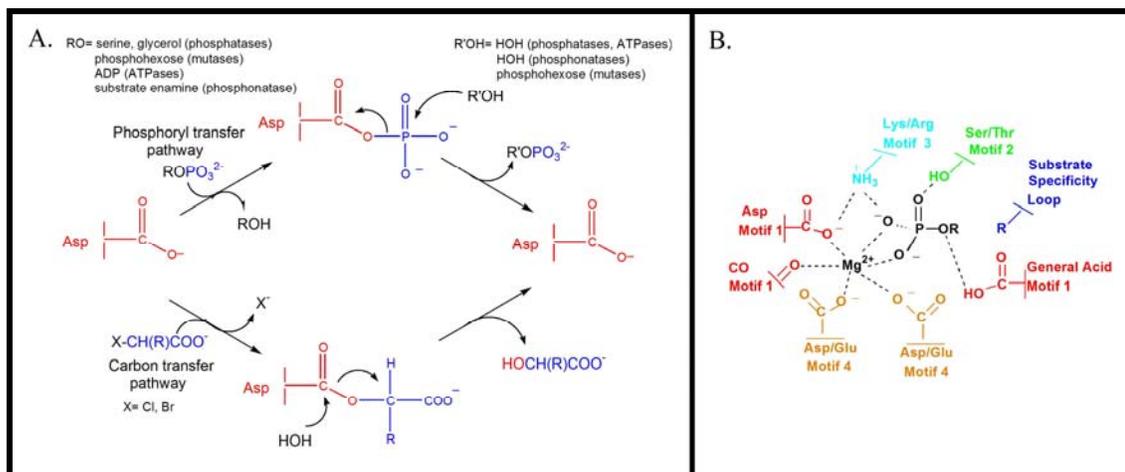


Fig 7. HAD reaction mechanisms

A schematic representation of the reaction pathway in carbon transfer and in phosphoryl transfer is depicted. (A) The 5 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. (B) Schematic of the active-site template for the phosphoryl transferases showing the interactions of substrate with the catalytic motifs (contributed from the core domain) and substrate specificity determinants (usually contributed by the cap domain). Residues contributed from each motif are color-coordinated; the substrate specificity component is colored in blue.

superfamily of phosphohydrolases[226], 3-layered α/β sandwiches, such as P-loop NTPases [181], HAD [227] and DHH [228], [229], α/β barrels such as the PHP phosphoesterases [230], and 4 layered α/β -sandwichessuch as the calcineurin-like [231] and synaptojanin-like phosphoesterases [232].

The HAD superfamily, named after the archetypal enzyme Haloacid Dehalogenase [219], includes enzymes catalyzing carbon or phosphoryl group transfer reactions on a diverse range of substrates, using an active site aspartate in nucleophilic catalysis (Fig. 7A). The majority of the enzymes in this superfamily are involved in phosphoryl transfer i.e. phosphate monoester hydrolases (phosphatases) or phosphoanhydride hydrolases P-type ATPases. These include

variations such as a phosphonoacetaldehyde hydrolase (phosphonatase) and phosphotransferases, such as β -phosphoglucomutase and β -mannophosphomutase. Each of the phosphotransferase enzymes requires a Mg^{2+} cofactor for catalysis [220] [219] (Fig 7B). The carbon group transfer reaction (Fig. 7A) catalyzed by haloalkanoic acid dehalogenase (HAD)[233] 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 were 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" [234]. 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, in this work 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.

Application of Methods

I performed all of the investigations contained within this study, with input and direction received at different steps of the analysis from Drs. Aravind, Allen, and Deborah Dunaway-Marino at the University of New Mexico. The non-redundant (NR) database of protein sequences

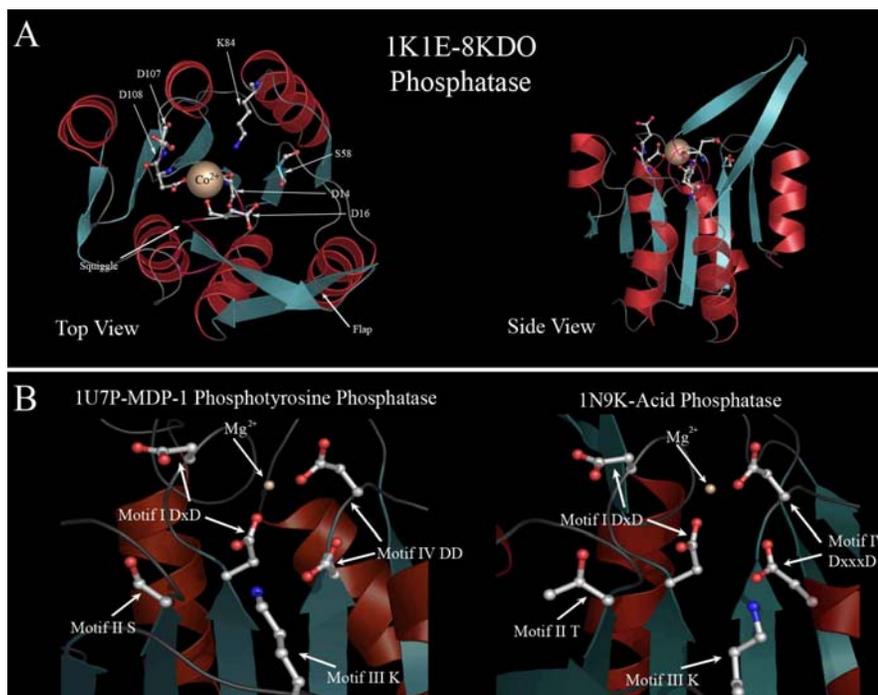
(National Center for Biotechnology Information, NIH, Bethesda, MD) was searched using the BLASTP program [38]. Iterative database searches were performed using the PSI-BLAST program with an alignment or a single sequence serving as the query, typically expectation value (E-value) of 0.01 for inclusion in the position-specific scoring matrix (PSSM); searches were iterated until convergence [38]. 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 [78] and/or the T-COFFEE [60] 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, LA unpublished). Transmembrane regions were predicted in individual proteins using the default parameters in the TMPRED (http://www.ch.embnet.org/software/TMPRED_form.html) and the TMMH2.0 [90] programs. Signal peptides in individual proteins were predicted using the SignalP program [91]. Protein structures were visualized and manipulated with the Swiss-PDB viewer [94] and PyMOL (<http://www.pymol.org>) 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 [89] program. The DALI program was used for structural comparisons [95] (See supplementary material for details). Similarity-based clustering of proteins was accomplished using BLASTCLUST (<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>).

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

Fig. 8. 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 8KDO phosphatase



of *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. (B) Close-up active site views of two HAD representatives with distinct motif IV signatures. The left panel has a motif IV DD signature while the right panel has a motif IV DxxxD signature. Panels are labeled with gene names and PDB identifiers.

(<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>) with an in-house PERL script.

Phylogenetic analysis was carried out using maximum-likelihood, neighbor-joining, and minimum evolution (least squares) methods (see supplementary material for details).

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 I first define its essential structural core, and compare it to other structurally related folds. The core catalytic domain of the HAD superfamily contains a 3-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 5 strands in a 54123 strand order (Fig. 8A, 9). These strands are hereafter referred to as S1-S5. The HAD fold is distinguished from all other related Rossmannoid folds by two key structural motifs (Fig. 9). First, immediately downstream of strand S1, is a unique ~ 6 residue 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 [212]. I term this motif the "squiggle". In some members of the HAD superfamily the squiggle forms hydrogen bonds between the i th and $i+5^{\text{th}}$ position resulting in the rare pi-helix conformation [235]. Second, downstream of the squiggle there is a β -hairpin turn formed by two strands projecting from the core of the domain (Fig. 8A). I 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 [220]. Sequence motif I corresponds to strand S1 and the DxD signature is present at the end of this strand (Fig. 7B, 10). The carboxylate group of the first Asp and the backbone C=O of the second Asp coordinate the Mg^{2+} cofactor (Fig. 7B). Additionally, the first Asp in motif I acts as a nucleophile that forms an aspartyl-intermediate during catalysis [236], [237], [238], [239], [240]. 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 [241]. 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 phosphonatasases, 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 (Fig. 7B, 10). Motif III is centered on a conserved lysine that occurs around the N-terminus of the helix located upstream of S4 (Fig. 8, 9, 10). 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 [242]. 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 3 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 [181]. Motif IV maps to strand S4 and the conserved acidic residues located at its end. These terminal acidic residues of Motif IV typically

Fig 10. Multiple sequence alignment of HAD-domain containing proteins

(From previous pages) The alignment shows only conserved structural regions. Unconserved regions, including cap regions, are replaced with numbers denoting the excised residue count. The top line of the alignment indicates the general 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. Otherwise, coloring is the same as outlined in Fig. 8. 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.

exhibit one of three basic signatures: DD, GDxxxD, or GDxxxxD (where x is any amino acid) (Fig. 7B, 8, 10). These acidic residues along with those in motif I are required for coordinating the Mg^{2+} ion in the active site [243], [241], [244], [236], [245], [246], [247], [248]. 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 (Fig. 7B). This binding cleft is partly covered by the β -hairpin flap occurring after S1 (Fig. 8A, 9). 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 catalytic functions, and play a central role in the reactions catalyzed by most HAD hydrolases [249], [237], [250] (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 3-layered α/β sandwich folds (Fig. 9). It shares with other Rossmannoid 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 (Fig. 9) [212]. 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 [251], Sir2-like deacetylases [252], the S-AdoMet-binding methyltransferases [253], [254], [255], the GTPase FtsZ [256], the ISOCOT fold [257] and the HUP superclass (Class I tRNA synthetases, HIGH nucleotidyltransferases, USPA, photolyase and electron transport flavoprotein) [212]. 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) [216], the receiver or CheY domain of the two-component signaling system [223], [224], the TOPRIM domain, which is the shared catalytic domain of the topoisomerases and DnaG-type primases [258], the PIN/5'-3' nuclease domain [259], the classical histone deacetylases/arginases [260] and the vWA (von Willebrandt factor A) domain [261] (Second division only depicted in Fig. 9). Most members of this division are also unified by a second acidic residue which is 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 Fig. 9). Like the HAD domains, the receiver domain forms an aspartyl phosphate intermediate [262],

which receives a phosphate from a histidinyl-phosphate on the histidine kinase [262], [263], [264]. Because of the mechanistic similarity, the receiver domain has previously been claimed to be a member of the HAD fold [265], [266]. However, a careful examination of the active site organization and sheet topology of the receiver domains (Fig. 9) shows that it does not share any 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 (Fig. 9).

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 (Fig. 9) and there is no evidence for any aspartylphosphate intermediate being formed [267],[268] [260]. In the PIN/5'-3' nuclease domains, a catalytic Mg^{2+} ion is chelated by the acidic residues including those occurring at the end of the S1 equivalent and the strand immediately to its left [259] 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 (Fig. 9) that acts as a general acid or base in the hydrolysis of the phosphoester bond or polynucleotide transfer [258], [269]. The DXD motif is instead borne at the end of strand left of S1 (Fig. 9) and coordinates an Mg^{2+} ion. In the vWA domain the first aspartate is part of the so called MIDAS metal-binding motif (DxSxS [261], [270]), which is critical for the 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 deacetylase/arginase and DHH superfamilies, suggesting that they might also form caps, which control 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 5-stranded central sheet (Fig. 9), to which a major division of the HAD superfamily appears to have added a C-terminal β - α unit after the 5th strand-helix unit (S6), extending the sandwich further (at the left side of the sheet in Fig. 9). The additional strand S6 was lost on rare occasions in members of this 6-stranded division, especially in the context of C-terminal domain fusions. Likewise, on the opposite side (right end of the sheet in Fig. 9) 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 (Fig. 9). 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 (Fig. 9, 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.

The most dramatic modification, however, is seen in the proteobacterial BcbF family of phosphatases, which exist as obligate dimers in the catalytic form (Fig. 9, 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 (Fig. 9). 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 motif I, II and III and the other monomer supplying motif IV associated with the swapped strand (Fig. 9). Given that this configuration has a very limited phyletic spread, this dramatic modification appears to have evolved rather recently through a relatively simple process. I suspect that the ancestral version was a 5-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 3 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 caps- these caps 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 (Fig. 9). 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 (Fig. 11). Slightly longer inserts are

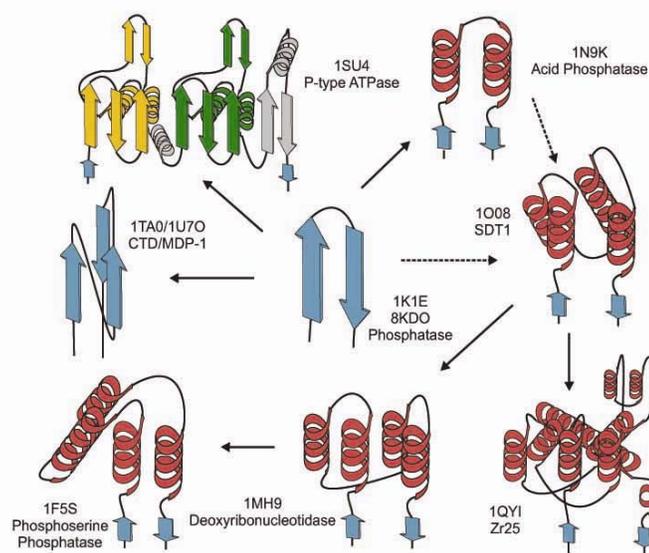


Fig 11. Topology diagrams of selected C0 and C1 cap HAD domains

Representatives are identified with PDB code followed by one or more HAD family name. With the exception of the P-type ATPases, 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. Arrows refer to the likely evolutionary

path that led to the diversification of C1 caps. Broken arrows reflect equally probable paths. 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 and are rendered in gray.

is elaborated further, with the addition of a strand between the 2 sheets forming the β -hairpin; resulting in a cap in the form of 3-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.

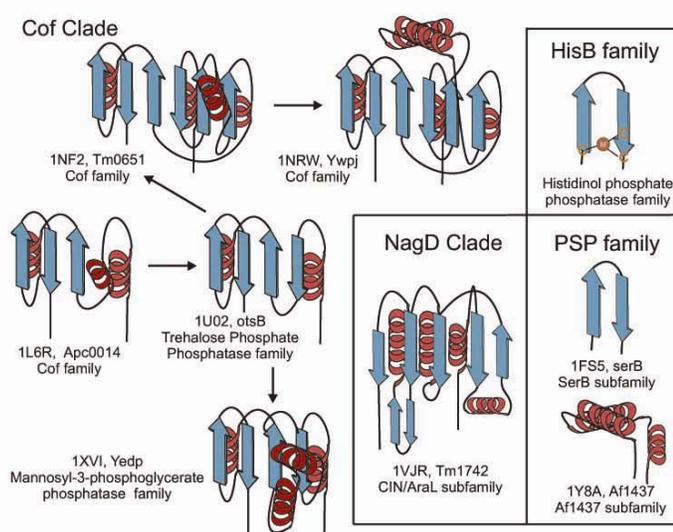
The classical C1 caps belong to two distinct structural classes, the α -helical C1 caps and the cap with the unique $\alpha+\beta$ fold seen in the P-type ATPases (Fig. 11). The most basic α -helical cap in the form of bi-helical α -hairpin is observed in the acid phosphatase and the cN-I nucleotidase families (Fig. 11). 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 8 prominent helices (Fig. 11). 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 [95],[96] 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 $\alpha+\beta$ unit, with a core sheet formed by a 3-stranded β -meander (Fig. 11). 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 3-stranded sheet. Subsequent duplication of this unit appears to have generated the C1 cap seen in extant P-type ATPases (Fig. 11). However, this 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

Fig 12. Topology diagrams of selected HAD C2 cap domains.

Representatives are identified with PDB code and family name. Strands are shown as arrows rendered in blue with the arrowhead on the C-terminal side. Helices are shown as red coils. The main section of the figure contains arrows that show the likely evolutionary progression of C2 caps in the Cof Hydrolase clade. 1L6R represents the most likely ancestral state. The boxes to the right show three other independent innovations of the C2 caps. The depiction of the HisB C2 cap is based on secondary structure predictions, as no structure is currently available. Orange-colored C's represent likely metal-chelating cysteine residues.



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 (Fig. 11).

There are two major unrelated types of classical C2 caps, respectively seen in the Cof-type phosphatases and the NagD-like phosphatases and its relatives (Fig. 12). Both these types of C2 caps are distinctly $\alpha+\beta$ with a core β -sheet containing at least 3 strands. However, in structural similarity searches with the DALI program [95], [96] and through manual examination of topologies, I was 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 I observed a remarkable diversification of the C2 cap through accretion of secondary structure elements to a basic unit with a 3-stranded anti-parallel β -sheet (Fig. 12). The most basic version, seen in the protein Ta0175 (PDB: 1L6R) from *Thermoplasma acidophilum* [271], contains a 3-

stranded anti-parallel 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 (Fig. 12). In the uncharacterized phosphatase Tm0651 from *Thermotoga maritima* (PDB:1NF2) [235], this trend is further exaggerated via the addition of 3 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.

The C2 cap of the NagD-like phosphatases is an α/β domain with a core 4-stranded parallel β -sheet, with an additional N-terminal anti-parallel strand. The parallel configuration of the sheet, combined with the lack of specific similarities to any other known domain, suggests that it might have possibly 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 3rd strand (Fig. 12). An examination of the sequence of the C2 caps of the histidinol phosphatase family reveals a conserved CxHx(6-13)Cx 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 (Fig. 12).

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 (Fig. 12). In an archaeal sub-family of the phosphoserine phosphatase, 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 (for e.g. [272], [273], [241], [236], [274], [246], [227]). 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+} ion 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 (Fig. 7A). As a result an acyl phosphate intermediate is formed with the carboxyl group of this aspartate [236], [237], [238], [239], [240]. Subsequently, the enzyme enters the open configuration again and allows the leaving group to escape (Fig. 7A). 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 [241]. 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 [275]. In the phosphonates and sugar phosphate mutases there are differences in the initial and the terminal stages of the reaction respectively

[236], [276], [277], [241], [247], [278], (Fig. 7A) 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 (Fig. 8A, 9). This differential winding in turn induces a movement in the flap immediately juxtaposed to the active site (Fig. 8A, 9) 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 [249], [237], [250].

The simplest structures add the cap to the flap motif itself, so as to completely seal the active site in the closed state (Fig. 13). 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 [276], [275], [241], [236], [279], [247]. 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 (Fig. 13). 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 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) (Fig. 13). 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 (Fig13). A similar strategy of occlusion via cooperation between two subunits is also seen in the aberrant BcbF family, which shows strand swapping

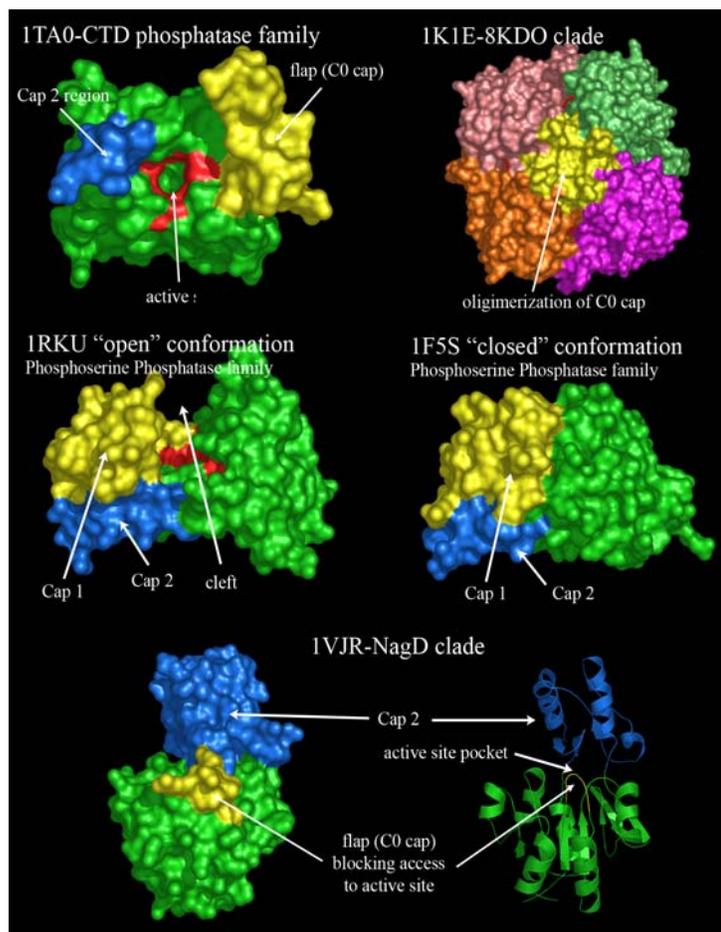


Fig 13. Interaction of cap modules with the active site in the HAD superfamily.

Molecular surface diagrams illustrate 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 dimerization 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. 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 and C2 cap regions are colored blue. The top right panel has cap domains colored yellow while core domains from each monomer are distinctly colored. Crystal structures are denoted by PDB identifiers followed by family names.

between adjacent subunits of the obligate dimer.

Natural classification of the HAD superfamily

Identification and clustering of the HAD superfamily enzymes

All available structures of the HAD superfamily were identified by using the DALI program [95] 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 material). I 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 (Fig. 10). Individual sequences from this alignment were used to initiate iterative PSI-BLAST searches [38] to identify all possible members of the HAD superfamily in the NR database (Materials and Methods and supplementary material). 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 sequences from the dehr family (gi: 691747, iteration 2, E-value: 9e-08), β -phosphoglucomutase family (gi: 1495997, iteration 2, E-value: 4e-04), phosphonate 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

material). Distinct clusters which fell out of this operation were aligned throughout their length and unique signatures beyond the 4 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 in the paper 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 Fig. 14 and the resultant natural classification is shown in Table 1 along with phyletic patterns, representatives in the PDB, and functional annotation while Fig. 15 depicts domain architectures observed within each family. The most basic split appears to separate a group of C0 cap proteins with a core 5-stranded sheet from the rest of the superfamily, which is unified by a 6-stranded core sheet. Within this 6-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 5 major radiations, namely the α -helical C1 cap assemblage, the P-type

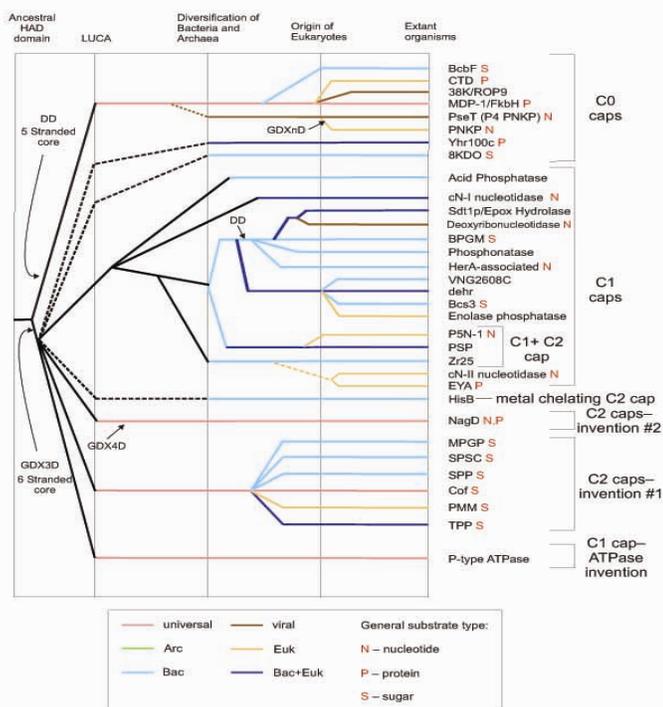


Fig 14. Reconstructed evolutionary scenario for the HAD superfamily.

The chart shows relative temporal eras that are demarcated by vertical black lines representing major evolutionary transitions. Individual HAD lineages are listed on the right side of the chart. Horizontal colored lines illustrate the maximum depth to which a HAD domain family can be currently traced relative to the temporal periods. Broken horizontal lines indicate that the lineage cannot be traced to a definite starting point. 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.

ATPases with their own C1 cap, and 3 distinct groups of dominant C2 cap proteins (Fig. 14). I

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 5-stranded core sheet and currently includes 5 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 (Fig. 14, Table 1).

The MDP-1/FkbH family

This family is prototyped by the eukaryotic MDP-1 type Mg(II)-dependent protein tyrosine phosphatases [243], [280], [243], which appears to be widely distributed in eukaryotes suggesting a basic cellular function. I also recovered a number of bacterial MDP-1-like proteins typified by

the 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* [281] and the bacterial symbiont *Candidatus Endobugula sertula* [282] respectively. The FkbH protein combines an N-terminal HAD domain with a C-terminal acetyltransferase domain (FkbH_Shy in Fig. 15), which contains 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 3 superkingdoms of life is indicative of their possible presence in the last universal common ancestor (LUCA) of cellular life (Fig. 14). 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 (Fig. 14, Table 1).

Table 1. Natural classification of HAD superfamily

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, phyletic distribution, and/or other distinct features are shown in parentheses underneath clade names or next to family/subfamily names. 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. '+' refers to a positive conserved residue (lysine or arginine) and a '-' refers to a negative conserved residue (aspartate, glutamate, or histidine).

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: 1U70

SS09580 subfamily (Several archaea) TWN motif II and DDR motif IV

RNA polymerase carboxyl terminal domain (CTD) phosphatase family

PerIp 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 (irdoviruses)

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: 1XPF

Polynucleotide kinase phosphatase (PNKP) family D in core helix 1, DxxK in core

helix 3, SGR motif II

Bacteriophage subfamily (bacteriophages)

PDB: 1LTQ

Eukaryotic 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, GGxGARRE motif at phosphatase C-term), tetramerizes

through flap strands

PDB: 1K1E

C. Yhr100e 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 B β -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

(Vertebrates, proteobacteria, *Thermosynechococcus*, *Arthrobacter*)

Tetra-helical C1 cap assemblage

C. Motif IV DD assemblage

Phosphotase family D β G motif I

Classic phosphotase 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-Epoxyde Hydrolase C-terminal domain family

SEHCT/Acad10 subfamily (animals, several α -proteobacteria)

PDB: 1S80, 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: 1Q08

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

dchr (dehalogenase-related) family

dchr subfamily I (Most bacteria, many archaea, fungi, animals, plants)

Isr subfamily (plants) EWE motif I, SNxxxE motif IV

dchr subfamily II (Most bacteria, with sporadic transfers to various eukaryotes and

archaea)

PDB: 1ZRN, 1Q05

Enolase-phosphatase family (Animals, Fungi, γ -proteobacteria, cyanobacteria, *Aquifex*,

Streptomyces) conserved FVxxxLTPY and DkxxxLxLQxxW regions in cap domain

Bes3 family (some proteobacteria and *Streptococcus*)

VNG268C family (cyanobacteria and some euryarchaea)

E. PSP-PSN-1 assemblage

PSN-1 (Pyrimidine S-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

Cic4 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 α + β cap with conserved K

Type I subfamily (Bacteria, Archaea, Eukaryotes) heavy metal, K⁺ transporting pumps

Type II subfamily (Bacteria, Eukaryotes) Ca²⁺, Na⁺/K⁺, H⁺/K⁺ transporters

PDB: 1SU4

Type III subfamily (Eukaryotes, Bacteria, Archaea) eukaryotic, archaeal proton pumps;

bacterial Mg²⁺ transporters

Type IV subfamily (Eukaryotes) aminophospholipid transporters

Type V subfamily (Eukaryotes) Ca²⁺ transporters

Type VI subfamily (Euryarchaea) soluble phosphatases

III. C2 CAP-CONTAINING HAD PROTEINS

C. HisB (Histidinol phosphatase) family

(Bacteria, only Thermoplasmatales amongst

archaea) metal-chelating C2 cap with conserved CxHxCxC region; histidine biosynthesis/ADP-D-8-D-heptose

synthesis/ADP-D- α -D-heptose synthesis

A. NagD family

GDXxxx motif IV, distinct α/β C2 cap

Aral subfamily (archaea, firmicutes, actinobacteria) conserved D in cap domain

PDB: 1VJR, 1YV9, 1WV1, 1YDF, 1YS9

chronophin (CTN) subfamily (Fungi, Animals, Slime molds, plants, kintoplastids)

conserved D in cap domain and glycine patch downstream of motif II; putative coflin-

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: 1XV1

Phosphomannosutase (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

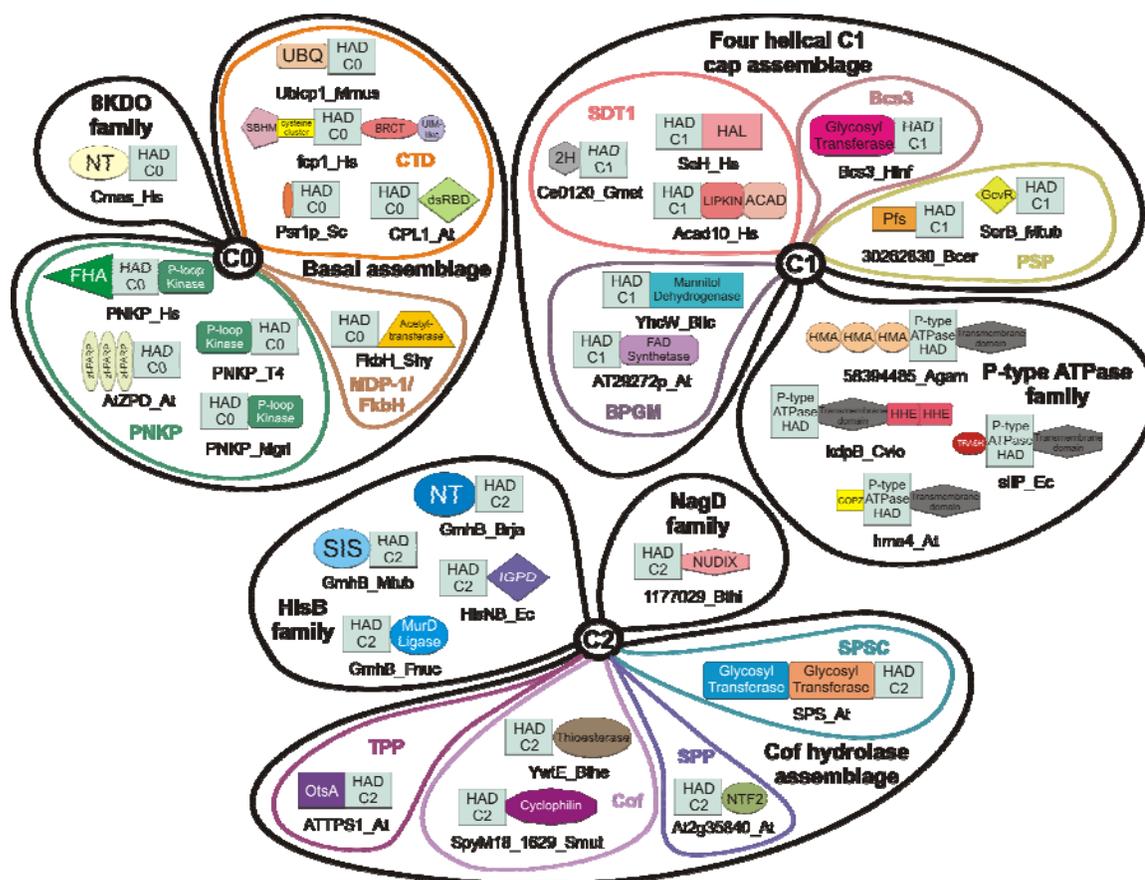


Fig. 15. Domain architectures of selected multidomain members of the HAD superfamily. The architectures are grouped around a central circle indicating principal cap type (as outlined in Table 1). Domain architectures are further grouped according to the higher order classification (Table 1), with clades 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 2 fold; 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.

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 [283], [284], [285], [286], [287], [288], [274], a feature essential for the reinitiation of transcription by the RNA polymerase [289]. This family has diversified into 7 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 [290], [291], and a conserved cysteine (Psr1p_Sc in Fig. 15). 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 [292]. The Tim50 subfamily, is also a membrane protein with a peculiar N-terminal membrane-spanning segment [276]. It associates with the mitochondrial inner membrane and regulates the translocation of internal mitochondrial proteins. Recently, the Tim50a isoform have 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 [293].

The remaining CTD phosphatase subfamilies are soluble proteins and include the Fcp1p-CPL subfamily, typified by the eponymous protein from *S. cerevisiae* [284]. 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 Fig. 15). The BRCT domain in these proteins has been implicated in recognizing the phosphorylated RNA-polymerase II substrate [294]. 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 [295], [296] 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 Fig. 15). 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 Fig. 15) [297] and might regulate RNA polymerase stability through the ubiquitin pathway [298].

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 [299], 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 [300], suggesting that it might act as phosphatase on one of building blocks of the polysaccharide.

Polynucleotide kinase phosphatase (PNKP) family

The next major lineage of basal C0 cap HADs is the PNKP family which play a role in both RNA and DNA repair [301] by removing 3'-terminal phosphate groups[302]. There are two subfamilies of these proteins (Table 1) with distinct motif IV sequence signatures (Fig. 14), the first being the bacteriophage subfamily (PseT in Fig. 14) with an N-terminal P-loop polynucleotide kinase domain (PNKP_T4 in Fig. 15). The second is the eukaryotic subfamily (PNKP in Fig. 14), which is seen in most major eukaryotic lineages and often contains a C-terminal polynucleotide kinase domain (PNKP_Mgri in Fig. 15). In plants, the phosphatase is fused to a Zn-finger found in Poly(ADP-ribose) polymerases and DNA-Ligases (AtZPD_At in Fig. 15) [303], [304]. Another previously uncharacterized in the eukaryotic subfamily is found in animals and is fused to the phosphopeptide-binding forkhead-associated domain (FHA) (PNKP_Hs in Fig. 15). 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 (Fig. 14). These enzyme removes 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 [305], [306] and bacterial capsular polysaccharides [307]. 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 nucleotidyltransferase transferase that potentially catalyzes the subsequent step in the biosynthesis pathways (Cmas_Hs in Fig. 15).

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 septins implying that these forms 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 (Fig. 14). The **acid phosphatase family** is characterized by an N-terminal signal peptide, which suggests that they are secreted proteins which 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 [308], [309], [310]. The **cN-I family** is a family of cytosolic 5'-nucleotidases found in vertebrates and several proteobacteria which regulate pyrimidine pools in the cytosol [311], [312].

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 **Phosphonate, SDT1-Epoxyde Hydrolase C-terminal domain, Deoxyribonucleotidase, HerA-associated (HA) and β -phosphoglucomutase (BPGM) families.**

The phosphonate family includes the phosphonoacetaldehyde phosphatases, which hydrolyze phosphonoacetaldehyde to orthophosphate and acetaldehyde [313], [314], [315], [316], [317], [318]. Experimental results suggest a role for cap residues in the catalytic activity of the classic phosphonates of this family [236]. 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 taken up a secondary binding function. **The Sdt1p-Epoxyde Hydrolase C-terminal domain family** is widely represented in both bacteria and eukaryotes and appears to have diversified into 4 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 Fig. 15). The animal enzyme has been shown to have hydroxyl lipid phosphate phosphatase activity in lipid degradation [319], [320], [321]. Some

animal members of this subfamily, like Acad10, are fused to two C-terminal domains (Acad10_Hs in Fig. 15); 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 [322] while its paralog Sdt1p has been shown to be a pyrimidine 5'-nucleotidase[323].

The Deoxyribonucleotidase family includes one of the major types of 5' (3')-deoxyribonucleotidases responsible for dephosphorylating uracil and thymine deoxyribonucleotides[324], [325], [326]. 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 the other similarities between their DNA replication processes [327], [177] 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 [158]. The contextual association 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 [158]. **The β -phosphoglucomutase (BPGM) family** is a large group which 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 α -D-glucose 1-phosphate and D-glucose 6-phosphate [328]. 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 [329]. In the related CbbY subfamily (typified by *Rhodobacter*

CbbY [330]; 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 [331], [332] and other fungal members of this subfamily have been characterized as glycerol 3-phosphatases[333]. 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 Fig. 15), which adenylates FMN to form FAD [334], [335]. 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 Fig. 15), 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 [250], [336], 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[337], [338]. 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 [339], [340]. Members of the restricted bacterial **Bcs3 family** are fused to an N-terminal glycosyltransferase domain (Bcs3_Hinf in Fig. 15) and might function as sugar phosphatases in the biosynthesis of capsular polysaccharides in certain pathogenic bacteria [341] (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 [342], [343]. 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 (Fig. 14).

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 [344], [345]. I also 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 Fig. 15). Given the connection between serine catabolism and glycine metabolism [346], [347], [348], 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 [349], [350]. The PHOSPHO1 subfamily contains a peculiar C2 cap which has 3 conserved cysteines suggesting that it is stabilized by metal chelation. The vertebrate versions of this subfamily are believed to mobilize inorganic phosphate for skeletal matrix mineralization through their action on phosphocholine and phosphoethanolamine [351], [352]. The fusion of this subfamily in some Gram positive bacteria to a nucleoside phosphorylase involved in methionine metabolism [353] 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 cytosolic 5'-nucleotidases [354], [355] 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 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 [356], with EYA itself and RNA polymerase II CTD repeats as its targets [307], [357]. 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 [245]. All except one subfamily of these proteins are fused to

membrane spanning regions and additional potential metal-ion binding domains (Fig. 15). As the P-type ATPase clade has previously been subjected to extensive phylogenetic analysis [358], [359], [360], I only briefly summarize the relationship 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 [358], [359], 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^+ [358], [359]. The type III subfamily includes eukaryotic and archaeal proton pumps and bacterial Mg^{2+} transporters [359]. Type IV ATPases are aminophospholipid transporters [359] and type V ATPases were recently characterized as eukaryotic Ca^{2+} transporters [361]. 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 experimentally studied [362] and proposed to be a phosphatase [363]. I 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 (Fig. 12, Table 1) and of these the HisB family shows the simplest version of a C2 cap. These caps contain a $CxHx_nCx_C$ 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 [364] (HisNB_Ec in Fig. 15). Other bacterial members of the HisB family, the GmhB proteins, catalyze the formation of D- α -D-heptose 1-P from an initial D- α -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 [365]. 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, Fig. 15).

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 (Fig. 14, 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* [366], although it is not required for the production of NAG [367]. 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 Fig. 15) which hydrolyzes a variety of substrates with a nucleoside diphosphate linked to another moiety [368], [369] 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 [370]. The Cut-1 subfamily (after the Cut-1 protein from *Neurospora*), is encoded in a predicted operon in α -proteobacteria with the bi-functional riboflavin kinase/FAD synthetase protein (RibF), and an adenylyltransferase that catalyzes the formation of FAD, and might function in co-factor biosynthesis. Except for the phosphohistidine/phospholysine phosphatase [371], [372] 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 (Fig. 12). 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 *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* [271], 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 synthesizes trehalose from glucose-6-phosphate and UDP-glucose [373], [374] (Table 1, Fig. 15) The broad phyletic pattern suggests the TPP-dependent trehalose biosynthesis or assimilation is one of the most prevalent of the three known catalytic pathways for trehalose biosynthesis [375], [376]. The *Mannosyl-3-phosphoglycerate phosphatase (MPGP) family* is a small family comprised of proteins catalyzing the dephosphorylation of mannosyl-3-phosphoglycerate to mannosylglycerate [377] 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[377],[378],[379].

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 [380] (Table 1). The *Sucrose phosphate synthase C-terminal domain (SPSC) family* are 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 Fig. 15) [381], [382], [381], [383]. 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 [384], [385]. The SPP plant versions additionally have a highly conserved C-terminal domain (At2g35840_At in Fig. 15), which I show belongs to the NTF2 class of $\alpha+\beta$ domains [386]. 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 [387], and probably binds a small molecule effector to function as an allosteric regulatory site. I 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 [381].

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 Rossmannoid 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 (Fig. 9). 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 5 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 sub-types, 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 (Fig. 14, 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 (Fig. 14, 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 4 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, I observe that at least 4 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 [339], [340], [350]. 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 inter-conversion 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 3 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 tetrahedral 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 5 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) phosphonate (Fig 7A). These reactions show mechanistic similarity [329] 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 [250], [336]) or additions from the cap (phosphonate). These observations suggest that the intricate active site of the HAD superfamily, with contribution from 4 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 an 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 [384], [385]) or recruited for new functions (e.g. the chronophin subfamily [370]). However, there also appear to be a few genuine innovations in the eukaryotes such as PMM and EYA protein phosphatases [388], [356]. 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 [224]. 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.

Supplementary material

A collection of the tree files in the Newick format of all the HAD families discussed in the text, along with the corresponding alignments and other documents referenced in the text as supplementary material can be accessed through the following website:

http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=ShowDetailView&TermToSearch=16889794&ordinalpos=4&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_RVDocSum.

The X-Ray Crystallographic Structure and Activity Analysis of a *Pseudomonas* Specific Subfamily of the HAD Enzyme Superfamily Evidences a Novel Biochemical Function

(based on reference [389])

Introduction

Phylogenetic analysis dividing the HAD superfamily into 33 families was reported in the previous section [213]. Although diversification of chemical function and substrate recognition is coupled with cap domain variation, the division of the families is not. Function assignment to the HAD families has been undertaken by numerous laboratories. In this study, computational methods were employed to identify novel sequence motifs that support a unique function, and having a novel protein targeted X-ray crystallographic analysis was carried out to define its three-dimensional structure. The electrostatic and steric features of the active site were thus defined along with information derived from genomic analysis to identify potential substrates. In the present work, we applied this approach to function identification of the targeted protein PSPTO_2114 (accession Q884H9; gi: 28869317) from *Pseudomonas syringae* pv. *Tomato*. PSPTO_2114 is a member of the recently described PA2803 subfamily found in *Pseudomonads* [213]. The PA2803 subfamily belongs to a larger family of the HAD superfamily that includes the well characterized enzyme phosphonoacetaldehyde hydrolase (phosphonatase), and is thus named the “phosphonatase family”. A predominant structural feature of the PA2803 subfamily in general, and PSPTO_2114 in particular, that sets it apart from the other members of the phosphonatase family, is the absence of the cap module insert within its amino-acid sequence. Phosphonatase, a C1 HAD superfamily member, utilizes the cap domain to desolvate the catalytic site as well as to position a Lys residue to activate the phosphonoacetaldehyde via Schiff

site is devoid of the HAD superfamily core catalytic residues and the metal-ion cofactor which typify the phosphotransferase branch of the HAD superfamily, and an elongated surface of conserved residues has been acquired. We hypothesize that the elongated surface conserved within the PA2803 subfamily is used in protein-protein interaction. Although the identity of the putative protein partner is unknown, a conserved gene context provides insight into the biochemical process in which PA2803 subfamily members might participate.

Materials and Methods

I performed the initial computational identification of the target, as well as all contextual gene neighborhood analysis. Crystallization-related experimentation was performed by Dr. Ezra Peisach, through which I was able to learn many of the experimental techniques involved in crystallization procedures. Substrate activity screens were performed by Dr. Liangbing Wang, and all phases of the project were again monitored by Drs. Allen, Dunaway-Marino, and Aravind.

Except where indicated, all chemicals were obtained from Sigma-Aldrich. Primers, T4 DNA ligase, and restriction enzymes were from Invitrogen. *Pfu* polymerase and the pET23b vector kit were from Stratagene. Host cells were purchased from Stratagene. Genomic DNA from *Pseudomonas syringae* pv. *tomato* (ATCC BAA-871) was purchased from ATCC.

Target Selection

The PA2803 subfamily was identified through BLASTP searches [38] performed on the non-redundant (NR) database of protein sequences using known phosphonate sequences as queries. For example, a search initiated with the *Bacillus cereus* phosphonate (gi: 18254515), recovers a PA2803 member from *Pseudomonas syringae* (e-value=0.009; gi: 28852558). The corruption or lack of sequence motifs essential to phosphonate enzymatic activity in the *P.*

syringae sequence initially suggested that the hit may have been a false positive. To safeguard against this possibility reverse searches were performed with the *P. syringae* sequence, and the recovery of the *B. cereus* phosphonate sequence at a significant expectation value (e-value=2e⁻⁰⁴) supported the existence of a genuine evolutionary relationship between the sequences. To further probe this relationship, a multiple alignment of the PA2803 subfamily was constructed using the T-COFFEE program [60] and the secondary structure of the proteins in the subfamily was predicted by feeding the alignment into the JPRED2 program [89]. The JPRED2 output revealed an alternating α/β pattern perfectly congruous with the Rossmann-like fold found in other members of the HAD superfamily, confirming the presence of the HAD fold in the PA2803-like proteins.

Cloning, Expression, and Purification

The DNA encoding the gene PSPTO_2114 from *Pseudomonas syringae pv. tomat* was amplified by PCR using the genomic DNA from *Pseudomonas syringae pv. tomat* (ATCC BAA-871), and *Pfu* DNA polymerase. Oligonucleotide primers (5'-GCTCGCGCCGCGCCATATGCCTTTACCAACC) and (5'-AGTGA^TCTCAAGACGGATCCAGTAAGCTCGC) containing restriction endonuclease cleavage sites *Nde*I and *Bam*HI (underlined), were used in the PCR reaction.

The pET-23b vector, cut by restriction enzymes *Nde*I and *Bam*HI, was ligated with the PCR product that had been isolated and digested with the same restriction enzymes. The ligation product was used to transform *Escherichia coli* BL21(DE3) competent cells (Stratagene) and the plasmid DNA produced was purified using a Qiaprep Spin Miniprep Kit. The gene sequence was confirmed by DNA sequencing carried out by Center for Genetics in Medicine at the Biochemistry and Molecular Biology Department, University of New Mexico. Transformed cells

(10 L) were grown at 25 °C with agitation at 200 rpm in Luria broth containing 50 µg/mL ampicillin to an OD₆₀₀ of 0.6-1.0 and induced for 4 h at 25 °C at a final concentration of 0.4 mM isopropyl α-D-thiogalactopyranoside (IPTG). The cells were harvested by centrifugation (7,855 g for 15 min at 4 °C) to yield 3 g/L of culture medium. The cell pellet was suspended in 1g wet cell/10 mL of ice-cold buffer A consisting of 50 mM K⁺-Hepes (pH 7.0 at 25 °C), 5 mM Mg²⁺ and 1 mM DTT. The cell suspension was passed through a French press at 1,200 PSIG before centrifugation at 48,384 g for 30 min at 4°C. The supernatant was loaded onto a 40cm × 5cm DEAE Sepharose column, which was eluted with a 2L linear gradient of KCl (0 to 0.5 M) in buffer A. The column fractions were analyzed by SDS-PAGE.

The desired protein fractions from DEAE were combined, adjusted to 15% (NH₄)₂SO₄ (W/V), and loaded onto a 18 ×3 cm butyl-Sepharose column pre-equilibrated with buffer A containing 15% (NH₄)₂SO₄. The column was eluted with a 0.5 L linear gradient of (NH₄)₂SO₄ (15% to 0%) in buffer A. The desired protein fractions shown to be homogeneous by SDS-PAGE were combined, dialyzed against buffer A, and concentrated at 4 °C using a 10K Amicon Ultra Centrifugal filter (Millipore) then stored at -80 °C. The protein concentration was determined by the Bradford method [390] and from the protein absorbance at 280 nm ($\epsilon=76335\text{M}^{-1}\text{cm}^{-1}$). The yield of PSPTO_2114 was 10mg/g wet cell. Selenomethionione (SeMet) substituted PSPTO_2114 was overexpressed in a methionine auxotroph of *E. Coli* BL21(DE3), grown in a defined media containing 40mg/ml SeMet and purified using the protocol described for the native protein.

Molecular Mass Determination

The theoretical subunit molecular mass of recombinant PSPTO_2114 was calculated from the amino acid composition, derived from the gene sequence, using the EXPASY Molecular Biology Server program Compute pI/MW [391]. The subunit size was determined by SDS-PAGE

analysis with molecular weight standards from Invitrogen, and the subunit mass was determined by MS-ES mass spectrometry (Mass Spectrometry Lab, University of New Mexico). The molecular size of native recombinant PSPTO_2114 was determined using gravity-flow gel-filtration techniques. PSPTO_2114 was subjected to chromatography at 4 °C on a calibrated (Pharmacia Gel Filtration Calibration Kit) 1.5 x 180 cm Sephacryl S-200 column (Pharmacia) using buffer A at a flow rate of 1 mL/min, as eluant. The molecular weight was determined from the elution volume using a plot of log(molecular weight) of a standard protein *vs.* elution volume as reference.

Protein Crystallization

Crystals were obtained via the vapor-diffusion method with hanging drop geometry at 4°C using a precipitant solution consisting of 15-20% PEG 3350, 150-200 mM sodium formate, 100 mM Hepes (pH 7.3 at ambient temperature) and 5 mM MgCl₂. Freshly thawed protein was diluted to 12 mg/ml in 2 mM DTT and 5 mM MgCl₂, combined in equal volumes (1.5 µl) of protein and precipitant and suspended over 0.5 ml of precipitant. It was critical to the appearance of single crystals to form an elongated drop. Failure to do so resulted in too many nucleation events along the entire circumference of the drop. Crystals are temperature sensitive and grow as rods 0.1 x 0.1 x 0.5 mm in size at 4 °C. Single crystals of SeMet protein were transferred to Paratone-N at 4 °C and frozen in liquid nitrogen. X-ray diffraction data were collected at beamline X12C at the National Synchrotron Light Source (NSLS), Brookhaven National Laboratory on an ADSC Q210 detector and processed with the HKL2000 [392] program suite. The crystal structure was solved by the multi-wavelength anomalous dispersion (MAD) method using the Selenium edge (Table 1). The programs SOLVE and RESOLVE (version 2.11)[393] were used to locate the eight selenium sites, automatically trace > 85% of the protein backbone, and

correctly identify the two monomers in the asymmetric unit. Refinement was carried out in CNS [394] and model rebuilding in COOT [395]. The final model, refined to 1.9Å resolution ($R_{\text{work}}=0.17$, $R_{\text{free}}=0.22$) consists of a dimer of 392 amino acid residues, 536 water molecules, 2 Hepes molecules, and 1 Mg^{2+} (found on the surface at the dimer interface) (see Table 1 for final model statistics). Residues 1-4 of chain A and the first residue of chain B within the dimer were not visible and were excluded from the final model.

Substrate-Activity Screens

The rate of p-nitrophenyl phosphate (PNPP) hydrolysis was determined by monitoring the increase in absorbance at 410 nm ($\Delta\epsilon = 18.4 \text{ mM}^{-1}\text{cm}^{-1}$) at 25 °C or 37 °C. The 0.5 ml assay mixtures contained 20 μM PSPTO_2114, 50 mM Hepes (pH =7.0), 5 mM MgCl_2 , and 1 mM PNPP.

Hydrolyses of all other phosphate esters were monitored using the Biomol green kit (Biomol International LP) to detect total phosphate release determined by monitoring the increase in absorbance at 360 nm at 25°C. The 1 ml assay mixture initially contained 50 mM Tris-HCl buffer (pH = 7.5), 1 mM MgCl_2 , 1 mM substrate and 20 μM PSPTO_2114, 0.2mM MESG (2-amino-6-mercapto-7-methylpurine ribonucleoside) and 1U purine nucleoside phosphorylase. In parallel, the background level of phosphate release was measured using a control reaction, which excluded the PSPTO_2114.

Phosphonate and phosphoglucomutase activities were tested using the reaction conditions listed above in conjunction with the published assay methods [237, 278].

Results and Discussion.

Identification of a Lineage-Specific Phosphonate Variant.

The PA2803 subfamily of the HAD superfamily was identified through searches performed on the non-redundant (NR) database of protein sequences using known phosphonate sequences as

queries. Multiple-sequence alignment and secondary-structure predictions revealed the lack of the helical “cap” domain characteristic of the phosphonatase proteins. These searches also revealed that the subfamily was restricted to various *Pseudomonas* strains (represented in Fig. 17), suggesting that the protein arose through a lineage-specific duplication of the phosphonatase protein; a scenario supported by the dual presence of phosphonatase and the variant with no cap in most *Pseudomonas* strains. In fact, of the strains with completely sequenced genomes carrying a PA2803-like protein, only *syringae pv. tomato* str. DC3000 and *syringae pv. syringae* B728a lack phosphonatase,

Datasets	Remote	Peak	Edge
Wavelength (Å)	0.9500	0.9791	0.9798
Space group		P2 ₁ 2 ₁ 2 ₁	
Unit-cell dimensions (Å, °)	$a = 48.61, b = 74.43, c = 106.35, \alpha = \beta = \gamma = 90$		
Resolution range (outer shell) (Å)		50-1.9 (1.97-1.9)	
Total Reflections	683234	654397	653527
Unique Reflections	31224	30874	30826
I/σ(I) (outer shell)	43.1 (8.1)	41.2 (7.3)	39.6 (6.9)
R _{sym} (outer shell) (%) ^a	5.9 (18.0)	6.5 (18.3)	5.7 (18.6)
Completeness (outer shell) (%)	99.0 (90.4)	98.0 (83.8)	98.2 (84.9)
Figure of merit (RESOLVE 2.1Å)		0.85	
f'/f'' (e)	-2.4/4.1	-5.8/5.5	-8.5/3.0
R _{work} (%) / R _{free} (%) ^b		0.17/0.22	
Monomers per asymmetric unit		2	
Atoms per asymmetric unit		3445	
Average B (Å ²)			
Protein		12.5	
Water		20.8	
Hepes and Mg ²⁺		15.1	
RMSD bonds(Å)/angles(°)		0.013/1.423	
Dihedrals (°)/improper(°)		0.005	
PDB ID		2ODA	

Table 2. Summary of Data Collection and Refinement Statistics

^a $R_{\text{sym}} = \sum_i |I_i - \langle I \rangle| / \sum \langle I \rangle$, where I_i is an individual intensity measurement and $\langle I \rangle$ is the average intensity for all measurements of the reflection i . ^b $R_{\text{work}}/R_{\text{free}} = \sum ||F_o| - |F_c|| / \sum |F_o|$, where F_o and F_c are the observed and calculated structure factors, respectively. R_{free} is calculated for 10% of reflections randomly and excluded from refinement.

indicating a lineage-specific loss of the phosphonate gene may have occurred in *syringae*. Cap-domain loss is consistent with the lineage-specific duplication scenario, as duplication is known to correspond with dramatic protein structural alterations [40]. Sequence motifs 1-4, which comprise the catalytic framework of not just the phosphonate family, but the entire HAD superfamily are not evident in the PA2803 subfamily (Fig. 17). These findings were verified by the X-ray crystal structure of the PA2803 subfamily member *P. syringae* PSPTO_2114 (*vide infra*).

Structure Determination of Recombinant P. syringae PSPTO_2114

Purification

Recombinant *P. syringae* PSPTO_2114 was purified to homogeneity using a column chromatography based protocol in an overall yield of 10 mg/g wet cells. The theoretical molecular weight of PSPTO_2114 minus the N-terminal Met as calculated from the amino-acid sequence is 20,704 Da. This value agreed with the experimental molecular weight of 20,702 Da measured by mass spectrometry, thus indicating that the N-terminal Met is removed by posttranslational modification. The subunit size of PSPTO_2114 estimated by SDS-PAGE analysis is ~21 kDa, whereas the native protein size determined by gel-filtration chromatography is ~47 kDa. These results are consistent with a homodimeric quaternary structure, which is also observed for the crystalline enzyme by X-ray crystallography.

Overall fold

The structure of PSPTO_2114 was determined to 1.9 Å using the MAD phasing method (Table 2). The native structure of this protein reveals an elliptical dimeric protein with overall dimensions of 44 Å x 45 Å x 83 Å (Fig. 18A). The monomer is a modified Rossmann fold consisting of a six-stranded parallel β-sheet surrounded by seven α-helices, characteristic of the core domain of HAD superfamily members. As predicted, the “cap” domain present in the related protein

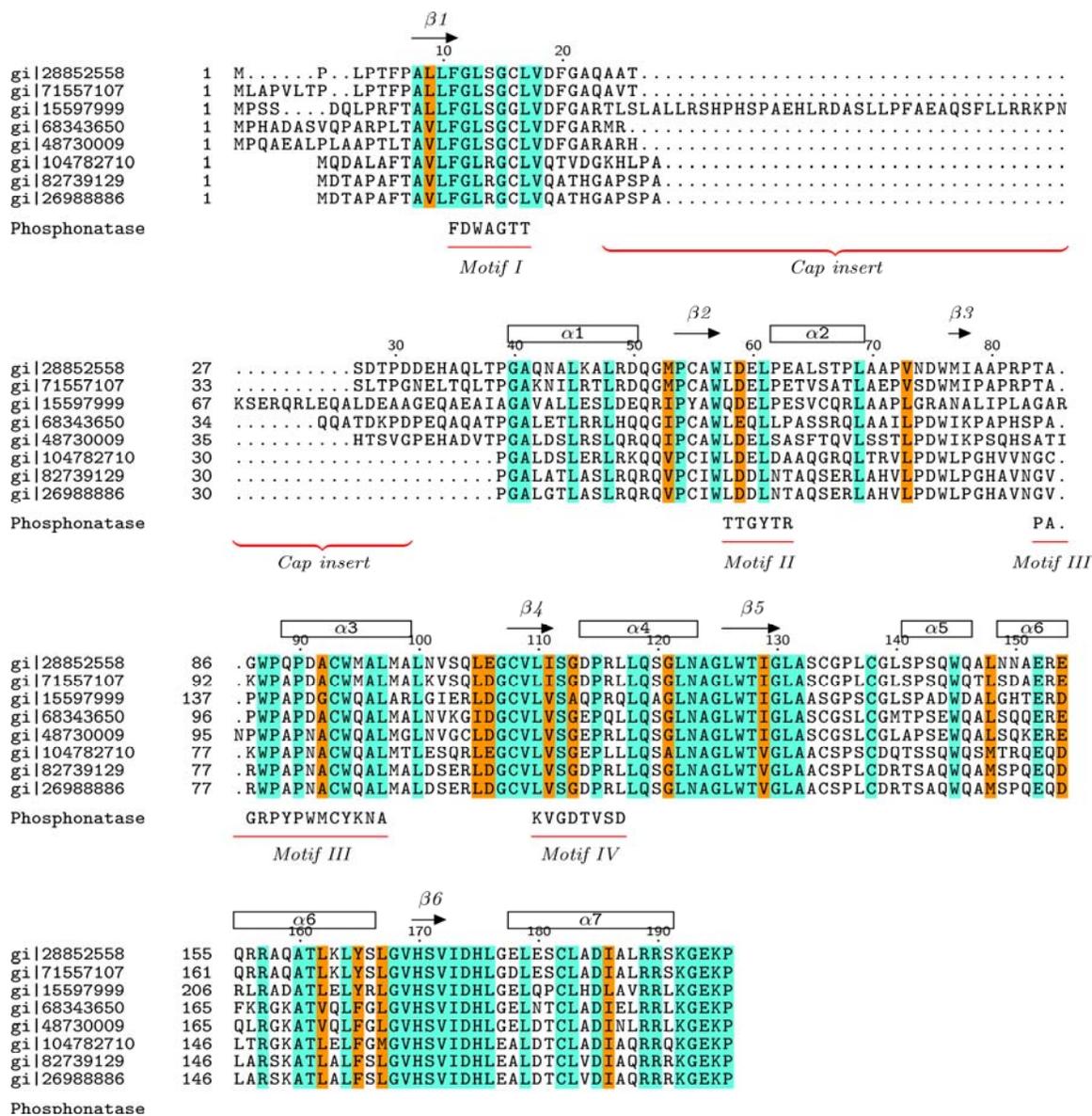


Fig. 17. Sequence alignment of members of the PA2803 subfamily.

