AAA Proteins and the Origins of Proteasomal Protein Degradation

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Moritz Ammelburg
aus Marburg

Tübingen
2011

Die vorliegende Arbeit wurde unter der Leitung von Herrn Prof. Dr. Andrei Lupas im Zeitraum von Oktober 2006 bis Februar 2011 in der Abteilung Protein Evolution am Max-Planck-Institut für Entwicklungsbiologie in Tübingen angefertigt.

Die Betreuung dieser Arbeit in der Mathematisch-Naturwissenschaftlichen Fakulät der Eberhard-Karls-Universitaet Tübingen wurde von Herrn Prof. Dr. Thilo Stehle übernommen.

Tag der mündlichen Qualifikation:	27.10.2011
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Andrei Lupas
2. Berichterstatter:	Prof. Dr. Thilo Stehle

DANKSAGUNG

Ich danke Andrei Lupas für die Möglichkeit meine Doktorarbeit in seinem Labor durchführen zu können. Ich danke ihm für seine großzügige Unterstützung sowie für seine äußerst lehrreiche und inspirierende wissenschaftliche Betreuung.

Ich danke Thilo Stehle für die Vertretung meiner Arbeit in der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Tübingen. Darüberhinaus danke ich den Professoren Gabriele Dodt, und Volkmar Braun, die gemeinsam mit den Professoren Thilo Stehle und Andrei Lupas an der Disputation dieser Arbeit teilgenmmen haben.

Bei Johannes Schiff, Bijan Mir-Montazeri, Cedric Hobel und ganz besonders Dara Forouzan möchte ich mich für eine freundschaftliche Zusammenarbeit bedanken, die für mich in vielerlei Hinsicht eine Bereicherung war.

Marcus Hartmann danke ich für die freundschaftliche Zusammenarbeit in den kristallographischen Aspekten dieser Arbeit, die von großer Wichtigkeit für ihren Fortschritt waren.

Ich danke Jörg Martin für zahlreiche Ratschläge in biochemischen Belangen dieser Arbeit, Heinz Schwarz für seine ständige Bereitschaft Proben elektronenmikroskopisch zu untersuchen, Guido Sauer für massenspektrometrische Analysen, Reinhard Albrecht und Kerstin Bär für Kristallisationsansätze. Dirk Linke, Vikram Alva und Volkmar Braun danke ich für hilfreiche Diskussionen. Besonders möchte ich mich bei Murray Coles für eine lehrreiche Zusammenarbeit in Teilprojekten dieser Arbeit bedanken.

Darüberhinaus bedanke ich mich bei den Kollegen der Abteilung Protein Evolution für eine angenehme Atmosphäre und ständige Hilfsbereitschaft: Karin Lehmann, Michael Hulko, Felipe Figueroa, Silvia Deiss, Avijit Pramanik, Ivan Kalev, Silvia Würtenberger, Michael Habeck, Thomas Arnold, Astrid Ursinus, Martin Schueckel, Martin Mechelke, Evgenia Afanasieva, Iwan Grin, Suat Özdirekcan, Kiri und Keanu von Klier, Johannes Söding, Stephanie Helbig, Christin Römer, Indronil Chaudhuri, Carolin Ewers, Markus Gruber, Birte Höcker, Birte Hernandez-Alvarez, Hedda Ferris, Silke Patzer, Vasuki Chellamuthu, Franka Scharfenberg und Nagarajan Paramasivam.

Für meine Familie

„Man kann so ein verwickeltes Gebilde (...) von vielen Seiten ansehn und im theoretischen Bild das oder jenes als Achse wählen; es entstehen Teilwahrheiten, aus deren gegenseitiger Durchdringung langsam die Wahrheit höher wächst: Wächst sie aber wirklich höher? Es hat sich noch jedes Mal gerächt, wenn man eine Teilwahrheit für das allein Gültige angesehen hat. Andererseits wäre man aber kaum zu dieser Teilwahrheit gelangt, hätte man sie nicht überschätzt."

Robert Musil, 1930: *Der Mann ohne Eigenschaften*, Zweites Buch, S. 1020

SUMMARY

AAA (+) proteins are ATPases associated with diverse cellular activities coupling ATP-hydrolysis to remodelling, disaggregation, and unfolding of a variety of substrates. The central ATPase domain functions as a molecular switch, which receives input from N-terminal substrate recognition domains, and which transfers the output to downstream effectors. AAA+ proteases recognize misfolded proteins with their N-domain, unfold and thread them through the pore of the hexameric ring, and feed them to effector proteases, either residing on the same polypeptide chain, or being contacted via (C-terminal) interaction motifs.

We have investigated the divergent evolution of N-domains of known and putative proteasomal ATPases and their C-terminal interaction motif, which is crucial for the regulation of the proteasome, a self-compartmentalizing protease involved in the degradation of unfolded substrates within a large cylinder-shaped architecture.

The first part comprises three case studies of hypothetical proteins, homologous to double-ψ barrel, β-clam and OB-fold N-domains of AAA proteins. We present the first characterization of a CTP-specific archaeal riboflavin kinase, which is homologous to the double-ψ barrel of AAA proteins of the CDC48 group, sharing a duplicated ββαβ-element in their common core. We show that archaeal riboflavin kinases provide an evolutionary bridge between highly symmetric RIFT-barrel transcription factors and ATP-specific bacterial/eukaryotic riboflavin kinases allowing us to describe an evolutionary trajectory from DNA-binding to enzymatic activity.

A β-clam domain, which is found in AAA proteins of the CDC48 and AMA groups, was detected in context of a C-terminal domain lacking significant similarity to known domains. We present the full-length structure of a member of this family, whose C-terminal domain forms a homohexameric twelve-bladed β-propeller (HP12). Each monomer accommodates two propeller-blades that have retained traces of a duplication event, suggesting that monomeric β-propellers evolved via oligomeric intermediates. We show that HP12 forms a ternary complex with a genetically coupled endonuclease III and DNA implying a function in base-excision DNA-repair.

We identified a protein family in methanogenic archaea that contains an OB-fold, similar to the N-domain of proteasome activating nucleotidases (PAN), and a proteasome-like Ntn-hydrolase domain. Our crystal structure of a member of this family reveals a monomeric proteasome-homolog of methanogens (MPM), which acquired an

OB-fold domain, probably functioning as a substrate recognition domain for the protease. The internal symmetry of the six-stranded OB-fold suggests that it evolved by duplication of an ancestral β-meander.