Secondary structure assignments (noted along the top) are from the current structure: gi28852558 (PSPTO_2114) *P. syringae* pv. tomato str. DC3000; gi71557107 *P. syringae* pv. phaseolicola 1448A; gi15597999 PA2803 from *P. aeruginosa* PAO1; gi68343650 *P. fluorescens* Pf-5; gi48730009 *P. fluorescens* PfO-1; gi104782710 PSEEN3700 from *P. entomophila* L48; gi82739129 *P. putida* F1; gi26988886 *P. putida* KT2440. Sequence Motifs 1-4 used to identify HAD superfamily members are underlined in red and labeled, and the corresponding motifs from *B. cereus* phosphonatase included for comparison (as well as the location of the insertion point of the phosphonatase cap (note that in phosphonatase the cap insert would be >60 amino acids and length this any insert in this region in the PA2803 subfamily is too small to form a typical cap).

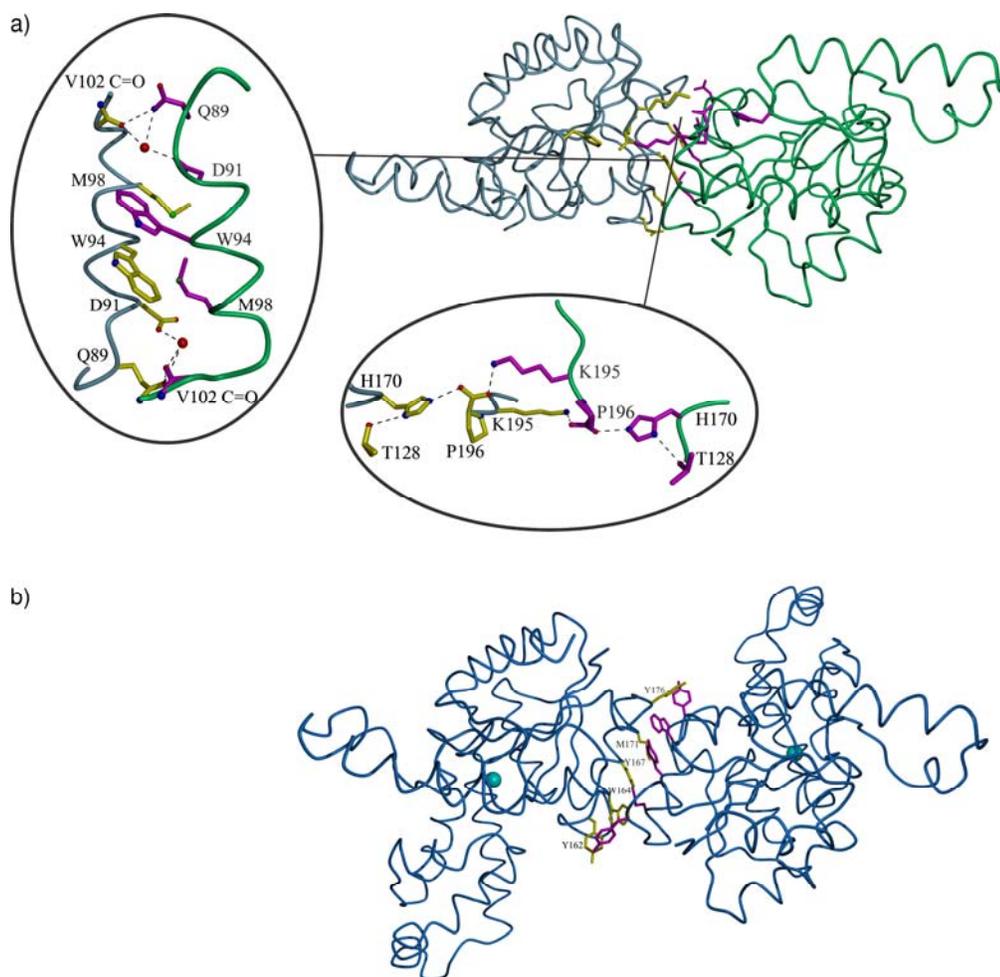


Fig. 18. Structures of PSPTO_2114 and phosphonate.

A. Structure of dimeric PSPTO_2114 with the residues lining the dimer interface in yellow (monomer A) and magenta (monomer B). The magnified regions show the dimer interface and the lys-lys stacking interactions making up the oligomeric contacts. B. Structure of dimeric *B. cereus* phosphonate shown in the same relative orientation as PSPTO_2114 in panel A with the residues lining the dimer interface in yellow (monomer A) and magenta (monomer B).

phosphonate (Fig. 18B) is absent. A multiple sequence alignment of the PA2803 subfamily shows that in some members additional sequence exists at the site of the phosphonate cap module insert (Fig. 17), however there is not enough sequence to form an entire cap domain. It is possible that this insertion represents the vestiges of the phosphonate cap domain.

A DALI [396] search for structural homologues using the PSPTO_2114 coordinates identifies *Bacillus cereus* phosphonate (class C1) as the closest match ($Z=21.3$) followed by *Lactococcus lactis* β -phosphoglucomutase ($Z=14.0$) (class C1). The RMSD between a single monomer of phosphonate (pdb accession 1fez) and PSPTO_2114 is 1.62 Å for 166 C α atoms.

Dimer Interface

Examination of the PSPTO_2114 subunit interface reveals an arrangement similar to that of phosphonate, with symmetrical packing of the two α helices and favorable amino-acid side chain interaction across the interface (Fig. 18A and 18B). The buried surface area in PSPTO_2114 is 5085 Å² compared to that of phosphonate at 6827 Å², confirming the similarity in the size of the dimer interface. The dimerization mode common to phosphonate and PSPTO_2114 has not been observed in other HAD superfamily members and thus, appears to be a specific trait of the phosphonate family.

Examination of the interacting residues at the respective phosphonate and PSPTO_2114 dimer interfaces, reveals both sites of conservation and sites of divergence. In *B. cereus* phosphonate conserved residues Met171, Trp164, Tyr162, Tyr167 and Tyr176 pair across the interface, and much of the driving force for subunit association is based on desolvation of the hydrophobic side chains. The phosphonate Met171 and Tyr167 positions correspond to Met98 and Trp94 in PSPTO_2114. The residues are conserved within the PA2083 subfamily, and are thus a shared dimerization motif across the phosphonate family. In contrast, the *B. cereus* phosphonate Trp164 is replaced by Asp91 in PSPTO_2114, and water molecule serves to bridge the two Asp91 residues across the dimer interface. Thus, while the amino-acid position for interaction is retained, the nature of the interaction has been altered. The Tyr162 of phosphonate is replaced with Gln89 in PSPTO_2114 (Fig. 18A) and with Ala in other PA2083

subfamily members. Although the Gln89 in PSPTO_2114 appears to interact with Asn101 (3.5 Å) and Val102 of the opposing subunit, the interaction distances are large (3.5 and 3.3 Å). It is unlikely that this particular position is important to dimerization in the PA2083 subfamily, and thus it constitutes divergence. The *B. cereus* phosphonate Tyr176 does not have a spatial counterpart in PSPTO_2114, also indicative of loss of utility and divergence.

In addition to these dimer contacts that are homologous to those of phosphonate, the PSPTO_2114 dimer is stabilized by contributions from Lys195 (inset, Fig. 18A). The hydrocarbon side-chains of the Lys195-Lys195 pair are stacked and desolvated while each N⁺-ammonium group engages in hydrogen-bond formation with the carbonyl oxygen of Pro196 on the opposing protomer. The conservation of the C-terminus sequence motif GEKP among PA2083 subfamily members suggests that this hydrogen-bond network serves a common function. The glycine caps the C-terminal helix whereas the Lys and Pro extend across the dimer interface. The C-terminal Pro196 is positioned in each protomer by hydrogen-bond interactions with Ne2 of His170 which in turn is oriented through a hydrogen bond from ND1 to the carboxylate side chain of Thr128 (Fig. 18A). Both His170 and Thr128 are absolutely conserved among the PA2083 subfamily. Among the *Pseudomonas* phosphonate sequences, the C-terminal sequence is GEMPXXX. This motif is equivalent to the PSPTO_2114 sequence GEKP, however the bridging Lys-Pro interaction is a unique feature of the PA2083 subfamily, and absent in phosphonate.

HAD superfamily Active Site

The active site of HAD superfamily members is comprised of residues from four conserved motifs (1-4) located within four loops of the Rossmannoid fold core domain. These

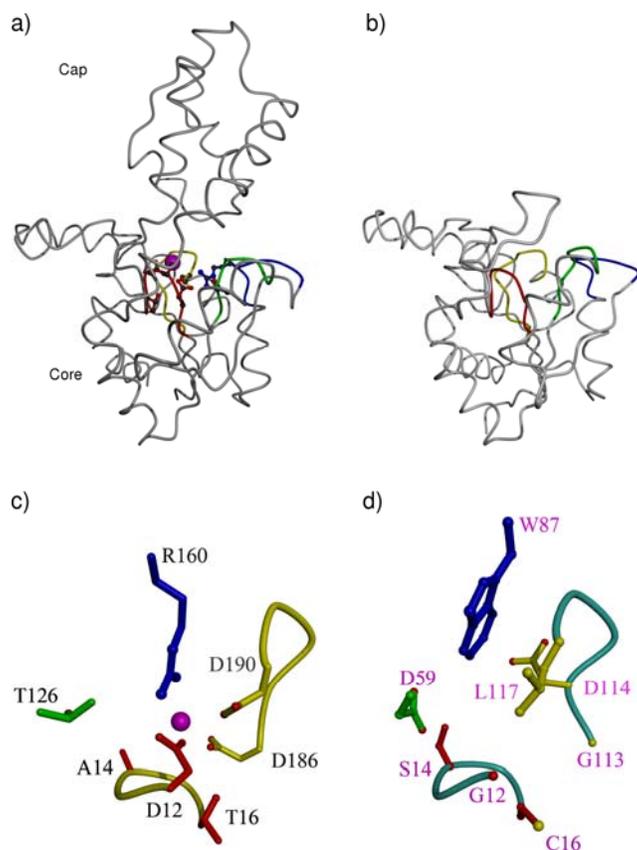


Fig. 19: Comparison of phosphonatae and PSPTO_2114.

A. Backbone trace of one subunit of phosphonatae (pdb accession code 1fez) and B. PSPTO_2114. The active site in phosphonatae is indicated by the cofactor Mg^{2+} (magenta sphere) with motifs 1-4 colored red, green, blue and yellow, respectively (as in Fig. 16). The corresponding structure of PSPTO_2114 using the same color scheme. There is no domain in PSPTO2114 corresponding to the mobile “cap” domain of phosphonatae. C. Close up of the active site of phosphonatae with coloring as in panel A (yellow backbone) and D. the corresponding structure of PSPTO_2114 (cyan backbone).

include the essential nucleophilic Asp located on loop I, a conserved Ser/Thr phosphate-binding residues on loop II, a conserved Lys/Arg phosphate-binding residue in loop III, and the Mg^{+2} cofactor Asp/Glu ligands located on loop IV (Fig. 16). These are highlighted in the active-site structure of phosphonatae (Fig. 19A and C). A comparison of the active sites of phosphonatae and PSPTO_2114 shows remarkable differences (Figures 19C-D). In PSPTO_2114, the nucleophile Asp 12 of phosphonatae has been replaced by the stringently conserved Gly12. Without the nucleophilic Asp, it is unlikely that PSPTO_2114 retains the phosphoryl-transfer activity of the HAD superfamily. This conclusion is supported by the activity screens for phosphonatae, phosphatase, and phosphomutase transfer activity wherein no activity above background was observed for any substrate tested. The panel of substrates included phosphonoacetaldehyde and

glucose 1-phosphate, the substrates for the closely related HAD superfamily members (*vide supra*) phosphonatase and β -phosphoglucomutase, and substrates of known HAD superfamily phosphatases viz. phosphoserine, phosphoglycolate, fructose 6-phosphate, N-acetyl glucoseamine 6-phosphate, and *p*-nitrophenyl phosphate.

Examination of the other catalytic HAD motifs reveals additional replacements of catalytic and cofactor-binding residues. Specifically, the hydrogen-bond donor of loop II, T126 of phosphonatase, is replaced by Asp59, which could assume the same phosphate binding role only if it were protonated. Loop III, Arg160 in phosphonatase, that provides electrostatic stabilization of the phosphoryl group and orients the Asp nucleophile, has been replaced with Trp87. The loop IV of phosphonatase stations two Asp residues for Mg^{+2} binding, consistent with the minimal requirement of two carboxylate residues for the HAD superfamily phosphotransferases [397]. PSPTO_2114 stations only one such residue, Asp114, and in this manner is similar to the HAD superfamily dehalogenases, which do not possess a metal-cofactor binding site [397]. Indeed, PSPTO_2114 crystallized from a solution containing 5 mM Mg^{2+} , yet the X-ray crystal structure evidences no Mg^{2+} bound to the active site.

The lack of catalytic residues and residues that form the metal-cofactor binding site, lead one to the conclusion that PSPTO_2114 does not catalyze phosphoryl transfer. The lack of the nucleophilic Asp also nullifies the possibility that PSPTO_2114 functions in carbon group transfer as exemplified by the dehalogenases of the HAD superfamily. In fact, examination of the conservation of the residues overlapping the canonical phosphatase active site (Fig. 17) shows that only Trp89 and Gly12 are absolutely conserved among PSPTO_2114 orthologs from different *Pseudomonads*. The lack of amino-acid conservation argues against use of this pocket as an enzymatic active site. A search carried out with ProFunc [398, 399] failed to identify a plausible

catalytic site in PSPTO_2114, and, notably, the vestigial phosphonate/HAD superfamily active site was not found.

Conservation of Binding Surfaces

If PSPTO_2114 is not an enzyme, what function does it provide to the *Pseudomonas*? Mean distance tests performed using the MEGA program [97] established the rate of sequence divergence within the PA2803 subfamily as significantly higher than the set of all classical phosphonate sequences found in *Pseudomonas*, a finding consistent with other catalytically inactive proteins which have assumed a secondary binding function [400]. Evolution of a binding function for small molecule ligands can easily be envisaged if the substrate-binding residues are separate in identity from catalytic residues. Selective replacement of the catalytic residues would produce a protein that could selectively bind but not transform the physiological substrate. Alternatively, if the protein scaffold is suited for presenting an extensive binding surface, the protein may be recruited to bind to a macromolecule (either protein or DNA/RNA). The patterns of conservation of residues in the PA2803 subfamily should allow the identification of binding surfaces critical to function. PSPTO_2114 is conserved in at least 8 different genomes in *Pseudomonas* with 50-92% identity between sequences (Fig. 17). By mapping the stringently conserved residues onto the PSPTO_2114 three-dimensional structure (Fig. 20) we identified one contiguous region of the protein dimer surface that is conserved while all other surface regions are not conserved. Calculation of the electrostatic potential surface shows a normal distribution of charges in the conserved patch. The total surface area of the conserved patch is 2432 Å² which is considerable larger than the 800 Å² per recognition patch buried in a single patch protein-

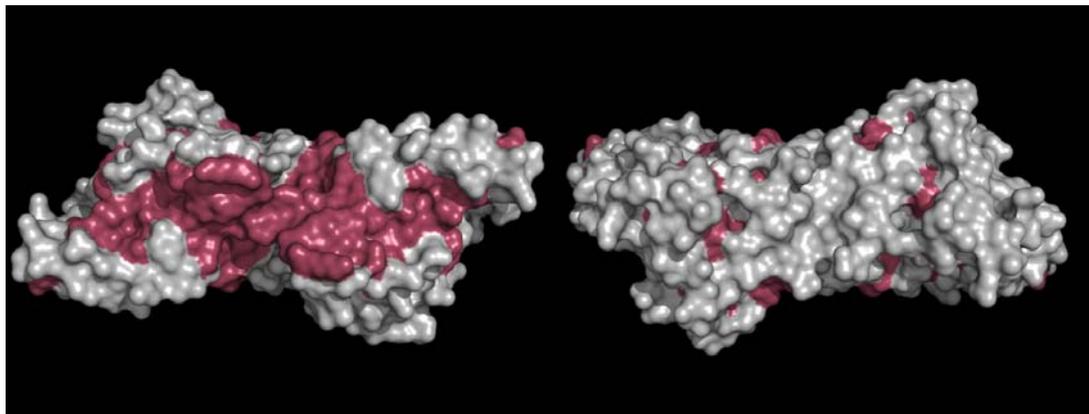


Fig. 20. Surface representation of PSPTO_2114.

Absolutely conserved residues identified in Fig. 17 are mapped onto the structure in dark pink. The top and bottom views are rotated 180° along the x axis.

protein interface [401]. Thus, it may be that the entire conserved patch is not involved in forming an interface or there may be two independent binding sites of half the area.

Because all 8 protein sequences in the multiple-sequence alignment of PSPTO_2114 derive from *Pseudomonads*, the validity of the analysis was tested by repeating the mapping exercise with an alignment of the four phosphonatase sequences derived from *Pseudomonads* and the X-ray crystal structure of the *Bacillus cereus* phosphonatase (PDB ID 1sww). For comparison, a member of a different fold family, serralyisin (an alkaline protease), from 6 *Pseudomonads* was similarly mapped onto the structure of the *P. aeruginosa* serralyisin (PDB ID 1akl). It is expected that the evolutionary drift among the *Pseudomonad* orthologs is similar and that the protein surface residues that do not perform specialized roles in function, folding or stability will diverge accordingly. The pattern of residue conservation on the surfaces of phosphonatase and serralyisin is one of well-dispersed small clusters (typical surface area is 1,200-2,000 Å²) in stark contrast to the pattern of residue conservation on the PSPTO_2114 surface. Thus, we conclude that the large

conserved surface region of the PSPTO_2114 orthologs identifies a site of function, as there is no apparent structural role that these residues might fill.

Functional Assessment Derived from Gene Context

As revealed by computational analyses, it appears that the PA2803 subfamily has undergone a secondary lineage-specific loss of the C1 cap domain. The former cap region as well as the vestigial active site is on the side of PSPTO_2114 dimer that is not conserved. This result can be contrasted to the results for the HAD member MDP-1, which has been identified as a protein sugar phosphatase[402], where the conserved surface encircles the active site, acting as docking surface for the protein substrate [243]. Because the electrostatic potential of the conserved surface is not positive, there is no evidence that the surface would be involved in DNA or RNA binding which usually takes place through conserved Arg/Lys residues [403]. Together, the sequence and structural evidence point to a non-catalytic role for PSPTO_2114 that involves the binding of a polar macromolecule.

In an attempt to further elucidate functional information for this subfamily of proteins we surveyed their gene neighborhoods. The products of genes co-occurring in the same neighborhood in multiple, sufficiently evolutionary diverse genomes tend to interact physically and functionally [6, 8, 404]. The PA2803 subfamily associates with several neighboring genes conserved across the *Pseudomonads* that encode proteins relating to the lipoprotein release pathway and the anti-anti sigma-stimulating stress response pathway (Fig. 21). However, given the close evolutionary distance of the genomes being investigated and the observation that many of these genes are transcribed in the opposite direction relative to the PSPTO_2114-like proteins, it seems unlikely these proteins form meaningful interactions. Nevertheless, a small, uncharacterized gene typically 90 residues in length tightly coupled immediately downstream

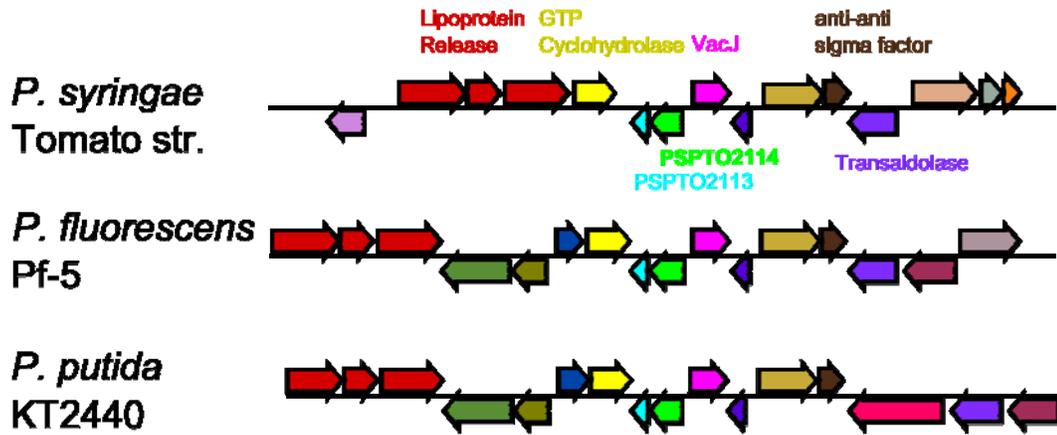


Fig. 21: Gene context of PSPTO_2114 in *Pseudomonads*.

Individual genes with similar function are identified.

(within ~100 base pairs and transcribed in the same direction) of the PSPTO_2114-like proteins was identified (PSPTO_2113). These two genes appear to form a small “gene island” likely comprising a co-regulated unit which in some cases is combined with a third member, a gene coding for a protein with an HD phosphohydrolase fold. Secondary structure predictions of the protein product of the previously uncharacterized PSPTO_2113 revealed a tetra- α -helical fold which stations a stringently conserved histidine residue between the third and fourth helices, reminiscent of certain heme-binding proteins of the HNOB domain [405]. This newly characterized protein is found in other bacteria as a solo domain; suggesting that instead of acting as a sensor for the activation of another domain (as is the case of the heme-binding HNOB domains which sense nitric oxide for soluble guanyl cyclases), these proteins may be involved in the simple transporting/transferring of heme or other prosthetic groups. The PA2803 subfamily of proteins could potentially bind to proteins or ligands, which would interact with and possibly receive or transfer a heme protein to/from this tetra-helical family of proteins. However, this

putative protein-protein binding function cannot currently be simulated because of the absence of a homologous structure on which to base a model of the tetra-helical PSPTO_2113. Therefore, creation of a docked complex of this protein with PSPTO_2114 is impossible. The nature of the hypothesized binding function must await future biochemical characterization of these proteins.

Evolutionary History of the E1-Like Fold and Architectural Themes Contributing to the Catalytic Roles of E1-Like Domains

Introduction

The ubiquitin protein modification cascade is initiated by the adenylation of the carboxy-terminal glycine residue of ubiquitin (Ub) or ubiquitin-like (Ubl) molecules by an E1 enzyme [406, 407]. Remarkably, a similar reaction occurs during the biosynthesis of thiamine and molybdenum/tungsten cofactor (MoCo/WCo), where a homolog of the E1 enzyme, ThiF or MoeB, adenylates the C-terminus of an ubiquitin-like ThiS/MoaD protein [408-413]. Upon modification, the trajectories of the ubiquitin-like proteins are very different in the two pathways. In the ubiquitin modification system, Ub /Ubl is covalently conjugated to protein substrates by a trimeric complex containing the E1, E2 and E3 enzymes. In contrast, in the cofactor biosynthesis pathways, the C-terminus of the ThiS or MoaD protein is further modified to a thiocarboxylate that serves as a sulfur donor during the biosynthesis of the cofactors. The evolutionary origins and diversification of the E1 superfamily are not completely understood. The thiamine and Mo/WCo biosynthesis pathways are present in both bacteria and archaea suggesting that both these pathways were present in the last universal common ancestor (LUCA) of life. Although their ubiquitin-like component, ThiS and MoaD, differentiated in LUCA, it is unclear if this is also reflected in their E1 proteins, ThiF and MoeB. E1-like enzymes are also detected in other pathways involved in sulfur metabolism or sulfur incorporation across a wide phyletic range of bacteria. Recently, our lab showed that the eukaryotic ubiquitin signaling system had its antecedents in the bacteria. In this study, the first prokaryotic homologs and contextual association of E1-like proteins with other core components of the eukaryotic signaling pathway

such as the E2 enzyme, ubiquitin and the JAB peptidase-like proteins was reported [414]. Although ubiquitin-like proteins are the most common and well recognized substrates of E1-like enzymes, there is a single example of a non-ubiquitin substrate in the microcin biosynthesis pathway. In this system, the microcin C7 polypeptide precursor is adenylated at its C-terminus by an E1 protein [415, 416]. It is unclear if the use of modified substrates is a rare occurrence or more common and poorly studied. Another conserved reaction performed by the E1 enzymes observed in the thiamine biosynthesis and ubiquitin signaling pathway is the transfer of the Ub-like substrate to an internal cysteine. Although this internal cysteine is homologous between the two systems, the adenylated Ub is linked to the E1 through a thioester linkage in ubiquitin signaling systems, whereas in the thiamine biosynthesis pathway, the ThiS protein is linked through a persulfide linkage.

Crystal structures of MoeB, ThiF and the E1-like enzymes revealed a shared core of a derived Rossmann-like fold that is responsible for both the principal reactions of adenylation and thiolation on the internal cysteine. The E1-like enzymes function as homo or hetero-dimers and the catalytic active site residues that perform these reactions are highly conserved between the ubiquitin-transferring and the ThiS/MoaD transferring E1-like enzymes. In addition to adenylation and thiolation, E1 enzymes may catalyze other enzymatic activities depending on their functional contexts and associations. In the ubiquitin transfer cascade, the E1 enzyme further transfers the Ub/Ubl linked to the internal cysteine to a conserved catalytic cysteine residue on the E2 enzyme. Structural studies on the UBA3-like E1 protein responsible for NEDD8 activation showed that they possess a C-terminal domain, the Ufd domain (a circularly permuted version of the β -grasp fold) [417, 418], that assists the E1 protein in catalyzing this transfer. Analogously some ThiF/MoeB/MOCS3-like E1 proteins possess a C-terminal rhodanese domain;

the adenylated Ub-like protein is transferred to the rhodanese domain through a persulfide linkage instead of the internal cysteine on the E1 domain. This suggests that a combination of core biochemical activities of the Rossmann fold domain and the presence of accessory domains influences the biochemical diversity of these enzymes.

Given the many biochemical reactions, interactions and associations, I was interested in exploring the evolutionary history of the E1-superfamily of enzymes. In particular, I wanted to address the following problems: 1) determining the relationships of the E1-like superfamily to other members of Rossmann fold domains and understanding the transitions, if any, that resulted in the evolution of its extant biochemical activities 2) Defining the conserved sequence and structural features common to all members of the fold and assessing how variations to this core set of features affect functional properties. 3) Cataloguing the complete set of domain fusions and exploring the functional roles, if any, these may play in the biochemical activities of the associated E1 domain. 4) Determining the evolutionary radiations of the fold, with an emphasis on the diversification of eukaryotic E1-like domains and establishing their evolutionary origins.

Application of Methods

I performed all analyses associated with this investigation and was assisted by Dr. Iyer in the lab for all structural analyses, who also aided in interpretation of the results. Dr. Aravind provided direction during the projects inception and guidance after it was initiated.

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) was searched using the BLASTP program [35]. Iterative database searches were performed using the PSI-BLAST program with an expectation value (E-value) of 0.01 used as the threshold for inclusion in the position-specific scoring matrix generated by the program [38]; searched were iterated until convergence. Multiple alignments

were constructed using the MUSCLE [78] and Kalign [419] programs, followed by manual correction based on BLAST high-scoring pairs and information derived from existing structures. All large-scale procedures were carried out using the TASS software package (Balaji S, Anantharaman V, Aravind L, unpublished results). Protein structures were visualized and manipulated using the Swiss-PDB [94] and PyMol (<http://www.pymol.org>) programs. Protein secondary structures were predicted by feeding multiple alignments into the JPRED2 program [89]. Clustering of proteins based on sequence similarity was accomplished using the BLASTCLUST program (<ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>).

Gene neighborhoods were obtained by isolating conserved genes immediately upstream and downstream of the gene in question showing separation of less than 70 nucleotides between gene termini. Neighborhoods were determined by searching NCBI PTT tables (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>) with a custom PERL script from the TASS package. Phylogenetic analysis was carried out using neighborhood-joining and minimum evolution-based methods as implemented in the MEGA program [420]. Additionally, maximum likelihood-based matrices were constructed using the Tree-Puzzle 5 program [65] and used as input for constructing trees with the Weighbor program [66].

Identification and Classification of E1-like Protein Families

In order to address the points made in the introduction and to ensure as complete a coverage as possible, a systematic and comprehensive analysis to retrieve all known sequences of the E1-like protein fold was undertaken. Primary sequences derived from crystal structures of known E1 proteins including the bacterial ThiF/MoeB, SUMO-activating UBA2, and NEDD8-activating UBA3 proteins were used as seeds to initiate sequence profile searches against the NCBI NR (non-redundant) database with the PSI-BLAST program and to initiate hidden Markov

model (HMM) searches using the HMMer package. In this manner, the set of known eukaryotic E1-like domain homologs was recovered including both active and inactive domains from the classical ubiquitin-activating UBA1, the ISG15-activating UBE1L, and the FAT10 and ubiquitin-activating UBA6 [421]; the Ufm1-activating UBA5, the Apg12-activating Apg7, and the AOS1 and APPBP1 inactive E1-like dimerization partners of UBA2 and UBA3, respectively. Additionally, the set of known prokaryotic E1-like homologs was recovered, including the HesA-like proteins involved in heterocyst formation [422], the MccB-like proteins involved in microcin production [415], the GodD-like proteins involved in goadsporin antibiotic production [416], and several previously described families that associate in conserved gene neighborhoods with other homologs of the ubiquitin modification system such as Ubls, E2-like proteins, and JAB domain metalloproteinases [414]. These searches also recovered one previously uncharacterized family of eukaryotic E1-like proteins typified by the *S. cerevisiae* protein YKL027W, and several novel bacterial E1-like families. For example, a search initiated with the primary sequence of the ThiF structure (PDB: 1ZKM [409], gi: 71042372) from *E. Coli* returns representatives of the HesA and MccB families in the first iteration and the eukaryotic Ub-adenylating versions within the first two iterations. It also recovers members of the YKL027W family (gi: 6322825, iteration 2, 7e-17), the 6c and 6e family of bacterial Ub-associating E1 domains (gi: 16519796, iteration 2, 2e-08 and gi: 120600775, iteration 2, 2e-10; respectively), and the GodD family (gi: 4218544, iteration 2, 5e-05).

Sequences that were obtained were clustered using BLASTCLUST and were further unified into families and subfamilies based on their residue conservation and associated domain contexts. As a result 13 families of E1-like proteins were obtained. Individual multiple alignments and predicted secondary structure of all families were prepared. Each family was closely

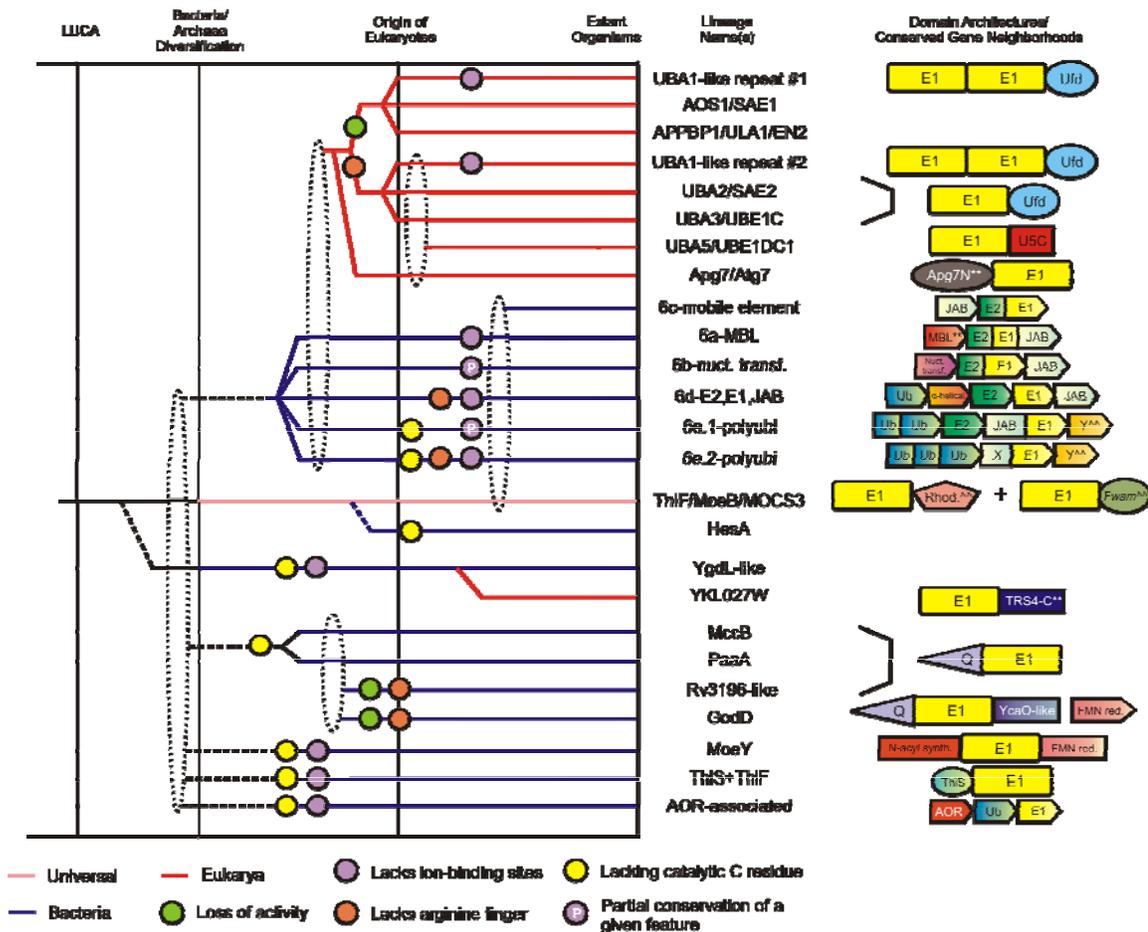


Fig. 22. Reconstructed evolutionary scenario and major architecture and gene neighborhood domain associations of the E1 domain.

The chart displays relative temporal eras representing major evolutionary transition events, demarcated by vertical black lines and labeled at the top of each line. E1 lineages are listed to the right of the chart, with horizontal lines (color-coded by phyletic distribution) extending to the left, tracing the maximal evolutionary depth to which a lineage can be confidently traced. Broken lines indicate uncertainty regarding point of emergence, and dotted ellipses indicate uncertainty over the precise phylogenetic affinities between groups of E1 lineages. Absence or partial absence of a given structural or sequence feature in a lineage is indicated by colored circles. Domain architecture(s) characteristic of a given residue are shown to the right of the lineage names with geometric shapes representing single domains. Conserved gene neighborhoods are always depicted by boxed arrows with the arrowhead pointing from the 5' to 3' direction. In both architectures and neighborhoods, the E1 domain is shaded yellow. Abbreviations: MBL, metallo- β -lactamase; X, β -strand rich globular domain; Y, cysteine-rich domain; Rhod., rhodanese domain; FMN red., FMN-like nitroreductase; N-acyl synth., N-acyl synthetase; AOR, aldehyde ferredoxin oxidoreductase.

inspected for congruence to the characteristic E1-like α/β fold (see SCOP database [423]); a super-alignment of the fold with representatives from each family is included in supplementary material. Individual alignments were also inspected for the presence or absence of key structural features and experimentally characterized residues essential for catalysis. A summary of the family-specific conservation is compiled in Table 3. Internal relationships between families were also determined through conventional phylogenetic analysis methods (Materials and Methods). Higher-order relationships between families were inferred through comparison of shared sequence and structural motifs (Table 3). Information derived from genome contextual information in the form of phyletic patterns, domain architectures, and predicted operon organization was also employed to clarify higher-order relationships. The reconstructed evolutionary history of the E1-like fold is illustrated in Fig. 22 and the natural classification of the fold is provided in detail in supplementary material.

The Origins of the E1 Superfamily

A comparison of all the available crystal structures, and secondary structure predictions of sequence families of E1-like enzymes suggests that their conserved core is composed of an N-terminal β - α units and a C-terminal α/β unit that form a central, eight-stranded β -sheet of order 87654123 (hereafter referred to as S1-S8). DALI searches with the E1-like proteins retrieved, in addition to other E1 proteins, several Rossmann fold proteins including distinct members of the FAD/NAD-dependent dehydrogenases and methylases (Z-scores: 8-9). The Rossmann-like fold of the E1 enzymes encompasses the region from S1 and S5 and the helices that follow S1-S4 (called H1 – H4). The C-terminal region of the sheet, from S6-S8, contains a distinct topology with S6 and S8 being anti-parallel to the other strands (Fig. 23) and is unrelated to any known fold. The

Rossmann fold is a 3-layered fold consisting of 4-7 α/β units (see SCOP) that form a central sheet and has a characteristic topology that distinguishes it from other α/β folds. Based on sequence and structural features Rossmann fold proteins are classified into the nucleotide binding domain division, with a nucleotide binding loop between strand 1 and the helix following it, and the phosphohydrolase or divalent cation-chelating division that contains a conserved acidic residue in the same loop. Members of the nucleotide binding division include the classic Rossmann NAD/FAD-dependent dehydrogenases, the *S*-AdoMet-binding methyltransferases, Sir2-like deacetylases, the FtsZ/tubulin-like GTPase, members of the ISOCOT fold, and the HUP superclass (HIGH nucleotidyltransferases, USPA, photolyase, class I tRNA synthetases, and electron transport flavoprotein). Proteins belonging to the latter division include the Haloacid dehalogenase (HAD) superfamily, the DHH phosphoesterases, the CheY-like receiver domains, the TOPRIM domain, the PIN/5'-3' nuclease domain, the arginase/classical histone deacetylases, and the von Willebrandt factor A (vWA) domain [212, 213, 258, 423]. The presence of a nucleotide binding loop between S1 and H1, places the E1-like proteins within the nucleotide binding division of the Rossmann fold. Within this group, the E1s are closer to the NAD/FAD-dependent dehydrogenases and the *S*-AdoMet-binding methyltransferases, as they specifically share all of the following four characters: 1) a sheet order of 54123, 2) a nucleotide/ligand binding glycine rich loop between S1 and H1, 3) A conserved aspartate residue at the end of strand 2, and 4) a characteristic aspartate or asparagine residue (D in E1, D/N in the other Rossmann fold proteins) at the end of S4 (Fig. 23).

Table 3. Secondary structure features of major E1 fold structural categories.

1. Abbreviations: assoc., associated; nuct. transf., Nucleotidyl transferase; MBL, Metallo- β -Lactamase domain; Ub, Ubiquitin.
2. Abbreviations: S, conserved strand; H, conserved helix; Nh, N-terminal conserved helix; Rf, arginine finger; Rc, conserved arginine in different position predicted to act as arginine finger; CC, extended coil region housing adenylation active site; +, conserved positively-charged Ub/Ubl-recognition residue; I1, insert region occurring before ascending arm; CxxC1, first Mg²⁺ chelating motif; CxxC2, second Mg²⁺ chelating motif; Ime, insert found in 6c lineage; I2, insert region occurring after ascending arm; e, insert in extended conformation (strand-like); h, insert in helical conformation; cc, long coil insert; *, marks residues essential for catalysis; --, indicates a feature is absent in the lineage. Capital letters in a column denote conserved residues; lower case letters under these columns indicates only partial conservation of the residue in the lineage.

The E1-like proteins further possess several unique sequence and structural features that distinguish them from other members of the Rossmann folds. N-terminal to the Rossmann fold domain is a set of three α -helices of which the central one, that is positioned roughly perpendicular to the central β -sheet, contains a highly conserved arginine residue. This arginine residue, called the arginine finger, contacts the γ phosphate of the ATP molecule present in the opposite unit of the E1 homo or heterodimer and is proposed to stabilize the negative charge on the β -phosphate of the pentavalent reaction intermediate formed during the adenylation of the Ub-like protein [412]. This feature in the E1 proteins is reminiscent of arginine fingers associated with a wide range of enzymes involved in phosphohydrolysis, such as members of the P-loop NTPase fold and the HAD phosphatases [213, 242]. The second feature unique to E1 proteins is an extended loop between S2 and H2. This loop contains several residues conserved across all major families of E1; including an aspartate, asparagine, arginine, and lysine. The asparagine residue lies on a highly conserved helix nested in the loop. Along with the arginine finger mentioned above, these comprise the residues necessary for catalyzing the adenylation reaction.

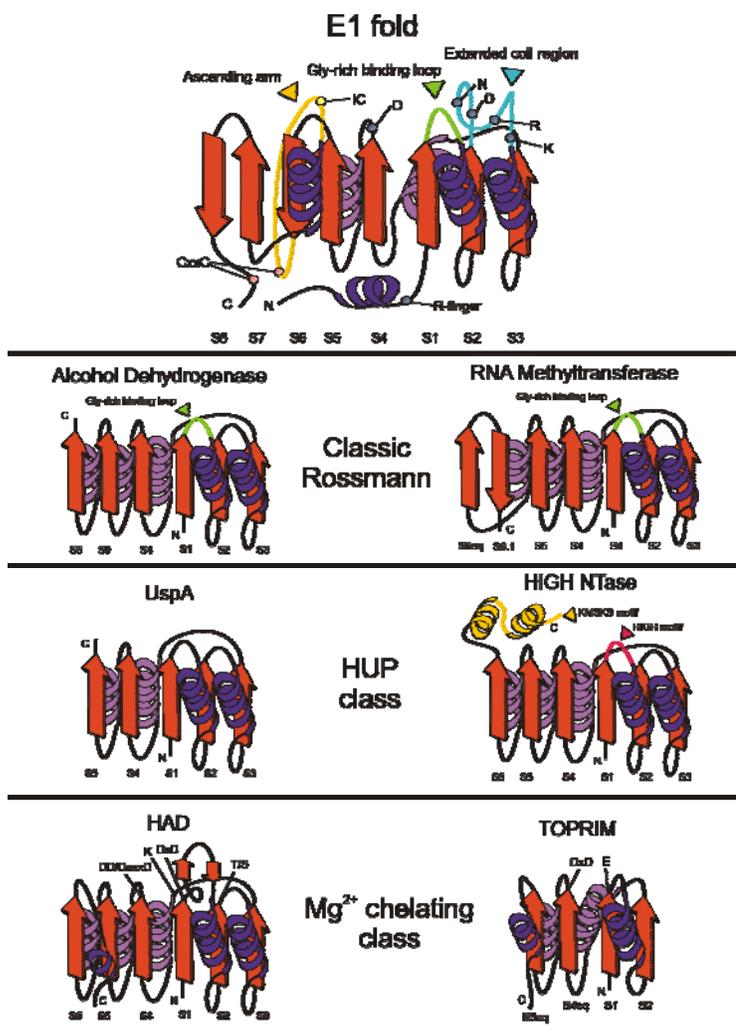


Fig. 23. Colored topology diagrams of representatives from the major divisions of Rossmannoid folds.

Located at the top is the E1 domain; strands are shown as arrows with the arrowhead on the carboxyl-terminal end and labeled below and colored in burnt orange while helices are colored in northwest purple, with helices below the plane of the central β -sheet shown in a lighter shade. Major structural features characteristic of the domain and other divisions are colored and labeled. Sequence features found in the domain are marked by small, colored circles and labeled. One circle, colored in yellow, represents the position of the internal cysteine residue. N- and C-termini are labeled accordingly.

Finally, the third distinguishing feature of the E1-like proteins is the C-terminal region encompassing strands S6-S8. This region contains a characteristic ascending arm between S6 and S7 that causes the polypeptide backbone to cross over to the opposite face of the protein. This ascending arm harbors the conserved cysteine to which the activated ubiquitin or ThiS is transferred in the ubiquitin and thiamine biosynthesis pathways respectively. Strands S6-S8 also make extensive contact with the exposed face of the ubiquitin, ThiS or Moad-like molecules, suggesting that they are involved in substrate selection.

The evolution of the E1-like enzymes can now be reconstructed as originating from an ancestral nucleotide binding Rossmann fold protein that passively bound a nucleotide as in the dehydrogenases and methyltransferases. Further developments in these proteins were aided by three principal lineage-specific structural elaborations, the N-terminal arginine finger containing helix, the extended loop between S2 and H2, and the C-terminal α/β extension that extensively contributed to the evolution of catalytic activity in the E1-like proteins.

Sequence and structural diversity within the E1 superfamily

Enzymes often attain diversity by exploring one or both of reaction space and substrate space. These are suggested by the residue conservation in the catalytic or substrate binding sites or by the presence of additional structural elements that have a potential to influence catalytic or substrate binding activity. Experimental studies and residue conservation suggest a conserved mechanism of ATP hydrolysis and adenylation across all E1 proteins. It is postulated that the C-terminus carboxyl residue of the Ub-like protein attacks the α -phosphate group of the bound ATP to generate a pentavalent intermediate that is stabilized by the arginine finger and the conserved residues in the loop between S2 and S3 resulting in the formation of adenylylated ubiquitin and inorganic pyrophosphate [424]. The residues responsible for E1 activity can be divided into four types: 1) residues that influence nucleotide binding and adenylyating activity; 2) residues that influence the thiolation activity; 3) residues that determine substrate choice and binding; and 4) residues that influence global structural properties and movement during catalysis.

Several residues contacting the ATP nucleotide influence adenylation activity. These include the arginine finger, the glycine rich loop between S1 and H1, the conserved aspartate, asparagine, arginine and lysine residues in the loop between S2 and S3, and a conserved

aspartate at the end of strand-4 that coordinates the Mg ion of Mg/ATP and shields the repulsion between the α -phosphate and the C-terminal carboxyl group of the Ub-like protein. Catalytically inactive members have previously been detected in the eukaryotic E1-like proteins such as the inactive N-terminal E1 domain in UBA1, the AOS1/SAE1 and the ULA1/ENR2/APPBP1-like families. These proteins form heterodimers with catalytically active members of the E1 superfamily, such as the C-terminal E1 domain in UBA1, UBA2/SAE2 and UBA3 respectively. The inactive members lack most of the conserved residues between S2 and H2 and only contain an arginine finger that contacts the active site of the active partner. Correspondingly, the active partners lack the arginine finger. The inactive versions also lack the aspartate after strand-4 that coordinates MgATP, which correlates with the absence of a nucleotide in their active site. On inspecting the active site of all E1 families, two E1 subfamilies were further recovered, the Rv3196-like and GodD, and one E1 family, MJ0693-like, with two or more modifications to the conserved residues between S2 and H2 (Table 3) and a missing arginine finger. The GodD-like family also lacks the glycine rich loop between S1 and H1 involved in binding nucleotides. The absence of critical polar residues in the loop between S2 and H2 suggests that these proteins are incapable of adenylation. Moreover, their genome or domain contexts do not hint the presence of any active member which would form a heterodimeric complex with them. The Rv3196-like and GodD-like proteins belong to the family of E1 proteins involved in peptide antibiotic biosynthesis (see below). Unlike microcin C7 which is adenylated at its C-terminus, the GodD peptide isn't adenylated and this corresponds to the lack of adenylating residues in the E1 enzyme. Interestingly, the Rv3196-like and the archaeal MJ0693-like families possess the Mg coordinating aspartate in S4. The MJ0693-like proteins also possess the nucleotide contacting aspartate in the loop between S2 and H2. It is possible that members of these families retain their nucleotide-

binding activity. The Rv3196-like family is also notable for the presence of a distinct constellation of highly conserved polar residues which when mapped to the E1 structure suggests the possibility of a secondary active site near the C-terminal region of the domain. The conserved residues do not resemble those present in the loop between S2 and H2 and may represent the presence of a novel enzymatic function in these proteins. Another family with substitutions to residues in the loop between S2 and H2 is the HesA family found in nitrogen-fixing cyanobacteria and actinobacteria. These substitutions, however, are highly conserved across all members of this family and include a basic residue instead of the aspartate, an aspartate instead of the asparagine and an arginine instead of the lysine. They also possess a highly conserved asparagine in a distinct position in the loop between S2 and H2 and also lack the Mg coordinating aspartate in S4. These replacements suggest that the HesA proteins probably have a distinct enzymatic activity or a variation in the biochemical enzymatic activity in comparison to the classical E1 proteins. It is possible that the lack of the Mg-coordinating aspartate and the presence of a new aspartate in the loop between S2 and H2 caused an altered mode of nucleotide binding or bound a modified nucleotide that may have contributed to their distinctness. Thus the E1 enzymes with modified residues in the loop between S2 and H2 may function as substrate binding scaffolds in larger complexes or perhaps have enzymatic activities distinct from adenylation.

A common modification observed in the E1 proteins was the loss of the N-terminal arginine finger (Table 3, Fig. 22). These fall into two distinct types. 1) Proteins that are catalytically inactive (such as the above described families) and heterodimer-forming eukaryotic E1 enzymes that contact an inactive partner always lack the arginine finger. 2) In several families that are catalytically active and lack the finger, the loss is compensated by lineage-specific

emergences of conserved arginine residues at other distinct positions either within the E1 fold, or in a contextually-associated domain. For example, the E1 domain of Apg7/Atg7, which is involved in autophagy, contains two possible candidates that may substitute the N-terminal arginine finger. One of them is a conserved arginine residue at the N-terminus that is present in a distinct position that may contact the active site of a dimeric partner. The other is an extended coil insert between S7 and S8 that harbors an absolutely conserved arginine residue that may fold over the S5/S6 hairpin and ascending arm (Fig. 23, supplementary material) to contact the catalytic site of the same protein. In the *Clostridium* pCPF5603-like family (labeled 6a in Fig. 23), the E1 domain is sandwiched between an E2 and a JAB domain and is in the neighborhood of an enzyme of the metalloβ-lactamase (MBL) fold. Sequence analysis of the MBL domain reveals a striking conserved arginine residue in the variable region connecting the two subdomains [217] of the fold. This residue, through association in a complex between the MBL and E1-like domains, may serve as the arginine finger in this family (Table 3, Fig. 22). Similarly in yet another family of prokaryotic E1-like proteins that are associated with a polyubiquitin and lack an arginine finger (labeled 6e.2 in Fig. 23), one of the conserved arginine residues of the poorly studied N-terminal domain (labeled X in the Fig. 23) could substitute as an arginine finger. The emergence of distinct arginine fingers is reminiscent of the P-loop ATPases, where arginine fingers have evolved on several occasions independently, and are either provided from within the protein such as in the helicases, other NTPase subunits in multimeric complexes such as in the AAA ATPases, or distinct domains associated with the P-loop NTPase, such as in the PilT NTPases [157].

The other major biochemical activity of the E1-like proteins is the transfer of the ubiquitin-like protein, through a thioester or persulfide linkage, to an internal cysteine in the

ascending arm. This thiolating cysteine residue is often found in catalytically active E1-like proteins that functionally or contextually associate with an Ub-like protein. Interestingly, these are also present in MoeB-like E1 proteins, where the ubiquitin-like MoaD is not thought to be transferred to the internal cysteine [425]. Nevertheless, the association of the thiolating cysteine with ubiquitin adenylating E1s appears to be an ancient one and it is possible that the transfer of an ubiquitin-like protein to an internal cysteine of an E1-like protein was lost in different lineages secondarily such as MoeB. Other families that lack the thiolating cysteine are catalytically inactive E1s, or families such as the Aldehyde ferredoxin-reductase (AOR) associated E1 where the cysteine appears to have been secondarily lost, or versions such as the MccB, PaaA and HesA where evidence suggests the presence of a distinct substrate other than ubiquitin (Table 3, see below). The region of the ascending arm, flanking the thiolating cysteine is often prone to structural elaborations or insertions. Both the active and inactive subunit of heterodimer forming E1s have helical inserts after the position corresponding to the thiolating cysteine (Table 3). Inserts in this region are also seen in several other families, such as a predicted beta hairpin in a prokaryotic Ub family associated with mobile elements, and helical inserts in the YdgL-like family and MccB and Rv3196 subfamilies. Similarly inserts are seen in the ascending arm prior to the thiolating cysteine in the inactive eukaryotic E1 proteins, the APG7 family, and at least three distinct prokaryotic Ub families involved in Ub signaling (Table 3). While the inserts in heterodimeric eukaryotic E1 proteins have a common origin, those in other families appear to have evolved independently of each other. An examination of the inserts in the heterodimeric UBA3-APPBP1 structure (PDB: 2nvu), shows that the insert after the thiolating cysteine of the active UBA3 interacts with the inserts prior to the cysteine in the inactive APPBP1. The insert in the inactive APPBP1 also interacts with the adenylated Ub substrate. Thus, it appears that these

inserts in the E1 proteins may play multiple roles including promoting efficient heterodimerization, creation of binding pockets for substrates and influencing access to the active site by ATP or the substrate, and have also recently been shown to be important in E2 -like protein interactions [426]. These inserts are reminiscent of the inserts in the flap motif of the Rossmann-like HAD superfamily that influence substrate recognition and access to the active site, thereby maximizing catalytic efficiency.

The only known substrates of the E1 proteins with detailed crystal structures are the Ub-like proteins. These contact S7 and S8 of E1 protein through the exposed face of their sheet. Interestingly, S7 and S8 are some of the most divergent sequences in the E1 proteins, which correlate with their involvement in substrate choice. Another residue that may influence the binding of protein tails is a conserved arginine residue in H4 of the E1 proteins. In crystal structures, this residue makes polar contacts with the backbone residues of the C-terminal tail of the ubiquitin-like protein perhaps directing the tail to the active site. This residue is conserved in catalytically active E1 proteins, and is also found in families that are predicted to bind non-ubiquitin substrates. A third set of conserved motifs that may influence substrate binding are a pair of CxxC motifs; one located in the ascending arm, and another in an unstructured region after S8. These conserved cysteines coordinate a zinc ion. An inspection of the structural context suggests that the coordination causes a widening of the angle between S6 and the helix following it, thereby accommodating the ubiquitin tail in the E1 catalytic active site. In contrast and corroborating this observation, the inactive chains in E1 heterodimers don't bind ubiquitin and lack this widening. This region may also help in the proper positioning of any C-terminal domain with respect to the rest of the structure. For example, experimental studies show a significant movement of the C-terminal Ufd domain during the transthiolation reaction. The zinc

coordination perhaps assists this process by holding the C-terminus in a way that does not affect the binding of the Ub substrate. All E1-like families bearing C-terminal domain fusions contain the CxxC motifs (Table 3). The zinc-coordinating motifs are sporadically present across many families suggesting that they were an ancient feature that were repeatedly lost across many lineages. Another ancient feature that is widespread and may influence dimer formation, and has not been studied previously, is the ExxK motif in the helix before S7 (Table 3, supplementary material). Crystal structures reveal that the glutamate and the lysine of the motif form a salt bridge and interact with the hairpin turn between S7 and S8 of the dimerizing partner. Thus this motif may be essential for stabilizing the conformation dimer interface. The widespread presence suggests that the E1-like proteins dimerize in a similar way in a wide range of contexts. Alternatively, the salt bridge may be involved in resetting any changes caused to the conformation of the E1 protein during the adenylation or thiolation reaction. Such conformation-specific salt bridges are also seen elsewhere. For example, in the ligand-gated ion channels a salt bridge in the 'outer sheet' of the ligand-binding β -sandwich domain is proposed to regulate the preferential movement of the sheets after ligand-binding [427]. However, loss of the ExxK motif in several lineages suggest that this can easily be replaced by lineage-specific innovation of stabilizing residues.

The evolution of the E1 proteins can now be construed to have emerged in distinct stages with the early innovation of the glycine rich loop after S1, the catalytic loop between S2 and S3, and the arginine finger for nucleotide hydrolysis and adenylation in a homodimeric context. These proteins also coordinated a zinc ion and contained a basic residue after H4 that perhaps regulated access to the active site by substrates. The next stage involved association with Ub/Ubl proteins in sulfur incorporating systems that led to the emergence of thiolating cysteine. Phyletic

distribution suggests that the earliest mode of transfer to the internal cysteine was perhaps through a persulfide linkage that changed to a thioester linkage as the Ub and E1 like proteins associated with signaling systems with the JAB and E2-like proteins. Finally, there were inactivations or modifications of the catalytic residues of E1-like proteins in different lineages and, distinct lineage-specific insertions that improved the catalytic efficiency of these proteins.

Diverse Domain Architectures in the E1 Superfamily

Contextual associations, such as fusion to additional domains, or conserved gene neighborhood associations in prokaryotes, are extensively used to gain functional insights into poorly studied protein domains and families. The catalytically active UBA3-like E1 protein, involved in ubiquitin signaling, was shown to contain a circularly permuted version of the β -grasp fold at its C-terminus; the Ufd domain. E1-like proteins are also fused to the JAB protease and E2 domains in the prokaryotic Ub signaling systems [414]. In biosynthetic pathways, E1-like proteins have been previously described to be fused to ThiS-like Ub and rhodanese domains [414]. These fusions suggest a significant role for context-specific fusions in expanding the activities of the E1-like proteins. In order to gain a comprehensive insight into the biochemical contexts and evolutionary histories of the fusion events, all possible contextual information available for the proteins in the E1 superfamily was analyzed.

Structural studies demonstrated that the Ufd domain binds the E2 enzyme and triggers a conformation change within the E1 enzyme, bringing the active site of the E2 protein in close contact with the Ub substrate that is linked to the internal cysteine residue [417]. Across the eukaryotes, the C-terminal Ufd domain is found in three E1 families, UBA1, UBA2 and UBA3. All three are catalytically active proteins and form heterodimers with catalytically inactive subunits which lack the Ufd domain. The remaining two eukaryotic E1 families also show very distinct

fusions. The Ufm-modifying UBA5-like proteins contain a distinct C-terminal domain (henceforth U5C domain from UBA5 C-terminal) predicted to contain three strands flanked by α -helices (see supplementary material). The secondary structure assignment suggests that the domain is distinct from the Ufd domain, or any domain of the β -grasp fold. Several polar residues are conserved in the UBA5C domain including a highly conserved [ED]a motif after the first strand (where a: aromatic). Given the position of the fusion at the C-terminus, and a size comparable to the Ufd domain, it is possible that the U5C domain functions similarly to the Ufd domain by binding an E2 enzyme. The other E1 family, Apg7 which is involved in the autophagy pathway, is fused to an N-terminal α/β domain (hereafter the Apg7N domain) with several highly conserved charged residues including four basic residues and a well-conserved DhKK motif (where 'h' is a conserved hydrophobic residue). This domain, like the U5C domain, lacks any observable homologs. The Apg7-like proteins interact with two distinct E2 enzymes, Apg3 and Apg10. Experimental studies have shown that an acidic insert present in the Apg3-like E2 protein, involved in Apg8-mediated lipidation, is essential for Apg7 interaction [428]. It is possible that the basic residues of Apg7N mediate interactions with the acidic insert of Apg3-like E2 enzyme. Alternatively, the highly conserved polar residues might imply an enzymatic function that supplements the E1 activity, or participates in the autophagy pathway in a separate capacity. The eukaryotic YKL027W family is related to the YgdL family of bacterial E1 proteins and represents an independent transfer of the E1 domain to eukaryotes (see below, Fig. 22). Members of this family lack the arginine finger and thiolating cysteine and are fused to the TRS4-C domain at their C-terminus. The TRS4-C domain is also fused to the TRS4N-like Ub domain and was speculated to play an enzymatic role as it has a highly conserved ExxH motif [429]. Their highly conserved arginine may serve as an arginine finger in these proteins as speculated above.

This network of contexts suggests an independent recruitment of the YKL027W proteins for adenylating the TRS4N-Ub domain. Thus accessory domains in eukaryotes appear to be involved in recruiting other members of the ubiquitin pathway such as the E2 domain or in context-specific enzymatic activities. Prokaryotic E1-like proteins involved in Ub signaling systems show very distinct fusions and contexts in comparison to the eukaryotic counterparts. In these, the multi-domain versions of the E1-like protein are either fused to an E2-like enzyme, the JAB metalloprotease or both. Genome neighborhoods of these proteins also showed distinct conserved associations that include other proteins that perhaps function as substrates or accessory proteins for the system. For example, as proposed above, the MBL domain present in some prokaryotic families might provide the missing arginine finger for their neighboring E1 proteins. Similarly, in some families with multiple fused ubiquitins (labeled 6e in Fig. 23), an uncharacterized domain (domain Y), which contains several conserved cysteine residues [414] might substitute for the thiolating cysteine absent in the some members of the E1 proteins.

Among the E1-like proteins of the ThiF/MoeB/MOCS3 family involved in sulfur metabolism, a widespread fusion is to the rhodanese domain that has an absolutely conserved cysteine residue that has been shown to function as a thiodonor and thioacceptor in some metabolic pathways [425, 430]. In the same family, a distinct fusion of a novel domain C-terminal to the E1 domain in the Low GC gram positive bacteria and sporadically in other bacterial and archaeal lineages was identified. Solo versions of the domain are present in euryarchaea (e.g. *Pyrococcus* Fwam). The domain (henceforth Fwam) is also fused to 4Fe-4S ferredoxins and HTHs in some euryarchaea and sporadically in bacterial lineages. Secondary structure predictions reveal an α/β topology with two strands followed by a helix, three conserved strands and a helix. The secondary structure progression, although reminiscent of the 5-stranded β -grasp fold,

appears to lack the characteristic connector arm [429]. This domain contains an absolutely conserved cysteine residue at the C-terminal end of the first predicted strand (supplementary material), suggesting that it may have a role in sulfur incorporation similar to the rhodanese domain.

Among families of E1-like proteins that have substrates distinct from Ub-like proteins are the MccB and GodD subfamilies, associated with the biosynthesis of the polypeptide antibiotics microcin C7 and goadsporin. Both of these share an uncharacterized domain of unknown function (henceforth Q) at their N-terminus. This domain is also present in the PaaA and Rv3196-like subfamilies suggesting that these four subfamilies are related and are most likely involved in the biosynthesis of polypeptide antibiotics. The Q domain is predicted to contain three β -strands followed by three α -helices (supplementary material). An examination of these sequences suggests that the sequence conservation patterns differ between different subfamilies. For example, the Q domain of MccB and PaaA families share conserved arginine and aspartate residues, while the one in the GodD family contains a strongly conserved glutamate residue after the second predicted strand (supplementary material). The GodD family is additionally fused at its C-terminus to the YcaO domain, an uncharacterized domain present in a wide range of bacteria and archaea with several absolutely conserved residues that may be involved in an enzymatic capacity during the formation of the modified goadsporin antibiotic. Microcin C7 and goadsporin are unrelated polypeptides and of the two only microcin C7 is adenylated, whereas goadsporin is modified such that residues within the sequence are heterocyclized to give thiazole and oxazole rings [416, 431, 432]. Moreover, as described above, versions of this family such as Rv3196-like and GodD are inactive for adenylation. These suggest a rapidly evolving system where components such as the polypeptide antibiotic and proteins that recognize and modify it

are under strong positive selection. Thus, the Q domain might be involved in recognizing the polypeptide antibiotics given its rapidly diverging sequences. An enigmatic family of proteins that may either be involved in a distinct metabolic pathway or in modifying polypeptides as above is the MoeY family. Members of this family are sporadically distributed across the bacteria and do not show any conserved gene neighborhood associations. They also lack the thiolating cysteine. A subset of them in *Mycobacteria*, bacteroidetes, and some proteobacteria are fused at their C-terminus to an FMN binding nitroreductase domain, of which the version in *Desulfovibrio* is further fused to an N-terminal N-acyl amino acid synthase domain. These contexts are suggestive of multiple modifications of the substrate post-adenylation and may represent a system similar to the polypeptide antibiotics.

Evolutionary Themes in the E1 Fold

The E1 superfamily can generally be divided into three major types based on their associated pathways and substrate affinities: (1) families involved as sulfur carriers in biosynthetic pathways present in a wide range of bacteria, archaea and eukaryotes, (2) families involved in ubiquitin signaling present in a wide range of bacteria and eukaryotes, and (3) families involved in polypeptide antibiotic biosynthesis, that are sporadic and widespread in the bacteria. These phyletic patterns suggest that the E1-like proteins involved as sulfur carriers are some of the most ancient versions from which the other versions evolved subsequently. However, an examination of the sulfur-carrier E1 proteins coupled with their gene neighborhoods and domain contexts presents an intriguing evolutionary history. Phylogenetic trees show that the E1 proteins of the thiamine and Moco biosynthetic pathways do not cluster within each other (Fig. 24). Instead, on three independent occasions in the β and γ proteobacteria, a sporadic set of Low GC gram positive bacteria and in *Corynebacterium*, the E1-like proteins

differentiated into a version that associated with Moco and another that associated with the thiamine biosynthesis pathways. Barring these three events, most prokaryotes typically possess a single E1-like protein, even though they may have other components of both the thiamine and Moco biosynthesis pathways. This is in contrast with the evolution of the ThiS and Moad Ub families that were previously shown to have diverged in the Last Universal Common Ancestor (LUCA) of life. This suggests that in most prokaryotes and eukaryotes, a single member of the ThiF/MoeB family performs a dual function in incorporating sulfur in the thiamine and Moco biosynthesis pathways. Another striking picture that emerges from the phylogenetic analysis is that in several distinct lineages, such as the Cyanobacteria, Actinomycetes and diverse proteobacteria, the E1 protein is either fused to or in the neighborhood of a rhodanese domain containing protein. In some sporadic lineages of bacteria, as described above, the E1 is fused to the Fwam domain that is predicted to have a function similar to the rhodanese domain. The lack of any similarity of the ThiF/MoeB/MOSC3 phylogenetic tree to the most commonly recovered bacterial species tree [433], and the sporadic distribution of associations are suggestive of widespread lateral transfer of these E1-like proteins between bacteria. Given this data, the lack of a Moad protein transfer event to the persulfide-forming cysteine residue in MoeB is a mystery [411, 434]. The independent evolution of Moco-associated E1s and the strong conservation of the

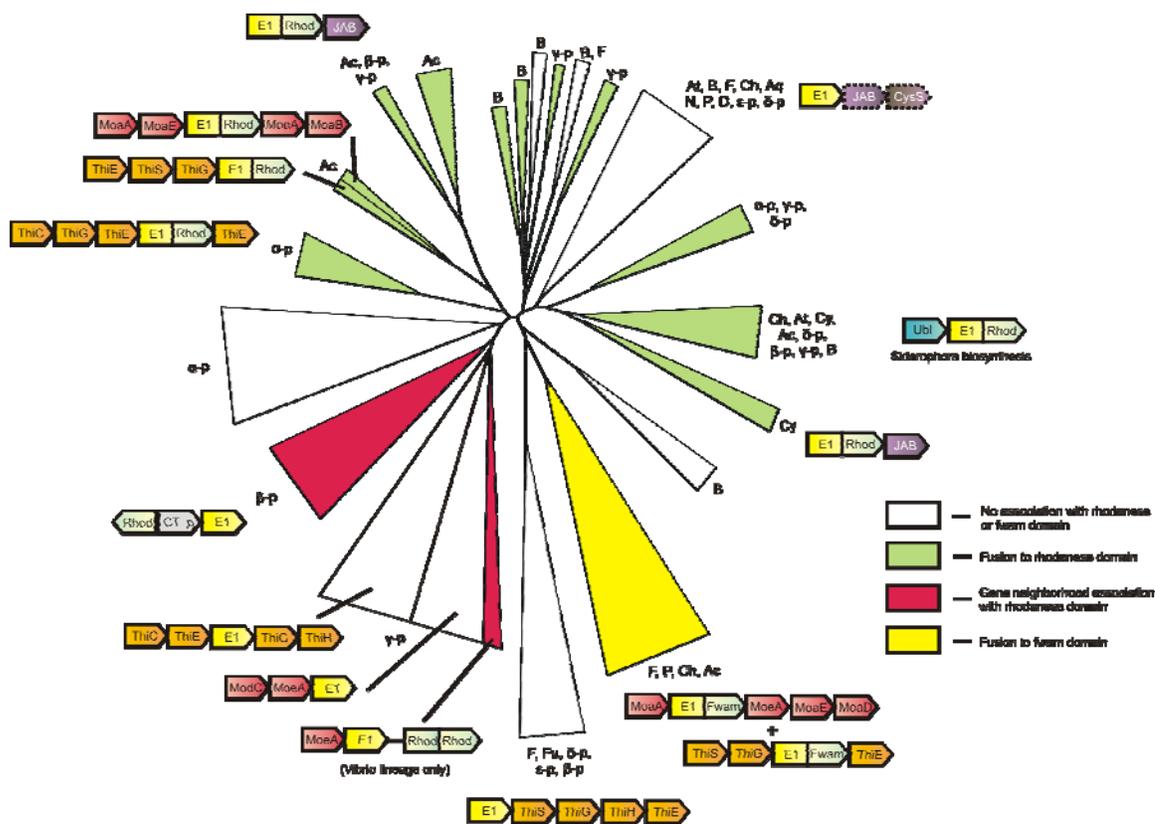


Fig. 24. Phylogenetic tree displaying interrelationships between lineages in bacterial This/MoeB E1 domains and the conserved gene neighborhoods characteristic of lineages.

Branches of the family linked to the fwam or rhodanese domains are color coded. The three branches with genomes that feature E1 domains differentiated into two distinct versions involved in thiamine and Mo-cofactor biosynthesis contain a solid black line dividing the branch. Branch sizes are roughly proportional to total number of representative proteins in the branch. Conserved gene neighborhoods are depicted in by boxed arrows with the arrowhead pointing from the 5' to 3' direction, the E1 domain is shaded in yellow. Domains in gene neighborhoods not always present in a given branch are outlined by a broken line.

thiolating cysteine, however, suggest that this may be a specialized feature restricted to the γ -proteobacterial E1 proteins.

As previously shown, the core of the ubiquitin signaling pathway evolved in the bacteria and seeded the eukaryotes early in their evolution, most likely during the primary endosymbiotic event that triggered eukaryogenesis. The distribution of the different E1 families suggest that last

eukaryotic common ancestor already possessed 7 families (Fig. 23) consisting of six families that formed three heterodimers with a catalytically active (UBA1 active E1, UBA2/SAE2 and UBA3), and inactive subunit (UBA1 inactive E1, AOS1/SAE1 and APPBP1), and the Apg7/Atg7 family involved in the autophagy pathway. The monophyly of the catalytically active and inactive subunits suggests that immediately after the primary endosymbiotic event there was a duplication of an ancestral E1 to give an active and inactive subunit that diversified into the six families before the last eukaryotic common ancestor. The distribution of the UBA5 family suggests an early evolution prior to the divergence of the kinetoplastids and heterolobosean lineages. The UBA5 family, however, was subsequently lost in several lineages [435] (supplementary material). An independent transfer from the YgdL-like bacterial proteins early in eukaryote evolution gave rise to the YKL027W family that seems to have secondarily associated with the Ub signaling pathway through association with the TRS4C domain. Although its precise role in the Ub signaling pathway is not identified, genome-scale analysis of protein complexes (as presented in the BioGRID database [34]) show that *S. cerevisiae* YKL027W associates with the RPN6 subunit of the 20S proteasome, suggesting that it may be involved in its modification. Although it should be noted that this modification may have some differences with the classical Ub pathway in that it definitely lacks a thiolating cysteine and may use the accessory TRS4C domain to supplement or bind the E2 protein. The four accessory domains fused to different eukaryotic E1s, Ufd, U5C, Apg7N and TRS4C appear to be eukaryote-specific innovations as they are missing in prokaryotes. Of these the Ufd and Apg7N fusions appear to have occurred in the last eukaryotic common ancestor and the Ufd was already fused to the catalytically active partner in the ancestor of the UBA1, UBA2/SAE1, UBA3/APPBP1- like heterodimeric E1s. The U5C and TRS4C were fused early in the divergence of the eukaryotes around the time the kinetoplastids

and heteroloboseans were diverging. These clearly evolved to attend to the specific functional needs of the eukaryotic E1s and may have replaced the simpler prokaryotic systems where the E1 enzyme was directly fused to E2 and other associated domains involved in the Ub-signaling pathway. In addition to these early events, there were several lineage-specific duplication events in diverse lineages, especially of the UBA-1 like proteins that spawned newer subfamilies. For example, UBA-1 like proteins gave rise to the ISG-15ylating UBE1L subfamily in vertebrates, the UBA6 subfamily in the animals, and paralogs in chromalveolates, *Trichomonas*, *Naegleria*, ciliates, and the apicomplexa. Lineage-specific expansions of UBA1 are also observed in the ciliates *Paramecium* and *Tetrahymena*. These duplication events suggest the colonization of functional niches that are unique to the different eukaryotic lineages.

Of the remaining families, most of the other families such as HesA and the polypeptide antibiotic synthesizing E1 proteins, though present in diverse lineages, show a sporadic distribution. This is suggestive of rampant lateral transfer of these genes between different bacterial lineages. The only other family with a widespread distribution is the YgdL-like E1 family. Contextual analysis provides few hints on its function. It is possible that they have their own distinct substrate or alternatively, since the related eukaryotic YKL027W associates with TRS4-N Ub-like domains, it is possible that these, too, associate with a ThiS-like protein in bacteria.

Conclusions and General Observations

Our analysis suggests that the E1-like proteins already attained their catalytic function in the LUCA and were probably involved in adenylating and thiolating a ThiS/MoaD-like protein. Subsequently, these proteins appear to have explored new functional niches by: 1) increased complexity of associations that led to its association with pathways involved in sulfur

metabolism, siderophore biosynthesis, and the ubiquitin signaling pathway in bacteria. 2) Changing the substrate types as in the polypeptide antibiotics and perhaps the HesA and YgdL families. These explorations went hand in hand with accommodations in the active site residues and associations with a variety of domains that provided context-specific activities, many which are described for the first time in this study. Our evolutionary reconstruction suggests that the earliest antecedents of the E1 domains were perhaps part of the Moco biosynthesis pathway with the MoaD/ThiS-like ubiquitin domains as substrate. Subsequent divergence involved the association with the Ub pathway, the polypeptide antibiotic pathways and other metabolic pathways in the bacteria. This was accompanied by rampant lateral transfer of the E1 proteins between species and between pathways, which may have had to do with their lack of strict specificity to a particular substrate. Finally, the eukaryotes acquired at least 2 distinct members of the E1 family during the primary symbiogenesis event of which one of them contributed to the many diverse E1 families associated with the Ub pathway in eukaryotes. In conclusion, the results presented in this research provide new leads into understanding the mechanism of Ubl transfer in various ubiquitin modification systems and the sulfur incorporation steps of diverse metabolic pathways. It also defines the complete structural space of the E1 fold and expands the scope of functional contexts occupied by the E1 fold. Further investigation into the less studied E1 families presented here may provide more insight into the biochemical diversity of the E1 proteins.

Supplementary Material

Supplementary material mentioned at various locations throughout the text above can be accessed at the following website:

http://www.ncbi.nlm.nih.gov/CBBresearch/Lakshmin/E1_supplement.txt.

INVESTIGATIONS RELATING TO THE β -GRASP FOLD

The β -grasp fold (β -GF), prototyped by ubiquitin (Ub), is a member of the $\alpha+\beta$ class of protein domains and has been recruited to a strikingly diverse range of biochemical functions. The domain consists of 4- or 5-stranded core β -sheet that appears to grasp a conserved helical segment and was first discovered in Ub-like proteins and later recognized as being present in a range of other proteins [423, 436, 437]. I was interested in determining the full scope of functional diversity within the β -GF across all three superkingdoms of life, as well as identifying and characterizing the different evolutionary adaptations resulting in the diversity observed in the fold.

The first study in this section was designed to determine the full functional and structural scope of the fold. To this end, all sequences belonging to the fold were collected, in the process discovering several previously unrecognized members of the fold. The higher-order evolutionary relationships of the fold were then constructed, resulting in the identification of many of the different evolutionary radiations and structural/sequence correlates of diversification. The second study describes a novel superfamily uncovered during the first study and characterizes it as a soluble ligand-binding superfamily, the first instance of a ligand-binding functionality for the fold. The third and final study specifically focuses on the early evolution of the Ub superfamily and expands into an investigation of the evolutionary origins of the entire eukaryotic Ub signaling system.

Small but Versatile: the Extraordinary Functional and Structural Diversity of the β -Grasp Fold

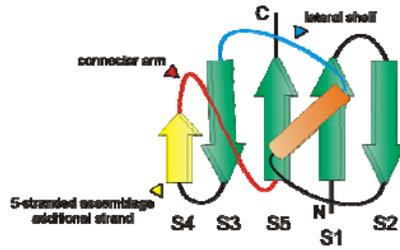
(based on reference [429])

Introduction

The discovery of covalent modification of eukaryotic proteins by the conjugation of ubiquitin to the ϵ -amino groups of target lysines has spawned some of the most exciting directions of research in current molecular biology [438-440]. Ubiquitin (Ub) itself is a small polypeptide of 76 residues, and its crystal structure revealed a distinctive fold dominated by a β -sheet with 5 anti-parallel β -strands and a single helical segment [441, 442] (Fig. 22A). Pioneering investigations of Kraulis, Overington and Murzin showed that this fold was not unique to Ub, but was also present in several other proteins with biologically distinct functions. These included the staphylococcal enterotoxin B, the streptococcal immunoglobulin (Ig)-binding protein G and 2Fe-2S ferredoxins [423, 436, 437]. The common fold shared by these proteins was termed the β -grasp, because the β -sheet appears to grasp the helical segment in this domain [436]. These early studies provided the first indications that, despite its small size, the β -grasp fold (β -GF) might serve as a multi-functional scaffold in diverse biological contexts.

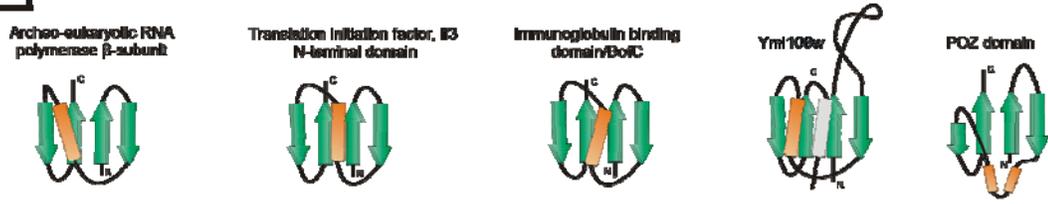
The centrality of Ub conjugation in eukaryotic molecular biology has led to numerous investigations on Ub and Ub-related domains [443, 444]. These studies have resulted in a large body of data on the properties of the Ub-like versions of the β -GF. The key emerging findings were that several other Ub-like proteins (Ubl), such as Urm1 [445], Apg12 [446], Nedd8 [447], and SUMO [448, 449] are also covalently linked to target polypeptides, just as Ub itself. In contrast, some Ub-related domains, like the Ubx domain or Ub-like domains of I κ B kinases, play adaptor roles in Ub-signaling [450-453]. These studies also showed that eukaryotes possess a distinctive enzymatic apparatus for Ub-modification, comprised of a cascade of three enzymes: E1, E2 and E3. These enzymes successively activated Ub/Ubls for transfer using the free energy derived from ATP hydrolysis, relayed it via thiocarboxylate linkages involving the C-terminal residue of

A.

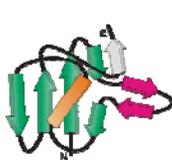


B.

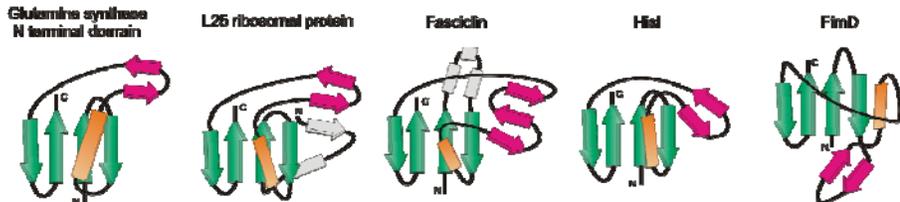
4-stranded barrel assemblage



Nudix (MufT) hydrolases



Four-stranded barrelizing variants



5-stranded Assemblage

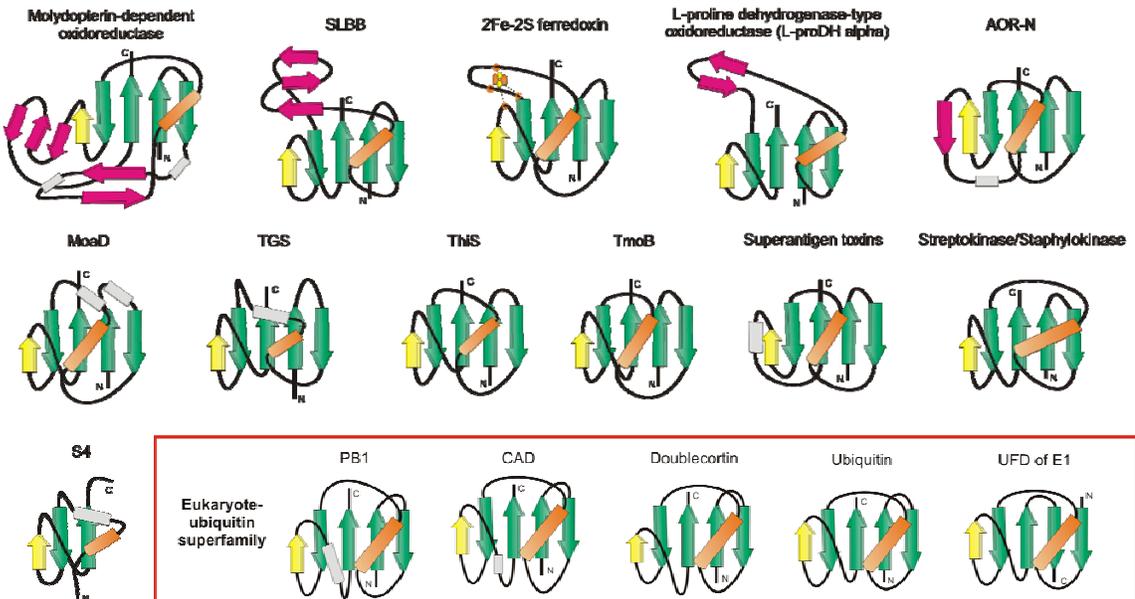


Fig. 25. Topology diagrams of selected β -GF members.

A generalized representative is shown in (A) with the key structural features found in certain lineages of the fold labeled, while (B) depicts idealized versions of specific lineages, the names of which are given above the diagrams. Strands are shown as arrows with the arrowhead at the C-terminal end. Strands belonging to the 4-stranded β -GF core are colored green, the additional strand found in the 5-stranded assemblage is colored yellow, strands forming a conserved insert within the β -GF scaffold are colored magenta, and other strands specific to a certain lineage are colored grey and outlined with a broken line. Helices are depicted as rectangles, with the core absolutely conserved helix colored orange and other helices specific to a certain lineage colored grey and outlined with a broken line. The diagrams are grouped and labeled in a manner consistent with the structural classes described in the text, with members of the eukaryotic Ub-like superfamily nested within other members of the 5-stranded assemblage. The 2Fe-2S cluster of the ferredoxins is shown as four small ovals bound to cysteine residues represented by the letter "C".

Ub/Ubls, and finally transferred it to lysines on target polypeptides [438, 444, 454-456].

Eukaryotes were also shown to contain an elaborate apparatus for removal of covalently linked Ub/Ubls and proteasomal degradation of Ub-modified proteins [457-461].

Concomitantly, structural studies also uncovered several new versions of the β -GF in a variety of domains, greatly widening its horizon of biological functions. Examples of such β -GF domains are: 1) the TGS domain, an RNA-binding domain found in aminoacyl tRNA synthetases and other translation regulators (PDB: 1QF6 [462, 463]). 2) The doublecortin (DCX) (PDB: 1MJD [464]), RA (PDB: 1C1Y [465]), PB1 (PDB: 1IPG [466]), and FERM N-terminal domains (PDB: 1EF1 [467]), which function as adaptors in animal signaling proteins and apoptosis regulators by mediating protein-protein interactions. 3) The soluble ligand-binding β -GF (SLBB) domain involved in binding vitamin B₁₂ and other solutes in animals and bacteria (PDB: 2BBC, 2FUGS [468-470]). 4) Various toxins related to the staphylococcal enterotoxin B including superantigens involved in the toxic shock syndrome (PDB: 1ESF [471]). 5) Functionally obscure subunits of various enzymatic complexes, like TmoB of the aromatic monooxygenase oxygenase complex (PDB: 1T0S [472]) and RnfH of the Rnf dehydrogenases [40]. 6) Conserved domains, perhaps

involved in RNA binding, in the archaeo-eukaryotic RNA polymerase RPB2 subunit [473] and bacterial translation initiation factor IF3 (PDB: 1TIF [474-476]). 7) Staphylokinases and streptokinases which are fibrinolytic enzymes of low GC Gram-positive bacteria (PDB: 2SAK [477]). 8) MutT/nudix enzymes- a group of phosphohydrolases acting on diverse substrates [478]. These observations suggested that the β -GF is indeed a widely utilized structural scaffold, with an underappreciated versatility and an evolutionary history rich in adaptive radiations.

One notable evolutionary question in this regard was the origin of eukaryotic Ub and its relationships to other domains with the β -GF. The first major advances in this direction came with the identification of the sulfur transfer proteins, ThiS and MoadD, respectively involved in thiamine and Molybdenum cofactor (MoCo) biosynthesis, which contained β -GFs closely related to Ub [412], [434]. Furthermore, it was demonstrated that their C-terminal residues formed thiocarboxylates, just like Ub, and this was catalyzed by enzymes (ThiF and MoeB), which are very similar to the E1 enzymes involved in Ub-conjugation [409, 410, 412, 413, 434]. More recently, research from our group showed that the Ub-conjugation systems might not be an exclusive feature of eukaryotes (see part 'C' of this section of the dissertation). Proteins with Ub-like β -GF domains, and functionally linked enzymes related to E1, E2 and deubiquitinating peptidases of the JAB domain superfamily were found in several, phylogenetically diverse bacteria. Evidence was presented that though some of these systems are likely to be involved in sulfur transfer reactions in metabolite biosynthesis, akin to ThiS and MoadD, others might potentially function as *bona fide* conjugation systems that transfer β -GF proteins to target polypeptides [40]. Hence, the eukaryotic Ub-conjugation system might have evolved from more ancient precursors that were present in bacteria prior to the origin of eukaryotes.

With some clarity emerging on issue of the origin of Ub/Ubls and the associated biochemical networks, I sought to investigate the broader issue of the adaptive radiations of the entire β -GF. In particular we were interested in a number of problems from structural and evolutionary stand points: 1) Establishing the entire gamut of structural and topological variations that have emerged in the β -GF. 2) Identifying any unifying structural themes that might exist across most or all functionally diverse versions of the fold. 3) Determination of the lineage-specific sequence-structure correlates for the varied functional adaptations of the β -GF. 4) Developing a higher order evolutionary classification for the β -GF and using it as a scaffold to identify the major temporal phases of adaptive radiation. 5) Identifying instances of drastic shifts in biological or biochemical functions in specific monophyletic lineages of the β -GF. One example of such a functional shift is seen in the evolution of the classical Ub-like proteins, where a unique post-translational modification system emerged from a core metabolic sulfur transfer system. 6) Identifying previously unrecognized members of the fold, if any, and thereby expanding the functional spectrum or providing a rationale for function prediction of uncharacterized members of the fold. 7) The lab was also hopeful that the β -GF might provide a model for understanding the more general problem of how certain small protein folds tend to be extensively deployed in a whole diversity of functional contexts.

In this article the results of our systematic analysis of the β -GF are presented, with the objective of addressing the above points.

Application of Methods

Initial DALI searches using β -GF members as queries were performed by Dr. S. Balaji in the lab, while the mapping of the contacts formed between β -GF members and their interacting partners were performed by Drs. Iyer and Balaji, with assistance from me (see below). I

performed the remainder of the analyses detailed below, including genome contextual, sequence, and phylogenetic analyses; as well as genome-relative fold complexity comparisons (see below). Drs. Iyer and Aravind provided input and guidance at various stages of the investigation.

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda, MD) was searched with the BLASTP program [38]. Profile searches were conducted using the PSI-BLAST program [38] with either single sequences or multiple alignments as queries, with a profile inclusion expectation (e) value threshold of 0.01; searches were iterated until convergence. Hidden Markov models (HMMs) built from alignments using the hmmbuild program were also employed in searches carried out using the hmmsearch program from the HMMER package [63]. For queries and searches containing compositionally biased segments, the statistical correction option built into the BLAST program was used [70]. Multiple alignments were constructed using the MUSCLE [78] and/or T-COFFEE programs [60], followed by manual adjustment based on PSI-BLAST hsp results and information provided by solved three-dimensional structures. All large-scale sequence and structure analysis procedures were carried out with the TASS software package (V. Anantharaman, SB and LA, unpublished results), a successor to the SEALS package [59]. Protein structures were visualized using the Swiss-PDB viewer [94] and cartoons were constructed with the PyMOL program [187]. Protein secondary structure predictions were made with the JPRED program [89], using multiple alignments as queries. Phylogenetic analysis was carried out using a variety of methods including maximum-likelihood, neighbor-joining, and minimum evolution (least squares) methods. Maximum-likelihood distance matrices were constructed using the TreePuzzle 5 program [65] and were used as input for the construction of neighbor-joining with the Weighbor program [66]. Additionally, trees were constructed using the neighbor-joining and minimum

evolution methods as implemented in the MEGA program [97] and the Bayesian inference method using Markov chain Monte Carlo simulations implemented by the MRBAYES program [479].

Structure similarity searches were conducted using the standalone version of the DALI program [95, 96] with the query structures scanned against local current version PDB that has all chains as separate entries. The structural hits for each query was collected, even if the DALI Z-score for the match was less than 2.0 and parsed for topological congruence to the β -GF template (Table 4) using a custom PERL script. To assess topologically congruence, coordinates of the matching regions detected by DALI searches using known β -GF domains as queries were extracted and analyzed for secondary structure using DSSP program. These secondary structure elements were then represented as a string (corresponding to a row in table 1) along with the polarity of the secondary structure element determined from the DALI match to the query structure. These strings were then matched with the equivalent secondary structure pattern strings constructed of *bona fide* β -grasp domains. If a complete match was obtained these structures were tagged as congruent, while those which were not were ranked in descending order of elements that did not match. This discrimination of the potential candidates was further confirmed by visual examination of each structure. The interacting residues of various proteins of β -grasp fold with their interacting molecules have been deduced using custom made PERL scripts. The scripts encode interacting distance cut-off values of 5.0 Å and 3.5 Å between appropriate atoms in 3-D for deducing the hydrophobic and polar interactions respectively. These inferred interactions were further examined manually using Swiss-PDB viewer for confirming the contacts between amino-acid residues of β -grasp fold proteins and atomic groups of interacting partners.

Higher-order Classification	Lineage Name	Secondary Structural Features Common to the β -GF Fold ¹												
		S1	L1	S2	L2	H	L3/LS	S3	L4	S4	L5/CA	S5	tail	notes
Basal 4-stranded versions of the β -GF	IF3-N	S1	--	S2	--	H	--	S3	--	O	O	S5	--	
	Archeo-eukaryotic RNA poly. β -subunit	S1	--	S2	--	H	--	S3	--	O	O	S5	--	
Sporadically-distributed 4-stranded versions	Yml108w	S1	cc	S2	--	H	--	S3	--	O	O	S5	h	
	BofC	S1	--	S2	--	H	--	S3	--	O	O	S5	--	
	Immunoglobulin-binding	S1	--	S2	--	H	--	S3	--	O	O	S5	--	
	POZ	S1	--	S2	--	H	h	S3	--	O	O	S5	--	
Nudix superfamily	Nudix (MutT)	S1	--	S(ee)2	--	H	*	S3	--	O	O	S5	e	
Fasciclin-like assemblage	L25	S1	--	S2	--	H	ee*	S3	--	O	O	S5	--	3
	glutamine synthetase N-terminal fasciclin	S1	--	S2	--	H	eee*	S3	--	O	O	S5	--	3
	phosphoribosyl AMP cyclohydrolase (HisI)	S1	hhh	S2	--	H	ee*	S3	--	O	O	S5	--	3
		S1	--	S2	--	H	ee*	S3	--	O	O	S5	--	3,4
5-stranded assemblage: classical 5-stranded clade	MoaD	S1	H	S2	--	H	h**	S3	--	S4	*	S5	--	
	ThiS	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
	TmoB	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
	Superantigen	S1	--	S2	--	H	*	S3	--	S4	h*	S5	--	
	Strepto/Staphylokinase	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
	YukD	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
	TGS	S1	--	S2	--	H	h*	S3	--	S4	*	S5	--	
Aldehyde OR ² N-terminal domain	S1	--	S2	--	H	*	S3	--	S4	eh*	S5	--		
5-stranded assemblage: Selected eukaryote UB-like superfamily members	classic UB-like	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
	PB1	S1	--	S2	--	H	*	S3	--	S4	h*	S5	--	
	CAD/Doublecortin (DCX)	S1	--	S2	--	H	*	S3	--	S4	[h]*	S5	--	6
	RA	S1	--	S2	--	H	*	S3	--	S4	h*	S5	--	
	Elongin	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
	UBX	S1	--	S2	--	H	*	S3	--	S4	*	S5	--	
5-stranded assemblage: soluble ligand binding or metal ion chelating clade	E1/UFD	O	--	S2	--	H	*	S3	--	S4	*	S5	S6	7
	molybdopterin-dependent oxidoreductase	S1	--	S2	hehee	H	*	S3	--	S4	eee*	S5	--	
	SLBB: Nqo1-type	S1	--	S2	--	H	*	S3	--	S4	hh*	S5	--	5
	SLBB: transcobalamin-type	S1	--	S2	--	H	eee*	S3	--	S4	*	S5	--	
	2Fe-2S ferredoxin	S1	--	S2	--	H	cc*	S3	--	S4	*	S5	--	
L-proline DH-like OR ² N-terminal domain	S1	--	S2	--	H	ee*	S3	--	S4	*	S5	--		
Miscellaneous	WWE	S1	--	S2	--	H	e*	S3	--	O	O	S5	e	8
	FimD N-terminal	S1	--	S2	ee	H	*	S3	--	O	O	S5	--	
	S4	O	O	O	O	H	h*	S3	--	S4	*	S5	--	

Table 4. Secondary structure features of major β -GF structural categories.

1. S: Strand, L: Loop, H: Helix, LS: Lateral Shelf, CA: Connector Arm, O: absence of given feature, --: presence of a loop feature, *: presence of LS or CA, h: insert in helical conformation, e: insert in extended conformation (strand-like), cc: long coil insert.

2. OR: oxidoreductase.

3. Versions form barrel through insertion of strands at the lateral shelf.

4. Barrel is less pronounced in this version; strands are inserted more upstream relative to the other 3 versions.

5. Two small helices are present in ascending arm.

6. Single helix found at ascending arm in several members.

7. Circular permutation results in new connections between strands; the S1 strand is found at C-terminus (See Fig. 25, 26).

8. Additional strand at tail inserted between S1 and S5; lateral shelf forms strand that also stacks with central sheet.

Results and Discussion

Identification of β -GF domains

As the β -GF is small in size and its representatives very divergent, it is not possible to exhaustively identify all members through sequence or structure similarity searches initiated from a single starting point. Accordingly, a multi-pronged strategy of sequence, structure, and topological similarity searches was used. All the currently available structures of β -GF proteins from the Protein Data Bank (PDB) [31] were used as a starting point. This set was compiled by collecting all structures already classified under the β -GF in the SCOP database [423], their relatives from the PDB database that are not present in SCOP, and new versions which were detected in our recent studies [468], [40]. These representatives were used as seeds for initiating sequence profile searches of the NCBI NR database with the PSI-BLAST program [38] (see materials and methods for details). Statistically significant hits ($e < 0.01$) recovered in these searches were used to generate alignments for further HMM searches of individual genome databases and representatives used for transitive PSI-BLAST searches of the NR database. All newly-identified clusters of domains distinct from previously identified sequence families containing the β -GF were aligned and used to predict secondary structure with the JPRED program [89]. The predicted secondary structure and the conservation pattern were superimposed onto the secondary structure and conservation patterns of the known β -GF sequence families to ascertain the validity of the newly-detected versions (see Additional file 1 for alignments and complete list of recovered sequences).

All available structures of *bona fide* β -GF domains were compared in order to establish a unique core template topology that discriminated the β -GF from all other folds (Fig. 25A; Table 4; see below for further details). Then the representative structures of β -GF domains were used as

queries to search a local current version of the PDB database for structurally similar domains using the DALILITE program [95, 96]. All hits were evaluated through reciprocal DALILITE searches of the PDB database to determine if their best matches included any known β -GF proteins. The hits were also further evaluated for congruence to the unique topological template. In addition to the match to the core structural template, all unique features of each newly-detected structure were systematically documented. Through these searches, around ten previously unknown families/superfamilies of domains containing the β -GF were identified, including certain structurally distinctive variants. Comparisons of the distributions of previously characterized globular domains in proteins from sequenced genomes suggests that the procedures have identified a major fraction of conserved lineages of the β -GF.

Core conserved topology, structural variation, and derivatives of the β -GF

A comparison of the available β -GF structures revealed a common core of 4 strands forming an anti-parallel sheet, and a single helical region (see Table 4, Fig. 25A). The characteristic topological feature is that the first and last strands are adjacent and parallel to each other, and the remaining two strands of the conserved core are anti-parallel and flank the former two strands on either side. The first and last strands are invariably located in the center of the sheet with a cross-over occurring via the single helical element. This helical region is packed against one face of the sheet, typically leaving the other face exposed. The chief interacting positions between sheet and the helical segment and the pattern of key stabilizing hydrophobic interactions are conserved throughout the fold, supporting its monophyletic origin. The β -GF domains found in IF3 and the second largest subunit (β -subunit orthologs) of the archaeo-eukaryotic RNA polymerase more or less correspond to this conserved core (Fig. 25B). Several β -GF domains display simple structural elaborations of this basic 4-stranded core. The simplest of

these is the seen in a small family of yeast proteins typified by Yml108w from *S. cerevisiae* (PDB: 1N6Z [480]). This version has a large insert between the first two strands and an additional helical extension at the C-terminus (Fig. 25B). Another notable variant of the basic 4-stranded form of the β -GF domain is seen in the catalytic domain of the NUDIX (MutT) hydrolases. Here, the middle of the second strand of the conserved core is interrupted by a peculiar insert that projects out to form a distinctive “outflow”. This outflow often assumes a hairpin-like configuration stabilized by hydrogen bonding between segments in an extended conformation (Fig. 25B).

All other versions of the β -GF are characterized by major modifications to the 4-stranded core in the form of distinct inserts that add new secondary structure elements. The first of these is a previously uncharacterized variation containing an insertion of one or more strands between the helical segment and strand 3. The conserved inserted strand seen in all domains with this version forms a hairpin with the connector segment between the helical segment and strand 3 which also assumes an extended conformation. This hairpin, together with any additional strands in the insert results in these versions of the fold assuming barrel-like structures with differing degrees of openness (Fig. 25, Table 4). Examples of this version of the β -GF domain are observed in the ribosomal protein L25 (PDB: 1B75 [481]), fasciclin (PDB: 1O70 [482]), and glutamine synthetase (PDB: 1LGR, 2GLS [483, 484]). Yet another novel variant of the β -GF was recovered in the N-terminal domain of the periplasmic pilus assembly protein FimD (PDB: 1ZE3, chain D [485]). This version is typified by a unique insert N-terminal to the helical segment which results in the formation of a barrel-like configuration comparable to the above structural variants.

The most common version of the β -GF is typified by the presence of an additional strand that packs against the conserved third strand at the margin of the core β -sheet. The acquisition of

this additional strand has resulted in the emergence of a connector arm that joins it to the terminal conserved strand of the core sheet (Fig. 25, Table 4). All ubiquitin-like β -GF domains, including sulfur carrier proteins like Moad and ThiS, contain this 5-stranded version of the fold. The connector arm is variable in structure and length and assumes a wide range of conformations ranging from coils to structured elements in different versions of the fold (Fig. 25B, Table 4). A derivative of this Ub-like 5-stranded version is found as a C-terminal domain (UFD) in most eukaryotic E1 Ub-conjugating enzymes [417, 418]—here a circular permutation appears to have displaced the N-terminus to the C-terminus. Given that the N- and C-terminal strands of the β -GF are adjacent to each other, the C-terminal strand in the permuted version occupies the same position as the N-terminal strand of the classical versions, but is oriented in the opposite direction (Fig. 25, Table 4).

The 5-stranded versions may show further variations due to inserts at different points in the conserved core. One prominent example is the 2Fe-2S ferredoxin, which contains an insert before the third conserved strand with conserved cysteines for chelating the Fe ion. Similarly, a long insert adopting an extended conformation is observed at a comparable position in several versions of the SLBB domain [468] and the molybdopterin-dependent oxidoreductases (Fig. 25B, Table 4). In the SLBB domain, the curved β -strands from the insert along with the strands of the β -GF domain core contribute to the formation of a barrel-like structure (PDB: 2BBC [468]). In the middle domain of molybdopterin-dependent oxidoreductases (PDB: 1SOX [486], chain A) there is an additional insert of 2 β -strands associated with the connector arm, which results in an even more complex 3-layered structure, with the two inserts forming a barrel-like element within it. Another previously unknown variant is seen in the N-terminal domain of the aldehyde oxidoreductases (AOR-N) (PDB: 1AOR [487]), wherein the connector arm assumes an extended

conformation and packs as an additional strand at the fringe of the β -sheet adjacent to the strand-4 which is a specific feature of the 5-stranded versions (Fig. 25B). In the AOR-Ns, two of these variant β -GF domains stack via the exposed surface of the β -sheet and form a 4-layered sandwich module.

Our structure similarity searches identified a few structures which, despite lacking the core conserved topology of the classical β -GF, aligned well with a part thereof. Reciprocal searches indicated that β -GF domains were the best hits for these structures. Additionally, these structures were not representatives of any other previously identified folds. These structures include the S4 RNA-binding domain (PDB: 1c05 [488]), the WWE domain (PDB: 2A90 [489]), and the POZ domain (PDB: 1BUO [490]). Previous structural studies had noted a region of local structural similarity, termed the α -L motif, between the S4 and the TGS domain [491]. Given the functional similarity (RNA-binding) and close structural congruence between the shared elements of these two domains, it is quite likely S4 domain is a degenerate variant of the 5-stranded TGS-like β -GF domain, which has emerged through partial loss of the N-terminal part of the domain including the first two strands. The WWE domain and the POZ domain are found only in eukaryotes [40], suggesting that they could have potentially emerged from pre-existing folds through rapid divergence. Given its general structural similarity with the β -GF domains, it is likely to have been derived from the 5-stranded version of this fold. The WWE domain appears to have acquired an additional strand after the terminal strand which is inserted in the middle of the core sheet. The pre-strand 3 region in this domain also adopts a peculiar structure which makes it appear very different from the classical β -GF domains. In contrast, the POZ domain appears to have been derived from a 4-stranded β -GF domain through different degrees of degradation of the penultimate strand on the fringe of the sheet.

Natural classification of β -GF domains

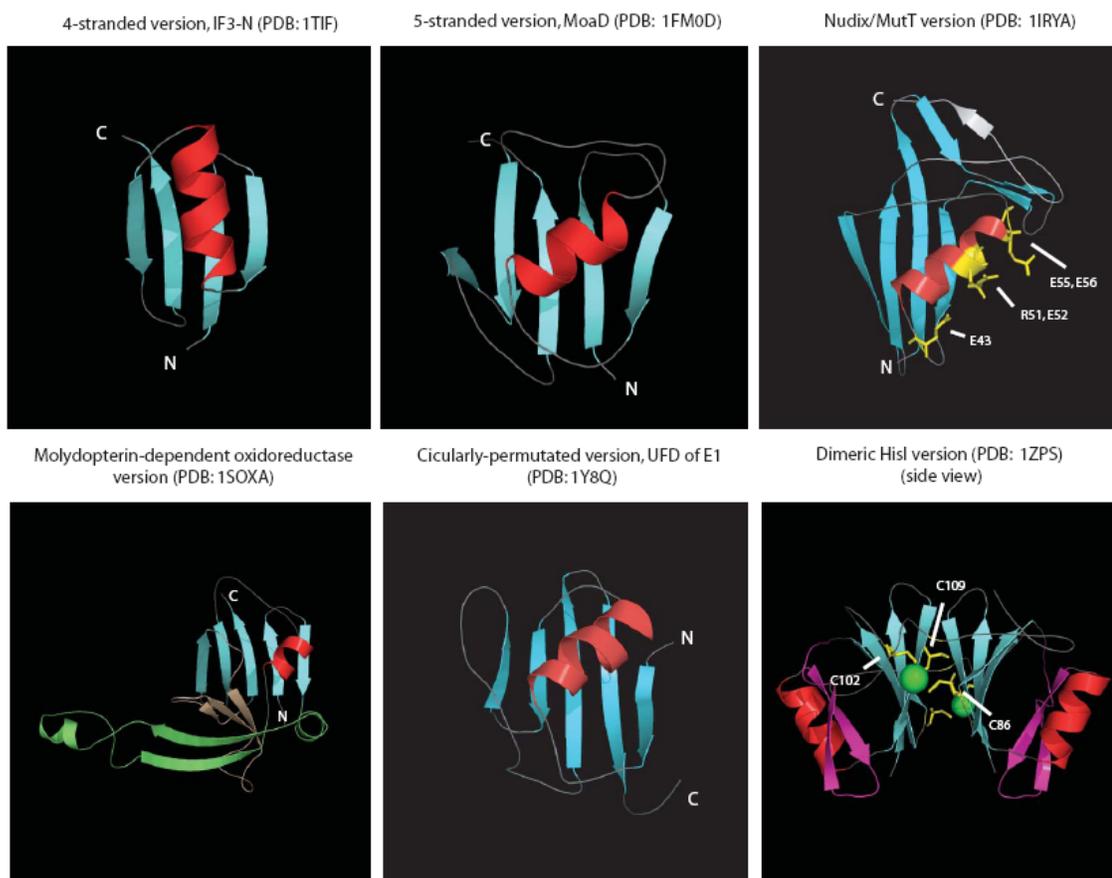
In order to address the prime evolutionary questions about the β -GF, I attempted to construct a classification that most closely approximates the higher-order evolutionary relationships of the members of this fold. The small size of the majority of the versions of this domain often precludes sufficient resolution of relationships using conventional phylogenetic tree methods, sometimes even within superfamilies that display significant sequence similarity. This difficulty is further compounded by the extreme sequence divergence even between versions having highly similar tertiary structures (e.g. ubiquitin and ThiS). Hence, I had to rely to a greater extent on structure similarity-based clustering, shared derived structural characters, and phyletic patterns of sequence superfamilies to reconstruct the evolutionary history. Thus, I produced the classification using the following general steps: 1) sequence similarity-based clustering with the BLASTCLUST program [186] helped in identifying the cores of all major sequence families of β -GF domains. 2) Subsequent comparison of the individual sequence conservation profiles led to the establishment of the most inclusive higher-order assemblages of these families (termed superfamilies) based on shared derived features. 3) The next level of relationships beyond what could be resolved through sequence comparisons was established using structural similarity. This was done both by constructing distance trees based on pairwise Z-scores for structure similarity and deriving the most parsimonious tree based on shared structural features (see Table 4 for major structural features). This procedure, while allowing reasonable resolution of the higher-order relationships, might on occasions produce relatively flat hierarchies for lower-level clusters where none of the methods offer reliable resolution of relationships. A summary of this classification is presented in Fig. 26, 27 and Additional file 2.

Basal versions and other sporadically distributed 4-stranded versions of the β -GF

The above analysis of the structural diversity of the fold suggests that the 4-stranded version is the simplest form from which all other versions could have been derived through accretion of inserts and additional secondary structure elements. Two structurally close superfamilies of the 4-stranded β -GF domain, namely the IF3-N and the archaeo-eukaryotic RNA polymerase domain, are respectively universally conserved in the bacterial and archaeal-eukaryotic branches of life. This, taken together with their shared general functional connection

Fig. 26. Cartoon representations of distinct β -GF domains.

Critical residues in MutT and HisI that are involved in enzyme catalysis are also shown.



representative of the β -GF is likely to represent one of the most basal lineages of the fold. The remaining sequence clusters BofC, yeast Yml108w, and immunoglobulin-binding proteins of low GC Gram-positive bacteria with structurally comparable, simple 4-stranded β -GF domains show extremely limited phyletic patterns (Additional file 2), suggesting a probable recent derivation from the more ancient versions. The versions in the Ig-binding proteins and BofC are restricted to Gram-positive bacteria, and the former might have been derived in pathogenic forms from BofC, which is a secreted developmental signaling molecule widely distributed in free-living Gram-positive bacteria [492, 493]. The eukaryote-specific POZ domain might represent another derivative of a more widely-distributed 4-stranded version, which has accreted an additional C-terminal helical bundle to form a distinctive globular structure (Fig. 25, 26 and Table 4).

The Nudix (MutT) superfamily

The remaining versions of the β -GF fold appear to form a monophyletic clade unified by the presence of an ancestral “lateral shelf” or “flange” that forms an extended connector between the helical segment and the remaining portion of the sheet after the topological cross-over (Fig. 27 and Table 4). Of these versions, the Nudix superfamily appears to be one of the early branches given that its β -sheet retains the ancestral 4-stranded core. All members of this superfamily share the above-described insert or “outflow” in the middle of strand 2 which forms a distinctive shelf for accommodating substrates. This superfamily is also unified by the presence of a conserved PXG motif in strand-2, immediately after the “outflow”, and a unique constellation of conserved residues in the helical segment which form the phosphohydrolase active site [369, 494]. The Nudix superfamily represents a rare instance of adaptation of the β -GF as a scaffold for catalytic activity. Its phyletic patterns suggest an ancestral presence in all three superkingdoms implying that it might have been present in the LUCA (Fig. 27, Additional file 2).

Fasciclin-like assemblage

A structurally distinct subgroup of β -GF domains which was uncovered as a result of our analysis unifies previously unrecognized versions of the fold, namely the fasciclin domain (PDB: 1O70 [482]), the ribosomal protein L25 (PDB: 1B75 [481]), and the phosphoribosyl AMP cyclohydrolase (HisI) (PDB: 1ZPS [495]) with the glutamine synthetase N-terminal domain. The unique insert and associated structural peculiarities such as the barrel-like configuration shared by these domains strongly suggests that they form a higher-order monophyletic cluster within the β -GF termed the fasciclin-like assemblage. The similar ligand-interaction patterns seen in most of these lineages also support the monophyly of this assemblage (see below for details). Most characterized sequence superfamilies within this assemblage appear to bind small molecules or soluble ligands. The fasciclin domain binds sugar moieties of cell-surface glycoproteins [482], the HisI domain binds phosphoribosyl AMP [496], and the glutamine synthetase N-terminal domain contributes to the substrate binding pocket of the enzyme [483]. The L25 domain binds 5S RNA (PDB: 1DFU [497]), although there is no evidence that it does so in a comparable manner as the other members of this assemblage. Given the above observations, it is possible that the ancestral version of this assemblage had small-molecule binding capabilities. Despite the distinctive structural innovations, the conserved core of the β -GF domain in this assemblage is a 4-stranded version with a “lateral shelf” suggesting that it represents an early branch of the clade unified by the latter derived feature (Fig. 27). Of the sequence superfamilies of this assemblage, the glutamine synthetase N-terminal domain is traceable to LUCA. Hence, the fasciclin-like version of the β -GF domain might have diverged from other major lineages of the fold prior to LUCA.

The FimD superfamily, while containing a unique structural variant of the fold, shows greatest structural similarity to the above assemblage. Its phyletic pattern is limited, being found only in proteobacteria and deinococci (Additional file 2). Thus, it could have been derived from the above assemblage in a lineage-specific manner.

The 5-stranded assemblage

The 5-stranded assemblage is unified by the addition of the fifth strand to the core sheet and the consequent emergence of the “connector arm” linking the additional strand to the terminal strand (Fig. 25A). The strong conservation of this unique structural feature, in conjunction with the exclusive grouping of these versions in structure similarity-based clustering, suggests that they form a monophyletic assemblage. This version of the fold is most prevalent, both in terms of number of distinct superfamilies contained within it and universal representation found across all life forms. At least 4 monophyletic lineages of this assembly, namely the TGS domain, the ThiS and MoaD proteins, and the 2Fe-2S ferredoxins can be traced to LUCA. Beyond these, there are several lineages that are conserved in a single superkingdom or distributed more sporadically within a superkingdom. On the whole, two major clades can be recognized within the 5-stranded assemblage. The first of these, termed *the classical 5-stranded clade*, unites the three ancient lineages TGS, ThiS, and MoaD and several other closely-related versions. This clade is also supported by the presence of a highly conserved alcoholic residue at the transition between the N-terminal hairpins and the helical segment of the fold [40]. The UB-like β -GF domains are derived from the ThiS and MoaD-like versions and comprise the most diverse superfamily within the classical 5-stranded clade.