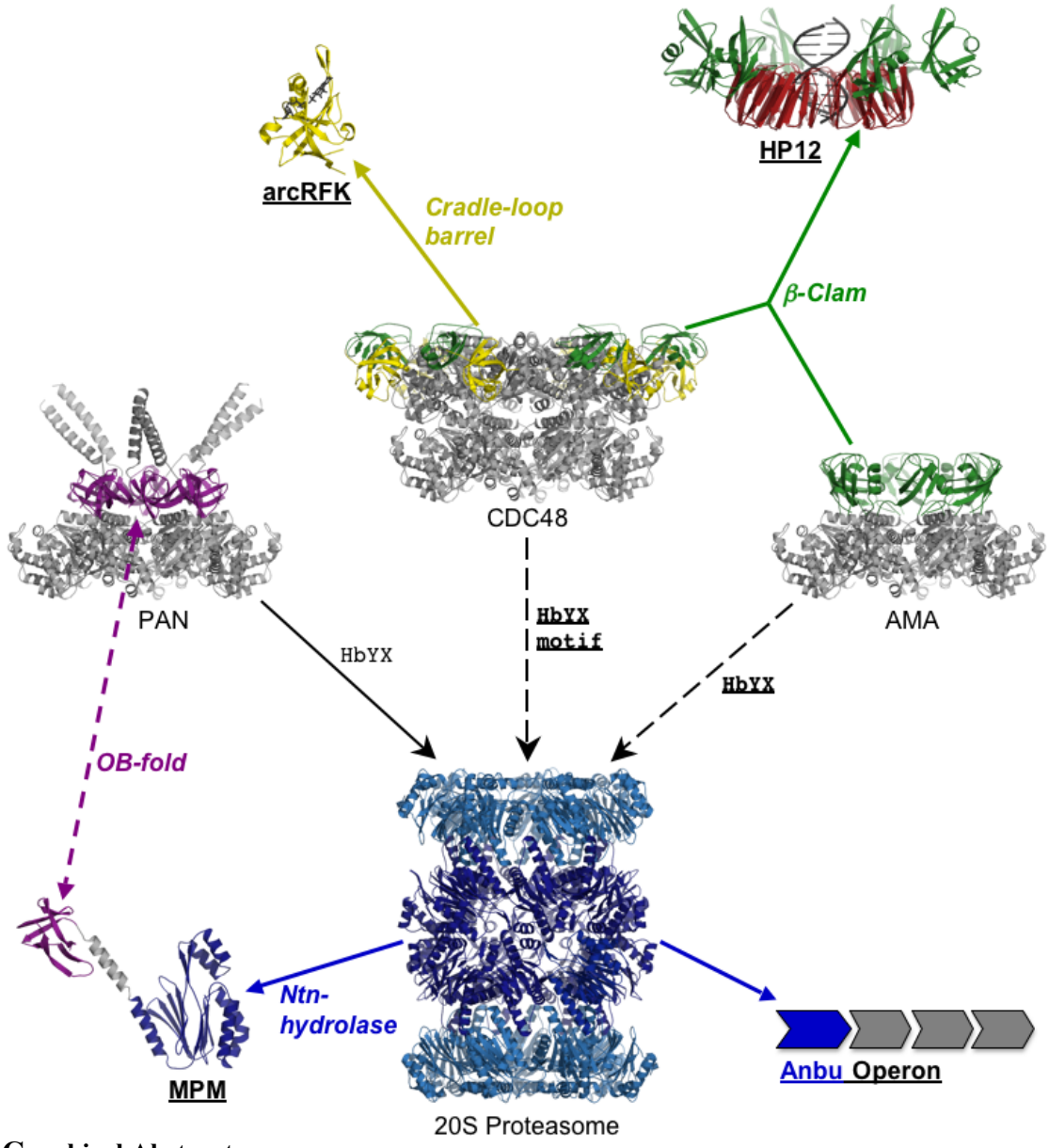
The ββαβ-element of double-ψ barrels and riboflavin kinases, the propeller blade of HP12, and the three-stranded β-meander of the OB-fold of MPM shed light on the evolution of autonomously folding domains through duplication and fusion of ancestral supersecondary structure elements, presumably via oligomeric intermediates.

In the second part, we trace the origins of proteasomal protein degradation. We present a systematic sequence analysis of the C-termini of archaeal AAA proteins uncovering the presence of the proteasome-interaction motif in AAA proteins of the CDC48 and AMA group in addition to known proteasome activating nucleotidases of the PAN group. Furthermore, we detect the absence of PAN proteins in major archaeal lineages supporting our hypothesis that kingdom-wide conserved CDC48 proteins function as regulatory ATPases of the proteasome. The presence of up to five putative proteasomal ATPases in certain archaea prompted us to predict a network of AAA ATPases, which regulates the archaeal proteasome. This network could increase the capabilities of proteasomal protein degradation in archaea through the participation of different N-terminal substrate recognition domains.

Analysis of the genetic context of the yet uncharacterized Anbu proteasome homolog revealed a conserved operon structure, widespread in proteobacteria and cyanobacteria. The components of the operon point to a peptide tagging system, remotely resembling ubiquitylation and sampylation, which target substrates for degradation by the proteasome in eukaryotes and archaea. Experimental evidence for this hypothesis as well as for the network hypothesis is pending.

Finally, we describe the global distribution of proteasome-like Ntn-hydrolases and putative proteasomal ATPases on the tree of life. This analysis supports the scenario that actinobacteria, the only bacterial taxon containing a proteasome, acquired it through lateral gene transfer. Among the large assemblies of proteasome-like Ntn-hydrolases the highly divergent monomeric proteasome-homolog of methanogens (MPM) is an exception, because this derived group has lost its ability to self-compartmentalize. In contrast to proteasome-ATPase complexes MPM including the OB-fold has evolved from an intricate towards a more simplified molecular phenotype.

GRAPHICAL ABSTRACT



Graphical Abstract

Underscored text denotes projects conducted in this work. **arcRFK**: Characterization of a CTP-specific **arc**haeal **R**ibo**F**lavin **K**inase, which is homologous to the double-ψ β-barrel domain (yellow) of AAA proteins like CDC48 (AAA+ modules in grey). **HP12**: Characterization of a **H**omohexameric β-**P**ropeller with **12** blades (red) involved in DNA-repair in pyrococci (model with DNA tentatively placed in the central pore); its N-domain is a β-clam (green) related to the N-domain of AAA proteins like AMA and CDC48. **MPM**: Characterization of a **M**onomeric **P**roteasome-homolog of **M**ethanogens containing a proteasome-like Ntn-hydrolase domain (blue) and a C-terminal OB-fold (magenta) similar to the N-domain of proteasome activating nucleotidases (PAN). **Anbu Operon**: Context analysis of the proteasome homolog Anbu (blue), predicting a peptide tagging system for targeted protein degradation. **HbYX-motif**: Systematic analysis of proteasome interaction motifs of archaeal AAA proteins, predicting a network that regulates the archaeal proteasome (blue); the network includes known PAN (continuous arrow), CDC48, and AMA proteins (dashed arrows).

Coordinates used: 1PMA (20S proteasome, [1]) 1E32 (CDC48, [2]) 2WG5 (PAN-N, [3]). Full-length structures of AAA proteins are schematically assembled; structures of arcRFK [4], full-length HP12, and MPM are described in this work.

Table of Contents

1 INTRODUCTION

Folding is an emergent property of sufficiently long polypeptide chains. In principle, the linear amino acid sequence encodes all information required to adopt a defined 3-dimensional structure [5]. Proteins reach their native conformation through the entropically driven hydrophobic collapse by which they exclude water from their interior [6]. This is a stochastic process, and misfolding happens under all conditions [7]. Within the essentially infinite sequence space, the fraction of polypeptide chains capable of folding is tiny. In a hypothetical library of random polypeptide chains presumably much less than one out of a billion instances would fold [8]. Although folding is physically sensitive and evolutionarily unlikely, proteins have to fold correctly in order to fulfil their functions. In fact, selection against the toxicity of misfolding is a dominant constraint on the evolution of protein-coding sequences [9, 10].

1.1 Protein Quality Control

Many proteins, especially enzymes, are intrinsically susceptible to non-native conformational changes, because dynamic access to conformational sub-states is intimately linked to a correct and reversible exertion of their physiological roles [11]. Additionally, stress conditions may perturb the folding state of proteins and enrich non-native conformations frequently resulting in the accumulation of misfolded and aggregated proteins. Since such non-native proteins are potentially harmful for the organism [12], all cellular life forms have developed a battery of molecular machines, molecular chaperones [13], which constantly monitor the folding state of the protein complement within the cell and respond to the arise of pathological proteins [14].

The principal components of this machinery are molecular chaperones and ATP-dependent proteases [13, 15]. Chaperones avoid aggregation and aid folding; proteases degrade irreversibly damaged proteins. The protein triage model describes the decision about the fate of a non-native protein as a matter of kinetic partitioning between (re)-folding and degradation (Figure 1) [16, 17]. Both responses, however, involve various ATPases that disaggregate, refold, or unfold non-native proteins. Therefore, post-translational quality control of proteins consumes a considerable amount of a cell's energy [17]. Moreover, the folding of essential cellular components depends on

chaperonin-mediaed folding [18]. Besides quality control, cellular organisms, especially eukaryotes, employ targeted protein degradation for regulatory purposes [19].

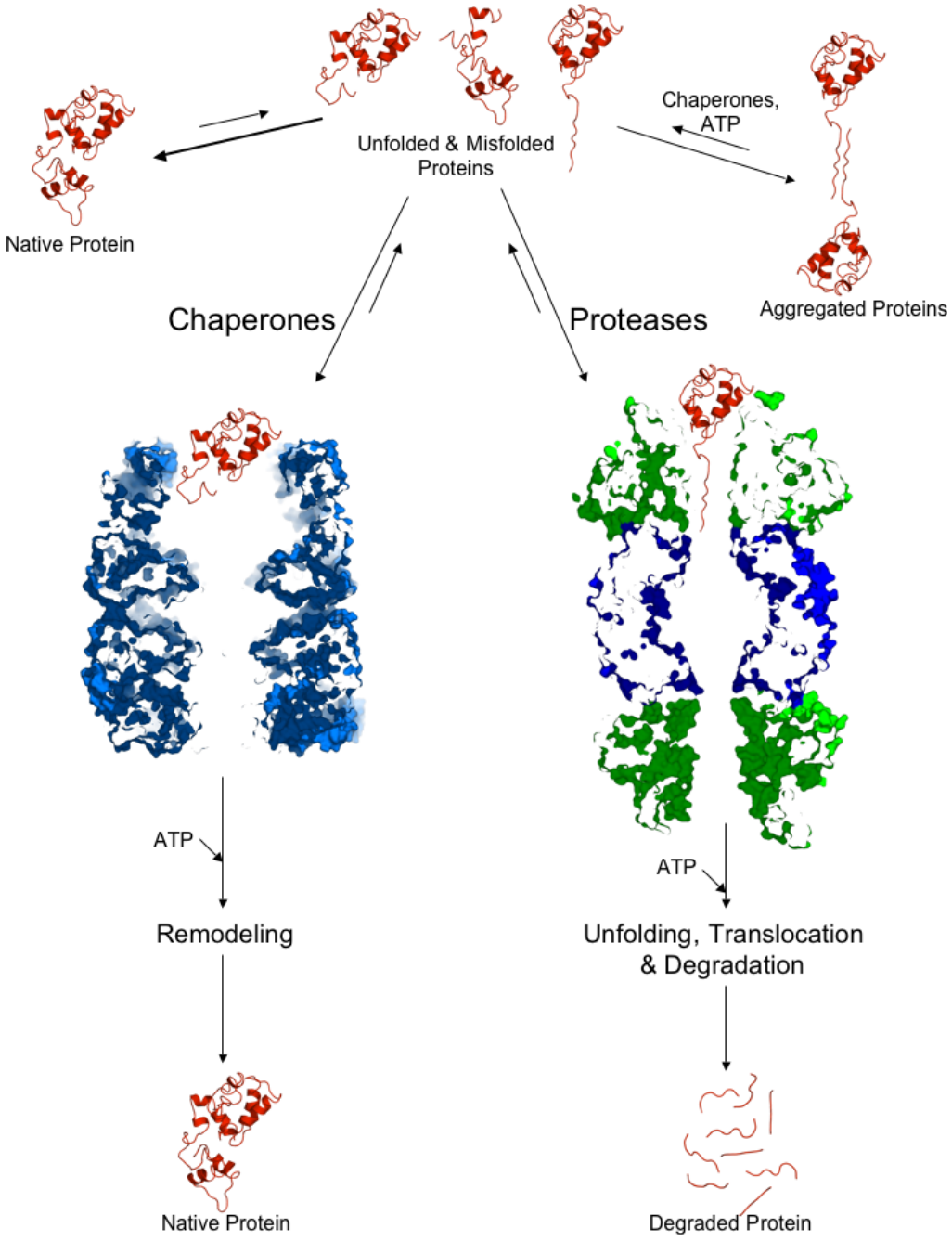


Figure 1. The Triage Model of Protein Quality Control

Unfolded or misfolded proteins expose hydrophobic regions to their surface, which are recognized by a variety of chaperones or energy –dependent proteases. The activity of chaperones helps proteins to reach their native conformation whereas proteases remove proteins from the pool of non-native proteins. Both systems lower the concentrations of non-native proteins in order to minimize the formation of potentially harmful aggregates, which in turn may be dissolved by disaggregation chaperones. The kinetics of partitioning between chaperones and proteases determines whether a protein is degraded before it folds properly, leading to a preferential degradation of proteins that do not readily reach the native state. The figure follows the model proposed by Wickner et al [17]. Proteins shown are the chaperonin GroEL (PDB-ID 1AON) and the AAA+ protease HslUV (1G3I).

In energy-dependent protein degradation, ATPases of the AAA+ superfamily play a central role [15, 20]. They sense proteins of aberrant conformation with their N-domains, unfold them through ATP-hydrolysis of their ATPase domains, and deliver them to proteases, either residing as C-terminal domains on the same polypeptide chain, or being contacted by C-terminal interaction motifs. In the following, we introduce the ATPase domain of AAA+ proteins (AAA+ module) as a molecular switch that enforces conformational changes in substrate proteins (1.2). Subsequently, we describe selected N-terminal substrate recognition domains (N-domains) providing input for the molecular switch (1.3), and the proteasome as a downstream effector that degrades the unfolded output of the ATPase (1.4).

1.2 AAA+ Proteins

AAA proteins are **A**TPases **a**ssociated with diverse cellular **a**ctivities [21, 22]. They belong to the superfamily of AAA+ proteins [20], which in turn share a common ancestry with P-loop nucleotidases [23], the largest monophyletic group of enzymes found in nature [24]. The unifying activity of AAA+ proteins is the coupling of ATP hydrolysis to disaggregation, unfolding, and remodelling of a wide variety of substrates including proteins and nucleic acids. The AAA+ module, the defining characteristic of all members of the superfamily, enables functional versatility by acting as a molecular switch that transmits ATP-driven conformational changes to a target macromolecule [25]. Additional domains and interaction motifs decorating the AAA+ module confer specificity to the particular reaction.

The AAA+ module is the structural hallmark of AAA+ proteins consisting of two domains, an N-terminal αβα sandwich and a C-terminal α-helical domain (Figure 2 A). The former adopts a P-loop NTPase fold accommodating the Walker A and B motifs crucial for ATP binding and hydrolysis, respectively [26]. The latter is a four-helical bundle contributing important residues to the ATP binding pocket and the oligomerization interface. This C-domain serves as an ancestral morphological trait used to stringently classify the AAA+ proteins within the P-loop NTPases [27]. Furthermore, it gave rise to the dimeric histones through a 3D domain-swapping event [28]. Some members of the superfamily like CDC48/p97/VAT (CDC48), NSF, or PEX1

contain two AAA+ modules in a tandem repeat referred to as D1 and D2, one of which is often degenerate, for instance D1 in PEX1 or D2 in NSF [29].