Eukaryotic representatives of the UB-like superfamily β -GF domains

In eukaryotes, this superfamily has undergone explosive diversification with at least 19-20 distinct families which can be traced back to the last eukaryotic common ancestor (LECA). These families include six conjugated versions (ubiquitin, Urm1, Apg8/Aut7, Apg12, Ufm1 and SUMO/SMT3) [498, 499] and several known or predicted to function as adapters in multi-domain proteins, like the tubulin cofactor B (TBCB) [500], Ub/Ubl conjugating E1 enzymes [417, 418] and phosphatidylinositol 3 kinase (PI3K) [501]. Overall, in the course of eukaryotic evolution, at least 67 distinct sequence families appear to have emerged within this superfamily with some restricted to particular eukaryotic kingdoms like animals or plants. Several previously uncharacterized eukaryotic families were identified such as NPL4p, the UB-like domains of the BMI1/Posterior Sex Combs family of chromatin associated E3 ligases, a family with the UB-like domain fused to a cytochrome b5 domain, and the auxin response factor (BIPOSTO) in plants (see Additional file 1 for alignments). On the whole, comparisons of sequence conservation profiles showed that β -GF domains related to the classical ubiquitin domain form a large monophyletic assemblage within the superfamily, including several distinct families such as Nedd8, SUMO, ubiquitin, NPL4, BAG, the Ubx domain, the tubulin co-factors or chaperones (TBCB and TBCE), Bat3/Dsk and Apg12/Gate16 (Fig. 27). The circularly permuted C-terminal UFD of eukaryotic E1s, which distinguishes them from the prokaryotic E1-related enzymes, was also likely derived from this lineage. Sequence comparisons also showed that the RA, FERM N-terminal module, and PI3K adapter domain families form another distinct higher-order monophyletic lineage. The remaining lineages typified by ECR1/UBA1 and BM-002, while structurally close to the rest, formed distinct sequence families that could not be placed into any of the above larger assemblages of families (see Additional file 2 for details).

Bacterial representatives of the UB-like superfamily and the classical 5-stranded assemblage

In bacteria, Ub-like superfamily includes several sporadically distributed UB-like families which have been previously described in considerable detail [40]. Several other sporadic bacterial lineages also belong to the classical 5-stranded clade, such as the fibrinolytic adapters of several Gram-positive bacteria (e.g. streptokinase), the superantigen/toxin domains, the RnfH proteins and subunits of aromatic compound monooxygenases like TmoB. Our searches also identified a previously unknown version of the classical 5-stranded clade in a group of bacterial flagellar assembly proteins typified by FliD, FlgL and FlgK, and related bacteriophage-tail proteins found in a range of Mu-like caudoviruses (see Additional file 1). Sequence searches indicate that RnfH is closest to the TGS domains and is likely to be an offshoot of that superfamily (Fig. 27). The superantigen/toxin versions and the streptokinase/staphylokinases appear to form a monophyletic cluster, as they are both secreted versions and interact with substrates similarly (See below). However, barring RnfH, the exact relationships of these more sporadic bacterial lineages to the more ancient lineages of the classical 5-stranded clade remain unclear.

The soluble ligand or metal-binding clade of the 5-stranded assemblage

The second major clade of the 5-stranded assemblage unifies a group of β -GF domains whose interrelationships were previously unknown. This clade is unified by the presence of a set of inserts that are associated with binding soluble ligands or chelating metal ions. While the inserts themselves are poorly conserved in sequence, their position, especially in relation to the bound ion or ligand, is well conserved. The main sequence superfamilies in this clade are the 2Fe-2S ferredoxins, the SLBB domains, and the molybdopterin-dependent oxidoreductase domains. As recently shown, the SLBB superfamily is of bacterial provenance [468]. The molybdopterin-dependent oxidoreductases, typified by the sulfite oxidase (SOX), are widely distributed in all the three superkingdoms but show no evidence in phylogenetic analysis for being present in LUCA.

Given that the eukaryotic versions localize to the mitochondrion [502], they appear to have probably been derived from the bacterial progenitor of the mitochondria. The N-terminal domain of the L-proline dehydrogenase-type oxidoreductase (PDB: 1Y56 [503]) is another family of proteins belonging to this clade of the 5-stranded assemblage. Sequence profile analysis showed a statistically significant relationship between these domains and the 2Fe-2S ferredoxins, suggesting that they belong to the same superfamily. They appear to have been derived from the more universally distributed 2Fe-2S ferredoxins through loss of the metal-chelating conserved cysteines relatively early in bacterial evolution.

The N-terminal module of the aldehyde oxidoreductases

A distinctive superfamily of the 5-stranded assemblage that was discovered in our analysis was the N-terminal module of the aldehyde oxidoreductase (AOR-N) (PDB: 1AOR [487]) that contains two tandem, distantly related copies of the β -fold. These are unified by the modified structure of their connector arm, ligand-binding and dimerization pattern. This structural modification makes it difficult to identify their affinities to other members of the 5-stranded assemblage. It should be noted that they lack any unique structure or sequence feature unifying them to the sulfite oxidase-like molybdopterin-binding β -GF domains. Hence, it is possible that they arose from a Moad-like precursor that evolved an ability to bind metallopterins specifically (See below). Phyletic patterns indicate a potential bacterial origin for this superfamily. The above-mentioned structural similarity of the universally distributed S4 RNA-binding domain with the TGS domain suggests that the former might be another highly divergent lineage that was derived from a TGS-like classical 5-stranded β -GF domain prior to LUCA.

The relative timeline of major adaptive radiations and functional transitions of the β -GF domains

The pre-LUCA phase and inference of the ancestral function of the β -GF

The inference of at least 7 β -GF or β -GF-derived (the S4 domain) lineages in LUCA suggests that there was a major diversification of the fold even before LUCA (Fig. 27). In structural terms, the inferred representatives in LUCA span all major variants of the fold, from the simplest 4-stranded versions to the barrel-like forms (GS-N domain) to simple and elaborated versions the 5-stranded form. This suggests that the major structural variations were already in place as a result of the early diversification events of the pre-LUCA phase. In functional terms, versions close to the primitive state of both the 4- and 5-stranded forms, the RNA polymerase/IF3-N domain and the TGS domain, respectively, as well as the possible β -GF derivative, the S4 domain, have functions related to RNA metabolism or RNA-binding [463, 476, 504]. Even members of the Nudix clade are known to interact with nucleic acids or chemically-related molecules such as nucleoside diphosphate derivatives [369]. RNA metabolism-associated functions are also sporadically observed in later-derived lineages such as the L25 ribosomal proteins in the fasciclin-like assemblage, the family of prokaryotic UB-related domains fused to the Mut-7C-like RNAses [40], and several eukaryotic UB-like domains like those found in eIF3 p135/Clu-1 (see Additional file 1 for an alignment), RBBP6 (DWNN domain) [505], and prp21/Splicing factor 3 [506]. Given that the at least 4 of the seven main lineages traceable to LUCA, including some of the inferred basal lineages, have a RNA/ribonucleoprotein associated role, it appears likely that the ancestral version of the β -GF was probably involved in RNA-binding. The distribution of RNA-related roles (Fig. 27, 28) implies that this function seems to have been retained or re-acquired in some sense in several later derived versions of the fold.

A corollary to the inference of the ancestral function of the fold is that there were major functional innovations even in the pre-LUCA period. These are most prominently seen in the 5-stranded assemblage, and appear to be associated with the emergence of distinctive roles in sulfur delivery and scaffolding of Fe-S clusters. Previous observations have shown biochemical links between the formation of metal-sulfur clusters and sulfur transfer, including pathways in which ThiS and MoaD-like proteins participate [507]. This observation raises the intriguing possibility that the earliest functional shift involved recruitment of a 5-stranded β -GF domain for a shared general role in both sulfur transfer and generation of Fe-S clusters. It is quite possible that the subsequent specialization of such a generic precursor spawned the two paralogous families of sulfur transfer proteins (MoaD and ThiS) on one hand and the 2Fe-2S ferredoxins on the other. The rise of the 2Fe-2S ferredoxins probably coincided with the emergence of the precursors of the electron transfer chains of respiratory metabolism. The early divergence of MoaD and ThiS suggests that some basic aspects of the biosynthetic pathways for complex sulfur-containing metabolites like molybdenum/tungsten cofactor and thiamine evolved prior to LUCA.

The post-LUCA phase: the prokaryotic superkingdoms

The emergence of the two prokaryotic superkingdoms, the archaea and bacteria, was marked by numerous superkingdom-specific innovations. Several of these innovations appear to have happened early in the history of the bacteria followed by multiple lateral transfers to the archaea. Likewise, innovations occurring in bacteria were also transferred to eukaryotes both during the primary endosymbiotic event and sporadically through later transfers. Members performing some form of most of the biochemical functions observed in extant representatives of the fold emerged in course of the post-LUCA diversification in bacteria. In certain cases there

were no major shifts in basic biochemical activity but only an expansion of the range of specific biological contexts in which these activities were deployed. These included new RNA-binding/ribonucleoprotein-related contexts emerging within diverse branches of the clade (e.g. L25 in the fasciclin-like assemblage and the prokaryotic UB-like family fused to Mut-7C RNase) or adaptation of ThiS/MoaD-type proteins in sulfur transfer systems related to synthesis of lineage-specific metabolites [508]. The principal, early functional innovations in the prokaryotic radiations were the independent acquisition of multiple small molecule/solute-binding capabilities across distant members of the fold, as seen in the SLBB, fasciclin, and AOR-N domains. Another notable feature of this evolutionary phase was the emergence of at least three catalytic versions amongst phylogenetically distant assemblages of the fold. The phosphoribosyl AMP cyclohydrolase of the fasciclin-like assemblage and molybdopterin-dependent oxidoreductase domain related to the 2Fe-2S ferredoxins and SLBB domains are sister groups of the small-molecule binding versions. This suggests that the transition to catalysis probably occurred from an ancestral soluble ligand-binding state. However, emergence of catalysis in the Nudix superfamily appears to be a likely extension of the original nucleic acid-binding properties of the fold.

This phase also saw the recruitment of several forms of the β -GF domain for mediating specific protein-protein interactions in the assembly or stabilization of multi-protein complexes. Different distantly related β -GF domains were recruited in the biogenetic systems of flagella and analogous structures, the pili. The FimD protein has an N-terminal β -GF domain fused to a C-terminal outer membrane-spanning domain [509]. This β -GF domain serves as an adapter to recruit the fimbrial subunit chaperone FimC while the C-terminal domain serves as a platform on which the fimbrial subunits assemble to form the pilus [510, 511]. Likewise, novel versions of the

classical 5-stranded β -GF domain, which was discovered in FliD and FlgL/FlgK, are likely to play roles in the assembly of flagellum (FliD) and its hook (FlgL/K), while their relatives in Mu-like bacteriophages might similarly help in assembly of the viral tail (see Additional file 1). Pathogenic bacteria appear to have sporadically adapted both 4- and 5-stranded versions in roles related to interaction with host proteins as a part of their virulence. The strepto/staphylokinases which interact with plasmin, and the superantigens which interact with vertebrate T-cell receptors [512] from the 5-stranded assemblage and the immunoglobulin-binding domains [513] of the 4-stranded assemblage appear to represent multiple convergent recruitments for virulence-related interactions. The classical 5-stranded clade in particular appears to have given rise to several lineages that seem to function as protein interaction adapters, assembly or stability factors in very different biochemical contexts. For example, the TmoB family might function in stabilizing the proteobacterial aromatic monooxygenase complex [472], different members of the RnfH family might play roles in protein stability or assembly of the Rnf oxidoreductase complex, and YukD in the assembly of the ESAT-type export systems of Firmicutes [40].

However, the most important innovation in the bacteria was the emergence of potential conjugation systems that covalently linked ubiquitin-like β -GF domains to other proteins (predecessors of the eukaryotic conjugation systems). In functional terms, this process represents a collusion of the sulfur-transfer aspect with the protein interaction function which was also widely emerging in members of the fold. The preliminary analysis of these bacterial UB-like systems suggests that they might have already acquired roles related to protein stability and signaling. The details of the bacterial antecedents of the eukaryotic UB-conjugation system have already been discussed in a recent work [40] and are not dwelt upon here.

The eukaryotic phase: expansion of the ubiquitin-like domains

Genomic and cell biological evidence suggests that the eukaryotes emerged as a result of a basic endosymbiotic event between a proteobacterium and an archaeon (most likely a euryarchaeon) [514-516]. Consequently, eukaryotes inherited several versions of the β -GF domain found in both their archaeal and bacterial (mitochondrial) precursors (see Fig. 26 and Additional file 2). The currently available data implies that in eukaryotes there was no diversification of the β -GF domain comparable to what happened in bacterial evolution that resulted in emergence of fundamentally new biochemical activities. Eukaryotes, however, showed an explosive development of the ubiquitin-like lineage resulting in forms that occupied biological functional niches across the entire cell. Most of these functions depend on the ancient property of the classical ubiquitin-like 5-stranded version to mediate protein-protein interactions, particularly in relation to the assembly or stabilization of complexes. These functions were performed either via conjugation of UB/UBLs to target proteins and phosphatidylethanolamine, or as domains within multi-domain proteins. The biochemical diversification of the UB-like clade to perform multiple biological roles appears to have been notable even in LECA (Fig. 28). These adaptations include:

- 1) conjugation to proteins destined for degradation (classical UB).
- 2) Tagging of proteins for altering interactions and localization (e.g. SUMO/SMT3) [448, 449]
- 3) conjugation to both a protein target (Apg5p) and the amino group of the lipid phosphatidylethanolamine (Apg8p/Aut7p) in regulation of the distinctly eukaryotic process of autophagy.
- 4) Possible recognition of proteins with conjugated UB moieties (e.g. NPL4) [517].
- 5) Binding of E2s to present them to the active site of E1s for conjugation of UB/UBLs (the UFD of E1s [417, 418]).
- 6) Assembly of tubulin polymers (TBCB) [500] and microtubule-binding (DCX domains [464]) .
- 7) Protein-protein interactions in Ub-modification (e.g. Ub-like domains in Ub-deconjugating

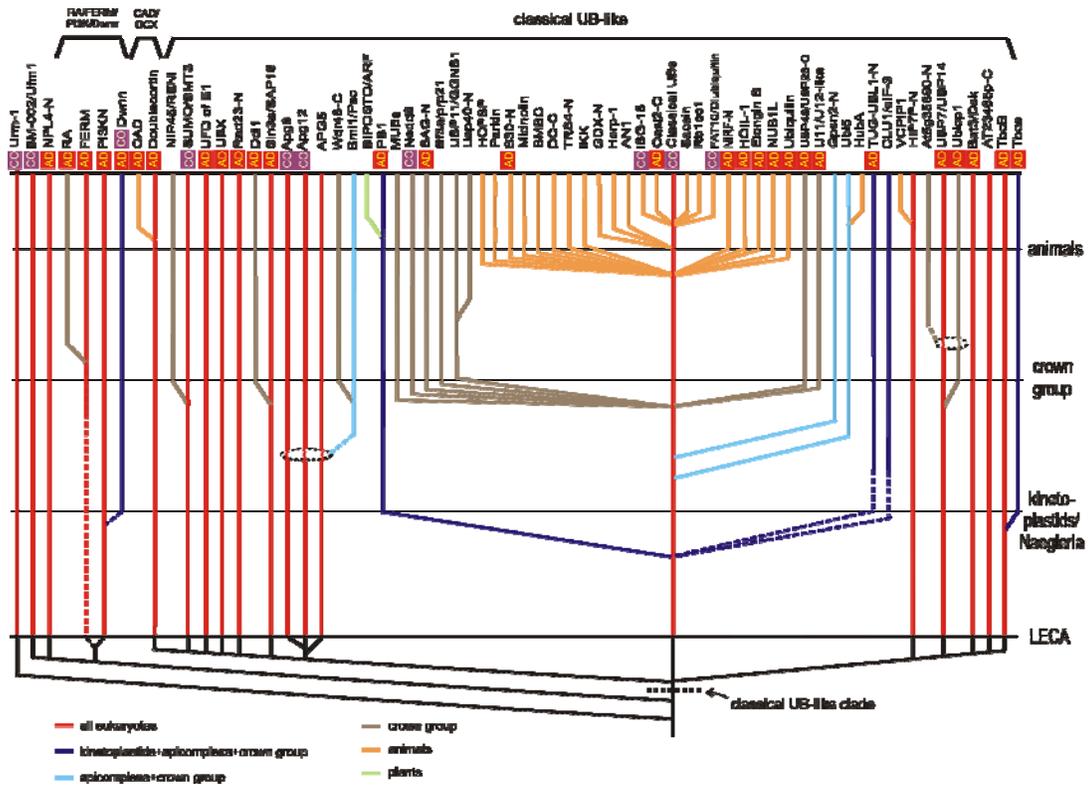


Fig. 28. Reconstructed evolutionary history of eukaryotic ubiquitin superfamily.

Similar to Fig. 27, however, major evolutionary transitions are now shown as horizontal lines and the maximum depth to which these individual lineages can be traced is now shown with solid vertical lines. Functional categories are the same as described in Fig. 27.

enzymes like Ubp7/Ubp14 and the Bmi1/Posterior Sex Combs-like E3s) and other signaling pathways (e.g. PI3 Kinase N-terminal domain) [501]. The ancestral eukaryotic member of the UB-like clade is likely to have been a conjugated version because: 1) conjugated forms are seen across the entire diversity of the eukaryotic UB-like clade, which includes at least 5 versions traceable to LECA and 2) they preserve the basic thiocarboxylate-forming chemistry seen in their even more ancient precursors like ThiS or MoaD. Given the inferred presence of multiple non-conjugated forms in LECA, multiple early functional shifts resulting in non-conjugated appear to have

occurred prior to the divergence of extant eukaryotes from LECA, but after the emergence of the first eukaryotic cell.

Subsequently in eukaryotic evolution, there appear to have been several innovations of non-conjugated versions. Many of these continued to function in contexts related to UB signaling, presumably by recognizing conjugated UB moieties (Fig. 28, Additional file 2). However, a few seem to have acquired entirely unrelated functions; for example, the RA domain in RAS signaling [465] and the CAD domain in apoptotic signaling [518-521]. In temporal terms, a major pre-LECA expansion resulted in at least 19-20 distinct families in the ancestor of extant eukaryotes, followed by new families like the PB1 domain sporadically appearing throughout subsequent eukaryotic evolution. A notable phase of new innovation through sequence diversification resulted in several new families (e.g. Nedd8) prior to the radiation of the eukaryotic crown group comprised of plants, slime molds, fungi, and animals. Interestingly, in the animal lineage alone, there appears to have been another massive round of diversification resulting in more than 10 distinct sequence families. The plants show a lineage expansion of a group of UB-like domains in the BIPOSTO/ARF transcriptional regulators (see Additional file 1) which emerged from the more ancient PB1 family. Thus, in general, there appears to be a correlation between the emergence of new UB-like families and that of multi-domain proteins in the signaling systems of crown group eukaryotes, especially animals [522]. Parallel to this expansion of UB-like domains in eukaryotes, there was also an expansion of other components of the UB-conjugation system such as E1, E2, and E3 enzymes, F-box and UBA domains, and deubiquitinating peptidases [454, 457, 460]. In the eukaryotes there also appears to have been a derivation of at least two domains, namely the POZ and WWE domain through major structural modification of the core β -GF domains.

Evolutionary trends in the domain architectures of β -GF domains

Previous studies on domains occurring in diverse architectural contexts in multi-domain proteins have hinted at a strong relationship between domain architectures and functional constraints [8]. The domain architectures of all β -GF domains were systematically analyzed and their conservation across evolution was used to identify constraints and any role they might have in predicting functions of uncharacterized versions of the domain. Both the sulfur-carrier function and conjugation to other proteins require the free carboxy-terminus of the standalone β -GF domain. As a result, the standalone copies of the 5-stranded UB-like version have been preserved across all three superkingdoms since LUCA. But an alternative strategy to this, observed primarily in eukaryotes, is the generation of free C-termini through post-translational proteolytic cleavage as seen in the polyubiquitins and APG8p (Aut7p). This raises that possibility that there might be other as yet undiscovered versions which are released for conjugation by proteolytic processing, as has been previously proposed for the DWNN domain [505]. In this context, it remains to be seen if the Ub-like domain in the eukaryotic DDI1p-like proteins [40], which is connected via a glycine-rich linker to the rest of the protein (Fig. 29) might be processed by the C-terminal aspartyl peptidase domain release a free UB-like polypeptide.

In contrast, versions involved in protein and nucleic acid interactions are under no major constraints to remain as standalone forms of the domain. Hence, numerous instances of β -GF domains involved in this function occur in multi-domain architectures. The ribosomal proteins tend to be small and usually one or two-domain proteins. Accordingly, there is not much architectural diversity seen in case of forms like L25. The forms found in the DNA-dependent RNA polymerase represent some of the most complex architectures wherein the β -grasp domain is inserted within an RRM-fold domain which in turn is inserted within a larger, multi-domain

scaffold [476]. In most cases, the multi-domain architectures of RNA metabolism-related proteins are well-conserved across entire superkingdoms or even the three superkingdoms of Life because of the universality of these functions in their respective phyletic ranges. Multi-domain architectures associated with signaling or small-molecule interactions are often more restricted in their phyletic range and show lineage-specific diversity [523, 524]. Consistent with this, considerable lineage-specific diversity is observed in prokaryotic β -GF domains involved in small molecule-binding like the cobalamin-binding SLBB domains and fasciclin domains and certain enzymes such as the molybdopterin-dependent oxidoreductases (Fig. 29). All these domains are typically encountered in secreted proteins and form highly variable multi-domain architectures in various bacteria. In some instances two distinct versions of the β -GF domain might occur in the same polypeptide: for example, the fasciclin domain and the molybdopterin-dependent oxidoreductase domains occur in certain secreted enzymes (Fig. 29). Conversely, the small molecule-binding β -GF in certain highly conserved intracellular enzymes like glutamine synthetase and aldehyde oxidoreductases do not show much diversity in domain architectures.

To objectively assess the trends in domain architectural complexity, I made use of the previously devised complexity quotient (CQ) [453]. The CQ provides a measure of the complexity of domain architectures in which a given domain occurs (Fig. 29). Specifically, it is defined as the product of the number of different types of domains that co-occur with β -grasp domain containing proteins and the average number of domains detected in these proteins. The complexity quotient was plotted against the total number of proteins containing β -GF domains in a given organism. This was done for 19 completely sequenced species of prokaryotes and 19 eukaryotic proteomes spanning the entire currently available phyletic spectrum of organisms with sequenced genomes. In the case of prokaryotes the plot reveals a more or less flat line with

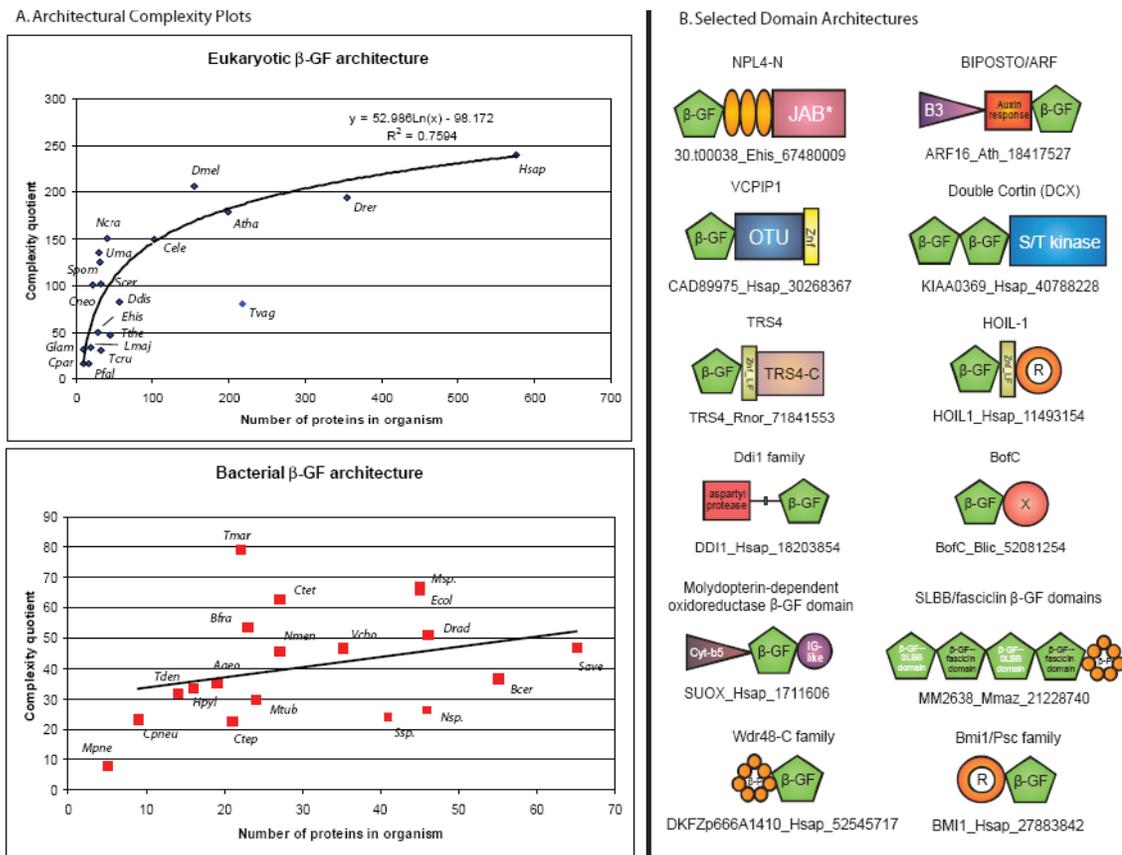


Fig. 29. Architectural complexity plot and novel domain architectures for β -grasp domains.

A) Architectural complexity plot of β -grasp domains found in eukaryotes and prokaryotes. The complexity quotient for a given species (y-axis) is plotted against the total number of β -grasp domain containing proteins in the same species. Names of species are given next to plot points. B) Domain architectures of β -grasp domains. Only a small sample of architectures is shown. These mainly represent novel or recently reported architectures that are described in the text. The TRS4 C-terminal domain, also found fused to certain E1-enzymes that lack the C-terminal UFD has a highly conserved ExxxH implying enzymatic function (see Additional file 1 for an alignment). Orange ellipses represent the conserved cysteine clusters observed in the NPL4-N family (see Additional file 1). A straight line with a small green box in the Ddi1 family architecture represents a possible cleavage site located between the domains. The proteins are not drawn to scale as only globular segments are shown. Explanation of abbreviations/domain names: B3, DNA-binding domain; Auxin response, auxin-responsive transcription factor domain; OTU, OTU-like family of cysteine proteases; Znf, zinc-finger; Znf_LF, little finger family of zinc finger domains; R, Ring-finger domain; β -P, β -propeller domain; X, previously uncharacterized BofC C-terminal domain also found fused to a serine/threonine phosphatase in actinobacteria (see Additional file 1 for alignment).

an approximately constant domain architectural complexity across all prokaryotes, irrespective of the number of β -GF proteins they possessed (Fig. 29). The plot only showed a few anomalous points: there was a greater than expected paucity of β -GF proteins in the highly reduced genome of *Mycoplasma* and an inexplicably high architectural complexity in *Thermotoga maritima*. Thus, barring very few exceptions, the main tendency in prokaryotes is a wide variability in the number of proteins with β -GF domains rather than any concerted increase in architectural complexity.

Eukaryotes not only have greater numbers of β -GF domain proteins, but also appear to display greater diversity of domain architectures relative to the prokaryotes. The complexity of the β -GF proteins as well as their numbers appear to increase throughout eukaryotic evolution with the highest figures observed in multicellular organisms of the eukaryotic crown group. However, the increase in architectural complexity is not linear across eukaryotes, with a tendency to plateau in animals. The only exception to the strong trend is *Trichomonas vaginalis*, a basal eukaryote, which appears to have undergone a massive, relatively recent proliferation across most protein families [525]. As a result it possesses an unexpectedly large number of β -GF proteins, but low architectural complexity comparable to other basal eukaryotes with similar numbers of β -GF-containing proteins (Fig. 29). In terms of actual architectures, the multicellular eukaryotes show numerous lineage-specific multi-domain proteins with different β -GF domains, which are often involved in specific signaling pathways that correspond to unique aspects of the biology of these organisms. For example, the programmed cell death pathways in animals and the auxin-response in plants contain representatives with such unique architectures (Fig. 29) [453].

Typically, many of the eukaryotic multi-domain architectures, both ancient and lineage-specific, tend to combine the UBL domains with other signaling domains, typically those involved in UB-signaling. These combinations include those with deubiquitinating peptidases (e.g. of the OTU superfamily), E3 ligases usually of the RING superfamily (Fig. 29), and other UB-binding domains like UBA, or other kinds of signaling domains like kinases as seen in the IKKs and Doublecortin. Another feature seen in eukaryotic architectures is the architectural variability through domain loss or accretion, even in the case of highly conserved orthologous proteins. For example, the Npl4p family [526] of Ubls is conserved throughout eukaryotes and might play a role as a novel E3 in degradation of proteins in the endoplasmic reticulum. It can be reconstructed as having an ancestral architecture that combined an N-terminal Ubl with a central region containing variable numbers of a novel Zn-chelating cysteine cluster domain and a C-terminal catalytically inactive version of the JAB peptidase domain (Fig. 29, see Additional file 1). In the plant lineage the central Zn-chelating cluster is lost, while in animals and fungi an additional Zn-finger domain is inserted N-terminal to the cysteine-rich Zn-cluster.

Structural correlates for functional diversity in the β -GF

The next step in the investigation was to decipher the relationship between functional diversification and structural elaborations of the fold. For this purpose, an idealized representation of the β -GF fold was created (Fig. 30), dividing the structural elements into equivalent zones that are comparable across available structures. Interactions were then mapped to ligands (see materials and methods for details) in all members for which this data is available onto the above scaffold to obtain an interaction map for the fold (Fig. 30). This interaction map was then used in conjunction with the above developed classification scheme and relative temporal pattern of diversification to explore the evolution of the structure-function

relationships. For the sake of convention, the exposed surface of the core β -sheet is referred to as the “exposed face” and the opposite surface of the sheet which might be obscured by the packing helical segment, the lateral shelf or flange, and the connector arm (in the 5-stranded versions) as the “obscured face”. The C-terminal most portion of the final strand is referred to as the “tail”.

Little is known of the exact mode of interactions of the basal 4-stranded versions of the fold. However, the apparent rarity of the simple 4-stranded versions suggests that there appears

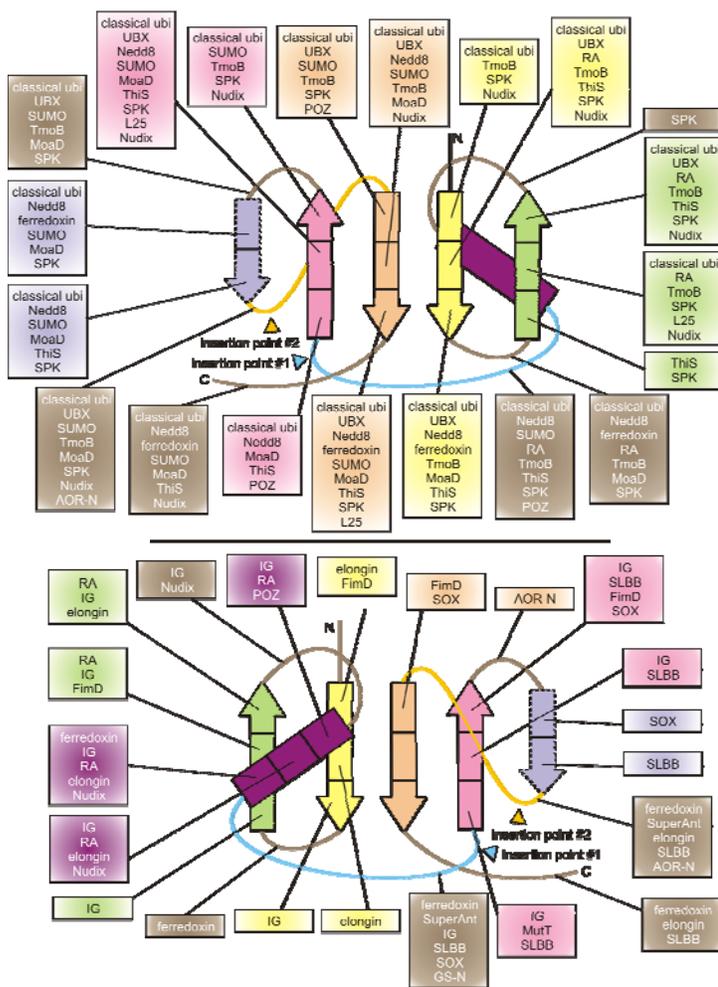


Fig. 30. Diagram of relative location of β -grasp interacting partners.

The strands and core helix of an idealized β -GF domain have been broken into interaction zones, and the names of representatives of the fold that interact using each of these zones is listed. The top view depicts the exposed face while the bottom view depicts the obscured face. Coloring of the boxes containing lists of specific β -GF domains interacting via a particular region correspond to coloring of structural elements (i.e. a particular strand or loop) involved in the interaction.

to be a tendency to elaborate the core sheet to provide an increased interface for interactions. On the

whole, the exposed face mediates more interactions across the β -GF fold compared to the obscured face. Thus, the proliferation and widespread utilization of the 5-stranded version might

be associated with the availability of a larger surface on the exposed face for mediating contacts. Another evolutionary trend is the formation of a barrel-like configuration through insertion of strands which on instances provides a classical interaction interface at the open end of the barrel. The more specific themes of interaction that were observed in multiple superfamilies of the fold are discussed below.

Solute interaction in the fasciclin-like assemblage

As discussed above, the prevalence of soluble ligands such as sugars, amino acids, and metabolic intermediates for different sequence superfamilies of the clade suggested an ancestral solute-binding role for these proteins. Analysis of the interactions with respect to the shared structural core of this assemblage suggests that the insert and the lateral shelf form an interface for soluble ligand interaction in fasciclin, GS-N, and phosphoribosyl-AMP cyclohydrolase domains [483, 496, 527]. Furthermore, in glutamine synthetase this interaction might indirectly contribute to catalysis via a conserved aspartate from this region that interacts with the substrate bound at the active site and helps in anchoring it there. This suggests that ancestral versions of this assemblage probably mediated a generic ligand interaction via a similar interface. The interactions of the L25 domain via this interface, if any, remain unknown. However, it is known to contact 5S rRNA via the exposed face [497]. FimD, which appears to be a distant relative of the fasciclin-like assemblage, assumes a classical barrel configuration, with the “open-end” of the barrel providing an interface for interacting with the FimC immunoglobulin domain [485]. Similar “open-ends” of topologically unrelated barrels like the OB fold, PRC, and SH3 barrels are known to mediate interactions with ligands in a like manner [528-530]. The loop between the penultimate two strands of the core FimD β -GF domain is one of the major determinants of the

interaction and this feature is comparable to certain interactions of the phosphoribosyl-AMP cyclohydrolase domain (see below).

Metal chelation, solute interaction, and prosthetic group attachment in the SLBB/ferredoxin/molybdopterin-dependent oxidoreductase clade

The unifying inserts of this clade typically occur in the region prior to strand 3 and in the region associated with the connector arm or the additional strand of the 5-stranded core. However, there is considerable diversity in the means by which these inserts mediate specific interactions, both between and within different superfamilies of this clade. The 2Fe-2S ferredoxins contain a characteristic set of four cysteines, three of which come from the pre-strand 3 insert and one from the connector arm-associated insert which help in coordination of the 2Fe-2S cluster [531]. The proline dehydrogenase N-terminal domain lineage of this superfamily lacks the cysteines but retains the inserts, suggesting that it might have been reused for interactions with as yet uncharacterized small molecule ligands. As previously shown, members of the SLBB superfamily typified by transcobalamin and related B₁₂-binding proteins contain a conserved aromatic residue in the pre-strand 3 insert which plays a central role in binding the ligand [468]. In the molybdopterin-dependent oxidoreductase superfamily, the barrel formed by the pre-strand-3 and the connector arm regions provides an open face for accommodating the molybdopterin ligand. Additionally, a conserved cysteine present in the pre-strand-3 insert is covalently linked to molybdopterin [486]. The above-described evolutionary history of this clade suggests that the ferredoxins were probably the most ancient versions. The subsequent diversification of this clade appears to have involved extensive adaptation of the binding site that originally contained the 2Fe-2S cluster for accommodating a diverse set of new ligands. Additionally, the exposed face in most of these cases remains available for interaction with other

domains or polypeptides to recruit the β -GF domain to larger complexes. This has been extensively demonstrated in the case of the 2Fe-2S ferredoxins [532, 533].

Interactions of AOR-N β -GF domains with metallopterins

The AOR-N domain represents the second independent case of a β -GF domain acquiring the capability to bind tungstoperin or molybdopterin and iron-sulfur clusters (4Fe-4S). Here, a head-to-tail dimer formed by the two tandemly repeated β -GF domains bind metallopterin via the unusually structured connector arms that form a strand at the fringe of the core 5-stranded sheet [487] (Fig. 30). While the two tandem repeats are very similar in structure, they are highly divergent in sequence, and contribute different sets of residues in the connector arm to contact the metallopterin. The N-terminal domain contributes an asparagine that directly interacts with the pterin moiety, whereas the C-terminal domain contributes a conserved arginine that interacts with the sulfoxide moiety that chelates the Tungsten or Molybdenum. The same arginine from the C-terminal domain also interacts with the 4Fe-4S metal cluster. Additionally, a threonine from the loop between strand-3 and stand-4 of the core β -grasp domain also interacts with the 4Fe-4S cluster [487] (Fig. 30). Thus, the 4-layered sandwich formed by the two derived β -GF domains help in positioning the metallopterin and 4Fe-4S for the C-terminal α -helical domain to catalyze the redox reactions on the substrates. It is possible that this neomorphic mode of substrate interaction arose from MoaD-like precursors that evolved the ability to recognize metallopterins as dimers, as an offshoot of their ancestral function in metallopterin biosynthesis.

Principal protein and nucleic acid interactions observed in the 5-stranded assemblage

A diverse range of protein-protein interactions are shown by both prokaryotic and eukaryotic members of the 5-stranded assemblage, including those with the E1, E2 and E3 enzymes or their prokaryotic counterparts. The recently published structure of the complex of

Nedd8 with its E1 and E2 enzymes [417], in conjunction with the data accumulated from several other structures and mutagenesis experiments helps in deciphering the key modes of interaction prevalent in the 5-stranded clade. Nedd8 interacts via the exposed face with the sheet of the Rossmann fold domain of the adenylating domain of the E1, as in the case of the ThiS/MoaD clade [409, 410]. Similarly the exposed face is also used by the β -GF of the C-terminal UFD of the E1 to recruit the E2. More generally, different parts of the exposed face of the sheet mediate interactions specific to particular representatives of the 5-stranded assemblage (Fig. 30). In particular, zones corresponding to the C-termini of the first and last strands which lie in the center of the sheet are utilized for protein interactions by all studied members of the classical 5-stranded clade. The structures of the eukaryotic members of the classical 5-stranded clade show that many of the interaction positions on the exposed face are shared, though the actual residues at those positions might not be conserved. Hence, the interaction specificity of different members has mainly arisen via sequence diversity at spatially congruent interaction sites, as opposed to acquisition of entirely new modes of interaction. The availability of the exposed face that provides an extended surface for interaction appears to be the primary factor for the pervasive use of this fold as mediator of protein-protein interactions across biologically disparate contexts. In a few instances, the obscured face of the RA (PDB: 1LFD [534]) and elongin domains (PDB: 1VCB [535]) might mediate specific interactions suggesting that their adapter function might depend on using both faces to mediate different sets of specific interactions.

In the complex of Nedd8 with its conjugating enzymes, the Nedd8 moiety covalently linked to the cysteine in the thioester-forming α -helical domain of the E1 protein also serves to recruit its specific E2 [417]. This occurs via a unique interaction involving the cleft formed between the sheet and the helix of the β -GF, which constitutes the “open-end” of the barrel-like

form of the fold in Nedd8. This is reminiscent of the interaction observed in the FimD-like versions of the β -GF. From the side of the E2, the interaction is mediated via the conserved C-terminal helix. The high diversity of the residues in the E2 helix as well as the cleft of the Ub/Ubls suggests that this interaction is required for the specificity of E2-Ubl association. This interaction is representative of the more generic tendency of peripheral locations on the fold to be deployed in specific interactions that might be required only for the unique function performed by a particular superfamily (Fig. 30). In the sulfur carrier and conjugated versions, the C-terminal tail plays a specific role in interaction with the active site of enzymes performing the adenylation or thioesterification [409, 410, 417, 434]. The role of the exposed face in protein-protein interactions appears to be a conservative aspect of the entire 5-stranded assemblage, which has been preserved from a period predating LUCA. Similarly, the mode of interaction with modifying enzymes via the tail appears to be an ancient conserved one. The apparently complex multiple protein-protein interactions in the eukaryotic Ub-conjugation process also appear to have emerged from the repeated use of the exposed face for interaction with E1, E2 and E3 partners.

In contrast to protein-protein interaction, little is known of the actual mode of RNA-interaction by versions of the classical 5-stranded assemblage like the TGS domain. While it is possible that the exposed face provides an interface as observed in the L25-5S rRNA-interactions [497], the conserved " α -L" structural motif shared by the S4 and TGS domains [491] suggests a role for the obscured face in RNA interactions.

Multiple "inventions" of enzymes in the β -GF

At least 3 distinct superfamilies of enzymes use the β -GF the primary scaffold for their active sites. None of the enzymatic forms can be confidently traced to LUCA. Instead they appear to have emerged early in bacterial evolution. Furthermore, the enzymatic versions are only

distantly related to each other suggesting that the β -GF fold has been convergently used to provide catalytic scaffolds. In the case of phosphoribosyl-AMP cyclohydrolase and the molybdopterin-dependent oxidoreductases like sulfite oxidase, the main pre-adaptation for catalysis stems from the ancestral ligand binding capabilities that emerged in the assemblages to which they belong. In the case of the Nudix superfamily, the pre-adaptation was possibly the nucleic acid-binding ability of the ancient 4-stranded versions of β -GF. Consistent with the convergent origins of catalysis in each instance emergence of enzymatic activity appears to have resulted in some distinctive structural changes that go beyond the above pre-adaptations shared with ligand-binding forms. In the molybdopterin-dependent oxidoreductases, the structural innovation is in the form of an elaboration of the ancestral ligand-binding features—the extensive inserts resulting in the 3-layered structure with an open barrel-like element provide a pocket for the cysteine-attached co-factor as well as the substrate.

In the phosphoribosyl-AMP cyclohydrolases a sequence alignment and super-position of the conservation onto the available structures indicates two basic active site elements (see Additional file 1). The first of these is a set of conserved residues from the insert and the lateral shelf, which appear to form the basic substrate binding site, as in other characterized members of the fasciclin-like assemblage. The second is the active metal-chelating site formed by 3 conserved cysteines, one from the loop between the last 2 strands of the core fold and two others from a C-terminal β -hairpin extension. Further, these proteins form obligate dimers on account of a strand-swapping interaction between the core β -GF and the C-terminal β -hairpin extension [495]. This juxtaposes the proposed substrate-binding site from one monomer with the metal-chelating site from the other monomer to form two distinct equivalent active sites at the dimer interface. Thus,

the emergence of a neomorphic metal-chelating site on the basic fasciclin-like β -GF scaffold appears to have been central to the origin of catalysis in this case.

The primary structural innovation in the Nudix hydrolases is the peculiar “outflow” that juts out of strand-2 on the exposed face (see above). The PXG motif in strand-2 immediately after this “outflow” also provides the space to accommodate substrates due to a lack of large side-chains. Additional substrate contacting positions also come from the rest of the exposed face. However, the conserved residues actually required for catalysis do not come from the exposed face, but from the helical segment and loop connecting strand-2 to it (Fig. 30). This loop contains a conserved motif of the form [DE]xxE (where x is any amino acid) and the helix itself contains the motif RExxEE [369, 536]. Of these, 4 acidic residues from the two motifs, excluding the last glutamate of the RExxEE motif, form a negatively-charged cloud around the arginine, and appear to be critical for positioning the active site and providing the right polar environment. The last conserved glutamate projects towards the exposed face and directs the attack on the scissile diphosphate linkage [536]. The “outflow” in strand-2 creates the necessary perforation in the sheet that allows the active glutamate to access the scissile bond of the substrate bound on the exposed face. These structural features suggest that in precursors of the Nudix enzymes the β -GF domain most probably bound the substrate via the exposed face, as is common in this fold. The emergence of the “outflow” in strand-2 would have provided further contacts for substrate interaction and at the same time created an aperture in the sheet allowing residues from the helix to form a unique active site.

Other atypical modes of interactions

A few modes of interactions thus far appear to be restricted to certain sporadic lineages or are only seen in certain highly derived forms of the domain. The superantigen/toxin-type and

strepto/staphylo-kinase-type β -GF domains share an unusual general mode of interaction with plasmin and the T-cell receptor β -chains, respectively. Both appear to contact partners “side-on” via the edge of the domain on which the additional strand of the 5-stranded versions is situated. A comparable “side-on” interaction is also seen in the evolutionarily distant 4-stranded form of the fold found in the Ig-binding domain. However, this latter interaction is distinct in having additional extensive contacts supplied by the helical segment from the obscured face of the domain (Fig. 30). It is unclear if this “side-on” interaction might be another less-known but ancient interaction feature of the β -GF domain or has convergently evolved in the superantigen/toxin-type and strepto/staphylo-kinase-type domains on one hand and in the 4-stranded Ig-binding domain on the other. The interactions of the POZ-domain are rather distinct on account of its extreme structural modification as well as acquisition of a unique C-terminal helical sub-domain. However, the structure of the POZ domain in the potassium-channel complexed with the oxido-reductase subunit (PDB: 1EXB [537]) shows that the surface equivalent to the exposed face of the classical β -GF domains plays an important role in the interactions of the POZ domain (Fig. 30). This suggests that the POZ domain probably retained at least in part the interaction of the ancestral β -GF domains.

General conclusions

While the β -GF has been thoroughly investigated in the context of the interactions of ubiquitin and UBLs in eukaryotes and their prokaryotic relatives like ThiS and Moad involved in sulfur transfer, the broader evolutionary history of the fold was poorly understood. We sought to redress this by developing a natural classification for the fold and using it as guide for exploring the tempo of its evolutionary radiations and details of its functional adaptations. As a result we identified several novel members of the fold, including some distinctive previously unidentified

modifications. The reconstruction of the evolution of the fold suggests that the major structural variants and some of the basic biochemical features and modes of interaction had emerged prior to LUCA. This suggests that even before the radiation of the extant lineage of Life there were several rounds of duplication followed by extensive divergence, including major structural changes.

The scenario emerging from our analysis also suggests that the earliest reconstructed function of the β -GF domain was in the context of ribonucleoprotein complexes, probably as an RNA-binding domain. Based on the functions of extant versions of the domain, like the TGS superfamily, the IF3-N domain, and early structural derivatives such as the S4 superfamily, it is quite possible that the earliest versions of the fold played a generic role in a primitive pre-LUCA translation system. Thus, the earliest diversification events of the β -GF fold likely occurred in the context of the RNA-world, probably with the acquisition of increasingly specialized roles in the evolving translation apparatus. Amongst the major pre-LUCA functional shifts were those relating to the biosynthesis of sulfur-containing compounds and scaffolding of Fe-S clusters. On the face, such functional shifts from earlier roles in translation-associated RNPs appear drastic and puzzling. However, it should be noted that there is a functional connection between the sulfur incorporation pathways of thiamine biosynthesis and thiouridine synthesis in RNA [507, 538]. Hence, it is possible that these shifts might have occurred in the context of 5-stranded versions of the β -GF providing scaffolds for the synthesis of thio-base containing RNAs. This reconstruction also implies that the versions of the β -GF associated with major metabolic functions, including respiratory metabolism, radiated from the ancestral RNA-binding versions.

The major post-LUCA phases of the evolutionary history of the β -GF fold saw two major spurts of innovation. The first, occurring primarily in the bacteria, was accompanied by an

extensive exploration of the biochemical function and interaction space by different versions of the fold. This was marked by the acquisition of diverse soluble ligand-binding capabilities through distinctive structural modifications as well as extensive deployment in different protein-protein interaction contexts. Most notably, the scaffold on at least 3 independent occasions acquired very different enzymatic activities even though the β -GF fold did not ancestrally support catalytic activities. The eukaryotic phase did not see extensive innovation in terms of fundamentally different biochemical functions, but the diversity of protein interactions within the ubiquitin-like superfamily of the 5-stranded assemblage was vastly expanded through extensive sequence divergence of the primary interaction surfaces of the superfamily. This phase was also accompanied by ongoing innovation of new multi-domain architectures associated with the eukaryotic expansions of Ub-like signaling domains (Fig. 29).

As has been suggested before, the β -GF shows several parallels with the RRM-like fold [468]. Both are relatively small folds, forming asymmetric 2-layered structures, with one face of their sheets exposed, and the other partially or wholly obscured by helical segments. Importantly, when centered on the β -hairpin element in core of their sheets, strands in both these domains and one helix show the same orientation. Both have evolved to provide scaffolds for a comparable set of diverse biochemical functions, including RNA-binding, small-molecule or solute recognition, protein-binding, supporting Fe-S clusters (ferredoxins) and providing the skeleton for active sites of very different enzymes. Both of these folds also appear to have undergone extensive adaptive radiation even prior to LUCA after starting off as domains with primitive roles related to RNA metabolism [234]. These two folds differ from the ancient α/β 3-layered sandwich domains like the Rossmannoid and P-loop NTPase domains, which appear to have begun their existence as enzymatic domains and more-or-less retained a conserved set of basic biochemical activities

throughout their evolution [155, 234]. The versatility of the β -GF and RRM-like folds in providing scaffolds for both enzymatic and diverse non-enzymatic function might be attributable in part to two major factors: 1) an issue of contingency- these folds simply arose very early in evolution and had the time to colonize numerous functional roles. 2) Favorable structures- their relatively large sheet that is exposed on one side provides an interface for diverse interactions, especially in the form of binding various substrates. Sequence alteration to this binding surface, without disrupting the overall scaffold, could easily allow the emergence of a great diversity of new interactions. Secondly, the presence of the large sheet with just a single helical segment also favors formation of barrel-like structures, thereby opening new faces for interactions.

Despite intense investigations the precise functions of several eukaryotic UbIs remain unclear. More generally, the functional details of the non-ubiquitin-like members of the fold remain less studied, as in the case of the β -GF domains in RNP complexes which are in need of more detailed investigations. In conclusion, we hope that our analysis of the β -GF domain provides a new framework for further systematic experimental exploration of the functions of this fold.

Additional Files /Supplementary Material

Additional files referred to throughout the text can be accessed at the following site:

<http://www.biology-direct.com/content/2/1/18/additional/>.

A Novel Superfamily Containing the β -Grasp Fold Involved in Binding Diverse Soluble

Ligands

(based on reference [539])

Introduction

The β -grasp fold (β -GF) was first recognized in ubiquitin and the immunoglobulin-binding (IG-binding) domains of Gram-positive cocci [437, 540]. Since then it has come to be known as a widespread fold, utilized in proteins performing a great diversity of cellular functions. These include regulation of protein stability and signal transduction through the ubiquitin-conjugation system [541], RNA-protein interactions as seen in the TGS domain of tRNA synthetases [463], and adaptor functions involving protein-protein interactions as seen in the FERM module [542]. Additionally, standalone β -GF domain proteins ThiS/MoaD function as sulfur carriers in molybdopterin and thiamine biosynthesis [413] and the fold also provides an effective scaffold for binding iron-sulfur clusters in the case of the 2Fe-2S ferredoxins involved in electron transport (see SCOP database [203])

As part of our larger effort to understand the evolutionary and structural basis for the functional versatility of this widespread fold [429] I was keen to determine if there were as yet uncharacterized representatives that might widen the functional horizon of the β -GF. In particular, I was interested in exploring the possibility of versions of the β -GF domains binding soluble ligands. Such a function was of interest because the presence of 2Fe-2S ferredoxins suggested that the β -GF domains could potentially provide a scaffold for binding a wider range of small molecules or other prosthetic groups. Accordingly, this was investigated further by applying a combination of sensitive structural comparisons and sequence profile analysis on members of the β -GF. As a result, the lab has identified a novel domain superfamily with the β -

GF fold and provide support that its members might be involved in binding different soluble ligands. Their genomic contexts, domain architectures and phyletic patterns are also studied to present evidence for their role in diverse metabolic networks, including those related to vitamin B12.

Application of Methods

Initial DALI searches that aided in the discovery of the fold were performed in the lab by Dr. Balaji. The initial characterization of the mode of binding in members of the fold was performed by Drs. Balaji and Aravind. I performed the remainder of the analyses, with input from Drs. Iyer and Aravind.

Searches of the PDB database with query structures were conducted using the DALI program [96]. Structural visualization and manipulations were performed using the Swiss-PDB viewer program [174]. Sensitive profile searches were conducted using the PSI-BLAST [38] and HMMER programs [63]. PSI-BLAST searches were performed against the nonredundant (NR) database of protein sequences (National Center for Biotechnology Information [NCBI], NIH, Bethesda, MA, USA), with either a single sequence or an alignment used as the query, with a default profile inclusion expectation (e) value threshold of 0.01 (unless specified otherwise), and was iterated until convergence. All sequences collected in these searches are made available in Additional file 2. The library of profiles for various domains was prepared by extracting all alignments from the PFAM database [543] and updating them by adding new members from the NR database. These updated alignments were then used to make HMMs with the HMMER package or PSSMs with PSI-BLAST. For all searches involving membrane-spanning domains a statistical correction for compositional bias was used to reduce false positives due to the general hydrophobicity of these proteins [544]. Signal peptide and transmembrane helices were predicted

using the SignalP [91] and TMHMM programs [90]. Multiple alignments were constructed using the T_Coffee [171] and MUSCLE programs [62] followed by manual adjustments based on PSI-BLAST results. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED program [173], with information extracted from a PSSM, HMM, and the seed alignment itself. Similarity-based clustering of proteins was carried out using the BLASTCLUST program [148]. Gene neighborhoods were determined using a custom script that uses completely sequenced genomes or whole genome shotgun sequences to derive a table of gene neighbors for a query gene. The BLASTCLUST program was then used to cluster the proteins sequences in the neighborhoods and establish conserved co-occurring genes. The KEGG database was used to identify key components of the B12 synthesis pathway [32]. Automation of all large-scale sequence analysis procedures were carried out using the in-house TASS package (Anantharaman V, Balaji S, Aravind L; unpublished), which operates similar to the previously published SEALS package [545].

Results and Discussion

Detection of Sequence and structure relationships

To identify potential novel versions of the β -GF that bind soluble ligands, comprehensive structural comparisons were initiated with various previously characterized members of the fold (see β -grasp fold in SCOP database) using the DALI program. Several of these searches retrieved the C-terminal domain of the transcobalamin protein with significant Z-scores. For example, searches initiated with MoaD proteins (PDB: 1v8c, 1vjk) retrieved the C-terminal domain of transcobalamin (PDB: 2bbc) with Z-scores of ~ 7 . Transcobalamin is an animal-specific protein that binds cobalamin (vitamin B12), and is involved in its uptake by animal cells [546]. Transcobalamin contains an N-terminal α/α toroid domain, and a C-terminal α/β domain [470]

that corresponded to the β -GF domain recovered in the above searches. Further, DALI searches initiated with the C-terminal domain of transcobalamin recovered a diverse set of previously known β -GF domains such as MoaD (PDB:1vjk), YukD (PDB: 2bps) 2Fe-2S ferredoxin (PDB:1feh) and the middle domain of the Nqo1 subunit of the bacterial and mitochondrial NADPH-quinone oxidoreductase I complex (PDB: 2fug. S chain) with Z-scores in the range of 5-7. The structural alignments generated by these searches showed that the transcobalamin C-terminal domain aligns completely with all core structural elements of the β -GF, including a β -sheet of 5 strands and a helix between strands 2 and 3. However, the transcobalamin C-terminal domains are distinguished by the presence of a unique β -hairpin after the conserved helix of the β -GF (Fig. 31A). The N-terminal α/α toroid domain and the C-terminal β -GF domain cooperate in ligand-binding by sandwiching a single B12 molecule between them [470]. Systematic searches for contacts between the B12 ligand and the C-terminal β -GF domain in transcobalamin showed that the unique insert plays a prominent role in binding the ligand by contributing several direct or solvent-mediated interactions [470]. Additional contacts with the ligand are also made by residues from the core β -GF such as those from strand 3, the end of strand 4 and the “ascending connector” between strand 4 and 5 (Fig. 31A). These observations suggest that the C-terminal domain of transcobalamin represents a novel adaptation of the β -GF for small-molecule ligand interactions.