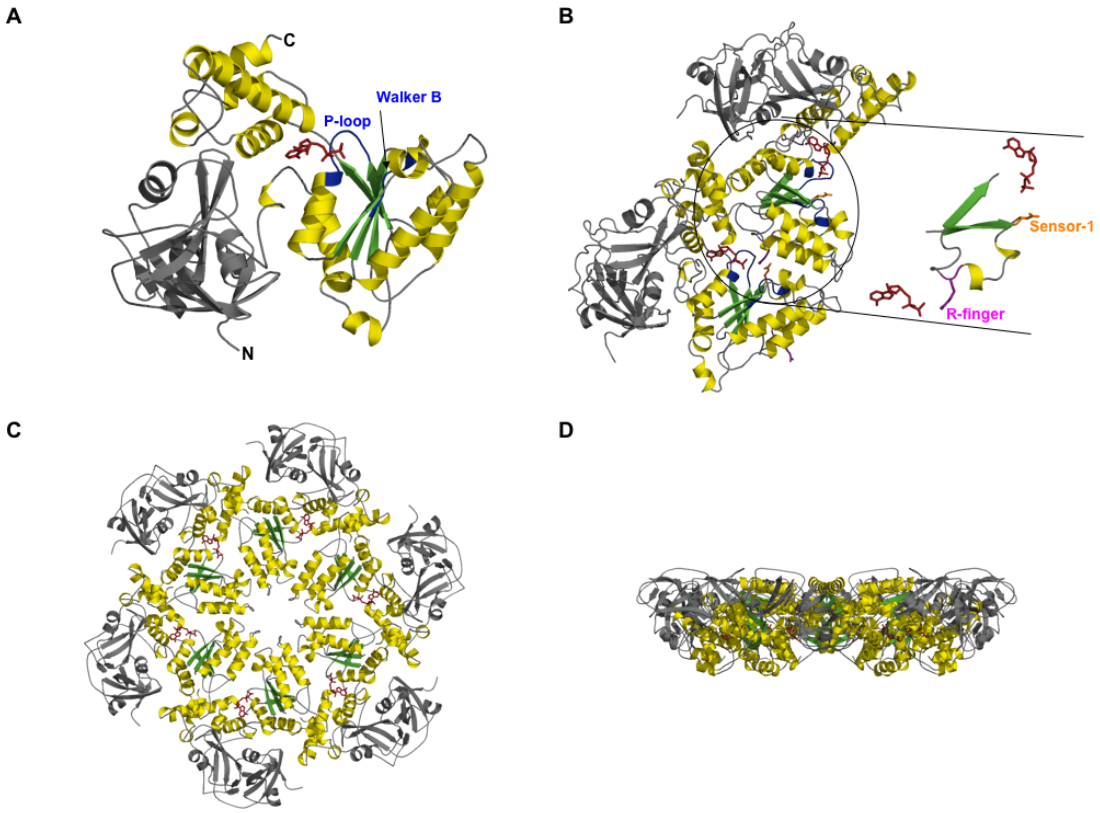


Figure 2. Architecture of a AAA Protein

(A) A typical AAA protein consists of an N-terminal substrate recognition domain (gray) and a AAA+ module. The latter contains two domains, an αβα sandwich of the P-loop NTPase fold, responsible for binding (P-loop/Walker A) and hydrolysis (Walker B) of ATP (red), and a C-terminal α-helical domain.

(B) AAA proteins are distinguished from other members of the AAA+ superfamily and of the P-loop NTPase fold by the second region of homology (SRH) harbouring the sensor-1 and arginine finger residues between β4 and β5 of the αβα sandwich. It has been suggested that both residues establish a path of communication between neighbouring subunits of a AAA hexamer [30].

(C and D) The N- and D1-ATPase-domain of CDC48 is depicted in a top (C) and a side view (D). The formation of oligomeric rings is a prevalent feature of P-loop unfoldases. AAA proteins typically form hexameric rings. A common mode of action is the ATP-dependent unfolding and threading of substrates through the central pore.

The CDC48 homolog from *M. musculus* (p97, 1E32 [2]) was used for structural representations. The second D2 ATPase domain is omitted for clarity.

The assembly of AAA+ proteins into oligomeric complexes is required for their function. Within the AAA+ superfamily various types of assemblies are known ranging from spiral-shaped pentameric complexes of the clamp loaders [31] to the cubic octamers of the CED-4 apoptosome [32]. The dominant oligomeric form within the superfamily appears to be the hexameric ring, which is supposed to be utilized by all members of the canonical AAA family (Figure 2 C and D). An important mode of

action among AAA proteins is the unfolding and simultaneous threading of the substrate protein through the central pore of the hexameric ring. It has been suggested that the sensor-1 and arginine finger residues of the ATPase domain sense the state of nucleotide hydrolysis in one subunit (sensor-1) and transmit this information to the neighbouring subunit (Figure 2 B) [30]. These residues are located in the second region of homology (SRH), which is important for the sequence-based distinction of AAA proteins from other members of the AAA+ superfamily.

Table 1. AAA+ Proteases

NAME	AAA	PROTEASE	CATALYSIS	OLIGOMER	BIND	SPECIES	NOTE
FtsH/ m-AAA			Zinc	hexamer	Covalent	Bacteria/ Mitoch*	Memb. bound
Lon			Ser-Lys	hexamer	Covalent	Bacteria/ Mitoch*.	
Archaeal Lon			Ser-Asp-Lys	hexamer	Covalent	Archaea	Memb. bound
ClpX/A-ClpP	ClpX/ ClpA	ClpP	Ser-His-Asp	14mer, 2 heptamers	Non-covalent	Bacteria/ Mitoch*	Symm. Mismat
HslUV	HslU	HslV	Thr(-Glu-Lys)	$\beta_6\beta_6$	Non-covalent	Bacteria/ Mitochond	
Protea-some	Rpt/ PAN/ ARC	20S Proteasome	Thr(-Glu-Lys)	$\alpha_7\beta_7\beta_7\alpha_7$	Non-covalent	Eukaryote/ Archaea/ Actinobac.	Symm. Mismat

*Indicates mitochondrial localization within eukaryotes, whereas the proteins are mostly encoded in the nucleus.
Note that this is not a comprehensive list of AAA+ proteases; catalysis and oligomerization refers to the protease not the ATPase; all ATPases belong to the AAA+ superfamily, but only proteasomal ATPases and FtsH belong to the AAA family.

AAA+ proteins perform a broad range of functions including DNA replication (Clamp loader, MCM), transcription activation (σ54 activator), metal insertion (Mg chelatase), protein disaggregation (ClpB/Hsp104), protein-complex disassembly (NSF), and protein degradation [33]. The cooperation of a AAA+ unfoldase with various types of proteases appears to be a successful one, because it has evolved several times independently within the AAA+ superfamily [34]. The group of AAA+ proteases comprises FtsH and Lon, which carry proteolytic domains on the same polypeptide chain, ClpA and ClpX, which interact with the proteases ClpP, and the proteasomal ATPases and HslU, which feed the 20S Proteasome and HslV (Table 1). In contrast to these ATPases, some of the proteases employ different proteolytic mechanisms and do

not share a common ancestry. Nevertheless, the operational principle they share is that the ATPase regulates access to the downstream effector protease, which sequesters the active site in the interior of a barrel shaped structure. Therefore, they are referred to as self-compartmentalizing proteases [35]. The number of AAA+ proteases is small in comparison to the vast number of mostly monomeric or dimeric proteases included in the MEROPS classification [36], but they carry much of the burden of intracellular stress responses.

Stress conditions perturb the structural integrity, especially of less stable proteins. As a result proteins expose hydrophobic residues to the surface that increase their tendency to form harmful aggregates notwithstanding the fact that they are rendered non-functional. AAA+ proteases and AAA proteins like CDC48 encompass substrate recognition domains at their N-termini that bind and hold misfolded proteins thereby preventing aggregation [37]. Just like the proteases that receive the output, the polyphyletic input domains have been independently recruited to the AAA+ module from different folds [38]. We have investigated the divergent evolution of selected N-domains (chapter 4) contained by those AAA proteins, for which we propose a function as proteasomal ATPases in archaea (chapter 5.1). These N-domains are introduced in the next section.

1.3 N-terminal Substrate Recognition Domains of AAA Proteins

A phylogenetic analysis subdivided the family of AAA proteins into six major clades: metalloproteinases, “meiotic”, D1 and D2 domains of proteins with two AAA domains, proteasome subunits, and BCS1 [38]. Despite the homology of AAA+ modules, their N-domains are of polyphyletic origin, which suggests that the wide variety of biological functions performed by AAA (+) proteins arises from the variety of N-domains. Substrate specificity is often conferred in conjunction with adaptor proteins, such as UBX proteins (CDC48), ClpS (ClpA) or SspB (ClpX).

A common theme shared across AAA N-domains is the ability to bind to proteins exposing hydrophobic residues on the surface and to prevent aggregation [37]. This points to the function of the corresponding full-length proteins in unfolding and disaggregation of misfolded proteins requiring energetic input provided by ATP-hydrolysis through the AAA+ module. Furthermore, several N-domains, like the β-clam of AMA or the tandem N-domain of CDC48, are negative regulators of ATPase activity

[39, 40]. This suggests that N-domains prevent unnecessary hydrolysis of ATP unless they are actively engaged in the processing of substrates.

In the following, three types of N-domains are described, the double-ψ barrel, the β-clam, and the OB-fold (Figure 3).

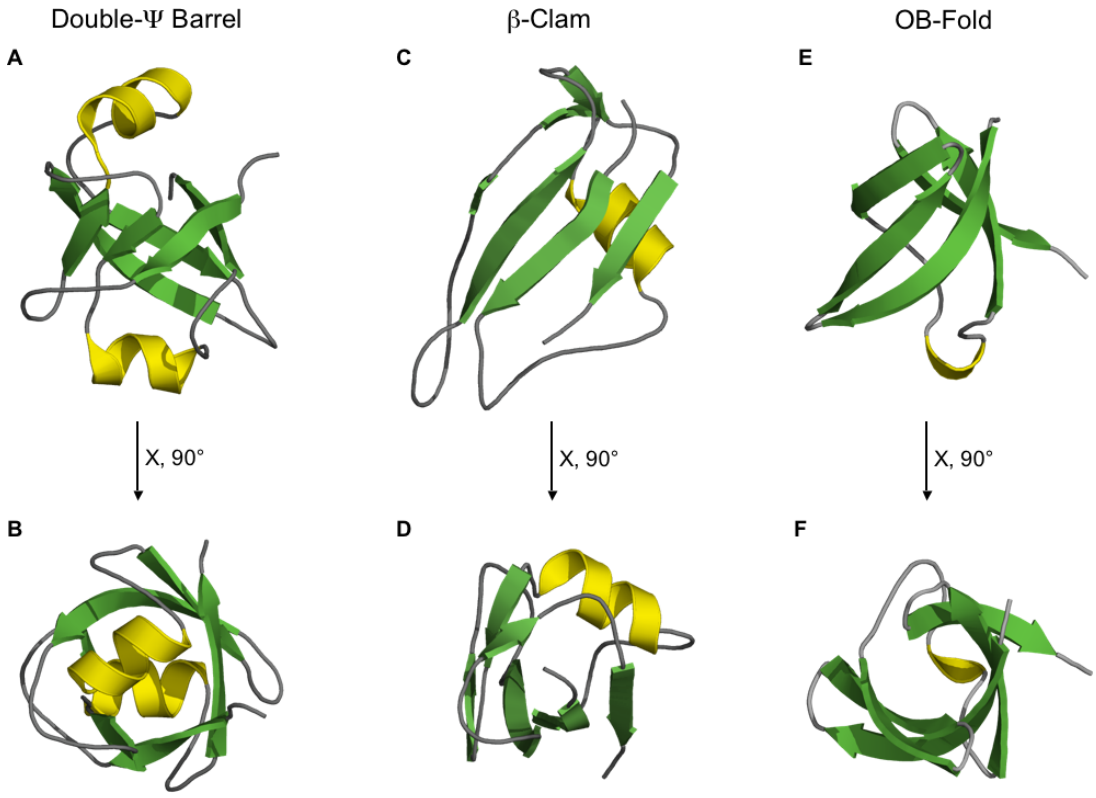


Figure 3. N-Domains of Selected AAA Proteins

(A and B) The Double-ψ barrel is a six-stranded β-barrel capped by short α-helices at both openings. It is found in AAA proteins like CDC48, NSF, or PEX1. The double-ψ barrel, the RIFT barrel, and the Swapped-hairpin barrel form the cradle-loop barrel metafold relating functionally diverse proteins by events of homologous fold change. (see 4.1)

(C and D) The β-Clam consists of a central β-sheet that forms an open clam-like structure embedding an α-helix. The β-clam mostly occurs in tandem with a double-ψ barrel at the N-terminus of AAA Proteins. The same tandem is also found in ubiquitin fusion and degradation (UFD1) proteins. The archaeal group of AMA proteins are also members of the AAA family but contain an isolated β-clam at their N-terminus (see 4.2).

(E and F) The Oligonucleotide/Oligosaccharide-binding fold is a five-stranded β-barrel often capped by an α-helix on one side. The OB-fold is a very common binding module, which recognizes a wide variety of substrates including proteins. The N-terminal portion of the proteasomal ATPases PAN (archaea), Rpt1-6 (eukaryotes), and ARC (actinobacteria) consists of a coiled coil and an OB-fold domain (see 4.3).

The proteins shown are 1CZ4 (residues 1-92), 1CZ4 (residues 94-176), and 2WG5 chain A (residues 63-120).

1.3.1 The Double-ψ Barrel Domain

AAA proteins like the peroxisome biogenesis factor 1 (PEX1), the membrane fusion protein NSF, and CDC48 contain a double-ψ barrel as the N-terminal half of their two-domain N-terminal substrate recognition part. In eukaryotes, CDC48 proteins perform a plethora of functions, and are generally implicated in the handling of ubiquitylated substrates en route to the proteasome [41], for instance the transport of proteins from the ER to the cytosol in the endoplasmic reticulum associated degradation pathway [42]. It was initially identified as Valosin-containing protein (VCP) in pig [43]; the homolog in yeast was named CDC48 (cell division control protein 48) [44], in mouse p97, and in *T. acidophlum* VAT (VCP-like ATPase of Thermoplasma) [45]. The latter functions as a general unfoldase, negatively regulated by its N-domain [40, 46]. In the following, we refer to proteins of this homologous group as CDC48.

The structure of a variety of protein domains contains closed β-sheets, in which the first strand is hydrogen-bonded to the last. They are described as β-barrels or orthogonally packed β-sheets [47, 48]. The double-ψ barrel is a six-stranded barrel capped with α-helices at both ends containing a pseudo two-fold internal symmetry (Figure 3 A and B) [49]. Each strand of one out of two ββαβ-elements forms hydrogen bonds only with strands of the other ββαβ-element, resulting in a distinctive and complex topology. This knotted appearance is enabled by elongated loops, the ψ-loops, between strands β1 and β2 that cross the second strand of the symmetry related element. The topological similarity to the Greek letter ψ inspired the name [50]. The two ψ-loops frame a cradle-shaped groove, which has been implicated in substrate binding [51].

Through its basic building block, the ββαβ-element, the double-ψ barrel is evolutionarily related to a divergent array of proteins comprising transcription factors as well as enzymes. On the structural level, the duplicated ββαβ-element shows a remarkable degree of plasticity and forms the homologous core of the cradle-loop barrel metafold, which groups at least three distinct folds [52] (see chapter 4.1). Whereas the fold is defined as a topologically distinct arrangement of secondary structure elements, the metafold groups folds that are related by an event of homologous fold change like a circular permutation or a strand swap [53].

1.3.2 The β-Clam Domain

The β-clam domain generally occurs C-terminally to the double-ψ barrel in the N-terminal substrate recognition part of PEX1, NSF and CDC48 [54]. This domain tandem is also present in Ubiquitin fusion degradation proteins (UFD1) [51] that have most likely lost the downstream AAA+ modules. The small, family of AMA proteins are an exception as they contain an isolated β-clam at the N-terminus [39]. AMA proteins are AAA **A**TPases found in **a**rchaeoglobales and **m**ethanogenic archaea. Our sequence analysis of AMA proteins suggests that they might function as proteasomal ATPases in archaea (5.1). Furthermore, the β-clam is detected in a protein family of the archaeal genus thermococci. In these proteins, a domain lacking any apparent sequence similarity follows the β-clam (4.2).