To better understand the diversity of this class of ligand-binding β -GF domains and their phyletic spread sequence profile and hidden Markov model (HMM) searches were initiated for homologs using PSI-BLAST and the HMMER package respectively. In addition to orthologs of transcobalamin, intrinsic factor and solo C-terminal domains from fishes, these searches retrieved numerous prokaryotic proteins, which were either present as stand-alone β -GF domains or in

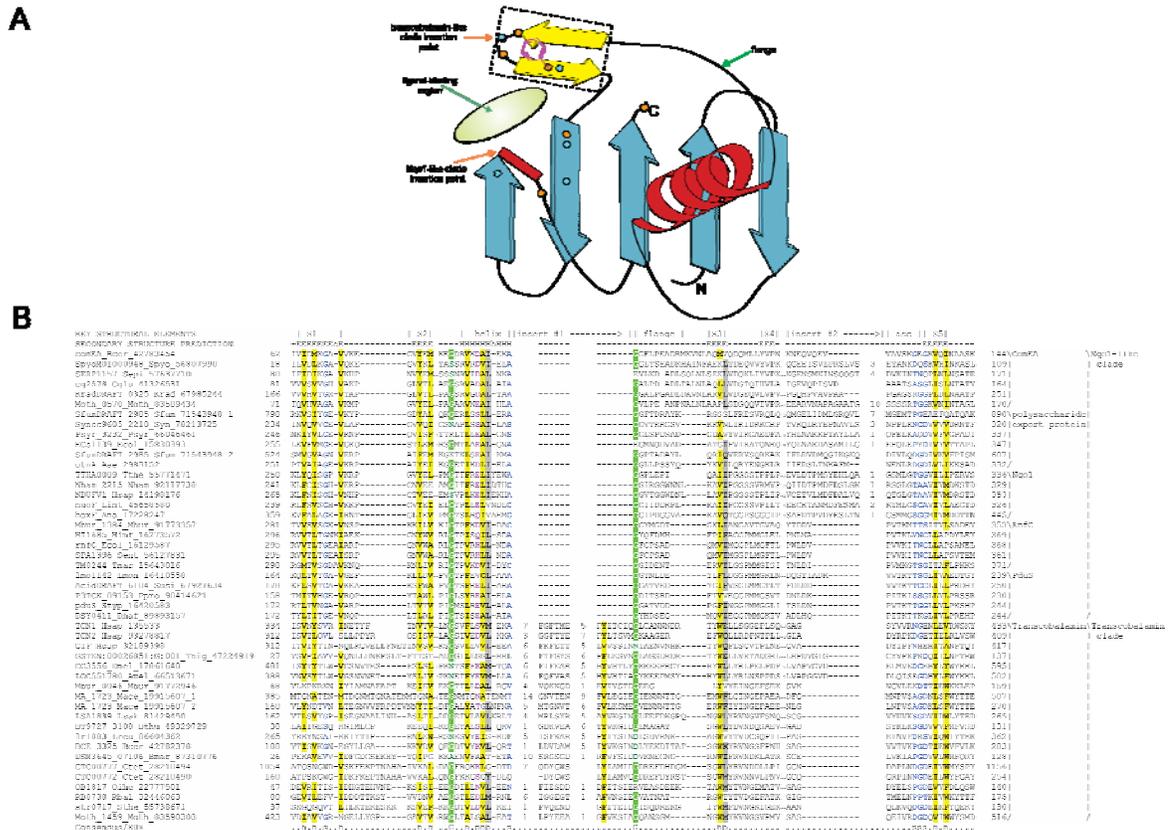


Fig. 31. Topology diagram of SLBB domain and multiple alignment of SLBB superfamily

(A) The five-stranded core (characteristic of all members of the β G-F) is shown with the helical face at the near side. β -strands are depicted as blue arrows, with the arrowhead at the C-terminus while the α -helix is shown in red. The two-strand insertion in the Transcobalamin-like clade is colored in yellow and enclosed in a dotted box. The insertion point for the Nqo1-like clade is marked by a red box. The approximate soluble ligand-binding spatial region is marked by a green oval. Residues known to contribute to cobalamin binding as derived from the crystal structure of Transcobalamin are shown as small circles. Orange circles indicate sidechain-mediated interactions while greenish blue circles indicate backbone or backbone and sidechain-mediated interactions. The conservation of an aromatic residue in Transcobalamin proteins is represented by a phenylalanine residue, rendered as a line drawing and colored purple. (B) Proteins are denoted by their gene names, species abbreviations, and gi numbers; demarcated by underscores. Amino acid residues are colored according to sidechain properties and degree of conservation within the alignment, set at 80% consensus. Consensus abbreviations are shown below the alignment. The secondary structure shown above the alignment is derived from the crystal structures of Transcobalamin and Nqo1 and secondary structure prediction programs. E and H denote β -strand and α -helix, respectively. Secondary structure elements conserved across the SLBB superfamily are labeled in the top line of the alignment. "Insert #1" refers to the Transcobalamin-like clade insert while "Insert #2" refers to the Nqo1-like clade insert. "asc" refers to the ascending connector between strands 4 and 5 often observed in the β -grasp fold. The consensus abbreviations and coloring scheme are as follows: h, hydrophobic residues

(ACFILMVWY) shaded yellow; s, small residues (AGSVCDN) colored blue; p, polar residues (STEDKRNQHC) colored purple; and b, big residues (LIYERFQKMW) shaded gray. The conserved glycine residues characteristic of this superfamily are shaded light green and colored white.

large multidomain proteins. For example, a search initiated with the β -GF domain of puffer fish transcobalamin (*Tetraodon nigroviridus*, gi: 47226456, region: 325-425) recovered closely related eukaryotic orthologs and paralogs fused to N-terminal α/α toroid modules (iteration 1), solo transcobalamin C-terminal domains with predicted signal peptides (e.g. XP_689937, *Danio rerio*, iteration 2, e-value: 3×10^{-12}), and several prokaryotic proteins (e.g. BAC13773, *Oceanobacterium theyensis*, iteration 3, e-value: 2×10^{-3}). In order to exhaustively recover all divergent homologs, transitive searches were conducted with all above-detected members and also evaluated all hits below the threshold of PSI-BLAST searches for the presence of potentially homologous domains. HMMs and PSSMs were also prepared from the alignment of this region of all proteins recovered with significant expect-values ($e < .01$ with statistical correction for compositional bias) and used these to search all completely sequenced genomes. These searches consistently retrieved hits to multiple sequence repeats in a group of bacterial cell-surface/secreted sugar-binding proteins involved in polysaccharide export with significant e-values (e.g. *Hahella* periplasmic protein, HCH_02380 residues 852-990). Inclusion of polysaccharide export proteins in profiles for further searches additionally recovered the N-terminal region of the ComEA family of DNA uptake receptors of Gram positive bacteria (e.g. *Clostridium* ComEA, gi: 67874543, iteration 2, $e = 10^{-6}$), PduS-like cobalamin reductases (e.g. *E. coli* cobalamin reductase iteration 7, $e = 10^{-3}$), the middle domain of the 51kD subunit (F chain) of the NADPH-quinone oxidoreductase complex I (Nqo1, E: 10^{-4} , iteration 11) and the RnfC subunit of the oxidoreductases encoded by the bacterial Rnf operons [547] (*Rhodobacter* RnfC, E: 10^{-6} ; iteration 14). This latter set of proteins was more similar

to the homologous region recovered in the polysaccharide export proteins than to the transcobalamin C-terminal domain (Fig. 31B). However, recovery of the middle domain of Nqo1 in sequence searches clearly confirms their relationship with transcobalamin C-terminal domains, because the former are also known, via structural analysis, to contain a similar β -GF domain [469] (See above and Additional file 1). This was additionally supported by separate secondary structure prediction for individual sub-groups with potentially homologous regions such as the ComEA N-terminal regions and the polysaccharide export proteins (Fig. 31B).

Hereafter, the homologous β -GF domains found in all these proteins are referred to as the Soluble-Ligand-Binding β -grasp (SLBB superfamily) as many members of this superfamily are known or predicted to bind soluble ligands (See below for further details).

Sequence and structure features of the SLBB superfamily

A comprehensive multiple alignment for the SLBB superfamily (Fig. 31B) was prepared by combining alignments for individual groups constructed using the T-Coffee program [50], based on the structural superposition of transcobalamin C-terminal domain (2bbc) and Nqo1 middle domain (2fug; chain S). Much of the conservation seen across the entire superfamily is in the form of hydrophobic residues forming the stabilizing core of the fold. However, there was a notable sequence feature in the form of two strongly conserved glycine residues, one in the turn leading into the horizontal flange preceding the third β -strand (Fig. 31A) of the β -GF and the other immediately downstream of the second conserved β -strand (Fig. 31). This conservation pattern is a unique feature of the SLBB superfamily that distinguishes them from all other previously characterized β -GF domains, supporting a common ancestry for this set of domains within the β -GF.

The alignment also helped us to classify the SLBB superfamily into several distinct families. The Transcobalamin C-terminal domain clade is unified by the presence of the above-described β -hairpin insert within the β -GF that plays an important role in contacting the ligand (Fig 31A, see Additional file 1). This β -hairpin contains a conserved hydrophobic position that makes a stacking interaction with the aromatic ring of the base in cobalamin. However, the rest of the sequence in this region is poorly conserved as most other interactions occur through backbone oxygen or nitrogen atoms [470] (Fig. 31B). Within animals, insects and most vertebrates have a single ortholog of the B12 binding protein, whereas the mammals have three distinct versions, transcobalamin I, transcobalamin II and the intrinsic factor. Besides animals, members of this clade are widely represented in Low GC Gram-positive bacteria and planctomycetes and less frequently in the euryarchaea.

The Nqo1-like clade includes at least two distinct families: 1) the first includes the NADPH-quinone oxidoreductase complex I subunit Nqo1 (51kD/F chain), the RnfC oxidoreductase subunit, and the PduS-like cobalamin reductases. 2) The second family contains polysaccharide export proteins and the DNA receptor ComEA. This clade is unified by the presence of a small, often α -helical insert, in the “connector arm” between the fourth and fifth strands of the domain (Fig. 31, see Additional file 1). In some cases, such as the ComEA protein, the helical segment is followed by a low complexity region; suggesting the presence of a disordered, extended loop. These proteins are also characterized by an sGG motif (where ‘s’ is any small residue) around the second conserved glycine of the superfamily (Fig. 31B). The Nqo1 subunit of the classical NADPH-quinone oxidoreductase complex I is present in all major bacterial lineages with well-developed electron-transport chains, in most mitochondriate eukaryotic lineages, and very rarely in euryarchaea. The RnfC proteins are strongly conserved in

γ -proteobacteria, but are also found in some representatives of Low GC Gram positive bacteria and the Bacteroidetes/Chlorobi assemblage. The PduS protein is restricted to the Low GC Gram-positive bacteria and certain γ and δ proteobacteria. The ComEA proteins and polysaccharide export proteins show a nearly mutually exclusive complementary distribution. The ComEA family is chiefly present in Low GC Gram-positive bacteria and actinobacteria, whereas the polysaccharide export family is more widespread and widely present in proteobacteria, cyanobacteria, acidobacteria, planctomycetes, bacteroidetes/chlorobi, and more sporadically in a few other groups.

While the interaction between B12 and the transcobalamin-like SLBB domain involves the unique β -hairpin insert, these key contacts are also contributed by the core fold (See above, Fig. 31A), and in general the position of the bound ligand is comparable to that of the bound metal-sulfur cluster in the β -GF ferredoxins. The Nqo1-like clade shows its distinctive innovation in the region between strands 4 and 5, which also corresponds to the same general spatial location where the ligands are bound in the transcobalamin-like clade and β -GF ferredoxins (Fig. 31B). This indicates that the structural innovation specific to the Nqo1-like clade might also be involved in binding a ligand at a similar position. This spatial location might thus represent a common site for soluble ligand interactions in the β -GF that is distinct from the C-terminal tail and the opposite protein surface that is key to the functional interaction of sulfur carriers like ThiS and MoaD and the ubiquitin-like proteins [548].

Contextual associations and inferences of possible functions for the SLBB

To investigate the functional diversification of the SLBB fold contextual analysis was used, which often provides insights into biochemical functions of poorly characterized protein domains or genes. Contextual analysis utilizes the information gleaned from the association of

Most members of the transcobalamin C-terminal domain clade of the SLBB superfamily contain signal peptides, and several also contain the C-terminal Gram-positive anchor motif [550], suggesting that they are secreted or cell-surface proteins. A common domain architecture in this clade encountered in both eukaryotes and bacteria is the fusion of the SLBB to an α/α toroid domain. In bacteria the toroid may be present either N-terminal (e.g. *Desulfotomaculum*, gi: 88945170) or C-terminal (e.g. *Bacillus*, gi: 42782379) to the SLBB (Fig. 32A). As the central cavity formed by the α/α toroid in transcobalamin plays a major role in binding B12 [470], it is likely that the two domains cooperate in binding B12 in all these proteins. Additional architectures include fusions to domains typically found in extracellular proteins, such as one or more immunoglobulin-fold domains (e.g. *Archaeoglobus*; gi: 11498993 and *Moorella*; gi: 83590303), the FIVAR (Pfam entry: PF07554) sugar-binding domain (*Clostridium*, gi: 28210467), the fasciclin domain (*Methanosarcina*; gi: 21228740) and a β -propeller domain (*Clostridium*, gi: 28210494). Given that many of these domains are often involved in interactions with polysaccharides, they might play a role in tethering these proteins to the cell surface by binding peptidoglycan or capsular polysaccharides [114, 482, 551-554]. Often these multi-domain SLBB proteins occur in conserved operons that might additionally code a second paralogous extracellular SLBB protein (Fig. 32). This might imply that different extracellular SLBB proteins interact together to form protein complexes on the cell surface. Interestingly, an analysis of the B12 biosynthesis pathways of all the bacteria that possess proteins with such SLBB domains showed they usually lacked key biosynthetic enzymes for B12. Furthermore, these SLBB proteins are generally encoded by predicted operons that also contain genes for CbiO-like ABC ATPase and the CbiQ-like integral membrane protein implicated in cobalt transport [555]. These observations suggest that the primary role of this clade of SLBB proteins might be to scavenge B12 or its precursors from the

environment. As the archaea which contain these SLBB proteins often possess an anaerobic B12 synthesis pathway, it is possible that these might instead be involved in scavenging a distinct metabolite. In *Syntrophomonas* the SLBB domain is found in putative extracellular enzymes fused to sulfite oxidase-like molybdopterin cofactor binding domain (e.g. gi: 71491441) [486]. It is likely that in these proteins the SLBB provides a B12 cofactor that might be required by these enzymes.

Intracellular versions of the transcobalamin-like clade show fusions of the SLBB domain with two distinct winged HTH domains, namely those of the ArsR-like (e.g. gi:72395507, *Methanosarcina*) and AraC-like families (E.g. gi: 86604362, *Lactobacillus*) (Fig. 32A). These proteins probably function as one-component transcription factors that respond to concentration of B12, its precursors or some other as yet unknown soluble ligands.

In the Nqo1-like clade, polysaccharide export proteins are predicted to be secreted or periplasmic proteins and contain an absolutely conserved N-terminal β -strand-rich domain followed by 1-8 repeats of the SLBB domain (Fig. 32A). They appear to be part of a larger complex that is involved in transport of polysaccharides to the cell surface and are believed to associate with the outer membrane and periplasmic space in proteobacteria [556]. Conserved gene-neighborhoods that encode these proteins are populated with proteins involved in the biosynthesis and modification of sugars or polysaccharides, which is consistent with their role in polysaccharide export (Fig. 32). The related ComEA proteins of Gram-positive bacteria also contain a signal peptide followed by an N-terminal SLBB domain that is always fused to a pair of DNA-binding Helix-hairpin-Helix domains at their C-terminus. This is consistent with the role of the ComEA protein as a non-specific DNA receptor in the transformation competence mechanism of Gram-positive bacteria [557, 558]. Prior studies suggest that this DNA receptor may be linked to the cell surface via the N-terminal region spanning the SLBB domain [557]. Taken together

these observations suggest that the SLBB domain in these proteins is likely to be critical for interaction with cell polysaccharides and/or sugars of the peptidoglycan. The complementary phyletic distribution of ComEA and polysaccharide export proteins is strongly correlated with the presence or the absence of the specialized Gram-positive cell wall (See above). This suggests that they probably diverged from a common ancestral polysaccharide/sugar-binding domain that was originally involved in uptake or extrusion of large molecules at the cell surface.

The remaining three groups of proteins, namely NqoI, RnfC and PduS, within the Nqo1-like clade of the SLBB superfamily share a common architectural core consisting of a fusion between an N-terminal Rossmannoid domain and an SLBB domain. Unlike classical Rossmann fold domains of oxidoreductases with 5-7 strands, the Rossmannoid domain of these proteins has a 4-stranded core in a 3214 order [469] coupled to a N-terminal two-stranded hairpin contributed by a module similar to the BBMs of RNA polymerases [476]. The SLBB and this Rossmannoid domain are additionally combined to variety of other domains in NqoI, RnfC and PduS proteins. The most common fusions seen in all three groups of proteins are those to 4Fe-S ferredoxin domains that flank the above two-domain core. The NqoI family also might contain a C-terminal tetrahelical bundle with an up-and-down topology that coordinates an Fe-S cluster via conserved cysteine residues, which in addition to the 4Fe-S ferredoxin is likely to provide an additional redox center for electron transport [469]. A biotin/lipoate carrier-like Sandwich Barrel Hybrid Motif (SBHM) domain [476] is found respectively at the C- and N- termini of members of PduS and RnfC families.

The roles of these versions of the SLBB remain enigmatic; however there is evidence that the PduS protein might bind soluble ligands. The PduS gene typically belongs to a large mobile operon coding proteins required for the biogenesis of carboxysome/polyhedral bodies, which

contains enzymes involved in propanediol degradation. The PduS has been shown to strongly bind cob(I)alamin and was characterized as a bifunctional cob(II)alamin and hydroxycobalamin (cob(III)alamin) reductase catalyzing the formation of cob(I)alamin. Cob(I)alamin is the immediate precursor of Ado-cobalamin, which serves as an essential coenzyme for the diol dehydratase in degradation of 1,2-propanediol [559, 560]. It is likely that the SLBB domain in PduS, like that in transcobalamin, binds cob(I)alamin or HO-cobalamin, while the N-terminal Rossmannoid domain binds the flavin nucleotide cofactor for the redox reaction. Such a function is also supported by the observation that cob(I)alamin is highly reactive and needs to be shielded from the environment [559]. The role of the fused SBHM domain seen in PduS proteins is less clear. However, given that the SBHM domain carries covalently associated ligands such as biotin/lipoate [524, 561], it might similarly carry cofactor ligands or intermediates in propanediol degradation such as 1,2-propanediol-1-yl radical [562]. There is currently no evidence for a soluble ligand interacting with the related SLBB domain in the RnfC and NqoI. Nevertheless, crystal structures indicate an exposed location for SLBB domain in these proteins, allowing the possibility that they might be allosterically regulated by hitherto unknown ligands interacting with this domain.

Evolutionary History of the SLBB Domain and General Conclusions

The phyletic and domain-architecture distributions show that the SLBB superfamily is well-represented and has diversified across the entire bacteria superkingdom. Their sporadic presence in archaea and the stronger affinity of the different eukaryotic versions to their bacterial counterparts suggests that the SLBB superfamily was derived early in evolution of bacteria, most probably from one of the many ancient β -grasp fold domains. The two major families in the NqoI-like clade appear to have a pan-bacterial distribution suggesting that this clade had already

differentiated into versions associated with cell wall related functions (ComEA-Polysaccharide export protein family) and intracellular oxido-reductase related functions (Nqo1, RnfC and PduS). Of the latter group the NqoI protein of the respiratory complex-I is seen across bacteria and was transferred to the eukaryotic lineage during the primary endosymbiotic event that generated the eukaryotic cell with mitochondria. RnfC and PduS proteins have more restricted phyletic distributions and are likely to be late derivatives of the more ancient Nqo1 lineage. The transcobalamin C-terminal clade is very divergent in sequence and appears to be a lineage-specific innovation in Gram positive bacteria that was recruited specifically for transporting extracellular B12-like cofactors or their precursors. Subsequently, a specific version that combined the α/α toroid domain with the SLBB domain appears to have been laterally transferred to the animal lineage early in its evolution. This event probably conferred on animals the ability to directly absorb cobalamin synthesized by bacteria in the gut. In parallel, there appear to have been sporadic transfers of large extracellular multidomain versions as well as intracellular versions fused to DNA-binding HTH domains to certain euryarchaeal lineages.

In this context, it is of interest to note that the discovery of soluble ligand binding versions of the β -GF points to a noteworthy structure-function analogy with the RNA-recognition motif (RRM)-like fold. In functional terms, the RRM-like fold has long been known to bind a range of soluble ligands such as amino acids, sugars and co-factors. Notable examples of these include the ACT domain superfamily and the amino acid-binding domain of the LRP-like transcription factors [114, 563, 564]. Both folds also provide scaffolds for iron-sulfur clusters, (4Fe-4S ferredoxins in the case of the RRM-like fold [565]) and are also involved in RNA-binding, as well as adaptor functions related to protein-protein interactions [524, 561]. These functional analogies in turn might be related to certain general organizational similarities seen in the two

domains: like the β -GF domain, the RRM-like fold domain is also a relatively small domain with an asymmetric two-layered structure. One surface of the core sheets is partially obscured by helical segments in both these folds, whereas the other is largely left exposed (see SCOP database [203]). Further study of these functional analogies might throw light on whether there exist certain general structural principles that have affected the recruitment of certain small ancient domains in similar contexts.

In conclusion, we show that the β -GF domains found in transcobalamin, polysaccharide export proteins, ComEA, PduS, and RnfC and Nqo1-like oxidoreductases define a novel superfamily, several of which might interact with different soluble ligands. This investigation provides the possible evolutionary scenario for the origin of the vitamin B12 uptake in animals via transcobalamin and intrinsic factor. It also provides leads for new investigations into B12 metabolism in bacteria and other aspects of protein-ligand interaction in competence, cell-surface biochemistry, and respiratory electron transfer.

Additional files/supplementary material

All additional files mentioned in the above text can be accessed at the following site:

<http://www.biology-direct.com/content/2/1/4/additional/>.

Experimental validation of work presented above

Dr. Naismith and colleagues at the University of St. Andrews recently published the crystal structure of Wza protein, the export protein that facilitates extracellular transport of polysaccharides in *E. coli* [566]. The Wza protein contains two SLBB repeats (see additional files). The solved structure reveals octameric symmetry, with the two SLBB domains bridging the periplasmic space in each monomer. A deep cleft rings the membrane protein at the SLBB-SLBB domain juncture. Interestingly, surface residues of this cleft correspond to residues of the

putative SLBB binding pocket. This appears to strengthen the contention above that the binding pocket is interacting with sugar moieties of the peptidoglycan, anchoring the transporter as it bridges the periplasmic space.

The Prokaryotic Antecedents of the Ubiquitin Signaling System and the Early Evolution of Ubiquitin-like β -Grasp Domains

(based on reference [414])

Introduction

The ubiquitin system is one of the most remarkable protein modification systems of eukaryotes, which appears to distinguish them from model prokaryotic systems. The modification of proteins by ubiquitin (Ub) or related polypeptides (Ubls) has been detected in all eukaryotes studied to date and is comprised of conserved machineries that both add ubiquitin and remove it [541, 567]. The Ub-conjugating system consists of a three-step cascade beginning with an E1 enzyme that uses ATP to adenylate the terminal carboxylate of Ub/Ubl and subsequently transfers this adenylated intermediate to a conserved internal cysteine in the form of a thioester linkage. The E1 enzyme then transfers this cysteine-linked Ub to the conserved cysteine of the E2 enzyme, which is the next enzyme in the cascade. Finally, the E2 enzyme transfers the Ub/Ubl to the target polypeptide with the help of an E3 enzyme [567, 568]. The E3 enzymes of the HECT domain superfamily contain a conserved internal cysteine, which accepts the Ub/Ubl through a thioester linkage and finally transfers it to the ϵ -amino group of a lysine on the target protein. The E3 ligases of the treble-clef fold, namely the RING and A20 finger superfamilies, appear to directly facilitate the transfer of Ub to the lysine of target protein, without forming a covalent link with Ub/Ubl (Fig. 33) [455, 569].

The proteins modified by ubiquitination might have different fates depending both on the specific Ub or Ubl used, and the type of modification they undergo [570, 571]. Monoubiquitination and poly-ubiquitination via G76-K63 linkages have regulatory roles in diverse systems such as signaling cascades, chromatin dynamics, DNA repair and RNA that

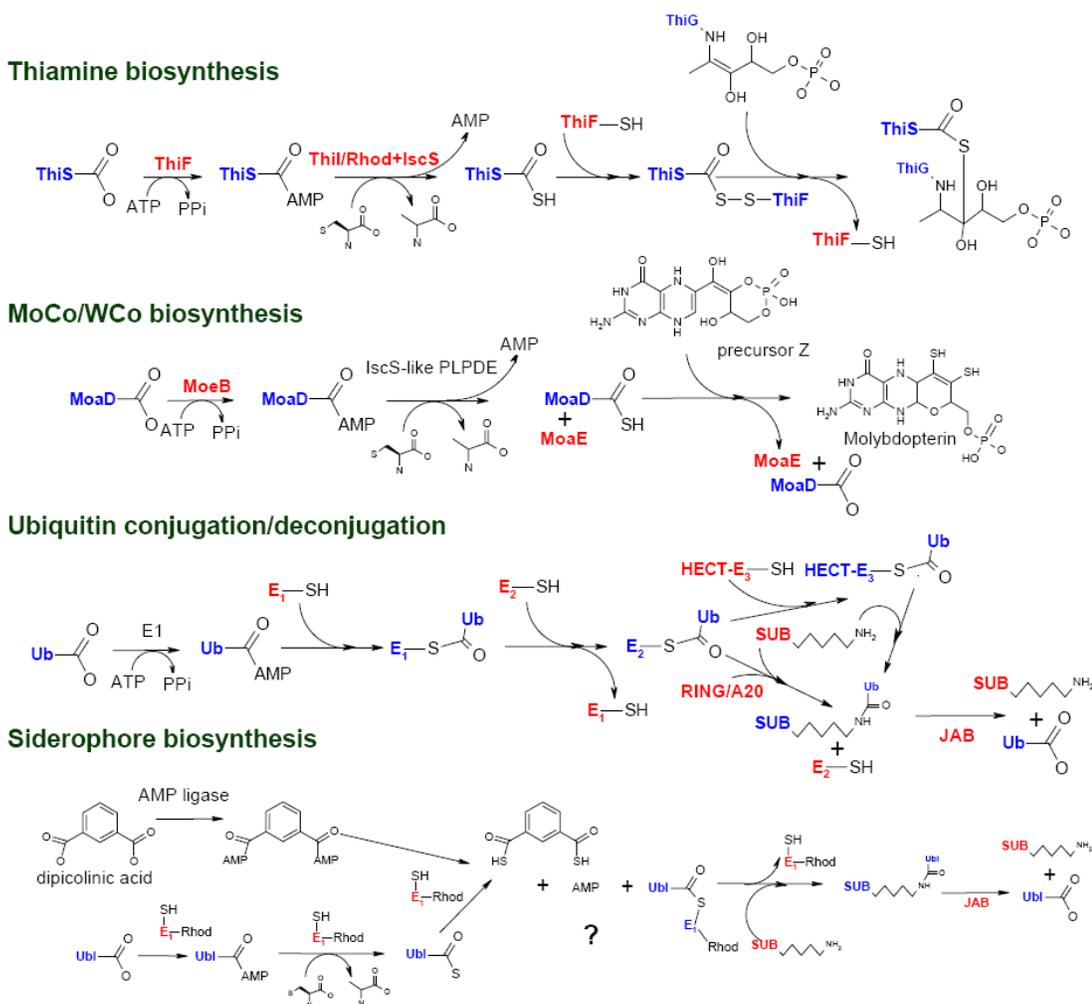


Fig. 33. ThiS/MoaD/Ubiquitin-based protein conjugation system.

The figure shows different themes by which a ThiS/MoaD/Ubiquitin-like polypeptide participates in thiamine biosynthesis, MoCo/WCo biosynthesis and the ubiquitin conjugation/deconjugation system and the siderophore biosynthesis pathways. The “?” refers to the speculated part of the pathway inferred from operon organization. SUB refers to the polypeptide/protein substrate.

results in targeting the polypeptide for proteasomal degradation [571]. Other polyubiquitin chains formed by linkages to K29, K6 and K11 are relatively minor species in model organisms and are poorly understood in functional terms. Similarly, modification by UbIs such as SUMO,

Nedd8, URM1, Apg8/Apg12 and ISG15 have specialized regulatory roles in the context of chromatin dynamics, RNA processing, oxidative stress response, autophagy and signaling [443, 572]. The Ub-modification is reversed by a variety of deubiquitinating peptidases (DUBs) belonging to various superfamilies of the papain-like fold and pepsin-like, JAB, and Zincin-like metalloprotease superfamilies [459, 573-578]. Of these the most conserved are certain versions of the papain-like fold and the JAB superfamily metallo-peptidases, which are components of the proteasomal lid and signalosome [460, 579-581]. The JAB peptidases are critical for removing the Ub chains before the targeted proteins are degraded in the proteasome [582, 583].

While the entire ubiquitin system with the apparatus for conjugation and deconjugation has only been observed in the eukaryotes, several structural and biochemical studies have thrown light on prokaryotic antecedents of this system. Most of these studies are related to the experimental characterization of the key sulfur incorporation steps in the biosynthetic pathways for thiamine and molybdenum/tungsten cofactors (MoCo/WCo). Both these pathways involve a sulfur carrier protein, ThiS or Moad, which is closely related to the eukaryotic URM1, and bears the sulfur in the form of a thiocarboxylate of a terminal glycine, just as the thioester linkages of Ub/Ubls formed in course of their conjugation [445, 584]. Furthermore, both ThiS and Moad are adenylated by the enzymes ThiF and MoeB respectively, prior to sulfur acceptance from the donor cysteine [409, 410, 412, 413, 434]. ThiF and MoeB are closely related to the Ub-conjugating E1 enzymes, and all of them display a characteristic architecture, with an N-terminal Rossmann-fold nucleotide-binding domain and a C-terminal β -strand-rich domain containing conserved cysteines [409]. Interestingly, in the case of the thiamine pathway, it has been shown that ThiS also gets covalently linked to a conserved cysteine in the ThiF enzyme, albeit via an acyl-persulfide linkage, unlike the direct thioester linkage of the E1-Ub covalent complex [410, 434]

(Fig. 33). However, no equivalent covalent linkage between MoeB and ThiS has been reported [411] (Fig. 33). There are other specific similarities between the eukaryotic Ub/Ubls and ThiS/MoaD, such as the presence of a conserved C-terminal glycine and the mode of interaction with their respective adenylating enzymes [409, 445]. These observations indicated that core components of the eukaryotic Ub-signaling system and the interactions between them were already in place in the prokaryotic sulfur transfer systems, and implied direct evolutionary connection between them [409, 585].

Homologs of other central components of the eukaryotic Ub-signaling pathway have also been detected in bacteria, such as the TS-N domain found in prokaryotic translation factors, which is the precursor of the helical ubiquitin-binding UBA domain [586-588]. Similarly, members of the papain-like fold, zincin-like metallopeptidases and the JAB domain superfamilies are also abundantly represented in prokaryotes [459, 573-578, 589]. However, to date there is no evidence for the functional interactions of any of the prokaryotic versions of these domains with endogenous co-occurring counterparts of Ub/Ubls and their ligases in potential pathways analogous to eukaryotic Ub signaling. Thus, despite the reasonably clear understanding on the possible precursors of Ub/Ubls and the E1 enzymes, the evolutionary process by which the complete eukaryotic Ub-signaling system as an apparatus for protein modification was pieced together remains murky. To address this problem we carried out a systematic comparative genomic analysis of the Ub-like (Also referred to as the β -grasp fold in the SCOP database [58]) fold in prokaryotes to decipher its early evolutionary radiations. We then utilized the vast dataset of contextual information derived from newly sequenced prokaryotic genomes to systematically identify the potential functional connections of the relevant members of the Ub-like fold and other functionally associated enzymes such as the E1/MoeB/ThiF (E1-like) family.

As a result of this analysis we were able to identify several new members of the Ub-like fold in prokaryotes as well as functionally associated components such as E1-like enzymes, JAB hydrolases and E2-like enzymes, which appear to interact even in prokaryotes to form novel pathways related to eukaryotic Ub signaling. We present evidence that not only are there multiple adenylating systems of Ub-related proteins in prokaryotes, but also predicted intricate pathways using JAB-like peptidases and E2-like enzymes in the context of diverse ubiquitin related proteins.

Application of Methods

I made the initial discovery of the presence of several previously undetected prokaryotic Ub-like families and their genome associations with other core components of the eukaryotic Ub-modification system through sequence and genome contextual analysis methods. Following this, Dr. Iyer and I worked closely to meticulously and thoroughly identify and classify all Ub-like homologues and their corresponding conserved gene neighborhoods of interest. Additional sequence and structural analyses were also split equally between Dr. Iyer and myself. Dr. Aravind provided input and was also involved in examining results, screening our results for potentially missed associations.

The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the BLASTP program [38]. A complete list of these genomes and the predicted proteomes of prokaryotes used in this analysis in fasta format can be downloaded from: <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>.

Additional sequences, from microbial genomes that have been sequenced but not completely assembled and submitted to the GenBank database were also used in this analysis. A list of these

prokaryotic genomes, from which sequences have been deposited in GenBank can be accessed from the following URL: http://www.ncbi.nlm.nih.gov/genomes/static/eub_u.html.

Gene neighborhoods were determined using a custom script that uses completely sequenced genomes or whole genome shot gun sequences to derive a table of gene neighbors centered on a query gene. Then the BLASTCLUST program is used to cluster the products in the neighborhood and establish conserved co-occurring genes. These conserved gene neighborhood are then sorted as per a ranking scheme based on occurrence in at least one other phylogenetically distinct lineage (“phylum” in NCBI Taxonomy database), complete conservation in a particular lineage (“phylum”) and physical closeness on the chromosome indicating sharing of regulatory -10 and -35 elements.

Profile searches were conducted using the PSI-BLAST program with either a single sequence or an alignment used as the query, with a default profile inclusion expectation (E) value threshold of 0.01 (unless specified otherwise), and was iterated until convergence. For all searches involving membrane-spanning domains we used a statistical correction for compositional bias to reduce false positives due the general hydrophobicity of these proteins [544]. The library of profiles for various signaling domains was prepared by extracting all alignments from the PFAM database (<http://www.sanger.ac.uk/Software/Pfam/index.shtml>) and updating them by adding new members from the NR database. These updated alignments were then used to make HMMs with the HMMER package [63] or PSSM with PSI-BLAST. Multiple alignments were constructed using the T_Coffee, MUSCLE and PCMA programs followed by manual adjustments based on PSI-BLAST results [62, 171, 172]. The GIBSS sampling method, as implemented in the MACAW program, was used for the identification and statistical evaluation of conserved motifs in multiple

protein sequences [590, 591]. All large-scale sequence analysis procedures were carried out using the TASS package (V.Anantharaman, S.Balaji and L.A unpublished). Structural manipulations were carried out using the Swiss-PDB viewer program [174]. Searches of the PDB database with query structures were conducted using the DALI program [95, 96]. Protein secondary structure was predicted using a multiple alignment as the input for the JPRED program, with information extracted from a PSSM, HMM and the seed alignment itself [173]. Similarity based clustering of proteins was carried out using the BLASTCLUST program: <ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html>. Phylogenetic analysis was carried out using the maximum-likelihood, neighbor-joining and least squares methods [100, 176, 592]. This process involved the construction of a least squares tree using the FITCH program or a neighbor joining tree using the NEIGHBOR program (both from the Phylip package) [100], followed by local rearrangement using the Protml program of the Molphy package [592] to arrive at the maximum likelihood (ML) tree. The statistical significance of various nodes of this ML tree was assessed using the relative estimate of logarithmic likelihood bootstrap (Protml REL-LBP), with 10,000 replicates. Text versions of all alignments reported in this study can be obtained in the Supplementary material.

Results and Discussion

Identification of novel prokaryotic Ub-related proteins

We investigated the origin of ubiquitin and the ubiquitin signaling system as a part of a comprehensive investigation on the evolutionary history of the ubiquitin-like (β -grasp) fold [429]. Earlier studies had shown that ThiS and MoaD are the closest prokaryotic relatives of the eukaryotic Ub/Ubls both in structural and functional terms [412, 434]. Structural similarity-based

clustering using the pair-wise structural alignment Z-scores derived from DALI program, as well morphological examination of the structures showed that several additional members of the β -grasp fold prevalent in prokaryotes are equally closely related to the eukaryotic Ub/Ubls. The most prominent of these was the RNA-binding TGS domain, which was previously reported by us as being found fused to several other domains in multidomain proteins such as the threonyl tRNA synthetase, OBG-family GTPases and the SpoT/RelA like ppGppp phosphohydrolases [463] (also see SCOP database). The β -grasp ferredoxin, a widespread metal-chelating domain, is also closely related, but is distinguished by the insertions of unique cysteine-containing flaps within the core β -grasp fold that chelate iron atoms [593]. Another version of the β -grasp fold closely related to the Ub-like proteins is subunit B of the toluene-4-monooxygenase system (e.g. PDB 1t0q) [472], which is sporadically encountered in several proteobacteria and actinobacteria (Table 5).

In order to identify novel prokaryotic Ub-related members of the β -grasp fold we initiated transitive PSI-BLAST searches, run to convergence, using multiple representatives from each of the above mentioned structurally characterized versions. Searches with the TGS domains and ThiS or Moad proteins were considerably effective in recovering diverse homologs with significant expect (e)-values ($e \leq 0.01$). Searches from these starting points were reasonably symmetric -- thus, searches initiated with various ThiS or Moad proteins detected eukaryotic URM1, representatives of the TGS domain, as also the β -grasp ferredoxins. Likewise, searches initiated with different representatives of the TGS domains also recovered ThiS, Moad and representatives of the β -grasp ferredoxins. These searches also recovered several previously uncharacterized prokaryotic proteins in addition to the above-stated previously known representatives of the Ub-like fold. These included several divergent small proteins equally

related to both ThiS and Moad, the N-terminal regions of a group of ThiF/MoeB-related (E1-like) proteins from various bacteria, the N-terminal regions of a family of bacterial RNAses with the Mut7-C domain, the N-terminal region of the family of tail assembly protein I of the lambdoid and T1-like bacteriophages, and the RnfH family that is highly conserved in numerous bacteria. For example, searches initiated with the *Escherichia coli* ThiS recovered the tail protein I of the phage lambda ($e=10^{-3}$; iteration 3) and T1 ($e=10^{-4}$; iteration 5) and the N-terminal domains of a *Thermotoga maritima* Mut7-C RNase ($e=10^{-5}$, iteration 5, gene TM_0779) and a E1-like protein from (e.g. $e=10^{-4}$; iteration 7 from *Shewanella oneidensis*, gene: SO1475, gi: domain of the *Bacillus subtilis* *Campylobacter jejuni* ($e=.01$; iteration 11, gi: 57166736). Likewise, a search initiated with the TGS 31340486). We prepared individual multiple alignments of all the novel families of proteins containing regions of similarity to the Ub-like β -grasp domains and predicted their secondary structures using the JPRED method that combines information from Hidden Markov models, PSI-BLAST profiles and amino acid frequency distributions derived from the alignments. In each case the predicted secondary structure of the region detected in the searches showed a characteristic pattern with two N-terminal strands, followed by a helical segment and another series of around 3 consecutive strands. This pattern is completely congruent with that observed in the Ub-like β -grasp proteins (See SCOP database), and was used as a guide along with the overall sequence conservation to prepare a comprehensive multiple alignment including all the major prokaryotic representatives of the Ub-like β -grasp domains (Fig. 34). Examination of the sequence across the different families revealed a similar pattern of hydrophobic residues that are likely to form the core of the β -grasp domain, as suggested by the structures of the ThiS, Moad and URM1 structures, and a highly conserved alcohol-group containing residue (serine or threonine) before

	Operon type	Phyletic pattern	Protein coded by conserved genes neighborhoods/ comments
1	Thiamine biosynthesis	All known bacterial lineages	ThiS, ThiG, ThiF, ThiC, ThiD, ThiE, ThiH and ThiO. - In many proteobacteria and actinobacterium RxyI, the ThiS is fused to a ThiG. In a subset of δ/ϵ proteobacteria and Low GC gram positive bacteria, the ThiS is fused to a ThiF and these operons also encode a second solo ThiS-like protein.
2	Molybdenum cofactor biosynthesis	All known bacterial and most archaeal lineages	MoaE, MoaC and MoaA. - In some rare instances, MoeB is present in the same operon as MoaD.
3	Tungsten cofactor biosynthesis	a) Euryarchaea: Mace, Mmaz, Paby, Pflur, Pflur, Phor, Tkod b) $\alpha, \beta, \gamma, \delta/\epsilon$ Proteobacteria: Aehr, Asp., Dace, Ddes, Dpsy, Dvul, Gmet, Gsul, Mmag, Pcar, Pnap, Ppro, Rfer, Rgel, Sfum, Wsuc c) Low GC gram positive: Chyd, Moth, Swol, Teth, The d) Actinobacteria: Sthe e) Other bacteria: Tth	MoaD, Aldehyde-ferredoxin oxidoreductase, Moeb, MoaE, MoaA, Pyridine disulfide oxidoreductase, and 4Fe-S ferredoxin. - In <i>Azoarcus</i> , the MoaD is fused to C-terminal to the AOR (Fig. 3)
4a	Siderophore biosynthesis	β and γ proteobacteria: Neur, Nmml, Rsol, Pflu, Hche, Pstu, Pput	ThiS/MoaD-like Ub (PdtH), E1-like enzyme fused to a Rhodanese domain (PdtF), JAB (PdtG), CaIB like coA transferase (PdtI) and AMP-acid ligase (PdtJ). - Experimentally characterized siderophores encoded by this pathway include PDTC and quinolobactin
4b	Uncharacterized operon encoding a ThiS/MoaD, a JAB peptidase and E1-like enzyme	a) $\gamma, \delta/\epsilon$ proteobacteria: Adeh*, Aehr*, Noce b) Cyanobacteria: Ana, Avar, Gvio*, Npun, Pmar Syn, Telo	E1 fused to a Rhodanese domain and JAB. - Only species marked with an * additionally possess a ThiS/MoaD-like Ub
4c	Uncharacterized operon with a ThiS/MoaD, E1-like enzyme, a JAB and a Cysteine synthase	a) α, γ proteobacteria: Paer, Rpal b) Acidobacteria: Susi c) Actinobacteria: RxyI d) Bacteroidetes/Chlorobi: Srub e) Chloroflexus: Caur	E1 is fused to a Rhodanese domain
4d	Uncharacterized operon with a ThiS/MoaD, JAB, Cysteine synthase and ClpS	Actinobacteria: Fsp., Mtub, Nfar, Nsp., Save, Scoe, Tfus	Additionally the operon encodes an uncharacterized conserved protein with an α -helical domain (Fig 3).
4e	Operons with genes for sulfur metabolism proteins	a) δ/ϵ proteobacteria: Gmet, Wsuc b) Low GC gram positive: Amet, Bcer, Chyd, Csc, Cthe, Dhaf c) Bacteroidetes/Chlorobi: Cpha d) Actinobacteria: Nsp., Acel e) Crenarchaea: Pyae	ThiS/MoaD-like protein, JAB, E1-like protein, SirA, sulfite/sulfate ABC transporters, PAPS reductase, ATP sulfurylase, sulfite reductase, O-acetylhomoserine sulfhydrylase and Adenylylsulfate kinase. - The ThiS/MoaD domain in Nsp and Acel are fused to a Sulfite reductase
5	Phage Tail assembly associated Ub	Lambdoid and T1 phages	Ub-like TAPI, TAPK protein with a JAB and NlpC domains, and TAPJ. - The TAPI proteins additionally have a C-terminal domain that is separated from the Ub-domain by a glycine rich region. In some prophages, TAPI is fused to the TAPJ protein. In one particular prophage of Ecol (Fig. 3) the TAPI is fused to the JAB. The NlpC domains of these versions almost always lack the JAB domain. These latter operons also encode a β -strand rich domain containing protein (labeled 'Z' in Fig. 4)
6a	Uncharacterized operon with a triple module protein containing an E2-like, E1-like and JAB domains	a) $\alpha, \beta, \gamma, \delta/\epsilon$ proteobacteria: gKT 71, Goxy, Maqu, Msp, Nwin, Obat, Pnap, Rmet, Rsph, Saci, Sdeg, Xaxo. b) Low GC gram positive: Cper	Triple module protein with E2 (UBC), E1-like domain and JAB, lined in a single polypeptide in that order. In most operons, these are almost always next to a Metallo- β -lactamase.
6b	Uncharacterized operon coding a multidomain protein with E2 and E1 domains	a) $\alpha, \beta, \gamma, \delta/\epsilon$ proteobacteria: Ecol, Elit, Gura, Obat, Parc, Pber, Reti, RhNGR234a, Rosp., Rusp., Shsp., Vcho. b) Actinobacteria: Asp. c) Low GC gram positive: Cper	Multi-domain protein with E2 and E1 domains, JAB and pol β superfamily nucleotidyl transferase. - Both the E2+E1 protein and the JAB are closely related to the corresponding sequences of the operons in the previous row of the table. Most of these operons are in ICE-like mobile elements and plasmids.
6c	Uncharacterized operon coding a distinctive multidomain protein with E2 and E1 related domains	α proteobacteria: Mlot, Mmag, Reti, RhNGR234, Rpal	Multi-domain E2+E1 protein, JAB and predicted metal binding protein. - In Mmag and Rpal, the E1 domain is fused to a distinct domain instead of E2. The E2-like domain has a conserved cysteine in place of the conserved histidine of the classical E2s
6d	Uncharacterized operon coding a Ub-like protein, a JAB, an E1-like protein and an E2-like protein	a) $\beta, \delta/\epsilon$ proteobacteria: Asp., Bvie, Cnc, Daro, Pnap, Ppro, Posp., Rfer, Rmet, Rsol b) Low GC gram positive: Bcer, Bthu c) Cyanobacteria: Ana, Avar d) Bacteroides: Bthe	Ub-like protein, JAB, E1-like, E2-like and novel α -helical protein. - The E2-like protein lacks the conserved histidine of the classical E2-fold. However, they have an absolutely conserved histidine C-terminal to the conserved cysteine. The rapidly diverging α -helical protein has several absolutely conserved charged residues suggesting that it may function as an enzyme. The JAB domains of this family additionally have an N-terminal $\alpha+\beta$ domain characterized by a conserved arginine and tryptophan residue.
6e	Uncharacterized operons coding a protein with tandem repeats of a ubiquitin-like domain (polyUbl)	a) $\alpha, \beta, \gamma, \delta/\epsilon$ proteobacteria: Amac, Bvie ^b , Mlot ^b , Nham ^b , Pnap ^b , Rmet ^b , Rpal ^b , Shsp. ^b , Vpar ^b b) Actinobacteria: Fsp. ^a c) Cyanobacteria: Ana, Syn	PolyUbl, inactive E2-/RWD like UBC fold domain, multi-domain protein with a JAB fused to an E1 domain, and a metal binding protein (labeled Y in Fig 3). - The polyUbls contain between 2-3 Ub-like domains (Fig. 3). Some versions of the E1 domain have a distinct domain in place of the JAB domain (marked with a ^a , Domain X in Fig. 3). In some species (marked with a ^{ab}), the poly Ubl is fused to an inactive E2-like domain. Amac has a solo Ub-like
7	Ubl fused to Mut7-C	a) Wide range of β proteobacteria and Avin b) Actinobacteria: Mtub, Scoe, Save, Mavi, Nfar, Tfus c) Acidobacteria: Susi d) Cyanobacteria: Npun e) Tmar	No conserved genome context.
8	Uncharacterized operon encoding a RnfH family protein	A wide range of β and γ proteobacteria and Mmag	Ub-like RnfH, a START domain containing protein, SmpA and SmpB.
9	Mobile RnfH operon	α, β, γ proteobacteria: Asp., Daro, Pstu, Rcap, Zmob	Ub-like RnfH, RnfB, RnfC, RnfD, RnfG and RnfE. - These components are part of an electron transport chain involved in reductive reactions such as nitrogen fixation
10	Toluene-O-Xylene Mono-oxygenase Hydroxylase	a) α, β and γ proteobacteria: Bcep, Bsp., Daro, Paer, Pmen, Psp, Reut, Rmet, Rpic, Xaut b) Actinobacteria: Rsp., Fsp.	Ub-like TmoB, Toluene-4-monoxygenase hydroxylase (TmoA), hydroxylase/monoxygenase regulatory protein (TmoD), Toluene-4-monoxygenase hydroxylase (TmoE), Rieske 2Fe-S protein (TmoC), NADH-ferredoxin oxidoreductase (TmoF), 4-oxalocrotonate decarboxylase (4OCCD), 4-oxalocrotonate tautomerase (4OCTT)

Table 5. Phyletic distributions and conserved gene neighborhoods of prokaryotic Ub-like families.

helix-1. A similar secondary structure and conservation pattern was also found in two additional Ub-related protein families that we recovered using contextual information from analysis of gene neighborhoods and domain fusions (Fig. 34; See next two sections for details). Taken together these observations strongly supported the presence of an Ub-related β -grasp fold in all the above-detected groups of proteins.

Like the ThiS, MoaD and URM1 proteins, the phage tail assembly protein I (TAPI) and one of the other newly detected Ub-related families also showed a highly conserved glycine at the C-terminus of the β -grasp domain, suggesting that they might participate in similar functional interactions with other proteins or undergo thiolation (Fig. 34). The remaining newly detected members, while showing a similar overall conservation as the above families, do not contain the glycine or any other highly conserved residue at the C-terminus of the domain.

Fig 34. Multiple alignment of ThiS/MoaD-like ubiquitin domain containing proteins.

Proteins are listed by gene name, species abbreviation, and gi number; separated by underscores. Amino acid residues are colored according to side chain properties and the extent of conservation in the multiple alignment. Coloring is indicative of 70% consensus, which is shown on the last line of the alignment. Consensus similarity designations and coloring scheme are as follows: h, hydrophobic residues (ACFILMVWY) shaded yellow; s, small residues (AGSVCDN) colored green; o, alcohol group containing residues (ST) colored aqua; b, big residues (EFHIKLMQRWY) colored purple and shaded in light grey. Secondary structure assignments are shown above the alignment where E represents a strand and H represents a helix. The families of the ubiquitin-related domains are shown to the right. Also shown to the right are the row numbers in Table 5, that describe a particular family.

Individual families also possess their own exclusive set of highly conserved residues, suggesting that each might participate in their own specific conserved interactions with other proteins or nucleic acids.

Identification of the contextual associations of the prokaryotic Ub-related proteins and their functional partners

Detection of architectures and conserved gene neighborhoods

Different types of contextual information can be obtained by means of prokaryotic comparative genomics and used to understand functionally uncharacterized proteins. 1) Fusions of uncharacterized domains or genes to functionally characterized domains or genes suggest participation of the former in similar processes as the latter. 2) Clustering of genes in operons usually implies coordinated gene expression and conserved prokaryotic gene neighborhoods are a strong indication of functional interaction especially through physical interactions of the encoded protein products. The power of contextual inference, especially for the less prevalent protein families, has been considerably boosted due to the enormous increase in the data from the various microbial genome sequencing projects [197, 549].

Accordingly, we set up a protocol to comprehensively identify the network of contextual connections centered on the prokaryotic Ub-related proteins detected in the above searches, and used it to infer the functional pathways in which they participate. We first determined the complete domain architectures of all the Ub-like proteins using a combination of case-by-case PSI-BLAST searches and searches against libraries of PSSMs or HMMs of previously characterized protein domains. We then established the gene neighborhoods (see Material and Methods) for these Ub-like proteins and found a number of conserved neighborhoods containing genes for specific protein families often co-occurring with the Ub-like proteins. Each of the

families belonging to the conserved neighborhoods were used as starting points for further PSI-BLAST searches to identify homologous proteins in prokaryotic genomes. These homologs were then used as foci to identify any conserved gene-neighborhoods occurring with them. This way we built up a comprehensive set of conserved gene neighborhoods for the Ub-like proteins as well as their putative functional partners and their homologs which were identified via contextual analysis. As a result we identified several persistent architectural and gene neighborhood themes associated with the prokaryotic Ub-like proteins. We discuss below the most prominent of these, especially those having relevance to the early evolution of the Ub-signaling related pathways.

Common architectural themes in prokaryotic Ub-like proteins

Several families of prokaryotic Ub-like proteins, namely ThiS, MoaD, RnfH, TmoB and a newly-detected family typified by *Ralstonia solanacearum* RSc1661 (gi: 17428677, see below) are characterized by a single stand-alone Ub-like domain. In several cases the ThiS and MoaD are fused to ThiG and MoaE (Fig. 35), which respectively are their functional partners in the transfer of sulfur to the substrates (Fig. 33). We also noted that a distinct version of ThiS is fused to the C-terminus of the sulfite reductase in certain actinobacteria (e.g. *Nocardiodetes* and *Acidothermus cellulolyticus*), while MoaD might be fused to aldehyde ferredoxin oxidoreductase (*Azoarcus*) (Fig. 35). Another newly characterized family of Ub-domains typified by the protein mlr6139 from *Mesorhizobium loti* (gi: 14025878) is characterized by three tandem repeats of the Ub-like domain (Fig. 35, see below for details). A family of Ub-like domains, distinct from ThiS, is found fused to the N-terminus of the adenylating Rossmann fold domain of certain ThiF proteins, such as that from *Campylobacter jejuni* (gi: 57166736) (Fig. 35). In the lambda and T1 phage TAPI proteins, the Ub-like domain is fused to another small globular C-terminal domain via a glycine rich low

complexity linker. In some cases the TAPI protein itself may be fused to the tail-assembly protein J (TAPJ) or K (TAPK), which contains two peptidase domains, namely the JAB domain and NlpC/P60 domain with the papain-like fold (Fig. 35) [575]. In the proteins typified by the

Fig 35. Domain Architectures of the ThiS/MoaD-like ubiquitin domains and functionally associated proteins.

Architectures belonging to a particular gene neighborhood or related pathway are grouped in boxes. Proteins are identified below the architectures by gene name, species abbreviation, and gi number demarcated by underscores. Proteins belonging to the classical thiamine and MoCo/WCo biosynthesis pathways are shown above the purple line. Abbreviations are as follows: Rhod: Rhodanese domain; X: β -strand rich, poorly conserved globular domain; ZnR: zinc-ribbon domain; TAPI-C: Domain found c-terminal to the phage λ - TAPI-like ubiquitin domain, JAB-N: An α + β domain found N-terminal to some JAB proteins.



Thermotoga maritima TM_0779, the N-terminal Ub-like domain is linked to a C-terminal Mut7-C RNase domain and a zinc ribbon domain (Fig. 35) [594]. Iterative sequence profile searches with the Mut7-C domain as a query recovered the previously characterized PIN (PiIT-N) RNase domains with significant e-values. The two domains share an identical pattern of conserved catalytic residues, suggesting a similar enzymatic mechanism [595]. The TGS domain, as previously reported, was almost always found in various RNA-binding multi-domain proteins; hence it is not discussed here in detail [463]. Likewise, the architectures of β -grasp ferredoxins, which are typically found as a part of multi-domain oxido-reductases, have been previously considered in depth, and are not dwelt upon in detail here [524].

Conserved gene neighborhoods related to the thiamine biosynthesis pathway

The multi-step biosynthetic pathways for the major co-factor thiamine is the experimentally best characterized of the prokaryotic systems involving Ub-like sulfur transfer proteins and associated E1-like enzymes. Furthermore, there has also been a comprehensive comparative genomics analysis of the components of the prokaryotic thiamine biosynthetic pathway [596]. In our current report we only focus on the associations in these systems pertinent to the evolution of the Ub-signaling related pathways and previously un-noticed features of the distribution and gene neighborhoods of the ThiS genes. The ThiS protein is highly conserved in all the major bacterial and archaeal lineages suggesting that it may be traced back to the last universal common ancestor (LUCA). In most bacterial lineages ThiS is encoded within a large operon including several other genes for thiamine biosynthesis. These include genes encoding proteins for both the major branches of the thiamine biosynthetic pathway, i.e. the aminoimidazole ribotide utilizing branch with ThiC and ThiD, and the sulfur transfer and

hydroxyl-ethyl-thiazole forming branch with ThiS, ThiG, ThiO, ThiH, as well as the stem combining the products of branches to form thiamine phosphate (ThiE) (Fig. 36) [596].