The structure of the β-clam consists of six β-strands and is organized around a long and sharply band N-terminal strand (Figure 3 C and D). This fold, however, is not a closed β-barrel, Instead, it resembles an open clam embedding the α-helix that follows the first beta strand [49].

1.3.3 The OB-fold Domain

Within the AAA family, canonical proteasomal ATPases – PAN (**p**roteasome **a**ctivating **n**ucleotidase) in archaea [55], ARC (**A**TPase forming **r**ing-shaped **c**omplexes) in actinobacteria [56], Rpt1-6 (**r**egulatory **p**article **t**riple-A ATPase) in eukaryotes [57] – contain an OB-fold domain, in case of ARC two of these. In all three subgroups, long α-helices precede the OB-fold at the N-terminus forming dimeric coiled coils such that these AAA proteins hexamerize as trimers of dimers. A short but flexible linker, including a conspicuous PP-motif, between the coiled coil and the OB-fold has been implicated in processing of substrates through the pore of the ATPase into the proteolytic chamber of the proteasome [3, 58, 59]. An OB-fold, remotely similar to the N-domain of PAN, is found in an archaeal protein family that contains a protease homologous to proteasome subunits (4.3).

The OB-fold consists of five rather coiled β-strands forming a closed β-sheet. The resulting β-barrel is capped by an α-helix inserted between strands β3 and β4 (Figure 3 E and F). The three loops connecting β1-β2, β3-α, and β4-β5 protrude from the barrel axis and mount a versatile binding site. The name OB-fold, an acronym for oligonucleotide- and oligosaccharide-binding fold that indicates this versatility [60], is

extended by the ability to bind proteins. A vast number of very different amino acid sequences adopt the OB-fold [61]. Whether all of them are homologous has not yet been clarified [62].

1.4 Proteasomal Protein Degradation

The Proteasome (multicatalytic proteinase) [63], formerly known as Macropain (high molecular weight proteinase) [64], is a large protease playing a crucial role in protein degradation in archaea, actinobacteria and eukaryotes [65]. The catalytic core particle, the 20S Proteasome, forms a cylindrical assembly consisting of 28 subunits based on the architectural principle of self-compartmentalization [35]. Four stacked heptameric rings of the stochiometry $\alpha_7\beta_7\beta_7\alpha_7$ sequester the active sites in order to confine access to proteins that present a degradation signal [1] (Figure 4 C and D). The β-subunits build the inner proteolytic chamber harbouring the catalytic nucleophile, a threonine residue [66], which is located at the very N-terminus after autoprocessing of a leader peptide [67]. The α-subunits provide regulated passage of substrates to the proteolytic core through flexible N-terminal tails lining the openings of the barrel [68, 69]. In eukaryotes, the heptameric rings are formed by seven different α- and seven different β-subunits [70]. Three additional β-subunits encoded on the MHC locus are incorporated in the immunoproteasome of infected cells for more efficient antigen presentation on MHC I complexes [71, 72]. In archaea and actinobacteria the rings are homooligomeric consisting of only one type of α- and β-subunits, although certain species like *Haloferax volcanii* (2 α, 1 β) or *Rhodococcus erythropolis* (2 α, 2 β) encode two α- and two β-subunits [73, 74].

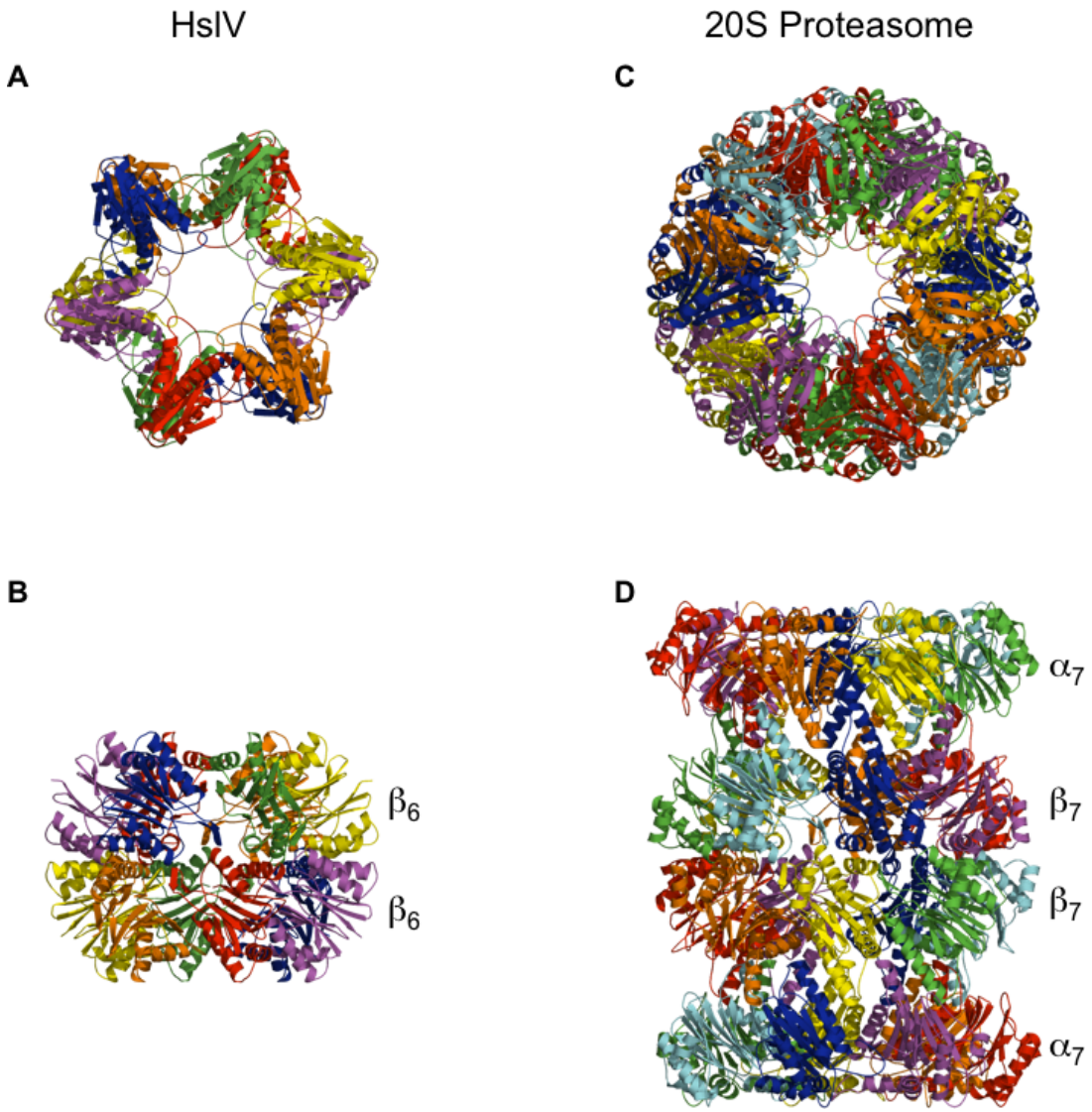


Figure 4. Architecture of HslV and the 20S Proteasome

HslV and the 20S proteasome are self-compartmentalizing proteases sequestering their active sites within a large barrel-shaped architecture. The monomers of both assemblies are homologous and belong to the Ntn-hydrolase fold. The proteins shown are HslV from *H. influenzae* (1G3K) and the proteasome from *T. acidophilum* (1PMA). Colouring is by chain.
(A and B) Top and side view of HslV. HslV forms a dodecamer consisting of two hexameric rings. It has been proposed that HslV gave rise to the more complex 20S proteasome by gene duplication [87].
(C and D) Top and side view of the 20S proteasome. Four heptameric rings compose the 20S proteasome. The two inner rings consist of proteolytically active β-subunits, whereas the outer rings are formed by inactive α-subunits. The latter interact with gate-keeping ATPases via the HbYX motif. In eukaryotes, the rings are heteroheptameric.

A simpler version of the proteasome is widespread in the bacterial lineage. The protease HslV (heat shock locus V) forms a dodecameric assembly, $\beta_6\beta_6$, subdivided into two hexameric rings [75] (Figure 4 A and B). The monomers of the 20S proteasome and HslV share significant sequence similarity and belong to the Ntn-hydrolases (**N**-terminal **n**ucleophile **hydrolases**) fold [76]. Both cleave a pro-peptide by

autoproteolysis and employ a similar proteolytic mechanism relying on a threonine as the catalytic nucleophile. Through autoprocessing the N-terminal α-amino group of the threonine is exposed and functions as a general base activating the alcohol group for catalytic attack of substrate carbonyl carbons [77]. Nevertheless, other residues contribute to the formation of the active site [78]. The Ntn-hydrolase fold is also adopted by enzymes like penicillin acylase, asparaginase, or γ-glutamyl transpeptidase [79]. The catalytic nucleophile of Ntn-hydrolases can be a serine, a threonine or a cysteine residue, which is a variation that is not observed in other protease families [80]. Although different families use different nucleophiles, the core architecture of the active site remains conserved, which is mounted onto the common scaffold of an αββα-sandwich. Despite structural and functional similarity, proteasome-like Ntn-hydrolases do not share significant sequence similarity with other non-self-compartmentalizing Ntn-hydrolases.

The AAA ATPases PAN, Rpt1-6, and ARC are the regulatory ATPases of the proteasome in archaea, actinobacteria and eukaryotes, respectively, which interact upon ATP-binding and stimulate proteolysis of substrate proteins [81, 82]. A conspicuous interaction motif at the C-terminus, the HbYX motif [83], of the ATPases enters defined pockets at the interface of two α-subunits and triggers conformational changes leading to the opening of the gates [84]. In contrast, HslU is a AAA+ protein that regulates HslV. Both of them are frequently encoded in an operon that is strongly induced upon heat stress [85]. This complex differs from the proteasome-ATPase complexes by the mode of interaction and the absence of the symmetry mismatch between the hexameric rings of the ATPase and the heptameric rings of the Proteasome [75].

The increased complexity of the eukaryotic 26S proteasome is also reflected by the large 19S regulatory particle and the heterohexameric Rpt-ATPase at its base [86] suggesting recurrent gene duplication of ATPase and α-/β-subunits in the eukaryotic ancestor [87]. PAN and ARC proteins of archaea and actinobacteria, however, are homohexamers. Furthermore, two other non-ATPase regulatory particles of the 11S and PA200 type diversify the functional repertoire of the eukaryotic proteasome [88, 89].

1.4.1 Tagging Systems for Targeted Protein Degradation

Another layer of regulation in the selection of substrates for targeted degradation by the proteasome is the covalent attachment of a degradation tag to a substrate protein (Figure 5, Table 2). The best-studied tagging system is the ubiquitin-conjugation pathway [90]. Ubiquitylation is the post-translational modification of target proteins with the small protein ubiquitin serving multiple purposes in eukaryotic organisms; among them is not only targeted protein degradation (mediated by lysine48-linked polyubiquitin chains), but also signal transduction, mebrane protein trafficking, and DNA repair (mediated by lysine63-linked polyubiquitin chains) [91]. The ubiquitin homologs sumo (sumoylation) and nedd8 (neddylation) and other ubiquitin-like proteins further diversify the capabilities of this tagging system [92]. For efficient degradation by the 26S Proteasome, however, a substrate needs to contain a two-part degron (degradation signal), which consists of a proteasome-binding signal (a lys48-polyubiquitin tag), and a degradation initiation site (an unstructured region) [93, 94].

Intricate enzymatic machinery catalyzes the conjugation of ubiquitin to substrate proteins. The three main components are the ubiquitin-activating enzyme (E1), the ubiquitin-conjugating enzymes (E2), and ubiquitin protein ligases (E3) [90]. E1 activates ubiquitin at the C-terminus through adenylation, E2 picks up ubiquitin via transthioylation from E1, and E3 ligates it to the substrate forming an isopeptide bond between the Cterminal carboxylate group of ubiquitin and a substrate lysine residue. Whereas most eukaryotes encode only one E1 enzyme, all contain multiple E2 and E3 isozymes. Especially the large number of E3 enzymes allows for the modification of a wide variety of substrates making ubiquitylation a highly specific tool for the regulation of crucial processes like cell cycle progression [95]. However, the evolutionary roots of the ubiquitin-conjugation systems are prokaryotic cofactor biosynthesis pathways, in which ubiquitin-homologs of the β-grasp fold, e.g. ThiS or MoaD, function as sulphur carriers activated by E1-like enzymes [96].

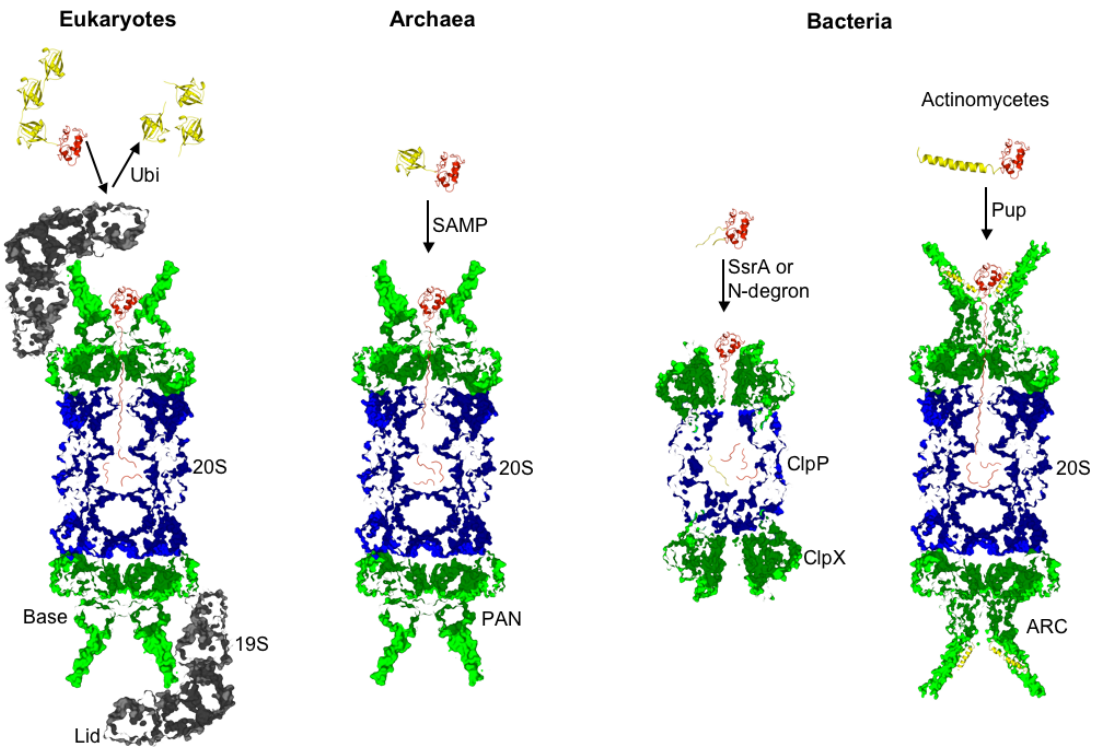


Figure 5. Tagging Systems for Targeted Protein Degradation

Tagging systems for targeted protein degradation exist in the three domains of life. Chaperones of the AAA+ superfamily (green) recognize, unfold and translocate substrates (red) into the inner of self-compartmentalizing proteases (blue) using the energy of ATP-hydrolysis. Eukaryotes employ the ubiquitin-conjugation system to mark proteins for degradation by the 26S proteasome. The polybiquitin is recognized and recycled by the highly complex 19S proteasome consisting of a lid (gray) and a base that includes the Rpt1-6 ATPase. Most archaea use SAMP (yellow), a homolog of ubiquitin, as a tag for degradation by the simpler PAN-proteasome machinery. Details of recognition and potential removal of SAMP are currently not known. The ATPase ARC regulates the 20S proteasome of actinomycetes binding Pup (yellow) with its N-terminal coiled coil domain. Bacteria including actinomycetes and the mitochondria of certain eukaryotes contain other AAA+ proteases like ClpA/X-P, which degrades SsrA-tagged substrates or substrates of the N-end rule pathway via various adaptor proteins.
These tagging systems for targeted protein degradation are composed by homologous and analogous components. All ATPases belong to the AAA+ superfamily, whereas the 20S proteasome and ClpP convergently evolved the feature of self-compartmentalization. Furthermore, ubiquitin and SAMP share a common ancestry, but Pup is a development specific for actinomycetes.
N-domains, ATPases, and Proteases were tentatively grouped intended as a schematic depiction. The concept of this figure is modified from Striebel et al [97].