Though the individual genes occurring in this conserved gene neighborhood shows some variability across different bacteria, ThiS is most strongly coupled with ThiG (~87%); its physically interacting functional partner within the operon. The next strongest coupling of ThiS in bacteria is with its other complex-forming partner, the adenylating enzyme ThiF (~17%). This is not surprising, given that ThiF and ThiG compete for ThiS to catalyze two successive steps in the sulfur incorporation process [409, 597]. Very rarely, ThiS may also be coupled with ThiC (e.g. *Cytophaga hutchinsonii*). The genes for the group of ThiF proteins containing a fused Ub-like domain at their N-termini (see above) typically co-occur in operons with standalone ThiS genes (Fig. 36). This suggests that their fused Ub-like domain plays a role different from the stand-alone ThiS protein. However, in a single case (*Pelobacter propionicus*), the Ub-like domain-ThiF fusion proteins do not occur in an operon with other thiamine biosynthesis genes, instead co-occurring with O-acetylhomoserine sulfhydrylase and cysteine synthase (Fig. 36). Similar operonic association of ThiS alone, or ThiS and ThiG with genes for cysteine biosynthesis like cysteine synthase, and sulfite transporter genes are also seen in *Pelodictyon* and *Chlorobium* (Fig. 36, Supplementary material). These represent multiple independent associations of thiamine biosynthetic genes with sulfur assimilation and cysteine biosynthesis genes, which is consistent with the fact that cysteine is the sulfur donor for the ThiS thiocarboxylate. The genes of the archaeal ThiS orthologs are not found in any conserved operons, and this is consistent with the previously noticed absence of ThiF and ThiG orthologs in the archaea, and the presence of an alternative branch for hydroxyl-ethyl-thiazole biosynthesis [596]. This observation suggests that

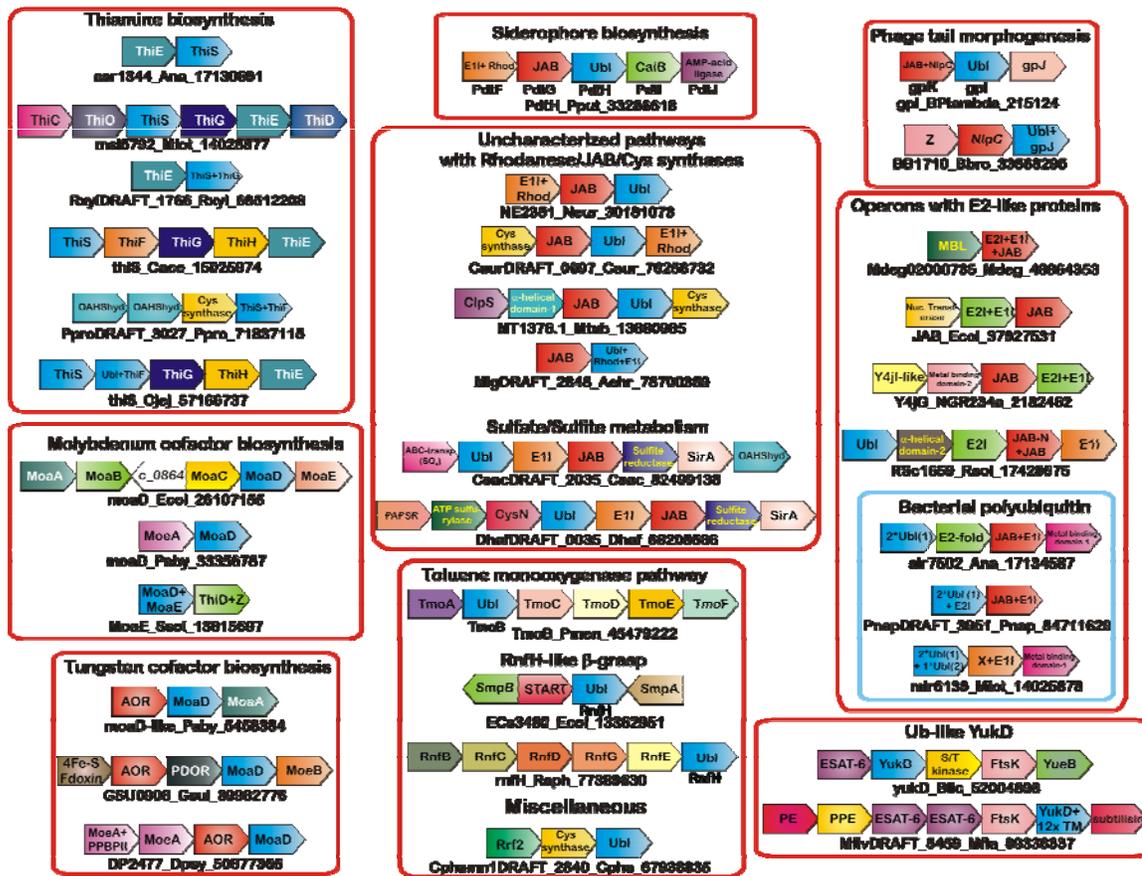


Fig 36. Gene neighborhoods of the prokaryotic ThiS/MoaD-like ubiquitin domains and functionally associated proteins.

Genes found in conserved neighborhoods are depicted as boxed arrows with the arrow head pointing from the 5' to the 3' direction. ThiS/MoaD-like proteins are shaded in blue. Other than in the classical ThiS and MoaD pathways, ThiS/MoaD/Ubiquitin-like proteins are labeled Ubi for ubiquitin-like domain. The ThiS/MoaD-like proteins in each operon are identified in black lettering below the neighborhood by gene name, species abbreviation, and gi number demarcated by underscores. In the instances where ThiS/MoaD-like domains are absent, the operons are identified by the JAB domain containing protein. Alternatively well-characterized genes are shown below the boxed arrows for that gene. Boxed arrows with no colors represent poorly conserved proteins. Conserved neighborhoods are clustered according to major assemblages of gene neighborhood as described in the text. Abbreviations are as follows: AOR: aldehyde ferredoxin oxidoreductase; Cys Synthase: Cysteine Synthase; Rhod: Rhodanese domain; Z: poorly characterized protein with an alpha+beta domain with several conserved charged residues; X: beta-strand rich globular domain.

the archaeal ThiS genes might have even been recruited for a sulfur transfer process distinct from thiamine biosynthesis.

Conserved gene neighborhoods related to molybdenum and tungsten cofactor biosynthesis

The MoaD-MoeB system in molybdenum and tungsten cofactor biosynthesis mirrors the ThiS-ThiF system in thiamine biosynthesis. MoaD is also conserved across all major archaeal and bacterial lineages suggesting that it existed in the LUCA. Unlike ThiS, MoaD is present in Mo/W cofactor biosynthesis operons in both bacteria and archaea. This implies that both ThiS and MoaD had probably diverged from each other by the time of the LUCA, but the recruitment of ThiS for a sulfur transfer system in thiamine biosynthesis emerged early in the bacterial lineage, only after it had split from the archaeal lineage. In contrast, the deployment of MoaD in Mo/W cofactor biosynthesis appears to have happened in the LUCA itself. The Mo/W cofactor biosynthesis operons from different bacteria encode a variety of proteins, including those involved in using the GTP precursor (MoaA and MoaC), the MoeB, MoaD and MoaE products, which are downstream of the former and involved in molybdopterin biosynthesis, and MoeE, MogA, MobD, and the MOSC domain proteins involved in formation of MoCo/WCo and its terminal derivatives [598-600]. While the operons show variability across prokaryotes in terms of the different genes included in them, the core conserved gene neighborhood in bacteria contains the genes for MoaD and MoaE, which together constitute the molybdopterin (MPT) synthase, that transfers the sulfur from the MoaD thiocarboxylate to the precursor Z (cyclic pyranopterin monophosphate) to form MPT [598, 601] (Fig. 33, 36). In a few cases MoaD might be adjacent to the gene for MoeA, which acts on the product downstream of the reaction catalyzed by the MPT synthase. MoaD, unlike ThiS, is rarely found immediately adjacent to the gene for its adenylating enzyme, MoeB (Fig. 36). This distinction may be related to experimental results which indicate

that MoaD and MoeB do not form a covalently-linked persulfide or thioester complex, unlike ThiS and ThiF or the Ub/Ubl and the E1s (Fig. 33) [411].

A distinct set of MoaD genes are found strictly adjacent to genes encoding an aldehyde ferredoxin oxidoreductase (AOR) in a sporadic group of phylogenetically distant archaea and bacteria (Table 5), suggesting that they might constitute a mobile gene cluster. Additionally, these gene neighborhoods often include MoeB and occasionally other cofactor biosynthesis genes such as MoaA and MoaE, and a pyridine disulfide oxidoreductase in the close vicinity to MoaD and the AOR genes (Fig. 36). In some organisms this MoaD containing gene cluster is distinct from the MoCo biosynthesis operon found elsewhere in the genome of the same organism. Experimentally characterized versions of these AORs have been shown to utilize a tungsten-containing variant of the cofactor [602]. Taken together, these observations suggest that these AOR linked MoaD genes might specifically participate in the synthesis of molybdopterin for WCo generation for the AORs. In *Sulfolobus* MoaD and MoaE are linked to ThiD, raising the intriguing possibility that they might have been secondarily recruited for an additional role in thiamine biosynthesis (Fig. 36).

Other potential novel pathways involving ThiS/MoaD-like proteins and E1-like enzymes

Beyond the above-stated operons, with the *bona fide* ThiS/MoaD and the ThiF/MoeB enzymes involved in conventional thiamine and MoCo/WCo biosynthesis, we also recovered several other predicted bacterial operons encoding homologous proteins. These gene clusters typically encode a ThiS/MoaD related protein and an E1-like enzyme related to ThiF/MoeB with a C-terminal rhodanese domain, but do not contain any genes encoding other components of the two co-factor biosynthesis pathways (Fig. 35, Fig. 36, Table 5). The bacteria which contain these predicted operons also contain independent thiamine or molybdenum operons, highlighting the

functional distinctness of the pathways coded by these operons (Table 5). Interestingly, this class of predicted operons also often contains a gene encoding a standalone version of the JAB metallopeptidase, which forms a monophyletic clade within the tree of all JAB domains (Fig. 36, 37). There are at least five distinct sub-types of this class of operons, which show a sporadic distribution across phylogenetically diverse bacteria suggesting possible dispersion through lateral gene transfer (Table 5, Fig. 36). One of these sub-types of gene clusters has been shown to encode components of the biosynthetic pathway for the siderophores and secreted protective compounds PDTC (pyridine-2,6-bis(thiocarboxylic acid)) and quinolobactin in *Pseudomonas stutzeri*/*P.putida* and *P. fluorescens*, respectively [508, 603]. Our analysis of gene neighborhoods showed that related operons are also found in several distantly related proteobacteria such as *Ralstonia solanacearum* and *Nitrosomonas europaea*, suggesting that such compounds might be widely produced (Table 5, Fig. 36).

There are considerable differences in the genes and corresponding biosynthetic pathways (related to amino acid biosynthetic pathways) producing the basic molecular skeleton of each of these metabolites. For example, in the case of quinolobactin a xanthurenic acid skeleton is used, whereas in the case of PDTC, a dipicolinic acid skeleton is used (Fig. 33) [508, 603]. However, all of these operons contain a conserved core of genes whose products catalyze the critical sulfurylation step required for the production of all of these compounds [508, 603]. This core group encodes a carboxylate AMP ligase, which adenylates a carboxylate group on the precursor, and proteins for a sulfur transfer system that forms a thiocarboxylate group from the carboxy adenylate produced by the AMP ligase (Fig. 33). The proteins of the sulfur transfer system include an E1-like protein with a C-terminal rhodanese domain, a ThiS/MoaD-like protein and a protein with a JAB metallopeptidase domain (Fig. 36). The first two enzymes are likely to

participate in a sulfur transfer pathway similar to those seen in the conventional thiamine and MoCo/WCo pathways, with the rhodanese domain probably abstracting the sulfur from a small molecule donor like cysteine (as in the case of ThiI), and the E1-like protein adenylyating and transferring the sulfur to the ThiS/MoaD-like protein to form a terminal thiocarboxylate (Fig. 33).

Most other predicted operon subtypes of this class appear to show different variants of the core sulfur transfer system seen in the above-described siderophore biosynthesis operons (Table 5, Fig. 36). A simple subtype seen in a wide range of bacteria contains just three genes encoding a ThiS/MoaD-like protein, a protein combining an E1-like module and a rhodanese domain and JAB domain peptidase. Derivatives of this basic subtype might simply contain genes for the JAB domain peptidase and E1+rhodanese protein (Table 5, Fig. 36). Another subtype additionally combines the cysteine synthase to the three genes of the basic operon, suggesting that they might couple sulfur transfer to the production of the major cellular sulfur donor cysteine. A variant of the cysteine synthase-containing operon subtype, which is particularly prevalent in the actinobacteria, includes ClpS which is involved in degradation of proteins through the Clp system and an uncharacterized helical protein that is almost exclusively encoded in this predicted operon subtype (Fig. 36). Other links to sulfur metabolism are hinted at by another major subtype of this class of operons, where genes for the ThiS/MoaD, JAB and E1-like proteins are combined with genes coding sulfite/sulfate ABC transporters, PAPS reductase, ATP sulfurylase, sulfite reductase, O-acetylhomoserine sulfhydrylase and Adenylylsulfate kinase. The E1-like protein of these predicted operons always lacks the C-terminal rhodanese-like domain. However, these operons always contain a SirA (Cysteine containing domain 1; CCD1) protein, which was predicted to play a role similar to rhodanese [604](Table 5, Fig. 36). These observations suggest that these operons are principally involved in the assimilation of sulfur from

sulfate/sulfite and this sulfur might be terminally transferred to the ThiS/MoaD-like proteins encoded by them.

The tail assembly operons of phages Lambdoid and T1-like phages

The genomes of lambdoid and T1-like phages are known to contain related tail assembly gene complexes [605]. In a large number of phages this complex encodes a protein TAPI that contains an Ub-like domain related to ThiS/MoaD (Fig. 34). The exact function of this protein tail assembly is unclear, but it is not incorporated into the mature tail. Analysis of the gene neighborhoods showed that TAPI is most often flanked by the genes of TAPK protein, with JAB and NlpC/P60 peptidase domains and the TAPJ protein, which is required for host specificity (Fig. 36). The JAB domains found in these gene associations are also a part of the monophyletic clade including those from the above-described class of operons. Variants of this organization, lacking either of the two flanking genes are seen in a few phages/prophages, and in a small group of phages TAPI is flanked by a version of TAPK containing only an NlpC/P60 peptidase domain (Fig. 36). It is possible that the latter versions are actually degenerate variants of the former versions and are typical of integrated prophages.

Predicted operons coding E1-like proteins, E2 (UBC)-like proteins, JAB peptidase and novel Ub-like proteins

A number of sets of predicted operons, each with a distinctive sporadic distribution across several phylogenetically distant bacteria and encoding proteins with JAB domain and E1-like enzymes, were recovered in our search for conserved gene neighborhoods. E1-like enzymes in these operons never contained a C-terminal rhodanese domain. However, they were typically fused, either at the N-terminus or the C-terminus, to the JAB domain. In the instances they were not fused to the JAB domain, there was always a JAB domain protein encoded by the

immediately adjacent gene in the predicted operon. One group of proteins, typified by an E1-like protein fused to a JAB domain at the C-terminus, also contained an additional conserved N-terminal domain, with a conserved histidine and cysteine (E.g. Mdeg02000735 from *Microbulbifer degradans*, gi:48864353). Iterative PSI-BLAST searches with the alignment of this domain as a seed recovered eukaryotic E2 (ubiquitin conjugating enzymes, UBC) enzymes as hits with significant e-values ($e=10^{-3}$, iteration 3). The predicted secondary structure of these domains was congruent with that of eukaryotic E2 domains, with a 4 strand β -meander and two flanking helices on either side [606]. Furthermore, the conserved histidine and cysteine of the bacterial proteins also precisely matched the cognate active site residues of the eukaryotic E2 enzymes, suggesting that the N-terminal domains of the bacterial domain are homologs of the E2 enzymes and likely to possess similar activity (Fig. 38).

In addition, each set of these predicted operons contain a distinct group of genes that almost exclusively co-occurred with a particular operon type. Based on the different groups of co-occurring genes we were able identify at least five major operon types (Table 5, Fig. 36). These groups of co-occurring genes coded for several conserved uncharacterized proteins, whose affinities we systematically investigated using sequence profile searches, secondary structure prediction and matches to libraries of profiles and HMMs for various previously characterized domains. The first of these operon types showed a very simple organization, usually with two genes: one of them encoded the triple module protein, with N-terminal E2-like and E1-like domains followed by a C-terminal JAB domain (Fig. 35). The second gene in the operon encoded a specialized version of the metallo- β -lactamase domain (Fig. 36). Another operon group typified by a conserved gene neighborhood from the *E.coli* integrative and conjugative element ICE [607] and related mobile elements was found to contain a nucleotidyl transferase of the polymerase β

fold [608], in addition to the genes encoding the E1-like and JAB domain proteins (Table 5, Fig. 36). Like the E1-like proteins from the first group of operons the E1-like proteins of this group also show a fusion to an E2-related domain with a conserved active site cysteine. Similarly, a conserved operon group prototyped by a gene neighborhood from the megaplasmid NGR234 of *Rhizobium sp* contains genes encoding two conserved uncharacterized proteins, one of which is predicted to contain a metal-binding domain based on the conserved pattern of two cysteines, a histidine and an acidic residue (Table 5, Fig. 36). We observed that the E1-like proteins encoded by both of these operon types contained an additional N-terminal domain with a conserved cysteine. Sequence searches with this N-terminal region recovered the UBC-like E2 domains from a variety of eukaryotes. The best hit to these domains was from a profile of the E2-like proteins and included a match to the conserved cysteine ($p < 10^{-5}$ match for this cysteine containing motif in a Gibbs sampling search including a wide range of known E2 domains). Secondary structure prediction for this conserved domain also showed complete congruence with the known structure of the E2 fold, suggesting that these N-terminal domains fused to the E1-like enzymes are also homologs of the eukaryotic E2 ubiquitin conjugating enzymes (Fig. 38).

A fourth operon type found in several diverse bacteria (Table 5) typically contained three additional genes in the conserved gene neighborhood, in addition to the genes of the JAB domain and E1-like proteins (Fig. 36). Furthermore, the JAB domain additionally has an N-terminal $\alpha+\beta$ domain that has a strictly conserved arginine and tryptophan residue (JAB-N, Fig. 35). These first of these encodes a small protein with a highly conserved glycine at their C-terminus. Secondary structure prediction revealed that this small protein has a progression of structural elements identical to that seen in the β -grasp fold (Fig. 34). The conservation pattern in this protein also strongly resembles that seen in the known β -grasp domains, and sequence-structure threading

using the PHYRE program also recovered β -grasp proteins (e.g. ThiS, PDB: 1tyg) as the best hits, suggesting that these are small stand-alone Ub-like proteins. The second protein encoded by this operon type was found to encode a largely α -helical protein with several absolutely conserved charged residues, suggesting that it might be an uncharacterized enzyme. The third conserved protein from these operons contained a conserved cysteine and gave significant hits to the profiles of the E2 Ub-conjugating enzymes, with the alignments spanning the conserved cysteine (Fig. 36). This relationship was also supported by their predicted secondary structure and general conservation pattern. While these proteins did not have the conserved histidine at the position often encountered in most E2 enzymes, they had an absolute conserved histidine further downstream (Fig. 38). Mapping of the sequences of representatives of this family of proteins on the structures of E2 enzymes showed that this downstream histidine from the helix would be positioned very close to the active site histidine of the classical E2 enzymes (Fig. 38). This would mean that these proteins are likely to effectively contain an active site similar to the classical E2 enzymes.

The fifth operon type is found sporadically in most proteobacterial lineages, cyanobacteria and certain actinobacteria (Table 5). Usually these operons contain two or three genes in addition to the central gene for an E1-like enzyme, which in most cases contains a JAB domain fused to the N-terminus of the E1-like module. However, in a subset of bacteria, the E1-like protein contains a fusion to an uncharacterized N-terminal domain in place of the JAB domain (Fig. 34). The conservation pattern of this domain is unrelated to that of the JAB domain, but it contains several conserved charged residues, making it tempting to speculate that it might perform a function analogous to the JAB domains. The other gene found in all operons of this type encodes a protein containing one to three repeats of an approximately 70-75 amino acid

domain. The conservation pattern is similar to that seen in UbIs, and the predicted secondary structure of this domain shows a progression completely congruent to other β -grasp fold domains (Fig. 34). Consistent with this, sequence-structure threading with the PHYRE program recovered the structures of the ThiS/MoaD proteins as the top hits (e.g. PDB: 1tyg). These observations strongly suggest that this group of proteins is comprised of one or more Ub-like domains (Table 5).

Furthermore, we noted that these predicted β -grasp domain proteins might also be fused with either of two unrelated C-terminal domains (Table 5). The first of these domains is a small domain of about 75 residues showing conservation pattern and secondary structure progression similar to the UbIs (Fig. 34). These domains also recovered ThiS/MoaD as their best hits in sequence-structure threading with the PHYRE program, implying that it might form the third Ub-like domain in a subset of these proteins. The second C-terminal domain found in a mutually exclusive subset of these proteins, also occasionally occurs as a standalone protein coded by a separate gene sandwiched between the genes for the multi- β -grasp domain protein and the JAB+E1 domain proteins (Fig. 35). Profile searches with an alignment of this domain recovered hits to the E2 enzymes and the eukaryotic RWD domain [606, 609], which contains a catalytically inactive version of the E2 fold as the best hits (e \sim .01-.005). This relationship was also supported by the congruence of the predicted secondary structure of these domains with that of the E2 and RWD domains [606]. Like the eukaryotic RWD domains, these bacterial domains also lacked the conserved cysteine residue, implying that they are likely to be catalytically inactive representatives of the E2-like fold (Fig. 38). The above operon type was also seen to encode another conserved protein with a C-x(3)-C-x(35-38)-H-x(2)-C signature (Fig. 36). The predicted

secondary structure of this potential metal-binding signature is consistent with proteins containing a Zn-finger domain, perhaps of the treble-clef fold.

The RnfH associated conserved gene neighborhoods and other miscellaneous operons

The RnfH protein is highly conserved across the β/γ proteobacteria (Table 5) and in each of these instances it occurs in a strongly conserved gene neighborhood also containing genes for a START domain protein, the tmRNA binding protein SmpB and a small membrane protein of unknown function SmpA. Within this conserved neighborhood the genes for the SmpB, the START domain protein and RnfH appear to share a common transcriptional regulatory region with the former gene being transcribed in the opposite direction to the latter two (Fig. 36). This neighborhood is of particular interest given that the SmpB-tmRNA complex is used in bacteria to tag proteins from mRNAs lacking stop codons with small peptide. This tag targets proteins for degradation analogous the eukaryotic Ub-system [610]. A second type of conserved gene neighborhood containing an RnfH gene is found sporadically in a few proteobacteria, where it is linked to group of Rnf genes whose products form a membrane associated complex involved in transporting electrons for various reductive reactions such as nitrogen fixation [611].

In addition to this, there other operons coding Ub related β -grasp domain proteins, such as the Tmo operon which encodes the toluene monooxygenase complex in several bacteria (Fig. 36, Table 5). TmoB, the Ub related protein of this complex, has been shown to be a subunit of monooxygenase, which binds a distinct conserved exposed ridge on the catalytic subunit [472]. However, it does not affect the activity of the enzyme *in vitro* and its exact role in the complex remains unknown.

Functional implications of the prokaryotic systems with components related to the eukaryotic Ub-signaling network

Much of the above-described diversity of prokaryotic functional systems involving Ub-signaling related proteins remains experimentally unexplored. However, the syntactical features of the domain architectures and conserved gene neighborhoods provide some hints regarding the general functional properties of these systems (Fig. 36, Fig. 39). One of the most striking features is the dichotomy in distribution, operon organization and domain architectures of the versions involved in thiamine and MoCo/WCo biosynthesis and majority of other operons (Table 5, Fig. 36). The former set of operons is highly conserved and is present across most bacterial and several archaeal lineages suggestive of a pattern of vertical inheritance from LUCA, or early in bacterial evolution. The other types of above-described operons are instead sporadic in their distribution and found patchily across phylogenetically unrelated bacteria (Table 5). The former types do not contain a single instance of a gene encoding a JAB domain protein or a fusion to a JAB domain. In contrast to the thiamine and MoCo/WCo operons, the majority of other operons code a JAB domain protein along with an E1-like enzyme and/or Ub-like protein (Fig. 36, Table 5). A subset of these, namely those involved in the biosynthesis of siderophore-like compounds and those associated with sulfur assimilation and cysteine synthase, are linked with genes encoding metabolic enzymes. This suggests a role for them in the biochemistry of sulfur transfer, albeit in pathways which are likely to be distinct from the thiamine and MoCo/WCo (Fig. 33). The other operons show no major links to metabolic enzymes suggesting that they might specify standalone regulatory pathways.

One of the most interesting features of these predicted functional systems is the presence of the JAB domain (Fig. 37), which is universally conserved in eukaryotes and is the primary

deubiquitinating peptidase associated with the proteasome [582, 583] (Fig. 38). The association of the JAB peptidase with just an Ub-like protein with a C-terminal glycine in the phage tail assembly operons strongly implies that the two domains form a functional unit even in the prokaryotes. It is quite probable that the phage TAPI is processed by the peptidase domains of TAPK, with the JAB probably releasing the Ub-like domain by cleaving at the point of the C-terminal-most glycine of the Ub-domain. A similar function may be envisaged for the JAB domain in the organisms where ThiS or MoeB is fused to some other proteins-- it might cleave off the Ub-like moiety and generate a free C-terminus for sulfur transfer. However, the strong association of the JAB with sporadically distributed operon types related to the *Pseudomonas* siderophore biosynthesis pathways is more mysterious. Based on complete absence of JAB proteins in the thiamine and MoCo/WCo pathways, we predict that in the pathways where the E1-like enzyme is found in association with the JAB domain it functions via a mechanism distinct from that used by classical ThiF or MoeB. This mechanism is likely to be closer to the Ub-transfer reaction of *bona fide* eukaryotic E1s, wherein the ThiS/MoeB or any other associated Ub-like protein is directly linked to a cysteine in the E1-like enzyme by a thioester linkage. In this situation, it is likely that the E1-like enzyme also transfers the covalently linked Ub-like protein to amino groups of lysines in particular target proteins. These linkages would then be cleaved by the associated JAB domain proteins (Fig. 33).

The potential regulatory pathways defined by operons that combine JAB and E1-like domain proteins often encode their own Ub domain proteins and also homologs of the eukaryotic Ub conjugating E2 enzymes. Given the presence of E2 homologs it is quite likely that these are indeed dedicated protein modifying systems that add the associated Ub-like proteins or the available ThiS/MoeB to target proteins. In these cases we predict that the JAB domain is likely to

Fig. 37. Multiple alignment of JAB domain-containing proteins.

Coloring is indicative of 80% consensus. The coloring scheme, consensus abbreviations and secondary structure representations are as in Fig. 34. The secondary structure, shown on the first line of the alignment, is derived from a JAB crystal structure whose primary sequence is found on the second line of the alignment, with PDB identifier shaded in gold. Conserved histidine and acidic residues (ED) are colored yellow and shaded in red. The conserved active site serine residue is colored light grey and shaded in teal. The conserved cysteine found in a subset of JABs (marked with a *) are shaded blue and colored white. The alignment is grouped according to families, with family names listed to the right. Also provided are references to the appropriate row on Table 5 that describe a particular JAB containing operon.

be important for both processing the Ub-like proteins and removing them from the target proteins, thus constituting a genuine bacterial version of the eukaryotic Ub-signaling system. The operon type prototyped by the *E. coli* ICE element also encodes a nucleotidyl transferase (Fig. 36), which might provide an additional protein modification like its homolog the uridylyl transferase, which modifies glutamine synthase [608, 612]. It is particularly interesting to note that some of these systems contain proteins with 2-3 tandem repeats of the Ub-like domain (reminiscent of the eukaryotic poly-ubiquitin) or RWD domain-like inactive versions of the E2-like fold, which probably bind the Ub moieties (Fig. 33, 37). Some of the other uncharacterized proteins encoded specifically by these operon sets, such as the Zinc finger protein (e.g. sll6052 from *Synechocystis*), might be involved in recognizing specific target proteins for modification by these systems. The high mobility of these operons in bacteria is illustrated by their differential presence or absence even within closely related strains of same organism and indeed, some of them are borne by conjugative mobile elements (Table 5). This pattern of mobility is reminiscent of some other conserved operon systems such as the restriction-modification operons, the toxin-antitoxin systems and the CRISPR system [158, 613-616]. It is quite possible that like the two former systems, these operons also maintain themselves by providing the host with oppositely directed

activities. Hence, we speculate that the JAB domain and the E1+E2 complex provides a system that uses an endogenous ThiS/MoaD protein or the distinct Ub-like protein encoded by the mobile operon to alternately modify or de-modify cellular target proteins. This system might provide a means of regulating target protein stability and maintains itself by either acting as an addiction system like the toxin-antitoxin systems or as a means of protection against invasive replicons as the restriction-modification systems.

Other tantalizing but uncertain links between components of the bacterial Ub-like systems and protein stability are suggested by some of the operons. The operon which encodes a JAB domain protein, an Ub-like protein related to ThiS/MoaD and ClpS is one such (Fig. 36, Table 5). The ClpS domain recognizes the N-terminal domain of proteins targeted for destruction and links them to the protein-degrading ClpAP machine in bacteria and the RING finger E3 ligase of the eukaryotic N-recognins [617, 618]. It is possible that this system may be involved in modification of proteins by an Ub-like modification prior to linkage by ClpS for degradation. A more enigmatic case is offered by the linkage between RnfH and SmpB – here apparently no Ub-like transfer system is involved. However, the tight neighborhood association with SmpB that RnfH could in principle, under as yet unstudied conditions, interact with the tmRNA and influence protein stability.

Evolutionary implications of prokaryotic cognates of the Ub-signaling system

The identification of numerous prokaryotic systems containing proteins related to ubiquitin, E1, E2, and the JAB domain, beyond the previously known versions found in the thiamine and MoCo/WCo biosynthesis operons throw considerable light on the emergence of the eukaryotic Ub signaling system (Fig. 39). Amongst the oldest versions of the Ub-fold are the TGS domains that are traced back to LUCA and bind RNA [462, 463]. This suggests that the Ub-like

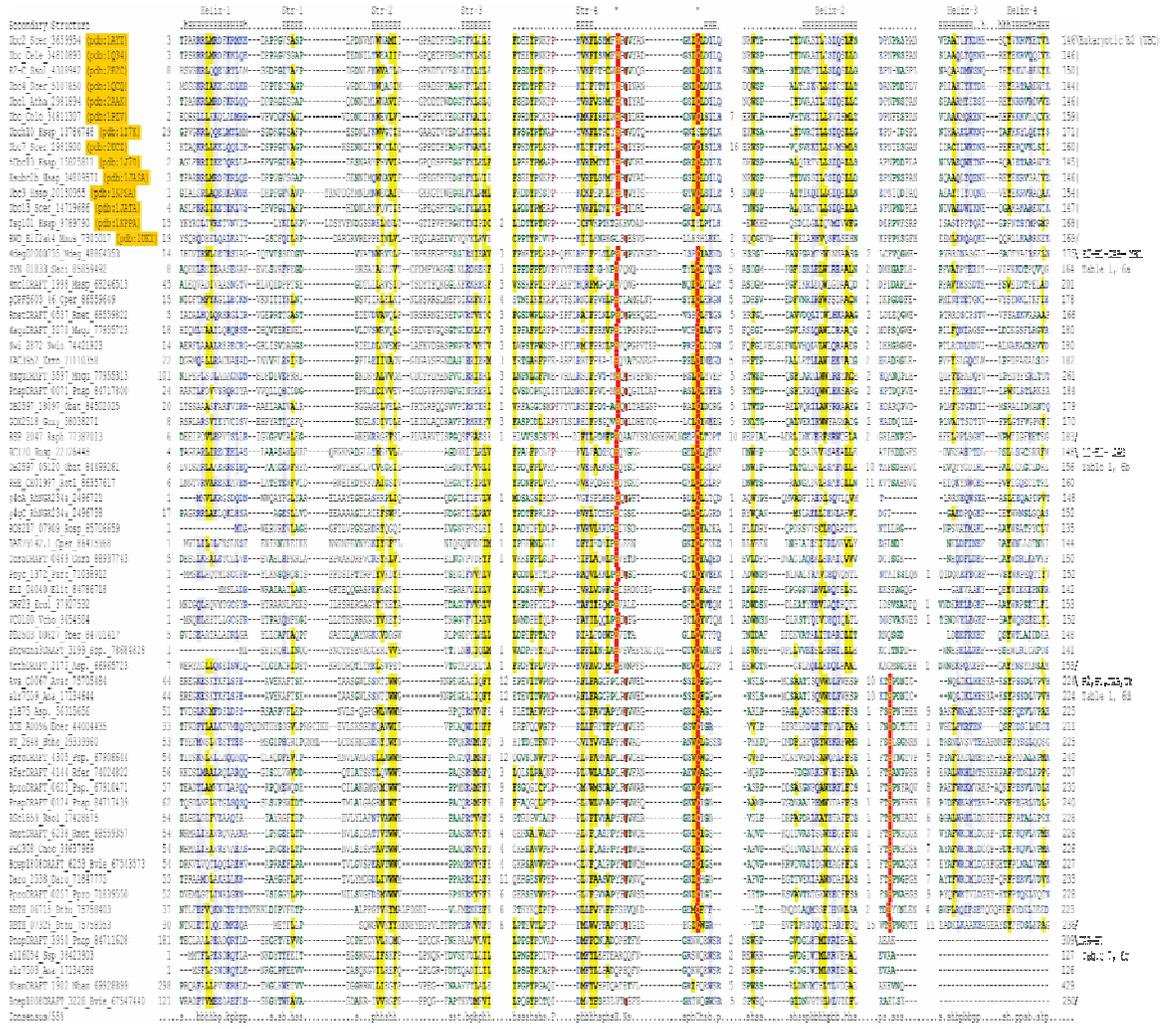


Fig. 38. Multiple alignment of E2 (UBC)-like proteins with a special emphasis on bacterial versions.

PDB identifiers of primary sequences derived from crystal structures are shaded in gold. Coloring is indicative of 55% consensus. The secondary structure, shown on the second line of the alignment, is derived from a general consensus of the secondary structure features from the different crystal structures shown in the alignment. Other features of the alignment are the same as Fig. 34, including coloring scheme, consensus abbreviations and secondary structure representations. Additionally, conserved polar residues (p) (CDEHKNQRST) are colored blue. The strongly conserved proline and asparagine residues are colored purple brown respectively. The major families of bacterial E2s are shown to the right. Also shown are the row numbers in Table 5, where a particular family is described.

versions of the β -grasp fold probably emerged prior to the LUCA as an RNA-binding domain. This is also supported by the observation that versions related to ThiS/MoaD, like the one fused to the Mut7-C RNase domain (Fig. 35), are also likely to participate in a RNA-binding function (Fig. 39). Such a function might also hold for the RnfH protein, which is most closely related to the TGS domains (Fig. 34). However, it is also clear that the MoaD and ThiS versions were also present in LUCA, implying that the divergence between sulfur carrier and RNA-binding versions occurred prior to the LUCA. The analysis of the phyletic patterns of the operons suggests that the sulfur carrier version was a part of molybdenum metabolism in LUCA itself, whereas its recruitment for thiamine biosynthesis happened at the base of the bacterial tree. Likewise, at least a single representative of the E1-like enzymes had differentiated from the remaining Rossmann-type folds, through the acquisition of a distinct C-terminal module, by the time of the LUCA. Even in these two ancient pathways there appears to have been a progressive increase in the complexity of the reaction catalyzed by the E1-like enzyme on the Ub-like protein. Originally, it appears to have been merely an adenylation reaction, as has been suggested for the MoeB-MoaD pair [411]. However, the ThiS-ThiF pair involved an additional formation of a covalent persulfide linkage between the E1-like enzyme the Ub-like protein (Fig. 33).

The operon and domain architecture evidence suggests that reaction mechanisms similar to the eukaryotic E1 enzymes emerged next in specialized versions of the E1-like-Ub-like protein pairs found in the prokaryotes. These systems also added a JAB domain protein, probably in a role similar to their eukaryotic counterparts. The sequence and organizational diversity of the E1-like, E2-like and Ub-like proteins from these remarkable bacterial systems is much higher than that seen in their eukaryotic cognates. This suggests that these systems probably first diversified in bacteria, and were acquired by the eukaryotes during their emergence via the symbiotic

process involving the α -proteobacterial precursor of the mitochondrion. This is consistent with the frequent presence of the more complex Ub-signaling related systems in α -proteobacteria (Table 5). At the face of it the E3 enzymes, such as the RING domain and the HECT domain, appear to be eukaryotic innovations. However, it cannot be ruled out that the additional uncharacterized proteins, such as the above-described Zn finger protein encoded in the bacterial operons (Fig. 36, Table 5), act as E3-like adaptors. However, it is clear that the core of the Ub-transfer system, as well as the main peptidase required for its removal – the JAB domain – were already linked as a functional complex in the bacteria, prior to the emergence of the eukaryotes. The bacteriophage tail assembly system contains an NlpC/P60 peptidase, typically fused to the JAB domain (Fig. 35), which might also be involved in processing the Ub-related protein. Given that the NlpC/P60 peptidase contains a papain-like fold also found in most of the eukaryotic DUBs, it is possible that the functional association between Ub-like domains and the papain-like peptidase emerged in the prokaryotic world. Links between these prokaryotic systems and protein degradation via ATP-dependent proteolytic machines are less clear, although there are some hints that the prokaryotic Ub-like domains might even have a role in such a process.

General conclusions

By performing a systematic search for Ub-like domains in bacteria we identified several novel domains with diverse domain architectures. We present evidence that there are several predicted bacterial operons, beyond those specifying the previously well-characterized thiamine and MoCo/WCo biosynthesis systems that encode Ub-related, JAB domain, E1-like and E2-like proteins. These operons show several distinct organizational themes, each of which is likely to specify a distinct functional system. Some of these systems are likely to possess the capacity to transfer Ub-like protein moieties on to target proteins via a relay of E1-like and E2-like proteins.

This is the first report of a genuine prokaryotic ubiquitin-like signaling system, and we suggest that these systems were the precursors to the eukaryotic Ub-signaling system. We hope this report may stimulate experimental analysis of these bacterial systems and thereby throw light on the emergence of a signaling system that was hitherto considered the unique property of the eukaryotes.

Supplementary Material

The complete list of alignments, conserved neighborhoods and architectures discussed in this article is available for download from:

<ftp://ftp.ncbi.nih.gov/pub/aravind/UB/>.

Experimental validation of work presented above

Begley and colleagues at Cornell University provided the first experimental confirmation of the results presented above, describing the interaction between a JAB deubiquitinase (erroneously referred to as a “JAMM motif protein”) and a Ub-like protein in thioquinolobactin siderophore biosynthesis [430]. They find that the Ub-like protein is pre-processed by the JAB domain before adenylation by an E1-like protein; removing the C-terminal pair of amino acid residues and exposing the conserved di-glycine motif conserved in this particular Ub-like family (see Supplementary Material). This stands as the first experimental confirmation of an ancestral association in prokaryotes between Ub-like proteins and JAB deubiquitinases.

CONCLUSIONS AND GENERAL OBSERVATIONS

Since Darwin published *The Origin of Species* [619], the recognition that selection for advantageous traits drives diversification has been a central theoretical tenant underlying biological research, from early research in taxonomy and paleontology observing and reporting shared morphological characteristics to recent research in molecular biology and biochemistry elucidating molecular adaptations in organisms. An additional tenant compatible with the work of Darwin, the neutral theory of molecular evolution, was formulated in the 1960s by Dr. Motoo Kimura which argued for a strong role of genetic drift in determining adaptive evolution [99]. Despite realizing the importance of the twin evolutionary forces of adaptive change and effective population size in shaping extant biological systems, the lack of a comprehensive dataset necessary for systematic application of evolutionary principles across all branches of biological research has long hampered efforts to view biological macromolecules and the systems they combine to form as products of natural selection; subsequently much information that could be derived from studying close and distant relationships between these macromolecules has gone undetected. The onset of the genomic age is at last providing the data needed to construct a comprehensive and unified picture of macromolecule and system development; allowing researchers to precisely define the pervasiveness of a discovery in the biological world and link that discovery to related biomolecules/biosystems in order to gain further insight. The studies in the preceding pages are examples of this kind of comprehensive analysis, illustrating the predictive power of studies based on large-scale genome data in deciphering functional roles for proteins and the evolutionary processes dictating functional divergence in related proteins.

One key objective of comparative evolutionary genome analysis is the search for conserved features, whether sequence, structure, or genome context in nature. Conservation of a feature across a broad range of representatives indicates a feature is under selective retentive pressure; as such, the feature is likely to contribute in some functional way to the biomolecule or biosystem to which it belongs and also provides a clear signal indicating a common evolutionary origin that can be used in elucidating higher-order relationships. Such conserved features were invaluable in deciphering function and evolution time and time again in this research. Several features contributed to a better understanding of function: the conserved flap and squiggle structural motifs of the HAD superfamily with important structural roles in the catalytic mechanism, the conservation of a 2-3 β -strand insert in members of the Fasciclin assemblage of the β -grasp fold important in mediating functional contacts with interacting partners, and the presence of a conserved gene neighborhood providing functional clues for the possible role of the PSPTO_2114 subfamily of HAD domains, to name a few. Conserved features also provide insight into evolution: the discovery of conserved gene neighborhoods in prokaryotes containing core components of the eukaryotic ubiquitin modification system suggests a bacterial provenance for the system, conserved sequence features in the portal proteins of DNA virus packaging systems indicate a monophyletic origin for these proteins, and the distribution of genome conservation in members of the transcobalamin-like family led to a novel proposal for the evolutionary mode of inheritance of the vitamin B12 receptor in animals. Nor are functional and evolutionary inferences mutually exclusive; the discovery of a positively-charged, absolutely conserved binding cleft in the ASCH domains contributed to the establishment of an ancient evolutionary link between the ASCH superfamily and the PUA fold and to its predicted functional role in RNA-binding.

Several theories of protein domain evolution have previously been described, which posit that protein domains often undergo functional diversification through utilization of the following strategies: accretion of insert modules into a conserved core scaffold, selective sequence divergence of primary interacting surfaces, and incorporation of an entire domain into diverse and novel architectures [620], [621]. The detailed research I have performed demonstrates that these strategies have been consistently applied to several folds, strengthening these theories of protein domain evolution. The particular evolutionary strategy or strategies of a given domain appear to depend on the functional niche to which a domain is initially recruited; domains associated with regulatory or signaling roles often diversify through multiple duplication events coinciding with sequence divergence and incorporation into new multi-domain architectures. This was observed in my research into the Ub superfamily, which has colonized a range of functional niches in the cell primarily through such sequence diversification and incorporation into novel domain architectures. Meanwhile, domains with more specific roles in the cell tend to conserve sequence and structure features critical to that role through strong purifying selection while undergoing significant divergence in other regions of the domain, driving the emergence of novel substrate specificities [621]. The HAD superfamily is an excellent example of this; the core Rossmannoid scaffold has incorporated a range of inserts providing novel surfaces that influence substrate specificity while retaining sequence conservation of residues critical to catalysis. A slight variation to this trend is observed in the SLBB superfamily and Fasciclin-like assemblage of the β -grasp fold which instead of conserving specific catalytic residues critical to functionality, acquired structural inserts early in their evolution which were retained through selection for function. The sequences of these domains and their respective structural inserts subsequently underwent lineage-specific diversification allowing for the exploration of interacting partner

space; in the case of the SLBB domains this entailed an exploration of small-molecule binding partners while the Fasciclin-like assemblage primarily explored novel protein-protein interaction partners.

Of all domains investigated in this dissertation, the E1-like superfamily of Rossmannoid domains has followed a pathway of more distinctive nature which led to functional diversification. Exploration of substrate space in this superfamily has been experimentally determined in some cases to be a function of association with distinct domains in multi-domain architectures [417] [425] rather than divergence in primary sequence of regions outside the central catalytic residues of the adenylation active site. In fact, my research supports other recent research [421] suggesting that sequence divergence of these fused domains may play an important role in substrate recognition in several E1 domain-containing proteins. Further research into protein fold evolution will help determine the extent to which E1-like superfamily represents a novel variation to the presently-accepted theories of protein domain evolution. It should be noted that lineage-specific inserts and distinct sequence features have emerged on occasion in the E1 superfamily; however, functional roles for these inserts are largely uncharacterized, presenting possible targets for future study.

In this research I have also proposed a novel evolutionary scenario describing the emergence of the primary vitamin B₁₂ receptor in eukaryotes. This also led to the discovery of a unique mode of structural adaptation in protein domain evolution wherein the ancestor of a domain acquires distinct inserts into distinct locations of the core scaffold with each insert facilitating the same functional adaptation. The recognition of this event enabled prediction of functional roles in contexts that otherwise would not have been possible. The extent of the

incidence of this type of functional diversification is not yet well-understood, but recognition in additional cases will likely aid the future elucidation of functional roles in diverse settings.

Another evolutionary observation that can be drawn from this research is the repeated recruitment of domains in the same superfamily or fold to similar functional niches in the cell. Examples of this phenomenon from this research include the repeated emergence of RNA or nucleotide-derived molecule-binding lineages in the β -grasp fold, the independent recruitment of β -grasp fold domains on at least two occasions as structural components of bacterial flagellum, emergence of different sugar moiety phosphatase activity in distant branches of the HAD superfamily, and perhaps most strikingly the multiple independent recruitments of different P-loop NTPase superfamilies to viral DNA packaging apparatuses. Although important, this observation is not altogether surprising; generic biochemical functions in ancient domains were likely established in the RNA world as these domains collaborated with catalytic RNAs [622]. As proteins diversified through employment of the evolutionary strategies discussed in the preceding paragraph, they likely began displacing RNAs in enzymatic functions, while still retaining a generic scaffold with inherent functional properties [621], properties that display an affinity for the functional state of common ancestor of the fold or superfamily. This evolutionary trend could be more comprehensively examined through an analysis of the folds populating conserved, complex pathways with distinct evolutionary origins; determining the extent of fold recruitment to similar roles across multiple pathways.

Data derived from genome sequencing projects will continue to pour in with the development of cheaper genome sequencing technologies [623], leading to expanded opportunities to understand protein domain function and evolution. The research contained in this dissertation demonstrates that computational tools designed for the analysis of genome data

and the formulation of sound molecular biological predictions based on solid statistical evidence have, for the large part, already been developed. The major challenge facing both computational and experimental researchers in the coming years is not necessarily in the development of new tools, but in the incorporation of existing tools into experimental designs that will expand the implications of their work beyond single model systems. This will significantly contribute to a stronger dependence of modern biological research on its founding tenants—the principles of evolution. In turn, this will lead to the ability to quickly and appropriately evaluate the significance of research results; propelling biological scientific research forward at a much faster pace.

BIBLIOGRAPHY

JOURNAL ABBREVIATIONS

Acta Crystallogr D Biol Crystallogr	Acta Crystallographica. Section D, Biological Crystallography
Adv Enzymol Relat Areas Mol Biol	Advances in Enzymology and Related Areas of Molecular Biology
Anal Biochem	Analytical Biochemistry
Anim Genet	Animal Genetics
Ann N Y Acad Sci	Annals of the New York Academy of Sciences
Ann Rev Plant Biol	Annual Review of Plant Biology
Annu Rev Biochem	Annual Review of Biochemistry
Annu Rev Genomics Hum Genet	Annual Review of Genomics and Human Genetics
Annu Rev Microbiol	Annual Review of Microbiology
Annu Rev Phytopathol	Annual Review of Phytopathology
Appl Environ Microbiol	Applied and Environmental Microbiology
Arch Biochem Biophys	Archives of Biochemistry and Biophysics
BMC Cell Biol	BMC Cell Biology
BMC Struct Biol	BMC Structural Biology
Biochem Biophys Res Commun	Biochemical and Biophysical Research Communications
Biochem Soc Trans	Biochemical Society Transactions
Biochim Biophys Acta	Biochimica et Biophysica Acta
Biol Direct	Biology Direct
Brief Bioinform	Briefings in Bioinformatics
Cell Mol Life Sci	Cellular and Molecular Life Sciences
Chem Biol	Chemistry & Biology
Comput Appl Biosci	Computer Applications in the Biosciences
Comput Sci Monogr	Monographs in Computer Science
Crit Rev Biochem Mol Biol	Critical Reviews in Biochemistry and Molecular Biology
Curr Biol	Current Biology
Curr Genet	Current Genetics
Curr Opin Chem Biol	Current Opinion in Chemical Biology
Curr Opin Genet Dev	Current Opinion in Genetics & Development
Curr Opin Struct Biol	Current Opinion in Structural Biology
Curr Protein Pept Sci	Current Protein & Peptide Science
Curr Top Cell Regul	Current Topics in Cellular Regulation
Embo J	The EMBO Journal
Environ Microbiol	Environmental Microbiology
Eukaryot Cell	Eukaryotic Cell
Eur J Biochem	European Journal of Biochemistry
FEBS Lett	FEBS Letters
FEMS Microbiol Lett	FEMS Microbiology Letters
Febs J	The FEBS Journal

Fed Proc	Federation Proceedings
Genes Dev	Genes & Development
Genome Biol	Genome Biology
Genome Res	Genome Research
Hum Mol Genet	Human Molecular Genetics
In Silico Biol	In Silico Biology
Infect Immun	Infection and Immunity
Int J Biochem Cell Biol	The International Journal of Biochemistry & Cell Biology
J Antibiot	The Journal of Antibiotics
J Bacteriol	Journal of Bacteriology
J Biochem	Journal of Biochemistry
J Biol Chem	Journal of Biological Chemistry
J Cell Biol	The Journal of Cell Biology
J Clin Invest	The Journal of Clinical Investigation
J Exp Bot	Journal of Experimental Botany
J Mol Biol	Journal of Molecular Biology
J Mol Evol	Journal of Molecular Evolution
J Mol Microbiol Biotechnol	Journal of Molecular Microbiology and Biotechnology
J Struct Biol	Journal of Structural Biology
J Struct Funct Genomics	Journal of Structural and Functional Genomics
J Theor Biol	Journal of Theoretical Biology
J Virol	Journal of Virology
Methods Enzymol	Methods in Enzymology
Mol Biochem Parasitol	Molecular and Biochemical Parasitology
Mol Biol Evol	Molecular Biology and Evolution
Mol Cell	Molecular Cell
Mol Cell Biol	Molecular and Cellular Biology
Mol Gen Genet	Molecular & General Genetics
Mol Microbiol	Molecular Microbiology
Nat Biotechnol	Nature Biotechnology
Nat Cell Biol	Nature Cell Biology
Nat Chem Biol	Nature Chemical Biology
Nat Genet	Nature Genetics
Nat Rev Mol Cell Biol	Nature Reviews. Molecular Cell Biology
Nat Struct Biol	Nature Structural Biology
Nucleic Acids Res	Nucleic Acids Research
PLoS Biol	PLoS Biology
PLoS Comp Biol	PLoS Computational Biology
Physiol Rev	Physiological Reviews
Plant Mol Biol	Plant Molecular Biology
Plant Physiol	Plant Physiology
Proc Int Conf Intell Syst Mol Biol	Proceedings of the International Conference on Intelligent Systems for Molecular Biology
Proc Natl Acad Sci U S A	Proceedings of the National Academy of Sciences of the United States of America

Protein Sci	Protein Science
Q Rev Biophys	Quarterly Reviews of Biophysics
Sci Am	Scientific American
Trends Biochem Sci	Trends in Biochemical Sciences
Trends Genet	Trends in Genetics
Vet Microbiol	Veterinary Microbiology
Virus Res	Virus Research

REFERENCES

1. Koonin EV, Mushegian AR: **Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics.** *Curr Opin Genet Dev* 1996, **6**:757-762.
2. Altschul SF, Koonin EV: **Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases.** *Trends Biochem Sci* 1998, **23**:444-447.
3. Thornton JW, DeSalle R: **Gene family evolution and homology: genomics meets phylogenetics.** *Annu Rev Genomics Hum Genet* 2000, **1**:41-73.
4. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context.** *Genome Res* 2001, **11**:356-372.
5. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci U S A* 1999, **96**:2896-2901.
6. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechnol* 2000, **18**:609-613.
7. Huynen M, and Snel, B.: **Gene and context: Integrative approaches to genome analysis.** In *Analysis of Amino Acid Sequences* Edited by Bork P. San Diego, California: Academic Press; 2000: 345-379
8. Aravind L: **Guilt by association: contextual information in genome analysis.** *Genome Res* 2000, **10**:1074-1077.
9. Hodgman TC: **A historical perspective on gene/protein functional assignment.** *Bioinformatics* 2000, **16**:10-15.
10. Doolittle RF, Blombaek B: **Amino-Acid Sequence Investigations of Fibrinopeptides from Various Mammals: Evolutionary Implications.** *Nature* 1964, **202**:147-152.
11. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history.** *J Theor Biol* 1965, **8**:357-366.

12. Doolittle RF, Singer SJ, Metzger H: **Evolution of immunoglobulin polypeptide chains: carboxy-terminal of an IgM heavy chain.** *Science* 1966, **154**:1561-1562.
13. Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214**:149-159.
14. Dayhoff MO: **Computer analysis of protein evolution.** *Sci Am* 1969, **221**:86-95.
15. Dayhoff MO: **Computer aids to protein sequence determination.** *J Theor Biol* 1965, **8**:97-112.
16. Niall HD: **Automated Edman degradation: the protein sequenator.** *Methods Enzymol* 1973, **27**:942-1010.
17. Chen HR, Barker WC: **Nucleic acid sequence database VI: Retroviral oncogenes and cellular proto-oncogenes.** *DNA* 1985, **4**:171-182.
18. Kneale GG, Bishop MJ: **Nucleic acid and protein sequence databases.** *Comput Appl Biosci* 1985, **1**:11-17.
19. Hodgman TC: **The elucidation of protein function from its amino acid sequence.** *Comput Appl Biosci* 1986, **2**:181-187.
20. Weber JL, Myers EW: **Human whole-genome shotgun sequencing.** *Genome Res* 1997, **7**:401-409.
21. Bennett ST, Barnes C, Cox A, Davies L, Brown C: **Toward the 1,000 dollars human genome.** *Pharmacogenomics* 2005, **6**:373-382.
22. Hattori M: **[High-throughput of DNA sequencing by automated machines].** *Nippon Rinsho* 1996, **54**:923-927.
23. Dovichi NJ: **DNA sequencing by capillary electrophoresis.** *Electrophoresis* 1997, **18**:2393-2399.
24. McPherson A: **Protein crystallization in the structural genomics era.** *J Struct Funct Genomics* 2004, **5**:3-12.
25. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier FW, Swaminathan S: **Structural genomics: beyond the human genome project.** *Nat Genet* 1999, **23**:151-157.
26. Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shirouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, et al: **Structural genomics projects in Japan.** *Nat Struct Biol* 2000, **7 Suppl**:943-945.

27. Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, Swindell CD, Tung CS, Bilofsky HS: **The GenBank nucleic acid sequence database.** *Comput Appl Biosci* 1985, **1**:225-233.
28. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007, **35**:D5-12.
29. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, et al: **EMBL Nucleotide Sequence Database in 2006.** *Nucleic Acids Res* 2007, **35**:D16-20.
30. Bairoch A, Boeckmann B: **The SWISS-PROT protein sequence data bank, recent developments.** *Nucleic Acids Res* 1993, **21**:3093-3096.
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
32. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34**:D354-357.
33. Hu Z, Ng DM, Yamada T, Chen C, Kawashima S, Mellor J, Linghu B, Kanehisa M, Stuart JM, DeLisi C: **VisANT 3.0: new modules for pathway visualization, editing, prediction and construction.** *Nucleic Acids Res* 2007, **35**:W625-632.
34. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**:D535-539.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
36. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227**:1435-1441.
37. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**:2444-2448.
38. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
39. Mans BJ, Anantharaman V, Aravind L, Koonin EV: **Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex.** *Cell Cycle* 2004, **3**:1612-1637.

40. Aravind L, Iyer LM, Koonin EV: **Comparative genomics and structural biology of the molecular innovations of eukaryotes.** *Curr Opin Struct Biol* 2006, **16**:409-419.
41. Lopez-Garcia P, Moreira D: **Metabolic symbiosis at the origin of eukaryotes.** *Trends Biochem Sci* 1999, **24**:88-93.
42. Zhang Y, Leaves NI, Anderson GG, Ponting CP, Broxholme J, Holt R, Edser P, Bhattacharyya S, Dunham A, Adcock IM, et al: **Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma.** *Nat Genet* 2003, **34**:181-186.
43. Ponting CP, Hutton M, Nyborg A, Baker M, Jansen K, Golde TE: **Identification of a novel family of presenilin homologues.** *Hum Mol Genet* 2002, **11**:1037-1044.
44. Dickens NJ, Beatson S, Ponting CP: **Cadherin-like domains in alpha-dystroglycan, alpha/epsilon-sarcoglycan and yeast and bacterial proteins.** *Curr Biol* 2002, **12**:R197-199.
45. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.
46. Ganley AR, Hayashi K, Horiuchi T, Kobayashi T: **Identifying gene-independent noncoding functional elements in the yeast ribosomal DNA by phylogenetic footprinting.** *Proc Natl Acad Sci U S A* 2005, **102**:11787-11792.
47. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
48. Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ: **Elucidation of the small RNA component of the transcriptome.** *Science* 2005, **309**:1567-1569.
49. McCutcheon JP, Eddy SR: **Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics.** *Nucleic Acids Res* 2003, **31**:4119-4128.
50. Mulkidjanian AY, Koonin EV, Makarova KS, Mekhedov SL, Sorokin A, Wolf YI, Dufresne A, Partensky F, Burd H, Kaznadzey D, et al: **The cyanobacterial genome core and the origin of photosynthesis.** *Proc Natl Acad Sci U S A* 2006, **103**:13126-13131.
51. Bork P, Koonin EV: **Predicting functions from protein sequences--where are the bottlenecks?** *Nat Genet* 1998, **18**:313-318.
52. Koonin EV, Mushegian AR, Galperin MY, Walker DR: **Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea.** *Mol Microbiol* 1997, **25**:619-637.

53. Dayhoff MO: **The origin and evolution of protein superfamilies.** *Fed Proc* 1976, **35**:2132-2138.
54. Rao ST, Rossmann MG: **Comparison of super-secondary structures in proteins.** *J Mol Biol* 1973, **76**:241-256.
55. Richardson JS: **Schematic drawings of protein structures.** *Methods Enzymol* 1985, **115**:359-380.
56. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233**:123-138.
57. Holm L, Sander C: **Mapping the protein universe.** *Science* 1996, **273**:595-603.
58. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 2004, **32**:D226-229.
59. Walker DR, Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:333-339.
60. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**:205-217.
61. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM: **MUSTANG: a multiple structural alignment algorithm.** *Proteins* 2006, **64**:559-574.
62. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
63. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
64. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:175-182.
65. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**:502-504.
66. Bruno WJ, Socci ND, Halpern AL: **Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction.** *Mol Biol Evol* 2000, **17**:189-197.
67. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci U S A* 1992, **89**:10915-10919.

68. Durbin R, Eddy, S., Krogh, A., and Mitchison, G.: *Biological Sequence Analysis*. Cambridge, United Kingdom: Cambridge University Press; 1998.
69. Sonnhammer EL, Hollich V: **Scoredist: a simple and robust protein sequence distance estimator**. *BMC Bioinformatics* 2005, **6**:108.
70. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements**. *Nucleic Acids Res* 2001, **29**:2994-3005.
71. Madera M, Gough J: **A comparison of profile hidden Markov model procedures for remote homology detection**. *Nucleic Acids Res* 2002, **30**:4321-4328.
72. Rollins BJ, Morton CC, Ledbetter DH, Eddy RL, Jr., Shows TB: **Assignment of the human small inducible cytokine A2 gene, SCYA2 (encoding JE or MCP-1), to 17q11.2-12: evolutionary relatedness of cytokines clustered at the same locus**. *Genomics* 1991, **10**:489-492.
73. Hannenhalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments**. *J Mol Biol* 2000, **303**:61-76.
74. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency**. *Bioinformatics* 2003, **19**:427-428.
75. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.
76. Edgar RC: **Local homology recognition and distance measures in linear time using compressed amino acid alphabets**. *Nucleic Acids Res* 2004, **32**:380-385.
77. Gotoh O: **Optimal alignment between groups of sequences and its application to multiple sequence alignment**. *Comput Appl Biosci* 1993, **9**:361-370.
78. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792-1797.
79. Rost B: **Review: protein secondary structure prediction continues to rise**. *J Struct Biol* 2001, **134**:204-218.
80. Heringa J: **Computational methods for protein secondary structure prediction using multiple sequence alignments**. *Curr Protein Pept Sci* 2000, **1**:273-301.
81. Simossis VA, Heringa J: **Integrating protein secondary structure prediction and multiple sequence alignment**. *Curr Protein Pept Sci* 2004, **5**:249-266.

82. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server.** *Bioinformatics* 1998, **14**:892-893.
83. Lin HN, Chang JM, Wu KP, Sung TY, Hsu WL: **HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence.** *Bioinformatics* 2005, **21**:3227-3233.
84. Rost B, Sander C, Schneider R: **Redefining the goals of protein secondary structure prediction.** *J Mol Biol* 1994, **235**:13-26.
85. King RD, Sternberg MJ: **Identification and application of the concepts important for accurate and reliable protein secondary structure prediction.** *Protein Sci* 1996, **5**:2298-2310.
86. Salamov AA, Solovyev VV: **Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments.** *J Mol Biol* 1995, **247**:11-15.
87. Frishman D, Argos P: **Seventy-five percent accuracy in protein secondary structure prediction.** *Proteins* 1997, **27**:329-335.
88. Granseth E, Viklund, H., and Elofsson, A.: **ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins.** *Bioinformatics* 2006, **22**:191-196.
89. Cuff JA, Barton GJ: **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins* 2000, **40**:502-511.
90. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
91. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340**:783-795.
92. Wootton JC: **Non-globular domains in protein sequences: automated segmentation using complexity measures.** *Comput Chem* 1994, **18**:269-285.
93. Altschul SF, Boguski MS, Gish W, Wootton JC: **Issues in searching molecular sequence databases.** *Nat Genet* 1994, **6**:119-129.
94. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18**:2714-2723.
95. Holm L, Sander C: **The FSSP database: fold classification based on structure-structure alignment of proteins.** *Nucleic Acids Res* 1996, **24**:206-209.

96. Holm L, Sander C: **Dali: a network tool for protein structure comparison.** *Trends Biochem Sci* 1995, **20**:478-480.
97. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5**:150-163.
98. Steel MA, Hendy MD, Penny D: **Loss of information in genetic distances.** *Nature* 1988, **336**:118.
99. Kimura M: **Evolutionary rate at the molecular level.** *Nature* 1968, **217**:624-626.
100. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266**:418-427.
101. Adachi J HM: **MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood.** *Comput Sci Monogr* 1996, **28**:1-150.
102. Swofford DL, G. J. Olsen, P. J. Waddell, and D. M. Hillis: **Phylogenetic inference.** In *Molecular Systematics* 2nd edition. Edited by D. M. Hillis CM, and B. Mable Sunderland, Massachusetts: Sinauer Associates; 1996: 407-514
103. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754-755.
104. Platnick NI: **Philosophy and the transformation of cladistics revisited.** In *Cladistics* 1. 1985: 87-94
105. Goldman N: **Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analysis.** *Systematic zoology* 1990:345-361.
106. Nei MaK, S.: *Molecular Evolution and Phylogenetics.* New York, New York: Oxford University Press; 2000.
107. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV: **Reconstruction of ancestral protosplice sites.** *Curr Biol* 2004, **14**:1505-1508.
108. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV: **Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution.** *Curr Biol* 2003, **13**:1512-1517.
109. Rogozin IB, Wolf YI, Carmel L, Koonin EV: **Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements.** *Mol Biol Evol* 2007, **24**:1080-1090.

110. Aravind L, Koonin EV: **Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system.** *Genome Res* 2001, **11**:1365-1374.
111. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci U S A* 1999, **96**:4285-4288.
112. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28**:33-36.
113. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285**:751-753.
114. Aravind L, Koonin EV: **Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.** *J Mol Biol* 1999, **287**:1023-1040.
115. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions.** *Nucleic Acids Res* 2007, **35**:D358-362.
116. Andorf C, Dobbs D, Honavar V: **Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach.** *BMC Bioinformatics* 2007, **8**:284.
117. Jothi R, Przytycka TM, Aravind L: **Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment.** *BMC Bioinformatics* 2007, **8**:173.
118. Krishna SS, Grishin NV: **Structural drift: a possible path to protein fold change.** *Bioinformatics* 2005, **21**:1308-1310.
119. Frickey T, Lupas AN: **Phylogenetic analysis of AAA proteins.** *J Struct Biol* 2004, **146**:2-10.
120. Aravind L, Koonin EV: **Novel predicted RNA-binding domains associated with the translation machinery.** *J Mol Evol* 1999, **48**:291-302.
121. Perez-Arellano I, Gallego J, Cervera J: **The PUA domain - a structural and functional overview.** *Febs J* 2007, **274**:4972-4984.
122. Iyer LM, Burroughs AM, Aravind L: **The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism.** *Bioinformatics* 2006, **22**:257-263.

123. Anantharaman V, Koonin EV, Aravind L: **Comparative genomics and evolution of proteins involved in RNA metabolism.** *Nucleic Acids Res* 2002, **30**:1427-1464.
124. Clissold PM, Ponting CP: **PIN domains in nonsense-mediated mRNA decay and RNAi.** *Curr Biol* 2000, **10**:888-890.
125. Cerutti L, Mian N, Bateman A: **Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain.** *Trends Biochem Sci* 2000, **25**:481-482.
126. Reid R, Greene PJ, Santi DV: **Exposition of a family of RNA m(5)C methyltransferases from searching genomic and proteomic sequences.** *Nucleic Acids Res* 1999, **27**:3138-3145.
127. Ishitani R, Nureki O, Fukai S, Kijimoto T, Nameki N, Watanabe M, Kondo H, Sekine M, Okada N, Nishimura S, Yokoyama S: **Crystal structure of archaeosine tRNA-guanine transglycosylase.** *J Mol Biol* 2002, **318**:665-677.
128. Korber P, Zander T, Herschlag D, Bardwell JC: **A new heat shock protein that binds nucleic acids.** *J Biol Chem* 1999, **274**:249-256.
129. Fatica A, Tollervey D, Dlakic M: **PIN domain of Nob1p is required for D-site cleavage in 20S pre-rRNA.** *RNA* 2004, **10**:1698-1701.
130. Kim HJ, Yi JY, Sung HS, Moore DD, Jhun BH, Lee YC, Lee JW: **Activating signal cointegrator 1, a novel transcription coactivator of nuclear receptors, and its cytosolic localization under conditions of serum deprivation.** *Mol Cell Biol* 1999, **19**:6323-6332.
131. Jung DJ, Sung HS, Goo YW, Lee HM, Park OK, Jung SY, Lim J, Kim HJ, Lee SK, Kim TS, et al: **Novel transcription coactivator complex containing activating signal cointegrator 1.** *Mol Cell Biol* 2002, **22**:5203-5211.
132. Mazumder R, Iyer LM, Vasudevan S, Aravind L: **Detection of novel members, structure-function analysis and evolutionary classification of the 2H phosphoesterase superfamily.** *Nucleic Acids Res* 2002, **30**:5229-5243.
133. Rost B, Sander C, Schneider R: **PHD--an automatic mail server for protein secondary structure prediction.** *Comput Appl Biosci* 1994, **10**:53-60.
134. Delano WL: *The PyMOL Molecular Graphics System.* San Carlos, CA, USA: DeLano Scientific; 2002.
135. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32**:D138-141.

136. Anantharaman V, Koonin EV, Aravind L: **SPOUT: a class of methyltransferases that includes spoU and trmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases.** *J Mol Microbiol Biotechnol* 2002, **4**:71-75.
137. Forouhar F, Shen J, Xiao R, Acton TB, Montelione GT, Tong L: **Functional assignment based on structural analysis: crystal structure of the yggJ protein (HI0303) of Haemophilus influenzae reveals an RNA methyltransferase with a deep trefoil knot.** *Proteins* 2003, **53**:329-332.
138. Hoang C, Ferre-D'Amare AR: **Cocrystal structure of a tRNA Psi55 pseudouridine synthase: nucleotide flipping by an RNA-modifying enzyme.** *Cell* 2001, **107**:929-939.
139. Pan H, Agarwalla S, Moustakas DT, Finer-Moore J, Stroud RM: **Structure of tRNA pseudouridine synthase TruB and its RNA complex: RNA recognition through a combination of rigid docking and induced fit.** *Proc Natl Acad Sci U S A* 2003, **100**:12648-12653.
140. Dowhan DH, Hong EP, Auboeuf D, Dennis AP, Wilson MM, Berget SM, O'Malley BW: **Steroid hormone receptor coactivation and alternative RNA splicing by U2AF65-related proteins CAPERalpha and CAPERbeta.** *Mol Cell* 2005, **17**:429-439.
141. Auboeuf D, Dowhan DH, Li X, Larkin K, Ko L, Berget SM, O'Malley BW: **CoAA, a nuclear receptor coactivator protein at the interface of transcriptional coactivation and RNA splicing.** *Mol Cell Biol* 2004, **24**:442-453.
142. Maniatis T, Reed R: **An extensive network of coupling among gene expression machines.** *Nature* 2002, **416**:499-506.
143. Lanz RB, McKenna NJ, Onate SA, Albrecht U, Wong J, Tsai SY, Tsai MJ, O'Malley BW: **A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex.** *Cell* 1999, **97**:17-27.
144. Zhao X, Patton JR, Davis SL, Florence B, Ames SJ, Spanjaard RA: **Regulation of nuclear receptor activity by a pseudouridine synthase through posttranscriptional modification of steroid receptor RNA activator.** *Mol Cell* 2004, **15**:549-558.
145. Shi Y, Downes M, Xie W, Kao HY, Ordentlich P, Tsai CC, Hon M, Evans RM: **Sharp, an inducible cofactor that integrates nuclear receptor repression and activation.** *Genes Dev* 2001, **15**:1140-1151.
146. Saraste M, Sibbald PR, Wittinghofer A: **The P-loop--a common motif in ATP- and GTP-binding proteins.** *Trends Biochem Sci* 1990, **15**:430-434.
147. Vetter IR, Wittinghofer A: **Nucleoside triphosphate-binding proteins: different scaffolds to achieve phosphoryl transfer.** *Q Rev Biophys* 1999, **32**:1-56.

148. Milner-White EJ, Coggins JR, Anton IA: **Evidence for an ancestral core structure in nucleotide-binding proteins with the type A motif.** *J Mol Biol* 1991, **221**:751-754.
149. Koonin EV, Wolf Y.I., and Aravind L.: **Protein fold recognition using sequence profiles and its application in structural genomics.** In *Advances in protein chemistry. Volume 54.* Edited by P. B: Academic Press; 2000: 245-275
150. Nandhagopal N, Simpson AA, Gurnon JR, Yan X, Baker TS, Graves MV, Van Etten JL, Rossmann MG: **The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus.** *Proc Natl Acad Sci U S A* 2002, **99**:14758-14763.
151. Hendrix RW: **Evolution: the long evolutionary reach of viruses.** *Curr Biol* 1999, **9**:914-917.
152. Newcomb WW, Juhas RM, Thomsen DR, Homa FL, Burch AD, Weller SK, Brown JC: **The UL6 gene product forms the portal for entry of DNA into the herpes simplex virus capsid.** *J Virol* 2001, **75**:10923-10932.
153. *Viral Genome Packaging: Genetics, Structure, and Mechanism.* New York, NY: Kluwer Academic / Plenum publisher; 2005.
154. Leipe DD, Wolf YI, Koonin EV, Aravind L: **Classification and evolution of P-loop GTPases and related ATPases.** *J Mol Biol* 2002, **317**:41-72.
155. Leipe DD, Koonin EV, Aravind L: **Evolution and classification of P-loop kinases and related proteins.** *J Mol Biol* 2003, **333**:781-815.
156. Aravind L, Iyer LM, Leipe DD, Koonin EV: **A novel family of P-loop NTPases with an unusual phyletic distribution and transmembrane segments inserted within the NTPase domain.** *Genome Biol* 2004, **5**:R30.
157. Leipe DD, Koonin EV, Aravind L: **STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer.** *J Mol Biol* 2004, **343**:1-28.
158. Iyer LM, Makarova KS, Koonin EV, Aravind L: **Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging.** *Nucleic Acids Res* 2004, **32**:5260-5279.
159. Burroughs AM, Iyer LM, Aravind L: **Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems.** In *Gene and Protein Evolution. Volume 3.* Edited by Volff J-N. Basel, Switzerland: Karger; 2007: 48-65.[Volff J-N (Series Editor): *Genome Dynamics*].

160. Wagner KE, Hewlett MJ: *Basic Virology*. second edn. Oxford, UK: Blackwell Publishers; 2003.
161. Haren L, Ton-Hoang B, Chandler M: **Integrating DNA: transposases and retroviral integrases**. *Annu Rev Microbiol* 1999, **53**:245-281.
162. Rao AL: **Genome Packaging by Spherical Plant RNA Viruses**. *Annu Rev Phytopathol* 2006.
163. Bernal RA, Hafenstein S, Esmeralda R, Fane BA, Rossmann MG: **The phiX174 protein J mediates DNA packaging and viral attachment to host cells**. *J Mol Biol* 2004, **337**:1109-1122.
164. Stromsten NJ, Bamford DH, Bamford JKH: **In vitro DNA packaging of PRD1: a common mechanism for internal-membrane viruses**. *J Mol Biol* 2005, **348**:617-629.
165. Catalano CE: **The terminase enzyme from bacteriophage lambda: a DNA-packaging machine**. *Cell Mol Life Sci* 2000, **57**:128-148.
166. Black LW: **DNA packaging and cutting by phage terminases: control in phage T4 by a synaptic mechanism**. *Bioessays* 1995, **17**:1025-1030.
167. Rentas FJ, Rao VB: **Defining the bacteriophage T4 DNA packaging machine: evidence for a C-terminal DNA cleavage domain in the large terminase/packaging protein gp17**. *J Mol Biol* 2003, **334**:37-52.
168. Goetzinger KR, Rao VB: **Defining the ATPase center of bacteriophage T4 DNA packaging machine: requirement for a catalytic glutamate residue in the large terminase protein gp17**. *J Mol Biol* 2003, **331**:139-154.
169. Ibarra B, Valpuesta JM, Carrascosa JL: **Purification and functional characterization of p16, the ATPase of the bacteriophage Phi29 packaging machinery**. *Nucleic Acids Res* 2001, **29**:4264-4273.
170. Koonin EV, Senkevich TG, Chernos VI: **Gene A32 product of vaccinia virus may be an ATPase involved in viral DNA packaging as indicated by sequence comparisons with other putative viral ATPases**. *Virus Genes* 1993, **7**:89-94.
171. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**:205-217.
172. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency**. *Bioinformatics* 2003, **19**:427-428.
173. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ: **JPred: a consensus secondary structure prediction server**. *Bioinformatics* 1998, **14**:892-893.

174. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling.** *Electrophoresis* 1997, **18**:2714-2723.
175. Felsenstein J: *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates; 2004.
176. Adachi J, Hasegawa M: *MOLPHY: Programs for Molecular Phylogenetics*. Tokyo: Institute of Statistical Mathematics; 1992.
177. Iyer LM, Aravind L, Koonin EV: **Common origin of four diverse families of large eukaryotic DNA viruses.** *J Virol* 2001, **75**:11720-11734.
178. Wuitschick JD, Gershan JA, Lochowicz AJ, Li S, Karrer KM: **A novel family of mobile genetic elements is limited to the germline genome in *Tetrahymena thermophila*.** *Nucleic Acids Res* 2002, **30**:2524-2537.
179. Zhang W, Imperiale MJ: **Requirement of the adenovirus IVa2 protein for virus assembly.** *J Virol* 2003, **77**:3586-3594.
180. Iyer LM, Balaji S, Koonin EV, Aravind L: **Evolutionary genomics of nucleocytoplasmic large DNA viruses.** *Virus Res* 2006, **117**:156-184.
181. Iyer LM, Leipe DD, Koonin EV, Aravind L: **Evolutionary history and higher order classification of AAA+ ATPases.** *J Struct Biol* 2004, **146**:11-31.
182. Neuwald AF, Aravind L, Spouge JL, Koonin EV: **AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes.** *Genome Res* 1999, **9**:27-43.
183. Kanamaru S, Kondabagil K, Rossmann MG, Rao VB: **The functional domains of bacteriophage t4 terminase.** *J Biol Chem* 2004, **279**:40795-40801.
184. Ponchon L, Boulanger P, Labesse G, Letellier L: **The endonuclease domain of bacteriophage terminases belongs to the resolvase/integrase/ribonuclease H superfamily: a bioinformatics analysis validated by a functional study on bacteriophage T5.** *J Biol Chem* 2006, **281**:5829-5836.
185. Holland IB, Blight MA: **ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans.** *J Mol Biol* 1999, **293**:381-399.
186. Lisal J, Kainov DE, Bamford DH, Thomas GJ, Jr., Tuma R: **Enzymatic mechanism of RNA translocation in double-stranded RNA bacteriophages.** *J Biol Chem* 2004, **279**:1343-1350.

187. Kainov DE, Pirttimaa M, Tuma R, Butcher SJ, Thomas GJ, Jr., Bamford DH, Makeyev EV: **RNA packaging device of double-stranded RNA bacteriophages, possibly as simple as hexamer of P4 protein.** *J Biol Chem* 2003, **278**:48084-48091.
188. Garcia AD, Aravind L, Koonin EV, Moss B: **Bacterial-type DNA holliday junction resolvases in eukaryotic viruses.** *Proc Natl Acad Sci U S A* 2000, **97**:8926-8931.
189. Aravind L, Makarova KS, Koonin EV: **SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.** *Nucleic Acids Res* 2000, **28**:3417-3432.
190. Simpson AA, Tao Y, Leiman PG, Badasso MO, He Y, Jardine PJ, Olson NH, Morais MC, Grimes S, Anderson DL, et al: **Structure of the bacteriophage phi29 DNA packaging motor.** *Nature* 2000, **408**:745-750.
191. Guasch A, Pous J, Parraga A, Valpuesta JM, Carrascosa JL, Coll M: **Crystallographic analysis reveals the 12-fold symmetry of the bacteriophage phi29 connector particle.** *J Mol Biol* 1998, **281**:219-225.
192. Mura C, Phillips M, Kozhukhovskiy A, Eisenberg D: **Structure and assembly of an augmented Sm-like archaeal protein 14-mer.** *Proc Natl Acad Sci U S A* 2003, **100**:4539-4544.
193. Perez-Romero P, Gustin KE, Imperiale MJ: **Dependence of the encapsidation function of the adenovirus L1 52/55-kilodalton protein on its ability to bind the packaging sequence.** *J Virol* 2006, **80**:1965-1971.
194. Gustin KE, Lutz P, Imperiale MJ: **Interaction of the adenovirus L1 52/55-kilodalton protein with the IVa2 gene product during infection.** *J Virol* 1996, **70**:6463-6467.
195. Mitchell MS, Matsuzaki S, Imai S, Rao VB: **Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases.** *Nucleic Acids Res* 2002, **30**:4009-4021.
196. Anantharaman V, Aravind L: **Novel conserved domains in proteins with predicted roles in eukaryotic cell-cycle regulation, decapping and RNA stability.** *BMC Genomics* 2004, **5**:45.
197. Huynen M, Snel B, Lathe W, 3rd, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10**:1204-1210.
198. Gual A, Alonso JC: **Characterization of the small subunit of the terminase enzyme of the *Bacillus subtilis* bacteriophage SPP1.** *Virology* 1998, **242**:279-287.

199. Lin H, Simon MN, Black LW: **Purification and characterization of the small subunit of phage T4 terminase, gp16, required for DNA packaging.** *J Biol Chem* 1997, **272**:3495-3501.
200. Bain DL, Berton N, Ortega M, Baran J, Yang Q, Catalano CE: **Biophysical characterization of the DNA binding domain of gpNu1, a viral DNA packaging protein.** *J Biol Chem* 2001, **276**:20175-20181.
201. de Beer T, Fang J, Ortega M, Yang Q, Maes L, Duffy C, Berton N, Sippy J, Overduin M, Feiss M, Catalano CE: **Insights into specific DNA recognition during the assembly of a viral genome packaging machine.** *Mol Cell* 2002, **9**:981-991.
202. Stiege AC, Isidro A, Droge A, Tavares P: **Specific targeting of a DNA-binding protein to the SPP1 procapsid by interaction with the portal oligomer.** *Mol Microbiol* 2003, **49**:1201-1212.
203. Droge A, Santos MA, Stiege AC, Alonso JC, Lurz R, Trautner TA, Tavares P: **Shape and DNA packaging activity of bacteriophage SPP1 procapsid: protein components and interactions during assembly.** *J Mol Biol* 2000, **296**:117-132.
204. Depping R, Lohaus C, Meyer HE, Ruger W: **The mono-ADP-ribosyltransferases Alt and ModB of bacteriophage T4: target proteins identified.** *Biochem Biophys Res Commun* 2005, **335**:1217-1223.
205. Dassa B, Yanai I, Pietrokovski S: **New type of polyubiquitin-like genes with intein-like autoprocessing domains.** *Trends Genet* 2004, **20**:538-542.
206. Iyer LM, Koonin EV, Aravind L: **Classification and evolutionary history of the single-strand annealing proteins, RecT, Redbeta, ERF and RAD52.** *BMC Genomics* 2002, **3**:8.
207. Liu J, Mushegian A: **Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages.** *J Bacteriol* 2004, **186**:4369-4375.
208. Rice PA, Baker TA: **Comparative architecture of transposase and integrase complexes.** *Nat Struct Biol* 2001, **8**:302-307.
209. Kapitonov VV, Jurka J: **RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons.** *PLoS Biol* 2005, **3**:e181.
210. Xiao F, Moll WD, Guo S, Guo P: **Binding of pRNA to the N-terminal 14 amino acids of connector protein of bacteriophage phi29.** *Nucleic Acids Res* 2005, **33**:2640-2649.
211. Sun S, Kondabagil K, Gentz PM, Rossmann MG, Rao VB: **The structure of the ATPase that powers DNA packaging into bacteriophage T4 procapsids.** *Mol Cell* 2007, **25**:943-949.

212. Aravind L, Anantharaman V, Koonin EV: **Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA.** *Proteins* 2002, **48**:1-14.
213. Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L: **Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes.** *J Mol Biol* 2006, **361**:1003-1034.
214. Vincent JB, Crowder MW, Averill BA: **Hydrolysis of phosphate monoesters: a biological problem with multiple chemical solutions.** *Trends Biochem Sci* 1992, **17**:105-110.
215. Bork P, Sander C, Valencia A: **An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins.** *Proc Natl Acad Sci U S A* 1992, **89**:7290-7294.
216. Aravind L, Koonin EV: **A novel family of predicted phosphoesterases includes *Drosophila* prune protein and bacterial RecJ exonuclease.** *Trends Biochem Sci* 1998, **23**:17-19.
217. Aravind L: **An evolutionary classification of the metallo-beta-lactamase fold proteins.** *In Silico Biol* 1999, **1**:69-91.
218. Aravind L, Koonin EV: **Phosphoesterase domains associated with DNA polymerases of diverse origins.** *Nucleic Acids Res* 1998, **26**:3746-3752.
219. Koonin EV, Tatusov RL: **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* 1994, **244**:125-132.
220. Aravind L, Galperin MY, Koonin EV: **The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold.** *Trends Biochem Sci* 1998, **23**:127-129.
221. Goldberg J, Huang HB, Kwon YG, Greengard P, Nairn AC, Kuriyan J: **Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1.** *Nature* 1995, **376**:745-753.
222. Whisstock JC, Romero S, Gurung R, Nandurkar H, Ooms LM, Bottomley SP, Mitchell CA: **The inositol polyphosphate 5-phosphatases and the apurinic/apyrimidinic base excision repair endonucleases share a common mechanism for catalysis.** *J Biol Chem* 2000, **275**:37055-37061.
223. Grebe TW, Stock JB: **The histidine protein kinase superfamily.** *Adv Microb Physiol* 1999, **41**:139-227.

224. Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR: **Evolution of two-component signal transduction.** *Mol Biol Evol* 2000, **17**:1956-1970.
225. Hogg T, Mechold U, Malke H, Cashel M, Hilgenfeld R: **Conformational antagonism between opposing active sites in a bifunctional RelA/SpoT homolog modulates (p)ppGpp metabolism during the stringent response [corrected].** *Cell* 2004, **117**:57-68.
226. Iyer LM, Aravind L: **The catalytic domains of thiamine triphosphatase and CyaB-like adenylyl cyclase define a novel superfamily of domains that bind organic phosphates.** *BMC Genomics* 2002, **3**:33.
227. Allen KN, Dunaway-Mariano D: **Phosphoryl group transfer: evolution of a catalytic scaffold.** *Trends Biochem Sci* 2004, **29**:495-503.
228. Yamagata A, Kakuta Y, Masui R, Fukuyama K: **The crystal structure of exonuclease RecJ bound to Mn²⁺ ion suggests how its characteristic motifs are involved in exonuclease activity.** *Proc Natl Acad Sci U S A* 2002, **99**:5908-5912.
229. Ahn S, Milner AJ, Futterer K, Konopka M, Ilias M, Young TW, White SA: **The "open" and "closed" structures of the type-C inorganic pyrophosphatases from Bacillus subtilis and Streptococcus gordonii.** *J Mol Biol* 2001, **313**:797-811.
230. Teplyakov A, Obmolova G, Khil PP, Howard AJ, Camerini-Otero RD, Gilliland GL: **Crystal structure of the Escherichia coli YcdX protein reveals a trinuclear zinc active site.** *Proteins* 2003, **51**:315-318.
231. Knofel T, Strater N: **X-ray structure of the Escherichia coli periplasmic 5'-nucleotidase containing a dimetal catalytic site.** *Nat Struct Biol* 1999, **6**:448-453.
232. Mol CD, Kuo CF, Thayer MM, Cunningham RP, Tainer JA: **Structure and function of the multifunctional DNA-repair enzyme exonuclease III.** *Nature* 1995, **374**:381-386.
233. Collet JF, Stroobant V, Pirard M, Delpierre G, Van Schaftingen E: **A new class of phosphotransferases phosphorylated on an aspartate residue in an amino-terminal DXDX(T/V) motif.** *J Biol Chem* 1998, **273**:14107-14112.
234. Anantharaman V, Aravind L, Koonin EV: **Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins.** *Curr Opin Chem Biol* 2003, **7**:12-20.
235. Shin DH, Roberts A, Jancarik J, Yokota H, Kim R, Wemmer DE, Kim SH: **Crystal structure of a phosphatase with a unique substrate binding domain from Thermotoga maritima.** *Protein Sci* 2003, **12**:1464-1472.
236. Morais MC, Zhang W, Baker AS, Zhang G, Dunaway-Mariano D, Allen KN: **The crystal structure of bacillus cereus phosphonoacetaldehyde hydrolase: insight into catalysis of**

- phosphorus bond cleavage and catalytic diversification within the HAD enzyme superfamily.** *Biochemistry* 2000, **39**:10385-10396.
237. Baker AS, Ciocci MJ, Metcalf WW, Kim J, Babbitt PC, Wanner BL, Martin BM, Dunaway-Mariano D: **Insights into the mechanism of catalysis by the P-C bond-cleaving enzyme phosphonoacetaldehyde hydrolase derived from gene sequence analysis and mutagenesis.** *Biochemistry* 1998, **37**:9305-9315.
238. Qian N, Stanley GA, Hahn-Hagerdal B, Radstrom P: **Purification and characterization of two phosphoglucomutases from *Lactococcus lactis* subsp. *lactis* and their regulation in maltose- and glucose-utilizing cells.** *J Bacteriol* 1994, **176**:5304-5311.
239. Collet JF, Gerin I, Rider MH, Veiga-da-Cunha M, Van Schaftingen E: **Human L-3-phosphoserine phosphatase: sequence, expression and evidence for a phosphoenzyme intermediate.** *FEBS Lett* 1997, **408**:281-284.
240. Seal SN, Rose ZB: **Characterization of a phosphoenzyme intermediate in the reaction of phosphoglycolate phosphatase.** *J Biol Chem* 1987, **262**:13496-13500.
241. Lahiri SD, Zhang G, Dunaway-Mariano D, Allen KN: **Caught in the act: the structure of phosphorylated beta-phosphoglucomutase from *Lactococcus lactis*.** *Biochemistry* 2002, **41**:8351-8359.
242. Ahmadian MR, Stege P, Scheffzek K, Wittinghofer A: **Confirmation of the arginine-finger hypothesis for the GAP-stimulated GTP-hydrolysis reaction of Ras.** *Nat Struct Biol* 1997, **4**:686-689.
243. Peisach E, Selengut JD, Dunaway-Mariano D, Allen KN: **X-ray crystal structure of the hypothetical phosphotyrosine phosphatase MDP-1 of the haloacid dehalogenase superfamily.** *Biochemistry* 2004, **43**:12770-12779.
244. Hisano T, Hata Y, Fujii T, Liu JQ, Kurihara T, Esaki N, Soda K: **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* 1996, **271**:20322-20330.
245. Toyoshima C, Nakasako M, Nomura H, Ogawa H: **Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution.** *Nature* 2000, **405**:647-655.
246. Wang W, Kim R, Jancarik J, Yokota H, Kim SH: **Crystal structure of phosphoserine phosphatase from *Methanococcus jannaschii*, a hyperthermophile, at 1.8 Å resolution.** *Structure (Camb)* 2001, **9**:65-71.
247. Lahiri SD, Zhang G, Dunaway-Mariano D, Allen KN: **The pentacovalent phosphorus intermediate of a phosphoryl transfer reaction.** *Science* 2003, **299**:2067-2071.

248. Rinaldo-Matthis A, Rampazzo C, Reichard P, Bianchi V, Nordlund P: **Crystal structure of a human mitochondrial deoxyribonucleotidase.** *Nat Struct Biol* 2002, **9**:779-787.
249. Olsen DB, Hepburn TW, Moos M, Mariano PS, Dunaway-Mariano D: **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* 1988, **27**:2229-2234.
250. Kurihara T, Liu JQ, Nardi-Dei V, Koshikawa H, Esaki N, Soda K: **Comprehensive site-directed mutagenesis of L-2-halo acid dehalogenase to probe catalytic amino acid residues.** *J Biochem (Tokyo)* 1995, **117**:1317-1322.
251. Rossmann MG, Moras D, Olsen KW: **Chemical and biological evolution of nucleotide-binding protein.** *Nature* 1974, **250**:194-199.
252. Zhao K, Chai X, Marmorstein R: **Structure of the yeast Hst2 protein deacetylase in ternary complex with 2'-O-acetyl ADP ribose and histone peptide.** *Structure* 2003, **11**:1403-1411.
253. Martin JL, McMillan FM: **SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold.** *Curr Opin Struct Biol* 2002, **12**:783-793.
254. Schubert HL, Blumenthal RM, Cheng X: **Many paths to methyltransfer: a chronicle of convergence.** *Trends Biochem Sci* 2003, **28**:329-335.
255. Sistla S, Rao DN: **S-Adenosyl-L-methionine-dependent restriction enzymes.** *Crit Rev Biochem Mol Biol* 2004, **39**:1-19.
256. Lowe J, Amos LA: **Crystal structure of the bacterial cell-division protein FtsZ.** *Nature* 1998, **391**:203-206.
257. Anantharaman V, Aravind L: **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* 2006, **356**:823-842.
258. Aravind L, Leipe DD, Koonin EV: **Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins.** *Nucleic Acids Res* 1998, **26**:4205-4213.
259. Clissold PM, Ponting CP: **PIN domains in nonsense-mediated mRNA decay and RNAi.** *Curr Biol* 2000, **10**:R888-890.
260. Finnin MS, Donigian JR, Cohen A, Richon VM, Rifkind RA, Marks PA, Breslow R, Pavletich NP: **Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors.** *Nature* 1999, **401**:188-193.

261. Whittaker CA, Hynes RO: **Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere.** *Mol Biol Cell* 2002, **13**:3369-3387.
262. Robinson VL, Buckler DR, Stock AM: **A tale of two components: a novel kinase and a regulatory switch.** *Nat Struct Biol* 2000, **7**:626-633.
263. Wolanin PM, Thomason PA, Stock JB: **Histidine protein kinases: key signal transducers outside the animal kingdom.** *Genome Biol* 2002, **3**:REVIEWS3013.
264. West AH, Stock AM: **Histidine kinases and response regulator proteins in two-component signaling systems.** *Trends Biochem Sci* 2001, **26**:369-376.
265. Ridder IS, Dijkstra BW: **Identification of the Mg²⁺-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* 1999, **339** (Pt 2):223-226.
266. Meng EC, Polacco BJ, Babbitt PC: **Superfamily active site templates.** *Proteins* 2004, **55**:962-976.
267. Merckel MC, Fabrichniy IP, Salminen A, Kalkkinen N, Baykov AA, Lahti R, Goldman A: **Crystal structure of Streptococcus mutans pyrophosphatase: a new fold for an old mechanism.** *Structure (Camb)* 2001, **9**:289-297.
268. Fabrichniy IP, Lehtio L, Salminen A, Zyryanov AB, Baykov AA, Lahti R, Goldman A: **Structural studies of metal ions in family II pyrophosphatases: the requirement for a Janus ion.** *Biochemistry* 2004, **43**:14403-14411.
269. Chen SJ, Wang JC: **Identification of active site residues in Escherichia coli DNA topoisomerase I.** *J Biol Chem* 1998, **273**:6050-6056.
270. Lee JO, Rieu P, Arnaout MA, Liddington R: **Crystal structure of the A domain from the alpha subunit of integrin CR3 (CD11b/CD18).** *Cell* 1995, **80**:631-638.
271. Kim Y, Yakunin AF, Kuznetsova E, Xu X, Pennycooke M, Gu J, Cheung F, Proudfoot M, Arrowsmith CH, Joachimiak A, et al: **Structure- and function-based characterization of a new phosphoglycolate phosphatase from Thermoplasma acidophilum.** *J Biol Chem* 2004, **279**:517-526.
272. Li YF, Hata Y, Fujii T, Hisano T, Nishihara M, Kurihara T, Esaki N: **Crystal structures of reaction intermediates of L-2-haloacid dehalogenase and implications for the reaction mechanism.** *J Biol Chem* 1998, **273**:15035-15044.

273. Ridder IS, Rozeboom HJ, Kalk KH, Janssen DB, Dijkstra BW: **Three-dimensional structure of L-2-haloacid dehalogenase from Xanthobacter autotrophicus GJ10 complexed with the substrate-analogue formate.** *J Biol Chem* 1997, **272**:33015-33022.
274. Calderone V, Forleo C, Benvenuti M, Cristina Thaller M, Maria Rossolini G, Mangani S: **The first structure of a bacterial class B Acid phosphatase reveals further structural heterogeneity among phosphatases of the haloacid dehalogenase fold.** *J Mol Biol* 2004, **335**:761-773.
275. Ridder IS, Rozeboom HJ, Kalk KH, Dijkstra BW: **Crystal structures of intermediates in the dehalogenation of haloalkanoates by L-2-haloacid dehalogenase.** *J Biol Chem* 1999, **274**:30672-30678.
276. Guo Y, Cheong N, Zhang Z, De Rose R, Deng Y, Farber SA, Fernandes-Alnemri T, Alnemri ES: **Tim50, a component of the mitochondrial translocator, regulates mitochondrial integrity and cell death.** *J Biol Chem* 2004, **279**:24813-24825.
277. Morais MC, Zhang G, Zhang W, Olsen DB, Dunaway-Mariano D, Allen KN: **X-ray crystallographic and site-directed mutagenesis analysis of the mechanism of Schiff-base formation in phosphonoacetaldehyde hydrolase catalysis.** *J Biol Chem* 2004, **279**:9353-9361.
278. Zhang G, Dai J, Wang L, Dunaway-Mariano D, Tremblay LW, Allen KN: **Catalytic cycling in beta-phosphoglucomutase: a kinetic and structural analysis.** *Biochemistry* 2005, **44**:9404-9416.
279. Wang W, Cho HS, Kim R, Jancarik J, Yokota H, Nguyen HH, Grigoriev IV, Wemmer DE, Kim SH: **Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic "snapshots" of intermediate states.** *J Mol Biol* 2002, **319**:421-431.
280. Selengut JD: **MDP-1 is a new and distinct member of the haloacid dehalogenase family of aspartate-dependent phosphohydrolases.** *Biochemistry* 2001, **40**:12704-12711.
281. Wu K, Chung L, Revill WP, Katz L, Reeves CD: **The FK520 gene cluster of Streptomyces hygroscopicus var. ascomyceticus (ATCC 14891) contains genes for biosynthesis of unusual polyketide extender units.** *Gene* 2000, **251**:81-90.
282. Hildebrand M, Waggoner LE, Liu H, Sudek S, Allen S, Anderson C, Sherman DH, Haygood M: **bryA: an unusual modular polyketide synthase gene from the uncultivated bacterial symbiont of the marine bryozoan Bugula neritina.** *Chem Biol* 2004, **11**:1543-1552.
283. Archambault J, Chambers RS, Kobor MS, Ho Y, Cartier M, Bolotin D, Andrews B, Kane CM, Greenblatt J: **An essential component of a C-terminal domain phosphatase that**

- interacts with transcription factor IIF in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci U S A* 1997, **94**:14300-14305.
284. Archambault J, Pan G, Dahmus GK, Cartier M, Marshall N, Zhang S, Dahmus ME, Greenblatt J: **FCP1, the RAP74-interacting subunit of a human protein phosphatase that dephosphorylates the carboxyl-terminal domain of RNA polymerase II.** *J Biol Chem* 1998, **273**:27593-27601.
285. Chambers RS, Dahmus ME: **Purification and characterization of a phosphatase from HeLa cells which dephosphorylates the C-terminal domain of RNA polymerase II.** *J Biol Chem* 1994, **269**:26243-26248.
286. Chambers RS, Kane CM: **Purification and characterization of an RNA polymerase II phosphatase from yeast.** *J Biol Chem* 1996, **271**:24498-24504.
287. Cho H, Kim TK, Mancebo H, Lane WS, Flores O, Reinberg D: **A protein phosphatase functions to recycle RNA polymerase II.** *Genes Dev* 1999, **13**:1540-1552.
288. Kobor MS, Archambault J, Lester W, Holstege FC, Gileadi O, Jansma DB, Jennings EG, Kouyoumdjian F, Davidson AR, Young RA, Greenblatt J: **An unusual eukaryotic protein phosphatase required for transcription by RNA polymerase II and CTD dephosphorylation in *S. cerevisiae*.** *Mol Cell* 1999, **4**:55-62.
289. Orphanides G, Reinberg D: **A unified theory of gene expression.** *Cell* 2002, **108**:439-451.
290. Siniosoglou S, Hurt EC, Pelham HR: **Psr1p/Psr2p, two plasma membrane phosphatases with an essential DXDX(T/V) motif required for sodium stress response in yeast.** *J Biol Chem* 2000, **275**:19352-19360.
291. Yeo M, Lin PS, Dahmus ME, Gill GN: **A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5.** *J Biol Chem* 2003, **278**:26078-26085.
292. Siniosoglou S, Santos-Rosa H, Rappsilber J, Mann M, Hurt E: **A novel complex of membrane proteins required for formation of a spherical nucleus.** *Embo J* 1998, **17**:6449-6464.
293. Xu H, Somers ZB, Robinson ML, 2nd, Hebert MD: **Tim50a, a nuclear isoform of the mitochondrial Tim50, interacts with proteins involved in snRNP biogenesis.** *BMC Cell Biol* 2005, **6**:29.
294. Yu X, Chini CC, He M, Mer G, Chen J: **The BRCT domain is a phospho-protein binding domain.** *Science* 2003, **302**:639-642.
295. Hugouvieux V, Kwak JM, Schroeder JI: **An mRNA cap binding protein, ABH1, modulates early abscisic acid signal transduction in *Arabidopsis*.** *Cell* 2001, **106**:477-487.

296. Xiong L, Lee H, Ishitani M, Tanaka Y, Stevenson B, Koiwa H, Bressan RA, Hasegawa PM, Zhu JK: **Repression of stress-responsive genes by FIERY2, a novel transcriptional regulator in Arabidopsis.** *Proc Natl Acad Sci U S A* 2002, **99**:10899-10904.
297. Zheng H, Ji C, Gu S, Shi B, Wang J, Xie Y, Mao Y: **Cloning and characterization of a novel RNA polymerase II C-terminal domain phosphatase.** *Biochem Biophys Res Commun* 2005, **331**:1401-1407.
298. Wu X, Chang A, Sudol M, Hanes SD: **Genetic interactions between the ESS1 prolyl-isomerase and the RSP5 ubiquitin ligase reveal opposing effects on RNA polymerase II function.** *Curr Genet* 2001, **40**:234-242.
299. Reichmann G, Dlugonska H, Fischer HG: **Characterization of TgROP9 (p36), a novel rhoptry protein of Toxoplasma gondii tachyzoites identified by T cell clone.** *Mol Biochem Parasitol* 2002, **119**:43-54.
300. Boyce JD, Chung JY, Adler B: **Genetic organisation of the capsule biosynthetic locus of Pasteurella multocida M1404 (B:2).** *Vet Microbiol* 2000, **72**:121-134.
301. Jilani A, Ramotar D, Slack C, Ong C, Yang XM, Scherer SW, Lasko DD: **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* 1999, **274**:24176-24186.
302. Soltis DA, Uhlenbeck OC: **Isolation and characterization of two mutant forms of T4 polynucleotide kinase.** *J Biol Chem* 1982, **257**:11332-11339.
303. Petrucco S, Volpi G, Bolchi A, Rivetti C, Ottonello S: **A nick-sensing DNA 3'-repair enzyme from Arabidopsis.** *J Biol Chem* 2002, **277**:23675-23683.
304. Betti M, Petrucco S, Bolchi A, Dieci G, Ottonello S: **A plant 3'-phosphoesterase involved in the repair of DNA strand breaks generated by oxidative damage.** *J Biol Chem* 2001, **276**:18038-18045.
305. Parsons JF, Lim K, Tempczyk A, Krajewski W, Eisenstein E, Herzberg O: **From structure to function: YrbI from Haemophilus influenzae (HI1679) is a phosphatase.** *Proteins* 2002, **46**:393-404.
306. Wu J, Woodard RW: **Escherichia coli YrbI is 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase.** *J Biol Chem* 2003, **278**:18117-18123.
307. Tzeng YL, Datta A, Strole C, Kolli VS, Birck MR, Taylor WP, Carlson RW, Woodard RW, Stephens DS: **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* 2002, **277**:24103-24113.

308. Leelapon O, Sarath G, Staswick PE: **A single amino acid substitution in soybean VSPalpha increases its acid phosphatase activity nearly 20-fold.** *Planta* 2004, **219**:1071-1079.
309. Utsugi S, Sakamoto W, Murata M, Motoyoshi F: **Arabidopsis thaliana vegetative storage protein (VSP) genes: gene organization and tissue-specific expression.** *Plant Mol Biol* 1998, **38**:565-576.
310. Gomez L, Faurobert M: **Contribution of vegetative storage proteins to seasonal nitrogen variations in the young shoots of peach trees (*Prunus persica* L. Batsch).** *J Exp Bot* 2002, **53**:2431-2439.
311. Hunsucker SA, Spsychala J, Mitchell BS: **Human cytosolic 5'-nucleotidase I: characterization and role in nucleoside analog resistance.** *J Biol Chem* 2001, **276**:10498-10504.
312. Sala-Newby GB, Skladanowski AC, Newby AC: **The mechanism of adenosine formation in cells. Cloning of cytosolic 5'-nucleotidase-I.** *J Biol Chem* 1999, **274**:17789-17793.
313. La Nauze JM, Rosenberg H: **The identification of 2-phosphonoacetaldehyde as an intermediate in the degradation of 2-aminoethylphosphonate by *Bacillus cereus*.** *Biochim Biophys Acta* 1968, **165**:438-447.
314. Dumora C, Lacoste AM, Cassaigne A: **Phosphonoacetaldehyde hydrolase from *Pseudomonas aeruginosa*: purification properties and comparison with *Bacillus cereus* enzyme.** *Biochim Biophys Acta* 1989, **997**:193-198.
315. Lee KS, Metcalf WW, Wanner BL: **Evidence for two phosphonate degradative pathways in *Enterobacter aerogenes*.** *J Bacteriol* 1992, **174**:2501-2510.
316. Jiang W, Metcalf WW, Lee KS, Wanner BL: **Molecular cloning, mapping, and regulation of Pho regulon genes for phosphonate breakdown by the phosphonatase pathway of *Salmonella typhimurium* LT2.** *J Bacteriol* 1995, **177**:6411-6421.
317. Ternan NG, Quinn JP: **In vitro cleavage of the carbon-phosphorus bond of phosphonopyruvate by cell extracts of an environmental *Burkholderia cepacia* isolate.** *Biochem Biophys Res Commun* 1998, **248**:378-381.
318. Parker GF, Higgins TP, Hawkes T, Robson RL: **Rhizobium (*Sinorhizobium*) meliloti phn genes: characterization and identification of their protein products.** *J Bacteriol* 1999, **181**:389-395.
319. Imig JD, Zhao X, Capdevila JH, Morisseau C, Hammock BD: **Soluble epoxide hydrolase inhibition lowers arterial blood pressure in angiotensin II hypertension.** *Hypertension* 2002, **39**:690-694.

320. Fang X, Kaduce TL, Weintraub NL, Harmon S, Teesch LM, Morisseau C, Thompson DA, Hammock BD, Spector AA: **Pathways of epoxyeicosatrienoic acid metabolism in endothelial cells. Implications for the vascular effects of soluble epoxide hydrolase inhibition.** *J Biol Chem* 2001, **276**:14867-14874.
321. Newman JW, Morisseau C, Harris TR, Hammock BD: **The soluble epoxide hydrolase encoded by EPXH2 is a bifunctional enzyme with novel lipid phosphate phosphatase activity.** *Proc Natl Acad Sci U S A* 2003, **100**:1558-1563.
322. Ogawa N, DeRisi J, Brown PO: **New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis.** *Mol Biol Cell* 2000, **11**:4309-4321.
323. Nakanishi T, Sekimizu K: **SDT1/SSM1, a multicopy suppressor of S-II null mutant, encodes a novel pyrimidine 5'-nucleotidase.** *J Biol Chem* 2002, **277**:22103-22106.
324. Fritzson P, Smith I: **A new nucleotidase of rat liver with activity toward 3'-and 5'-nucleotides.** *Biochim Biophys Acta* 1971, **235**:128-141.
325. Rampazzo C, Johansson M, Gallinaro L, Ferraro P, Hellman U, Karlsson A, Reichard P, Bianchi V: **Mammalian 5'(3')-deoxyribonucleotidase, cDNA cloning, and overexpression of the enzyme in *Escherichia coli* and mammalian cells.** *J Biol Chem* 2000, **275**:5409-5415.
326. Rampazzo C, Gallinaro L, Milanese E, Frigimelica E, Reichard P, Bianchi V: **A deoxyribonucleotidase in mitochondria: involvement in regulation of dNTP pools and possible link to genetic disease.** *Proc Natl Acad Sci U S A* 2000, **97**:8239-8244.
327. Leipe DD, Aravind L, Grishin NV, Koonin EV: **The bacterial replicative helicase DnaB evolved from a RecA duplication.** *Genome Res* 2000, **10**:5-16.
328. Levy HR: **Glucose-6-phosphate dehydrogenases.** *Adv Enzymol Relat Areas Mol Biol* 1979, **48**:97-192.
329. Lahiri SD, Zhang G, Dai J, Dunaway-Mariano D, Allen KN: **Analysis of the substrate specificity loop of the HAD superfamily cap domain.** *Biochemistry* 2004, **43**:2812-2820.
330. Gibson JL, Tabita FR: **Analysis of the cbbXYZ operon in *Rhodobacter sphaeroides*.** *J Bacteriol* 1997, **179**:663-669.
331. Sanz P, Randez-Gil F, Prieto JA: **Molecular characterization of a gene that confers 2-deoxyglucose resistance in yeast.** *Yeast* 1994, **10**:1195-1202.
332. Randez-Gil F, Blasco A, Prieto JA, Sanz P: **DOGR1 and DOGR2: two genes from *Saccharomyces cerevisiae* that confer 2-deoxyglucose resistance when overexpressed.** *Yeast* 1995, **11**:1233-1240.

333. Norbeck J, Pahlman AK, Akhtar N, Blomberg A, Adler L: **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* 1996, **271**:13875-13881.
334. Coquard D, Huecas M, Ott M, van Dijl JM, van Loon AP, Hohmann HP: **Molecular cloning and characterisation of the ribC gene from *Bacillus subtilis*: a point mutation in ribC results in riboflavin overproduction.** *Mol Gen Genet* 1997, **254**:81-84.
335. Mack M, van Loon AP, Hohmann HP: **Regulation of riboflavin biosynthesis in *Bacillus subtilis* is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC.** *J Bacteriol* 1998, **180**:950-955.
336. Hill KE, Marchesi JR, Weightman AJ: **Investigation of two evolutionarily unrelated halocarboxylic acid dehalogenase gene families.** *J Bacteriol* 1999, **181**:2535-2547.
337. Murdiyatmo U, Asmara W, Tsang JS, Baines AJ, Bull AT, Hardman DJ: **Molecular biology of the 2-haloacid halidohydrolase IVa from *Pseudomonas cepacia* MBA4.** *Biochem J* 1992, **284 (Pt 1)**:87-93.
338. Tsang JS, Pang BC: **Identification of the dimerization domain of dehalogenase IVa of *Burkholderia cepacia* MBA4.** *Appl Environ Microbiol* 2000, **66**:3180-3186.
339. Myers RW, Wray JW, Fish S, Abeles RH: **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* 1993, **268**:24785-24791.
340. Balakrishnan R, Frohlich M, Rahaim PT, Backman K, Yocum RR: **Appendix. Cloning and sequence of the gene encoding enzyme E-1 from the methionine salvage pathway of *Klebsiella oxytoca*.** *J Biol Chem* 1993, **268**:24792-24795.
341. Satola SW, Schirmer PL, Farley MM: **Complete sequence of the cap locus of *Haemophilus influenzae* serotype b and nonencapsulated b capsule-negative variants.** *Infect Immun* 2003, **71**:3639-3644.
342. Valentine WN, Fink K, Paglia DE, Harris SR, Adams WS: **Hereditary hemolytic anemia with human erythrocyte pyrimidine 5'-nucleotidase deficiency.** *J Clin Invest* 1974, **54**:866-879.
343. Paglia DE, Valentine WN: **Characteristics of a pyrimidine-specific 5'-nucleotidase in human erythrocytes.** *J Biol Chem* 1975, **250**:7973-7979.
344. Borkenhagen LF, Kennedy EP: **The enzymic equilibration of L-serine with O-phospho-L-serine.** *Biochim Biophys Acta* 1958, **28**:222-223.

345. Neuhaus FC, Byrne WL: **O-Phosphoserine phosphatase**. *Biochim Biophys Acta* 1958, **28**:223-224.
346. Schirch L, Gross T: **Serine transhydroxymethylase. Identification as the threonine and allothreonine aldolases**. *J Biol Chem* 1968, **243**:5651-5655.
347. Ulevitch RJ, Kallen RG: **Purification and characterization of pyridoxal 5'-phosphate dependent serine hydroxymethylase from lamb liver and its action upon beta-phenylserines**. *Biochemistry* 1977, **16**:5342-5350.
348. Szebenyi DM, Musayev FN, di Salvo ML, Safo MK, Schirch V: **Serine hydroxymethyltransferase: role of glu75 and evidence that serine is cleaved by a retroaldol mechanism**. *Biochemistry* 2004, **43**:6865-6876.
349. Patte JC, Clepet C, Bally M, Borne F, Mejean V, Foglino M: **ThrH, a homoserine kinase isozyme with in vivo phosphoserine phosphatase activity in Pseudomonas aeruginosa**. *Microbiology* 1999, **145 (Pt 4)**:845-853.
350. Singh SK, Yang K, Karthikeyan S, Huynh T, Zhang X, Phillips MA, Zhang H: **The thrH gene product of Pseudomonas aeruginosa is a dual activity enzyme with a novel phosphoserine:homoserine phosphotransferase activity**. *J Biol Chem* 2004, **279**:13166-13173.
351. Houston B, Seawright E, Jefferies D, Hoogland E, Lester D, Whitehead C, Farquharson C: **Identification and cloning of a novel phosphatase expressed at high levels in differentiating growth plate chondrocytes**. *Biochim Biophys Acta* 1999, **1448**:500-506.
352. Houston B, Paton IR, Burt DW, Farquharson C: **Chromosomal localization of the chicken and mammalian orthologues of the orphan phosphatase PHOSPHO1 gene**. *Anim Genet* 2002, **33**:451-454.
353. Beeston AL, Surette MG: **pfs-dependent regulation of autoinducer 2 production in Salmonella enterica serovar Typhimurium**. *J Bacteriol* 2002, **184**:3450-3456.
354. Allegrini S, Scaloni A, Ferrara L, Pesi R, Pinna P, Sgarrella F, Camici M, Eriksson S, Tozzi MG: **Bovine cytosolic 5'-nucleotidase acts through the formation of an aspartate 52-phosphoenzyme intermediate**. *J Biol Chem* 2001, **276**:33526-33532.
355. Oka J, Matsumoto A, Hosokawa Y, Inoue S: **Molecular cloning of human cytosolic purine 5'-nucleotidase**. *Biochem Biophys Res Commun* 1994, **205**:917-922.
356. Rebay I, Silver SJ, Tootle TL: **New vision from Eyes absent: transcription factors as enzymes**. *Trends Genet* 2005, **21**:163-171.

357. Tootle TL, Silver SJ, Davies EL, Newman V, Latek RR, Mills IA, Selengut JD, Parlikar BE, Rebay I: **The transcription factor Eyes absent is a protein tyrosine phosphatase.** *Nature* 2003, **426**:299-302.
358. Moller JV, Juul B, le Maire M: **Structural organization, ion transport, and energy transduction of P-type ATPases.** *Biochim Biophys Acta* 1996, **1286**:1-51.
359. Axelsen KB, Palmgren MG: **Evolution of substrate specificities in the P-type ATPase superfamily.** *J Mol Evol* 1998, **46**:84-101.
360. Fagan MJ, Saier MH, Jr.: **P-type ATPases of eukaryotes and bacteria: sequence analyses and construction of phylogenetic trees.** *J Mol Evol* 1994, **38**:57-99.
361. Cronin SR, Rao R, Hampton RY: **Cod1p/Spf1p is a P-type ATPase involved in ER function and Ca²⁺ homeostasis.** *J Cell Biol* 2002, **157**:1017-1028.
362. Ogawa H, Haga T, Toyoshima C: **Soluble P-type ATPase from an archaeon, Methanococcus jannaschii.** *FEBS Lett* 2000, **471**:99-102.
363. Bramkamp M, Gassel M, Herkenhoff-Hesselmann B, Bertrand J, Altendorf K: **The Methanocaldococcus jannaschii protein Mj0968 is not a P-type ATPase.** *FEBS Lett* 2003, **543**:31-36.
364. le Coq D, Fillinger S, Aymerich S: **Histidinol phosphate phosphatase, catalyzing the penultimate step of the histidine biosynthesis pathway, is encoded by ytvP (hisJ) in Bacillus subtilis.** *J Bacteriol* 1999, **181**:3277-3280.
365. Kneidinger B, Marolda C, Graninger M, Zamyatina A, McArthur F, Kosma P, Valvano MA, Messner P: **Biosynthesis pathway of ADP-L-glycero-beta-D-manno-heptose in Escherichia coli.** *J Bacteriol* 2002, **184**:363-369.
366. Plumbridge JA: **Sequence of the nagBACD operon in Escherichia coli K12 and pattern of transcription within the nag regulon.** *Mol Microbiol* 1989, **3**:505-515.
367. Peri KG, Goldie H, Waygood EB: **Cloning and characterization of the N-acetylglucosamine operon of Escherichia coli.** *Biochem Cell Biol* 1990, **68**:123-137.
368. Perraud AL, Fleig A, Dunn CA, Bagley LA, Launay P, Schmitz C, Stokes AJ, Zhu Q, Bessman MJ, Penner R, et al: **ADP-ribose gating of the calcium-permeable LTRPC2 channel revealed by Nudix motif homology.** *Nature* 2001, **411**:595-599.
369. Bessman MJ, Frick DN, O'Handley SF: **The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes.** *J Biol Chem* 1996, **271**:25059-25062.

370. Gohla A, Birkenfeld J, Bokoch GM: **Chronophin, a novel HAD-type serine protein phosphatase, regulates cofilin-dependent actin dynamics.** *Nat Cell Biol* 2005, **7**:21-29.
371. Hiraishi H, Ohmagari T, Otsuka Y, Yokoi F, Kumon A: **Purification and characterization of hepatic inorganic pyrophosphatase hydrolyzing imidodiphosphate.** *Arch Biochem Biophys* 1997, **341**:153-159.
372. Yokoi F, Hiraishi H, Izuhara K: **Molecular cloning of a cDNA for the human phospholysine phosphohistidine inorganic pyrophosphate phosphatase.** *J Biochem (Tokyo)* 2003, **133**:607-614.
373. Vandercammen A, Francois J, Hers HG: **Characterization of trehalose-6-phosphate synthase and trehalose-6-phosphate phosphatase of *Saccharomyces cerevisiae*.** *Eur J Biochem* 1989, **182**:613-620.
374. Kaasen I, Falkenberg P, Styrvold OB, Strom AR: **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* 1992, **174**:889-898.
375. De Smet KA, Weston A, Brown IN, Young DB, Robertson BD: **Three pathways for trehalose biosynthesis in mycobacteria.** *Microbiology* 2000, **146 (Pt 1)**:199-208.
376. Wolf A, Kramer R, Morbach S: **Three pathways for trehalose metabolism in *Corynebacterium glutamicum* ATCC13032 and their significance in response to osmotic stress.** *Mol Microbiol* 2003, **49**:1119-1134.
377. Empadinhas N, Marugg JD, Borges N, Santos H, da Costa MS: **Pathway for the synthesis of mannosylglycerate in the hyperthermophilic archaeon *Pyrococcus horikoshii*. Biochemical and genetic characterization of key enzymes.** *J Biol Chem* 2001, **276**:43580-43588.
378. Borges N, Marugg JD, Empadinhas N, da Costa MS, Santos H: **Specialized roles of the two pathways for the synthesis of mannosylglycerate in osmoadaptation and thermoadaptation of *Rhodothermus marinus*.** *J Biol Chem* 2004, **279**:9892-9898.
379. Empadinhas N, Albuquerque L, Henne A, Santos H, da Costa MS: **The bacterium *Thermus thermophilus*, like hyperthermophilic archaea, uses a two-step pathway for the synthesis of mannosylglycerate.** *Appl Environ Microbiol* 2003, **69**:3272-3279.
380. Tomavo S, Dubremetz JF, Schwarz RT: **Biosynthesis of glycolipid precursors for glycosylphosphatidylinositol membrane anchors in a *Toxoplasma gondii* cell-free system.** *J Biol Chem* 1992, **267**:21446-21458.
381. Lunn JE: **Evolution of sucrose synthesis.** *Plant Physiol* 2002, **128**:1490-1500.

382. Langenkamper G, Fung RW, Newcomb RD, Atkinson RG, Gardner RC, MacRae EA: **Sucrose phosphate synthase genes in plants belong to three different families.** *J Mol Evol* 2002, **54**:322-332.
383. Castleden CK, Aoki N, Gillespie VJ, MacRae EA, Quick WP, Buchner P, Foyer CH, Furbank RT, Lunn JE: **Evolution and function of the sucrose-phosphate synthase gene families in wheat and other grasses.** *Plant Physiol* 2004, **135**:1753-1764.
384. Hawker JS, Hatch MD: **A specific sucrose phosphatase from plant tissues.** *Biochem J* 1966, **99**:102-107.
385. Lunn JE, ap Rees T: **Apparent equilibrium constant and mass-action ratio for sucrose-phosphate synthase in seeds of *Pisum sativum*.** *Biochem J* 1990, **267**:739-743.
386. Murzin AG: **Structural classification of proteins: new superfamilies.** *Curr Opin Struct Biol* 1996, **6**:386-394.
387. Lunn JE: **Sucrose-phosphatase gene families in plants.** *Gene* 2003, **303**:187-196.
388. Bonini NM, Leiserson WM, Benzer S: **The eyes absent gene: genetic control of cell survival and differentiation in the developing *Drosophila* eye.** *Cell* 1993, **72**:379-395.
389. Peisach E, Wang L, Burroughs AM, Aravind L, Dunaway-Mariano D, Allen KN: **The X-ray crystallographic structure and activity analysis of a *Pseudomonas*-specific subfamily of the HAD enzyme superfamily evidences a novel biochemical function.** *Proteins* 2007.
390. Bradford MM: **A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.** *Anal Biochem* 1976, **72**:248-254.
391. Appel RD, Bairoch A, Hochstrasser DF: **A new generation of information retrieval tools for biologists: the example of the ExPASy WWW server.** *Trends Biochem Sci* 1994, **19**:258-260.
392. Otwinowski Z, Minor W: **Processing of X-ray Diffraction Data Collected in Oscillation Mode.** *Methods Enzymol* 1997, **276**:307-326.
393. Terwilliger TC, Berendzen J: **Automated MAD and MIR structure solution.** *Acta Crystallogr D Biol Crystallogr* 1999, **55 (Pt 4)**:849-861.
394. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al: **Crystallography & NMR system: A new software suite for macromolecular structure determination.** *Acta Crystallogr D Biol Crystallogr* 1998, **54 (Pt 5)**:905-921.

395. Emsley P, Cowtan K: **Coot: model-building tools for molecular graphics.** *Acta Crystallogr D Biol Crystallogr* 2004, **60**:2126-2132.
396. Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25**:231-234.
397. Zhang G, Morais MC, Dai J, Zhang W, Dunaway-Mariano D, Allen KN: **Investigation of metal ion binding in phosphonoacetaldehyde hydrolase identifies sequence markers for metal-activated enzymes of the HAD enzyme superfamily.** *Biochemistry* 2004, **43**:4990-4997.
398. Laskowski RA, Watson JD, Thornton JM: **From protein structure to biochemical function?** *J Struct Funct Genomics* 2003, **4**:167-177.
399. Laskowski RA, Watson JD, Thornton JM: **ProFunc: a server for predicting protein function from 3D structure.** *Nucleic Acids Res* 2005, **33**:W89-93.
400. del Sol A, Fujihashi H, Amoros D, Nussinov R: **Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families.** *Protein Sci* 2006, **15**:2120-2128.
401. Chakrabarti P, Janin J: **Dissecting protein-protein recognition sites.** *Proteins* 2002, **47**:334-343.
402. Fortpied J, Maliekal P, Vertommen D, Van Schaftingen E: **Magnesium-dependent phosphatase-1 is a protein-fructosamine-6-phosphatase potentially involved in glycation repair.** *J Biol Chem* 2006, **281**:18378-18385.
403. Jones S, Shanahan HP, Berman HM, Thornton JM: **Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins.** *Nucleic Acids Res* 2003, **31**:7189-7198.
404. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context.** *Curr Opin Struct Biol* 2000, **10**:366-370.
405. Iyer LM, Anantharaman V, Aravind L: **Ancient conserved domains shared by animal soluble guanylyl cyclases and bacterial signaling proteins.** *BMC Genomics* 2003, **4**:5.
406. Alberts B JA, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell*. 4th edn. New York, NY: Garland Science Publishing; 2002.
407. Ciechanover A, Orian A, Schwartz AL: **Ubiquitin-mediated proteolysis: biological regulation via destruction.** *Bioessays* 2000, **22**:442-451.
408. McGrath JP, Jentsch S, Varshavsky A: **UBA 1: an essential yeast gene encoding ubiquitin-activating enzyme.** *Embo J* 1991, **10**:227-236.

409. Duda DM, Walden H, Sfondouris J, Schulman BA: **Structural analysis of Escherichia coli ThiF.** *J Mol Biol* 2005, **349**:774-786.
410. Lehmann C, Begley TP, Ealick SE: **Structure of the Escherichia coli ThiS-ThiF complex, a key component of the sulfur transfer system in thiamin biosynthesis.** *Biochemistry* 2006, **45**:11-19.
411. Leimkuhler S, Wuebbens MM, Rajagopalan KV: **Characterization of Escherichia coli MoeB and its involvement in the activation of molybdopterin synthase for the biosynthesis of the molybdenum cofactor.** *J Biol Chem* 2001, **276**:34695-34701.
412. Lake MW, Wuebbens MM, Rajagopalan KV, Schindelin H: **Mechanism of ubiquitin activation revealed by the structure of a bacterial MoeB-MoaD complex.** *Nature* 2001, **414**:325-329.
413. Rudolph MJ, Wuebbens MM, Rajagopalan KV, Schindelin H: **Crystal structure of molybdopterin synthase and its evolutionary relationship to ubiquitin activation.** *Nat Struct Biol* 2001, **8**:42-46.
414. Iyer LM, Burroughs AM, Aravind L: **The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains.** *Genome Biol* 2006, **7**:R60.
415. Gonzalez-Pastor JE, San Millan JL, Castilla MA, Moreno F: **Structure and organization of plasmid genes required to produce the translation inhibitor microcin C7.** *J Bacteriol* 1995, **177**:7131-7140.
416. Onaka H, Nakaho M, Hayashi K, Igarashi Y, Furumai T: **Cloning and characterization of the goadsporin biosynthetic gene cluster from Streptomyces sp. TP-A0584.** *Microbiology* 2005, **151**:3923-3933.
417. Huang DT, Hunt HW, Zhuang M, Ohi MD, Holton JM, Schulman BA: **Basis for a ubiquitin-like protein thioester switch toggling E1-E2 affinity.** *Nature* 2007, **445**:394-398.
418. Lois LM, Lima CD: **Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1.** *Embo J* 2005, **24**:439-451.
419. Lassmann T, Sonnhammer EL: **Kalign--an accurate and fast multiple sequence alignment algorithm.** *BMC Bioinformatics* 2005, **6**:298.
420. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24**:1596-1599.
421. Jin J, Li X, Gygi SP, Harper JW: **Dual E1 activation systems for ubiquitin differentially regulate E2 enzyme charging.** *Nature* 2007, **447**:1135-1138.

422. Borthakur D, Basche M, Buikema WJ, Borthakur PB, Haselkorn R: **Expression, nucleotide sequence and mutational analysis of two open reading frames in the nif gene region of *Anabaena* sp. strain PCC7120.** *Mol Gen Genet* 1990, **221**:227-234.
423. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540.
424. Huang DT, Walden H, Duda D, Schulman BA: **Ubiquitin-like protein activation.** *Oncogene* 2004, **23**:1958-1971.
425. Matthies A, Nimtz M, Leimkuhler S: **Molybdenum cofactor biosynthesis in humans: identification of a persulfide group in the rhodanese-like domain of MOCS3 by mass spectrometry.** *Biochemistry* 2005, **44**:7912-7920.
426. Wang J, Hu W, Cai S, Lee B, Song J, Chen Y: **The intrinsic affinity between E2 and the Cys domain of E1 in ubiquitin-like modifications.** *Mol Cell* 2007, **27**:228-237.
427. Tasneem A, Iyer LM, Jakobsson E, Aravind L: **Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels.** *Genome Biol* 2005, **6**:R4.
428. Yamada Y, Suzuki NN, Hanada T, Ichimura Y, Kumeta H, Fujioka Y, Ohsumi Y, Inagaki F: **The crystal structure of Atg3, an autophagy-related ubiquitin carrier protein (E2) enzyme that mediates Atg8 lipidation.** *J Biol Chem* 2007, **282**:8036-8043.
429. Burroughs AM, Balaji S, Iyer LM, Aravind L: **Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold.** *Biol Direct* 2007, **2**:18.
430. Godert AM, Jin M, McLafferty FW, Begley TP: **Biosynthesis of the thioquinolobactin siderophore: an interesting variation on sulfur transfer.** *J Bacteriol* 2007, **189**:2941-2944.
431. Onaka H, Tabata H, Igarashi Y, Sato Y, Furumai T: **Goadsporin, a chemical substance which promotes secondary metabolism and morphogenesis in streptomycetes. I. Purification and characterization.** *J Antibiot (Tokyo)* 2001, **54**:1036-1044.
432. Igarashi Y, Kan Y, Fujii K, Fujita T, Harada K, Naoki H, Tabata H, Onaka H, Furumai T: **Goadsporin, a chemical substance which promotes secondary metabolism and Morphogenesis in streptomycetes. II. Structure determination.** *J Antibiot (Tokyo)* 2001, **54**:1045-1053.
433. Krepinsky K, Leimkuhler S: **Site-directed mutagenesis of the active site loop of the rhodanese-like domain of the human molybdopterin synthase sulfurase MOCS3. Major differences in substrate specificity between eukaryotic and bacterial homologs.** *Febs J* 2007, **274**:2778-2787.

434. Xi J, Ge Y, Kinsland C, McLafferty FW, Begley TP: **Biosynthesis of the thiazole moiety of thiamin in Escherichia coli: identification of an acyldisulfide-linked protein--protein conjugate that is functionally analogous to the ubiquitin/E1 complex.** *Proc Natl Acad Sci U S A* 2001, **98**:8513-8518.
435. Komatsu M, Chiba T, Tatsumi K, Iemura S, Tanida I, Okazaki N, Ueno T, Kominami E, Natsume T, Tanaka K: **A novel protein-conjugating system for Ufm1, a ubiquitin-fold modifier.** *Embo J* 2004, **23**:1977-1986.
436. Overington JP: **Comparison of three-dimensional structures of homologous proteins.** *Curr Opin Struct Biol* 1992, **2**:394-401.
437. Kraulis PJ: **Similarity of protein G and ubiquitin.** *Science* 1991, **254**:581-582.
438. Glickman MH, Ciechanover A: **The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction.** *Physiol Rev* 2002, **82**:373-428.
439. Goldstein G, Scheid M, Hammerling U, Schlesinger DH, Niall HD, Boyse EA: **Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells.** *Proc Natl Acad Sci U S A* 1975, **72**:11-15.
440. Wilkinson KD: **The discovery of ubiquitin-dependent proteolysis.** *Proc Natl Acad Sci U S A* 2005, **102**:15280-15282.
441. Vijay-Kumar S, Bugg CE, Cook WJ: **Structure of ubiquitin refined at 1.8 A resolution.** *J Mol Biol* 1987, **194**:531-544.
442. Vijay-Kumar S, Bugg CE, Wilkinson KD, Cook WJ: **Three-dimensional structure of ubiquitin at 2.8 A resolution.** *Proc Natl Acad Sci U S A* 1985, **82**:3582-3585.
443. Schwartz DC, Hochstrasser M: **A superfamily of protein tags: ubiquitin, SUMO and related modifiers.** *Trends Biochem Sci* 2003, **28**:321-328.
444. Weissman AM: **Themes and variations on ubiquitylation.** *Nat Rev Mol Cell Biol* 2001, **2**:169-178.
445. Furukawa K, Mizushima N, Noda T, Ohsumi Y: **A protein conjugation system in yeast with homology to biosynthetic enzyme reaction of prokaryotes.** *J Biol Chem* 2000, **275**:7462-7465.
446. Mizushima N, Noda T, Yoshimori T, Tanaka Y, Ishii T, George MD, Klionsky DJ, Ohsumi M, Ohsumi Y: **A protein conjugation system essential for autophagy.** *Nature* 1998, **395**:395-398.

447. Kamitani T, Kito K, Nguyen HP, Yeh ET: **Characterization of NEDD8, a developmentally down-regulated ubiquitin-like protein.** *J Biol Chem* 1997, **272**:28557-28562.
448. Dohmen RJ: **SUMO protein modification.** *Biochim Biophys Acta* 2004, **1695**:113-131.
449. Hay RT: **SUMO: a history of modification.** *Mol Cell* 2005, **18**:1-12.
450. May MJ, Larsen SE, Shim JH, Madge LA, Ghosh S: **A novel ubiquitin-like domain in IkappaB kinase beta is required for functional activity of the kinase.** *J Biol Chem* 2004, **279**:45528-45539.
451. Neuber O, Jarosch E, Volkwein C, Walter J, Sommer T: **Ubx2 links the Cdc48 complex to ER-associated protein degradation.** *Nat Cell Biol* 2005, **7**:993-998.
452. Schuberth C, Buchberger A: **Membrane-bound Ubx2 recruits Cdc48 to ubiquitin ligases and their substrates to ensure efficient ER-associated protein degradation.** *Nat Cell Biol* 2005, **7**:999-1006.
453. Aravind L, Dixit VM, Koonin EV: **Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons.** *Science* 2001, **291**:1279-1284.
454. Tanaka K, Suzuki T, Chiba T: **The ligation systems for ubiquitin and ubiquitin-like proteins.** *Mol Cells* 1998, **8**:503-512.
455. Ardley HC, Robinson PA: **E3 ubiquitin ligases.** *Essays Biochem* 2005, **41**:15-30.
456. Pickart CM: **Mechanisms underlying ubiquitination.** *Annu Rev Biochem* 2001, **70**:503-533.
457. Soboleva TA, Baker RT: **Deubiquitinating enzymes: their functions and substrate specificity.** *Curr Protein Pept Sci* 2004, **5**:191-200.
458. Guterman A, Glickman MH: **Deubiquitinating enzymes are IN/(trinsic to proteasome function).** *Curr Protein Pept Sci* 2004, **5**:201-211.
459. Iyer LM, Koonin EV, Aravind L: **Novel predicted peptidases with a potential role in the ubiquitin signaling pathway.** *Cell Cycle* 2004, **3**:1440-1450.
460. Nijman SM, Luna-Vargas MP, Velds A, Brummelkamp TR, Dirac AM, Sixma TK, Bernards R: **A genomic and functional inventory of deubiquitinating enzymes.** *Cell* 2005, **123**:773-786.
461. Wing SS: **Deubiquitinating enzymes--the importance of driving in reverse along the ubiquitin-proteasome pathway.** *Int J Biochem Cell Biol* 2003, **35**:590-605.
462. Sankaranarayanan R, Dock-Bregeon AC, Romby P, Caillet J, Springer M, Rees B, Ehresmann C, Ehresmann B, Moras D: **The structure of threonyl-tRNA synthetase-**

- tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site.** *Cell* 1999, **97**:371-381.
463. Wolf YI, Aravind L, Grishin NV, Koonin EV: **Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events.** *Genome Res* 1999, **9**:689-710.
464. Kim MH, Cierpicki T, Derewenda U, Krowarsch D, Feng Y, Devedjiev Y, Dauter Z, Walsh CA, Otlewski J, Bushweller JH, Derewenda ZS: **The DCX-domain tandems of doublecortin and doublecortin-like kinase.** *Nat Struct Biol* 2003, **10**:324-333.
465. Nassar N, Horn G, Herrmann C, Scherer A, McCormick F, Wittinghofer A: **The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue.** *Nature* 1995, **375**:554-560.
466. Ito T, Matsui Y, Ago T, Ota K, Sumimoto H: **Novel modular domain PB1 recognizes PC motif to mediate functional protein-protein interactions.** *Embo J* 2001, **20**:3938-3946.
467. Pearson MA, Reczek D, Bretscher A, Karplus PA: **Structure of the ERM protein moesin reveals the FERM domain fold masked by an extended actin binding tail domain.** *Cell* 2000, **101**:259-270.
468. Burroughs AM, Balaji S, Iyer LM, Aravind L: **A novel superfamily containing the β -grasp fold involved in binding diverse soluble ligands.** *Biology Direct* 2007, **2**:4.
469. Sazanov LA, Hinchliffe P: **Structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*.** *Science* 2006, **311**:1430-1436.
470. Wuerges J, Garau G, Geremia S, Fedosov SN, Petersen TE, Randaccio L: **Structural basis for mammalian vitamin B12 transport by transcobalamin.** *Proc Natl Acad Sci U S A* 2006, **103**:4386-4391.
471. Fraser JD, Urban RG, Strominger JL, Robinson H: **Zinc regulates the function of two superantigens.** *Proc Natl Acad Sci U S A* 1992, **89**:5507-5511.
472. Sazinsky MH, Bard J, Di Donato A, Lippard SJ: **Crystal structure of the toluene/o-xylene monooxygenase hydroxylase from *Pseudomonas stutzeri* OX1. Insight into the substrate specificity, substrate channeling, and active site tuning of multicomponent monooxygenases.** *J Biol Chem* 2004, **279**:30600-30610.
473. Gnatt AL, Cramer P, Fu J, Bushnell DA, Kornberg RD: **Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution.** *Science* 2001, **292**:1876-1882.

474. Biou V, Shu F, Ramakrishnan V: **X-ray crystallography shows that translational initiation factor IF3 consists of two compact alpha/beta domains linked by an alpha-helix.** *Embo J* 1995, **14**:4056-4064.
475. Kycia JH, Biou V, Shu F, Gerchman SE, Graziano V, Ramakrishnan V: **Prokaryotic translation initiation factor IF3 is an elongated protein consisting of two crystallizable domains.** *Biochemistry* 1995, **34**:6183-6187.
476. Iyer LM, Koonin EV, Aravind L: **Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases.** *BMC Struct Biol* 2003, **3**:1.
477. Rabijns A, De Bondt HL, De Ranter C: **Three-dimensional structure of staphylokinase, a plasminogen activator with therapeutic potential.** *Nat Struct Biol* 1997, **4**:357-360.
478. Weber DJ, Abeygunawardana C, Bessman MJ, Mildvan AS: **Secondary structure of the MutT enzyme as determined by NMR.** *Biochemistry* 1993, **32**:13081-13088.
479. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19**:1572-1574.
480. Pineda-Lucena A, Liao JC, Cort JR, Yee A, Kennedy MA, Edwards AM, Arrowsmith CH: **A novel member of the split betaalpha fold: Solution structure of the hypothetical protein YML108W from *Saccharomyces cerevisiae*.** *Protein Sci* 2003, **12**:1136-1140.
481. Stoldt M, Wohnert J, Gorchach M, Brown LR: **The NMR structure of *Escherichia coli* ribosomal protein L25 shows homology to general stress proteins and glutaminyl-tRNA synthetases.** *Embo J* 1998, **17**:6377-6384.
482. Clout NJ, Tisi D, Hohenester E: **Novel fold revealed by the structure of a FAS1 domain pair from the insect cell adhesion molecule fasciclin I.** *Structure* 2003, **11**:197-203.
483. Almasy RJ, Janson CA, Hamlin R, Xuong NH, Eisenberg D: **Novel subunit-subunit interactions in the structure of glutamine synthetase.** *Nature* 1986, **323**:304-309.
484. Janson CA, Kayne PS, Almasy RJ, Grunstein M, Eisenberg D: **Sequence of glutamine synthetase from *Salmonella typhimurium* and implications for the protein structure.** *Gene* 1986, **46**:297-300.
485. Nishiyama M, Horst R, Eidam O, Herrmann T, Ignatov O, Vetsch M, Bettendorff P, Jelesarov I, Grutter MG, Wuthrich K, et al: **Structural basis of chaperone-subunit complex recognition by the type 1 pilus assembly platform FimD.** *Embo J* 2005, **24**:2075-2086.

486. Kisker C, Schindelin H, Pacheco A, Wehbi WA, Garrett RM, Rajagopalan KV, Enemark JH, Rees DC: **Molecular basis of sulfite oxidase deficiency from the structure of sulfite oxidase.** *Cell* 1997, **91**:973-983.
487. Chan MK, Mukund S, Kletzin A, Adams MW, Rees DC: **Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase.** *Science* 1995, **267**:1463-1469.
488. Davies C, Gerstner RB, Draper DE, Ramakrishnan V, White SW: **The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: one domain shows structural homology to the ETS DNA-binding motif.** *Embo J* 1998, **17**:4545-4558.
489. Zweifel ME, Leahy DJ, Barrick D: **Structure and Notch receptor binding of the tandem WWE domain of Deltex.** *Structure* 2005, **13**:1599-1611.
490. Ahmad KF, Engel CK, Prive GG: **Crystal structure of the BTB domain from PLZF.** *Proc Natl Acad Sci U S A* 1998, **95**:12123-12128.
491. Staker BL, Korber P, Bardwell JC, Saper MA: **Structure of Hsp15 reveals a novel RNA-binding motif.** *Embo J* 2000, **19**:749-757.
492. Gomez M, Cutting SM: **BofC encodes a putative forespore regulator of the Bacillus subtilis sigma K checkpoint.** *Microbiology* 1997, **143 (Pt 1)**:157-170.
493. Wakeley P, Hoa NT, Cutting S: **BofC negatively regulates SpoIVB-mediated signalling in the Bacillus subtilis sigmaK-checkpoint.** *Mol Microbiol* 2000, **36**:1415-1424.
494. Koonin EV: **A highly conserved sequence motif defining the family of MutT-related proteins from eubacteria, eukaryotes and viruses.** *Nucleic Acids Res* 1993, **21**:4847.
495. Sivaraman J, Myers RS, Boju L, Sulea T, Cygler M, Jo Davison V, Schrag JD: **Crystal structure of Methanobacterium thermoautotrophicum phosphoribosyl-AMP cyclohydrolase HisI.** *Biochemistry* 2005, **44**:10071-10080.
496. D'Ordine RL, Klem TJ, Davison VJ: **N1-(5'-phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and characterization of a unique metalloenzyme.** *Biochemistry* 1999, **38**:1537-1546.
497. Lu M, Steitz TA: **Structure of Escherichia coli ribosomal protein L25 complexed with a 5S rRNA fragment at 1.8-A resolution.** *Proc Natl Acad Sci U S A* 2000, **97**:2023-2028.
498. Mahajan R, Delphin C, Guan T, Gerace L, Melchior F: **A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2.** *Cell* 1997, **88**:97-107.

499. Matunis MJ, Coutavas E, Blobel G: **A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex.** *J Cell Biol* 1996, **135**:1457-1470.
500. Grynberg M, Jaroszewski L, Godzik A: **Domain analysis of the tubulin cofactor system: a model for tubulin folding and dimerization.** *BMC Bioinformatics* 2003, **4**:46.
501. Walker EH, Perisic O, Ried C, Stephens L, Williams RL: **Structural insights into phosphoinositide 3-kinase catalysis and signalling.** *Nature* 1999, **402**:313-320.
502. Wattiaux-de Coninck S, Wattiaux R: **Subcellular distribution of sulfite cytochrome c reductase in rat liver tissue.** *Eur J Biochem* 1971, **19**:552-556.
503. Tsuge H, Kawakami R, Sakuraba H, Ago H, Miyano M, Aki K, Katunuma N, Ohshima T: **Crystal structure of a novel FAD-, FMN-, and ATP-containing L-proline dehydrogenase complex from *Pyrococcus horikoshii*.** *J Biol Chem* 2005, **280**:31045-31049.
504. Aravind L, Koonin EV: **Novel predicted RNA-binding domains associated with the translation machinery.** *J Mol Evol* 1999, **48**:291-302.
505. Pugh DJ, Ab E, Faro A, Luty PT, Hoffmann E, Rees DJ: **DWNN, a novel ubiquitin-like domain, implicates RBBP6 in mRNA processing and ubiquitin-like pathways.** *BMC Struct Biol* 2006, **6**:1.
506. Arenas JE, Abelson JN: **The *Saccharomyces cerevisiae* PRP21 gene product is an integral component of the prespliceosome.** *Proc Natl Acad Sci U S A* 1993, **90**:6771-6775.
507. Mueller EG: **Trafficking in persulfides: delivering sulfur in biosynthetic pathways.** *Nat Chem Biol* 2006, **2**:185-194.
508. Matthijs S, Baysse C, Koedam N, Tehrani KA, Verheyden L, Budzikiewicz H, Schafer M, Hoorelbeke B, Meyer JM, De Greve H, Cornelis P: **The *Pseudomonas siderophore* quinolobactin is synthesized from xanthurenic acid, an intermediate of the kynurenine pathway.** *Mol Microbiol* 2004, **52**:371-384.
509. Klemm P, Christiansen G: **The *fimD* gene required for cell surface localization of *Escherichia coli* type 1 fimbriae.** *Mol Gen Genet* 1990, **220**:334-338.
510. Saulino ET, Bullitt E, Hultgren SJ: **Snapshots of usher-mediated protein secretion and ordered pilus assembly.** *Proc Natl Acad Sci U S A* 2000, **97**:9240-9245.
511. Saulino ET, Thanassi DG, Pinkner JS, Hultgren SJ: **Ramifications of kinetic partitioning on usher-mediated pilus biogenesis.** *Embo J* 1998, **17**:2177-2185.

512. Papageorgiou AC, Tranter HS, Acharya KR: **Crystal structure of microbial superantigen staphylococcal enterotoxin B at 1.5 Å resolution: implications for superantigen recognition by MHC class II molecules and T-cell receptors.** *J Mol Biol* 1998, **277**:61-79.
513. Derrick JP, Wigley DB: **The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab.** *J Mol Biol* 1994, **243**:906-918.
514. Cavalier-Smith T: **The origin of eukaryotic and archaebacterial cells.** *Ann N Y Acad Sci* 1987, **503**:17-54.
515. Margulis L: *Symbiosis in Cell Evolution.* New York: WH Freeman; 1993.
516. Zillig W: **Comparative biochemistry of Archaea and Bacteria.** *Curr Opin Genet Dev* 1991, **1**:544-551.
517. Bruderer RM, Brasseur C, Meyer HH: **The AAA ATPase p97/VCP interacts with its alternative co-factors, Ufd1-Npl4 and p47, through a common bipartite binding mechanism.** *J Biol Chem* 2004, **279**:49609-49616.
518. Enari M, Sakahira H, Yokoyama H, Okawa K, Iwamatsu A, Nagata S: **A caspase-activated DNase that degrades DNA during apoptosis, and its inhibitor ICAD.** *Nature* 1998, **391**:43-50.
519. Halenbeck R, MacDonald H, Roulston A, Chen TT, Conroy L, Williams LT: **CPAN, a human nuclease regulated by the caspase-sensitive inhibitor DFF45.** *Curr Biol* 1998, **8**:537-540.
520. Liu X, Li P, Widlak P, Zou H, Luo X, Garrard WT, Wang X: **The 40-kDa subunit of DNA fragmentation factor induces DNA fragmentation and chromatin condensation during apoptosis.** *Proc Natl Acad Sci U S A* 1998, **95**:8461-8466.
521. Mukae N, Enari M, Sakahira H, Fukuda Y, Inazawa J, Toh H, Nagata S: **Molecular cloning and characterization of human caspase-activated DNase.** *Proc Natl Acad Sci U S A* 1998, **95**:9123-9128.
522. Anantharaman V, Iyer LM, Aravind L: **Comparative Genomics of Protists: New Insights on Evolution of Eukaryotic Signal Transduction and Gene Regulation.** *Annu Rev Microbiol* 2006.
523. Anantharaman V, Balaji S, Aravind L: **The signaling helix: a common functional theme in diverse signaling proteins.** *Biol Direct* 2006, **1**:25.
524. Anantharaman V, Koonin EV, Aravind L: **Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains.** *J Mol Biol* 2001, **307**:1271-1292.

525. Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UC, Besteiro S, et al: **Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis***. *Science* 2007, **315**:207-212.
526. Bays NW, Wilhovsky SK, Goradia A, Hodgkiss-Harlow K, Hampton RY: **HRD4/NPL4 is required for the proteasomal processing of ubiquitinated ER proteins**. *Mol Biol Cell* 2001, **12**:4114-4128.
527. Kim JE, Kim SJ, Lee BH, Park RW, Kim KS, Kim IS: **Identification of motifs for cell adhesion within the repeated domains of transforming growth factor-beta-induced gene, betaig-h3**. *J Biol Chem* 2000, **275**:30907-30915.
528. Anantharaman V, Aravind L: **The PRC-barrel: a widespread, conserved domain shared by photosynthetic reaction center subunits and proteins of RNA metabolism**. *Genome Biol* 2002, **3**:RESEARCH0061.
529. Arcus V: **OB-fold domains: a snapshot of the evolution of sequence, structure and function**. *Curr Opin Struct Biol* 2002, **12**:794-801.
530. Murzin AG: **OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences**. *Embo J* 1993, **12**:861-867.
531. Muller JJ, Muller A, Rottmann M, Bernhardt R, Heinemann U: **Vertebrate-type and plant-type ferredoxins: crystal structure comparison and electron transfer pathway modelling**. *J Mol Biol* 1999, **294**:501-513.
532. Truglio JJ, Theis K, Leimkuhler S, Rappa R, Rajagopalan KV, Kisker C: **Crystal structures of the active and alloxanthine-inhibited forms of xanthine dehydrogenase from *Rhodobacter capsulatus***. *Structure* 2002, **10**:115-125.
533. Yankovskaya V, Horsefield R, Tornroth S, Luna-Chavez C, Miyoshi H, Leger C, Byrne B, Cecchini G, Iwata S: **Architecture of succinate dehydrogenase and reactive oxygen species generation**. *Science* 2003, **299**:700-704.
534. Huang L, Hofer F, Martin GS, Kim SH: **Structural basis for the interaction of Ras with RaIGDS**. *Nat Struct Biol* 1998, **5**:422-426.
535. Stebbins CE, Kaelin WG, Jr., Pavletich NP: **Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function**. *Science* 1999, **284**:455-461.
536. Mildvan AS, Xia Z, Azurmendi HF, Saraswat V, Legler PM, Massiah MA, Gabelli SB, Bianchet MA, Kang LW, Amzel LM: **Structures and mechanisms of Nudix hydrolases**. *Arch Biochem Biophys* 2005, **433**:129-143.
537. Gulbis JM, Zhou M, Mann S, MacKinnon R: **Structure of the cytoplasmic beta subunit-T1 assembly of voltage-dependent K⁺ channels**. *Science* 2000, **289**:123-127.

538. Palenchar PM, Buck CJ, Cheng H, Larson TJ, Mueller EG: **Evidence that ThiI, an enzyme shared between thiamin and 4-thiouridine biosynthesis, may be a sulfurtransferase that proceeds through a persulfide intermediate.** *J Biol Chem* 2000, **275**:8283-8286.
539. Burroughs AM, Balaji S, Iyer LM, Aravind L: **A novel superfamily containing the beta-grasp fold involved in binding diverse soluble ligands.** *Biol Direct* 2007, **2**:4.
540. Murzin AG: **Familiar strangers.** *Nature* 1992, **360**:635.
541. Hershko A, Ciechanover A: **The ubiquitin system.** *Annu Rev Biochem* 1998, **67**:425-479.
542. Chishti AH, Kim AC, Marfatia SM, Lutchman M, Hanspal M, Jindal H, Liu SC, Low PS, Rouleau GA, Mohandas N, et al: **The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane.** *Trends Biochem Sci* 1998, **23**:281-282.
543. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**:D247-251.
544. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.** *Nucleic Acids Res* 2001, **29**:2994-3005.
545. Walker DR, Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5**:333-339.
546. Moestrup SK: **New insights into carrier binding and epithelial uptake of the erythropoietic nutrients cobalamin and folate.** *Curr Opin Hematol* 2006, **13**:119-123.
547. Schmehl M, Jahn A, Meyer zu Vilsendorf A, Hennecke S, Masepohl B, Schuppler M, Marxer M, Oelze J, Klipp W: **Identification of a new class of nitrogen fixation genes in *Rhodobacter capsulatus*: a putative membrane complex involved in electron transport to nitrogenase.** *Mol Gen Genet* 1993, **241**:602-615.
548. Mossessova E, Lima CD: **Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast.** *Mol Cell* 2000, **5**:865-876.
549. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context.** *Genome Res* 2001, **11**:356-372.
550. Fischetti VA, Pancholi V, Schneewind O: **Conservation of a hexapeptide sequence in the anchor region of surface proteins from gram-positive cocci.** *Mol Microbiol* 1990, **4**:1603-1605.

551. Williams RJ, Henderson B, Sharp LJ, Nair SP: **Identification of a fibronectin-binding protein from *Staphylococcus epidermidis*.** *Infect Immun* 2002, **70**:6805-6810.
552. Aravind L, Anantharaman V, Iyer LM: **Evolutionary connections between bacterial and eukaryotic signaling systems: a genomic perspective.** *Curr Opin Microbiol* 2003, **6**:490-497.
553. Hofmann BE, Bender H, Schulz GE: **Three-dimensional structure of cyclodextrin glycosyltransferase from *Bacillus circulans* at 3.4 Å resolution.** *J Mol Biol* 1989, **209**:793-800.
554. Ulstrup JC, Jeansson S, Wiker HG, Harboe M: **Relationship of secretion pattern and MPB70 homology with osteoblast-specific factor 2 to osteitis following *Mycobacterium bovis* BCG vaccination.** *Infect Immun* 1995, **63**:672-675.
555. Rodionov DA, Hebbeln P, Gelfand MS, Eitinger T: **Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters.** *J Bacteriol* 2006, **188**:317-327.
556. McNulty C, Thompson J, Barrett B, Lord L, Andersen C, Roberts IS: **The cell surface expression of group 2 capsular polysaccharides in *Escherichia coli*: the role of KpsD, RhsA and a multi-protein complex at the pole of the cell.** *Mol Microbiol* 2006, **59**:907-922.
557. Inamine GS, Dubnau D: **ComEA, a *Bacillus subtilis* integral membrane protein required for genetic transformation, is needed for both DNA binding and transport.** *J Bacteriol* 1995, **177**:3045-3051.
558. Provvedi R, Dubnau D: **ComEA is a DNA receptor for transformation of competent *Bacillus subtilis*.** *Mol Microbiol* 1999, **31**:271-280.
559. Sampson EM, Johnson CL, Bobik TA: **Biochemical evidence that the pduS gene encodes a bifunctional cobalamin reductase.** *Microbiology* 2005, **151**:1169-1177.
560. Bobik TA, Havemann GD, Busch RJ, Williams DS, Aldrich HC: **The propanediol utilization (pdu) operon of *Salmonella enterica* serovar Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B(12)-dependent 1, 2-propanediol degradation.** *J Bacteriol* 1999, **181**:5967-5975.
561. Perham RN: **Swinging arms and swinging domains in multifunctional enzymes: catalytic machines for multistep reactions.** *Annu Rev Biochem* 2000, **69**:961-1004.
562. Yamanishi M, Ide H, Murakami Y, Toraya T: **Identification of the 1,2-propanediol-1-yl radical as an intermediate in adenosylcobalamin-dependent diol dehydratase reaction.** *Biochemistry* 2005, **44**:2113-2118.

563. Leonard PM, Smits SH, Sedelnikova SE, Brinkman AB, de Vos WM, van der Oost J, Rice DW, Rafferty JB: **Crystal structure of the Lrp-like transcriptional regulator from the archaeon *Pyrococcus furiosus***. *Embo J* 2001, **20**:990-997.
564. Chipman DM, Shaanan B: **The ACT domain family**. *Curr Opin Struct Biol* 2001, **11**:694-700.
565. Sticht H, Rosch P: **The structure of iron-sulfur proteins**. *Prog Biophys Mol Biol* 1998, **70**:95-136.
566. Dong C, Beis K, Nesper J, Brunkan-Lamontagne AL, Clarke BR, Whitfield C, Naismith JH: **Wza the translocon for *E. coli* capsular polysaccharides defines a new class of membrane protein**. *Nature* 2006, **444**:226-229.
567. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: *Molecular Biology of the Cell*. 4th Bk&Cdr edition edn. New York, NY: Garland Science Publishing; 2002.
568. Ciechanover A, Orian A, Schwartz AL: **Ubiquitin-mediated proteolysis: biological regulation via destruction**. *Bioessays* 2000, **22**:442-451.
569. Wertz IE, O'Rourke KM, Zhou H, Eby M, Aravind L, Seshagiri S, Wu P, Wiesmann C, Baker R, Boone DL, et al: **De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling**. *Nature* 2004, **430**:694-699.
570. Pickart CM: **Mechanisms underlying ubiquitination**. *Annu Rev Biochem* 2001, **70**:503-533.
571. Weissman AM: **Themes and variations on ubiquitylation**. *Nat Rev Mol Cell Biol* 2001, **2**:169-178.
572. Hochstrasser M: **Biochemistry. All in the ubiquitin family**. *Science* 2000, **289**:563-564.
573. Aravind L, Ponting CP: **Homologues of 26S proteasome subunits are regulators of transcription and translation**. *Protein Sci* 1998, **7**:1250-1254.
574. Hofmann K, Bucher P: **The PCI domain: a common theme in three multiprotein complexes**. *Trends Biochem Sci* 1998, **23**:204-205.
575. Anantharaman V, Aravind L: **Evolutionary history, structural features and biochemical diversity of the NlpC/P60 superfamily of enzymes**. *Genome Biol* 2003, **4**:R11. Epub 2003 Feb 2003.
576. Anantharaman V, Koonin EV, Aravind L: **Peptide-N-glycanases and DNA repair proteins, Xp-C/Rad4, are, respectively, active and inactivated enzymes sharing a common transglutaminase fold**. *Hum Mol Genet* 2001, **10**:1627-1630.

577. Makarova KS, Aravind L, Koonin EV: **A superfamily of archaeal, bacterial, and eukaryotic proteins homologous to animal transglutaminases.** *Protein Sci* 1999, **8**:1714-1719.
578. Makarova KS, Aravind L, Koonin EV: **A novel superfamily of predicted cysteine proteases from eukaryotes, viruses and Chlamydia pneumoniae.** *Trends Biochem Sci* 2000, **25**:50-52.
579. Guterman A, Glickman MH: **Deubiquitinating enzymes are IN/(trinsic to proteasome function).** *Curr Protein Pept Sci* 2004, **5**:201-211.
580. Soboleva TA, Baker RT: **Deubiquitinating enzymes: their functions and substrate specificity.** *Curr Protein Pept Sci* 2004, **5**:191-200.
581. Wing SS: **Deubiquitinating enzymes--the importance of driving in reverse along the ubiquitin-proteasome pathway.** *Int J Biochem Cell Biol* 2003, **35**:590-605.
582. Cope GA, Suh GS, Aravind L, Schwarz SE, Zipursky SL, Koonin EV, Deshaies RJ: **Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1.** *Science* 2002, **298**:608-611. Epub 2002 Aug 2015.
583. Verma R, Aravind L, Oania R, McDonald WH, Yates JR, 3rd, Koonin EV, Deshaies RJ: **Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome.** *Science* 2002, **298**:611-615. Epub 2002 Aug 2015.
584. Goehring AS, Rivers DM, Sprague GF, Jr.: **Attachment of the ubiquitin-related protein Urm1p to the antioxidant protein Ahp1p.** *Eukaryot Cell* 2003, **2**:930-936.
585. Singh S, Tonelli M, Tyler RC, Bahrami A, Lee MS, Markley JL: **Three-dimensional structure of the AAH26994.1 protein from Mus musculus, a putative eukaryotic Urm1.** *Protein Sci* 2005, **14**:2095-2102.
586. Hofmann K, Bucher P: **The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway.** *Trends Biochem Sci* 1996, **21**:172-173.
587. Hofmann K, Falquet L: **A ubiquitin-interacting motif conserved in components of the proteasomal and lysosomal protein degradation systems.** *Trends Biochem Sci* 2001, **26**:347-350.
588. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV: **Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9**:608-628.
589. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, et al: **Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis.** *Science* 1998, **282**:754-759.

590. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4**:1618-1632.
591. Schuler GD, Altschul SF, Lipman DJ: **A workbench for multiple alignment construction and analysis.** *Proteins* 1991, **9**:180-190.
592. Hasegawa M, Kishino H, Saitou N: **On the maximum likelihood method in molecular phylogenetics.** *J Mol Evol* 1991, **32**:443-445.
593. Vriend G, Sander C: **Detection of common three-dimensional substructures in proteins.** *Proteins* 1991, **11**:52-58.
594. Aravind L, Koonin EV: **A natural classification of ribonucleases.** *Methods Enzymol* 2001, **341**:3-28.
595. Anantharaman V, Aravind L: **The NYN domains: Novel predicted RNAses with a PIN Domain-like fold.** *RNA Biology* 2006, *In Press*.
596. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS: **Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms.** *J Biol Chem* 2002, **277**:48949-48959.
597. Settembre EC, Dorrestein PC, Zhai H, Chatterjee A, McLafferty FW, Begley TP, Ealick SE: **Thiamin biosynthesis in *Bacillus subtilis*: structure of the thiazole synthase/sulfur carrier protein complex.** *Biochemistry* 2004, **43**:11647-11657.
598. Schwarz G, Mendel RR: **Molybdenum Cofactor Biosynthesis and Molybdenum Enzymes.** *Annu Rev Plant Biol* 2006.
599. Schwarz G: **Molybdenum cofactor biosynthesis and deficiency.** *Cell Mol Life Sci* 2005, **62**:2792-2810.
600. Anantharaman V, Aravind L: **MOSC domains: ancient, predicted sulfur-carrier domains, present in diverse metal-sulfur cluster biosynthesis proteins including Molybdenum cofactor sulfurases.** *FEMS Microbiol Lett* 2002, **207**:55-61.
601. Rajagopalan KV: **Biosynthesis and processing of the molybdenum cofactors.** *Biochem Soc Trans* 1997, **25**:757-761.
602. Johnson JL, Rajagopalan KV, Mukund S, Adams MW: **Identification of molybdopterin as the organic component of the tungsten cofactor in four enzymes from hyperthermophilic Archaea.** *J Biol Chem* 1993, **268**:4848-4852.
603. Cornelis P, Matthijs S: **Diversity of siderophore-mediated iron uptake systems in fluorescent pseudomonads: not only pyoverdines.** *Environ Microbiol* 2002, **4**:787-798.

604. Koonin EV, Aravind L, Galperin MY: **A comparative-genomic view of the microbial stress response.** In *Bacterial stress response*. Edited by Storz G, Hengge-Aronis R. Washington DC: ASM Press; 2000
605. Wietzorrek A, Schwarz H, Herrmann C, Braun V: **The genome of the novel phage Rtp, with a rosette-like tail tip, is homologous to the genome of phage T1.** *J Bacteriol* 2006, **188**:1419-1436.
606. Nameki N, Yoneyama M, Koshiha S, Tochio N, Inoue M, Seki E, Matsuda T, Tomo Y, Harada T, Saito K, et al: **Solution structure of the RWD domain of the mouse GCN2 protein.** *Protein Sci* 2004, **13**:2089-2100.
607. Schubert S, Dufke S, Sorsa J, Heesemann J: **A novel integrative and conjugative element (ICE) of Escherichia coli: the putative progenitor of the Yersinia high-pathogenicity island.** *Mol Microbiol* 2004, **51**:837-848.
608. Aravind L, Koonin EV: **DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history.** *Nucleic Acids Res* 1999, **27**:1609-1618.
609. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P: **Systematic identification of novel protein domain families associated with nuclear functions.** *Genome Res* 2002, **12**:47-56.
610. Karzai AW, Roche ED, Sauer RT: **The SsrA-SmpB system for protein tagging, directed degradation and ribosome rescue.** *Nat Struct Biol* 2000, **7**:449-455.
611. Jouanneau Y, Jeong HS, Hugo N, Meyer C, Willison JC: **Overexpression in Escherichia coli of the rnf genes from Rhodobacter capsulatus--characterization of two membrane-bound iron-sulfur proteins.** *Eur J Biochem* 1998, **251**:54-64.
612. Rhee SG, Park SC, Koo JH: **The role of adenylyltransferase and uridylyltransferase in the regulation of glutamine synthetase in Escherichia coli.** *Curr Top Cell Regul* 1985, **27**:221-232.
613. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.** *Biol Direct* 2006, **1**:7.
614. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes.** *PLoS Comput Biol* 2005, **1**:e60.
615. Anantharaman V, Aravind L: **New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system.** *Genome Biol* 2003, **4**:R81.

616. Roberts RJ, Vincze T, Posfai J, Macelis D: **REBASE--restriction enzymes and DNA methyltransferases.** *Nucleic Acids Res* 2005, **33**:D230-232.
617. Lupas AN, Koretke KK: **Bioinformatic analysis of ClpS, a protein module involved in prokaryotic and eukaryotic protein degradation.** *J Struct Biol* 2003, **141**:77-83.
618. Erbse A, Schmidt R, Bornemann T, Schneider-Mergener J, Mogk A, Zahn R, Dougan DA, Bukau B: **ClpS is an essential component of the N-end rule pathway in Escherichia coli.** *Nature* 2006, **439**:753-756.
619. Darwin C: *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: John Murray; 1859.
620. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307**:1113-1143.
621. Aravind L, Mazumder R, Vasudevan S, Koonin EV: **Trends in protein evolution inferred from sequence and structure analysis.** *Curr Opin Struct Biol* 2002, **12**:392-399.
622. Crick FH: **The origin of the genetic code.** *J Mol Biol* 1968, **38**:367-379.
623. Service RF: **Gene sequencing. The race for the \$1000 genome.** *Science* 2006, **311**:1544-1546.

CURRICULUM VITAE

Alexander Maxwell Burroughs
October, 2007

CONTACT INFORMATION

School: Boston University
Bioinformatics Department
44 Cummington St
Boston, MA 02215
e-mail aburroug@bu.edu

Research Lab: Protein and Genome Evolution Research Group
Computational Biology Branch
National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
8600 Rockville Pike, Building 38A, 8N811M
Bethesda, MD 20894
voice: (301) 496-9313
fax: (301) 480-4637
email burrough@ncbi.nlm.nih.gov

EDUCATION

Boston University, Bioinformatics program.

PhD candidate. September 2003-present. Projected graduation date: December 2007.

Brigham Young University-Hawaii.

B.S. Summa Cum Laude 2003. Biochemistry major. Mathematics, Chemistry Minor.

EXPERIENCE

Research

National Center for Biotechnology Information, National Library of Medicine, National Institute of Health. Dr. Aravind Iyer, advisor.

Research for Ph.D. dissertation, 09/04-present. Comparative evolutionary genomics at the fold and superfamily level of protein domain organization.

Center for Bioinformatics and Computational Biology, University of Maryland, Dr. Steven Salzberg, Director.

Member of *C. papaya* genome assembly and analysis project, 02/07-present.

Cellular and Plasticity Section, Neurobiology and Development Group, National Institute of Drug Abuse, National Institute of Health, Dr. Elin Lehmman.

Summer Research Assistant, 2004. Design and implementation of a curated database housing gene expression patterns across different forms of drug abuse.

Boston University Medical Center, Department of Pulmonary Medicine, Dr. Avi Spira and Dr. Jerome Brody, advisors.

Research assistant, 09/03-01/04. Gene expression analysis of COPD development between smokers, non-smokers, and former smokers.

Brigham Young University-Hawaii, Laie, Hawaii, Department of Biochemistry, Dr. Darren Heaton, chair of Biochemistry Department.

Research assistant, 01/02-06/03. Characterization of copper-ligand binding properties in Cox17 proteins.

Stanford Linear Accelerator Center (SLAC), Stanford Synchrotron Radiation Laboratory (SSRL), Dr. Graham George, advisor.

Summer internship, 2002. Research in characterizing iron-sulfur centers of Biotin Synthase enzyme using EXAFS-acquired data.

Teaching

Brigham Young University-Hawaii, Laie, Hawaii, Sciences Division, Dr. Gary Frederick, Chair of Sciences Division, Teacher.

Teaching Assistant for Organic Chemistry Course, 2001-2002 academic year.

RESEARCH SKILLS/SPECIAL SKILLS

General Research Skills

1) Computational Biology

- Extensive familiarity with comparative sequence/structure analysis techniques.
- Experience with large-scale genomic data network analysis.
- Experience with gene microarray preparation and statistical analysis.
- Implementation and working knowledge of theory behind various phylogenetic analyses including distance, maximum-likelihood, and Bayesian-based techniques.

2) Computers

- Operating systems: Unix and Windows
- Database management systems: Microsoft SQL, MySQL
- Languages: Perl, SQL

3) Molecular biology

- DNA/RNA extraction, cloning, and expression
- Protein isolation and characterization
- EXAFS experimental preparation and analysis.

Language Skills

Fluent in English, conversational Japanese.

HONORS AND AWARDS

1) Recipient of the inaugural Cecile M. Pickart Student Travel Award for outstanding research by a student in Ubiquitin-related research, The Ubiquitin Family: Cold Spring Harbor Laboratory Meeting, April 2007.

2) Predoctoral Intramural Research Training Award (Pre-IRTA), National Institutes of Health. 2003-present.

3) AAAS/Science Program for Excellence in Science, Boston University Bioinformatics Program Nominee. 2006.

4) Selected for participation in Boston University's Graduate Partnerships Program with the National Institutes of Health, 2003.

5) Salutatorian, Brigham Young University-Hawaii, Class of 2003.

6) David O. McKay Scholar, Brigham Young University-Hawaii, 1998, 2001-2003.

7) Phi Kappa Phi Outstanding Student in Biochemistry, Class of 2003.

- 8) American Chemical Society Hawaii Chapter Most Outstanding Research Associate and Analytical Chemistry Student, BYU-Hawaii, 2003.
- 9) American Chemical Society Hawaii Chapter Most Outstanding Biochemistry Student, BYU-Hawaii, 2002.
- 10) Phi Kappa Phi Inductee, 2001.

PROFESSIONAL ACTIVITIES

Journal Publications

- Burroughs AM, Iyer LM, L Aravind. Evolutionary history of the E1-like fold and architectural themes contributing to the catalytic roles of the E1-like protein. (in preparation).
- Burroughs AM, Jaffee M, Iyer LM, L Aravind. Elucidating the mechanism of E2 protein conjugation: remarkable variation of active site residues in a conserved structural scaffold. (*Journal of Structural Biology*, in review).
- Salzberg SL, Schatz M, Nagarajan N, Delcher AL, Burroughs AM [**only members of University of Maryland group shown in published author order]. Genome of the transgenic tropical fruit tree papaya (*Carica papaya* L.). (*Nature*, in review).
- Burroughs AM, Iyer LM, L Aravind. Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems. Volff J-N (editor): *Gene and Protein Evolution*, Genome Dynamics series. Basel, Karger, 2007, vol. 3, pp 48-65.
- Peisach E, Wang L, Burroughs AM, L Aravind, Dunaway-Mariano D, Allen KN. The X-ray crystallographic structure and activity analysis of a Pseudomonas-specific subfamily of the HAD enzyme superfamily evidences a novel biochemical function. *Proteins*. 2007 Jul 24; Epub.
- Burroughs AM, Balaji S, Iyer LM, L Aravind. Small but versatile: the extraordinary functional and structural diversity of the β -grasp fold. *Biology Direct*. 2007 Jul 2;2(1):18.
- Burroughs AM, Balaji S, Iyer LM, L Aravind. A novel superfamily containing the β -grasp fold involved in binding diverse soluble ligands. *Biology Direct*. 2007 Jan 24;2:4.
- Burroughs AM, Allen KN, Dunaway-Mariano D, and L Aravind. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *Journal of Molecular Biology*. 2006 Sep 1;361(5):1003-34.
- Burroughs AM*, Iyer LM*, L Aravind. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like β -grasp domains. *Genome Biology*. 2006 Jul 19;7(7):R60.
- Iyer LM, Burroughs AM, L Aravind. The ASCH superfamily: novel domains with a fold related to the PUA domain and a potential role in RNA metabolism. *Bioinformatics*. 2006 Feb 1;22(3):257-63.
- Burroughs AM. The Medical Examination in United States Immigration Applications: The Potential Use of Genetic Testing Leads to Heightened Privacy Concerns. *Journal of Biolaw and Business*. 2005;8(4):22-32

Invited Talks

- *Evolutionary adaptations and the discovery of a novel ligand-binding superfamily in the β -grasp fold*. Computational Biology Branch Seminar, NCBI/NLM/NIH, Oct. 2, 2007.

-*The origin of ubiquitin signaling pathways: comparative genomics uncovers evidence of prokaryotic homologs.* NIH Lambda Lunch special interest group hosted by Dr. Susan Gottesman, September 13, 2007.

-*Genomic analysis uncovers novel prokaryotic ubiquitin-signaling systems.* The Ubiquitin Family: Cold Spring Harbor Laboratory Meeting, April 2007.

-*Bioinformatics and computational techniques.* Guest lecture for Foundation for Advanced Education in the Sciences (FAES) Introduction to Laboratory Techniques course, October 2006.

-*The evolutionary history of the ubiquitin signaling system and the discovery of a novel superfamily of β -grasp-like proteins.* NIH Graduate Student Seminar Series, November 2006.

-*Cloning mutant proteins.* Annual Undergraduate Research Seminar, BYU-Hawaii campus, May 2003.

Poster Sessions

-*A novel superfamily binding diverse soluble ligands.* NIH Graduate Student Research Symposium, May 2007.

-*Prokaryotic antecedents of the ubiquitin signaling system and the early evolution of ubiquitin-like β -grasp domains.* NIH Research Festival, October 2006.

-*Prokaryotic origins of the ubiquitin-signaling system.* The Sixth International Workshop on Bioinformatics and Systems Biology, Boston, August 2006.

Other Professional Activities

-Referee for Gene journal, 2007.

-Grant review panel member: Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, Alaska Native-Serving and Native Hawaiian-Serving Institutions Education Grants Program, 2006.

REFERENCES

Dr. Aravind Iyer (L Aravind), Ph.D., Investigator
Protein and Genome Evolution Research Group
Computational Biology Branch
National Center for Biotechnology Information, National Library of Medicine, NIH
Bethesda, MD 20894, USA
+1 (301) 594-2445
aravind@ncbi.nlm.nih.gov

Dr. David Landsman, Ph.D., Senior Investigator
Chief, Computational Biology Branch
National Center for Biotechnology Information, National Library of Medicine, NIH
Bethesda, MD 20894, USA
+1 (301) 435-5981
landsman@ncbi.nlm.nih.gov

Dr. Karen Allen, Ph.D., Professor
Department of Physiology and Biophysics
Boston University School of Medicine
715 Albany Street
Boston, MA 02118, USA

+1 (617) 638-4398

allen@med-xtal.bu.edu

Dr. M. Madan Babu, PhD, Schlumberger Research Fellow at Darwin College

Group Leader/Investigator

MRC-Laboratory of Molecular Biology

Hills Road, Cambridge CB2 2QH, United Kingdom

University of Cambridge

+44 - (0)1223 - 402041

madanm@mrc-lmb.cam.ac.uk