Tagging mechanisms also regulate the archaeal and the actinobacterial proteasome (Table 1). In archaea, the tag is homologous to ubiquitin and called **s**mall **a**rchaeal **m**odifier **p**rotein (SAMP) [73]. The enzymes catalyzing addition (and removal) of SAMP have not been described yet. How specificity is conferred to the modification is also still elusive, and homologs of E3 enzymes have not been detected [98]. Actinobacteria have developed an alternative degradation tag that does not share a common ancestry with ubiquitin. Pup, **p**rokaryotic **u**biquitin-like **p**rotein [99], is

directly recognized at the proteasomal ATPase ARC, which induces folding of the otherwise unstructured Pup [100]. Furthermore, Pup is ligated to substrates through simpler machinery, and there is only one Pup ligase, PafA, characterized so far [101, 102]. Interestingly, the ligation of Pup occurs through the very C-terminal residue, which is also preceded by a GG-motif, as it is the case for ubiquitin. However, it is not the C-terminal carboxylate, which is participating in the isopeptide bond, but the γ-carboxylate of the ultimate glutamate residue [103]. In many Pup orthologs the deamidation of a glutamine residue catalyzed by dop (**d**eamidase **o**f **Pup**) is required to generate the glutamate [102].

Table 2. Tagging Systems for Targeted Protein Degradation

TAGGING	STRUCTURE	ATTACHMENT	LIGASE	PROTEASE	SPECIES
N-end rule (Leu/Phe/Arg	Single amino acid	Peptide bond: N-terminus of substrate	L/F-or R-Transferase (charged tRNA)[A]	ClpAP (adaptor ClpS)	Bacteria Eukaryotes[C]
SsrA-tagging (SsrA)	No fold (11 aa)	Peptide bond: C-terminus of substrate	Ribosome, (tmRNA) [A]	ClpX/AP (adaptors: SspB/ClpS)	Bacteria
Pupylation (Prokaryotic ubi-like protein)	Unfolded/helica upon binding to ARC coiled coil (GGE-motif)	Isopeptide bond: γ-COOH of Glu of Pup -> substrate lysine	PafA (ATP hydrolysis) [A]	Proteasome	Actinobact.
Sampylation (Small archaeal modifier)	β-Grasp fold (GG-motif)	Isopeptide bond: Ubi-COOH -> ε-NH2 of substrate lysine	?	Archaeal Proteasome	Archaea
Ubiquitylation (Ubiquitin)	β-Grasp fold (GG-motif)	Isopeptide bond: Ubi-COOH -> ε-NH2 of substrate lysine	E1, E2, E3 (adenylation of ubi) [A]	26S Proteasome (Polyubi-lysine48)[B]	Eukaryotes

A: The energy source for (iso)-peptide bond formation is denoted within brackets
B: Polyubiquitin chains linked via lysine48 target substrates to the proteasome; lysine63 linkages serve other purposes.
C: In eukaryotes, the N-end rule is part of the ubi-proteasome pathway; N-recognins, a class of E3 ligases confer specificity.

Analogous and homologous elements are used by different lineages in order to accomplish the functional task of targeted protein degradation. This is further illustrated by the ClpA/XP system that uses a AAA+ protein, ClpA or ClpX, as unfoldase, but

interacts with a downstream protease, ClpP, that is not evolutionarily related to the proteasome [34] (Figure 5). ClpP, however, shares multiple features with the proteasome like self-compartmentalization and the formation of heptameric rings [104]. Apart from misfolded proteins that are directly recognized, ClpA/XP, assisted by adaptor proteins, specifically degrades proteins carrying an SsrA tag that was added to translationally stalled proteins [105, 106] or an N-degron that was added via the N-end rule pathway (ClpA) [107] (Table 2). In eukaryotes, the N-end rule pathway is part of the ubiquitin system [108]. The variety of analogous degradation tags or the employment of different types of proteases point to a selective advantage, which is conferred by generalized proteolytic systems. Reuse of homologous elements in different machineries, like the AAA+ domain, exemplifies the principle of modularity. Nevertheless, de novo invention is rare, and reuse is the rule, highlighting modularity as a general attribute of living matter across levels of organization [109, 110].

1.5 Protein Evolution

Central aspects of Darwin's theory of evolution are [111]: Random heritable variation provides the material for evolution; rare beneficial changes are fixed by natural selection, the main driving force of evolution, leading to increasingly complex adaptive traits and a general progress of evolution; progress occurs gradually by infinitesimally small steps that become fixed by positive selection; evolution moves on in a uniform manner. The "Modern synthesis in the light of genomics" updates the theory with respect to molecular data [112]. The range of random variation does not only comprise point mutations but also duplication of genes and entire genomes, loss of genes, lateral gene transfer and endosymbiosis accompanied by massive gene flux. Therefore, gradualism is not a general principle of evolution. The neutral theory of evolution by Kimura describes random drift as another fundamental evolutionary force [113]. It states that the vast majority of constantly occurring mutations, which become fixed, are selectively neutral. Subsequent purifying selection removes deleterious mutations. Whereas positive selection of rare beneficial variation remains important, random drift quantitatively dominates. Furthermore, there is no general trend towards increased complexity [112].

1.5.1 Molecular Evolution

A consequence of the neutral theory is the existence of a clock that determines the rate at which sequences of genes evolve. The earlier developed concept of a molecular clock posits that a protein-coding gene evolves at a characteristic rate, which remains constant over long evolutionary time spans in the absence of functional change [114]. These characteristic rates have been derived from sets of orthologous genes (homologs separated by speciation events). They vary by four orders of magnitude. Two reasons were considered to be the cause of this variation. First, intrinsic structural-functional constraints that influence a particular protein, and second the biological importance of the given protein within the organism were supposed to determine the evolutionary rate. With the availability of system-wide expression data from various model organisms it became apparent that the rate is strongly affected by the expression level of a protein from bacteria to mammals [115]. The general observation is that highly expressed proteins evolve slowly.

The Mistranslation-induced Misfolding hypothesis attributes the covariation of evolutionary rate and translation rate to selection for robustness to misfolding, because the expression level amplifies the fitness cost of misfolded and toxic proteins [9]. This hypothesis initially considered mistranslation ("phenotypic mutations") as the primary source of misfolding and connected it to selection for the usage of the most efficient synonymous codons, which is correlated to tRNA abundance for a particular codon, especially in highly expressed genes. Nevertheless, apart from mistranslation stochastic misfolding of the native sequence takes place as well. The frequency of stochastic misfolding, however, is also affected by the intrinsic robustness and plasticity of a given protein [116]. Therefore, protein evolution seems to be primarily constrained by the maintenance of native folding [10]. Unexpectedly, essentiality or the numbers of interaction partners do not correlate with the evolutionary rate of an orthologous group.

Studies investigating stability effects of mutations pointed out that most mutations impose an effective fitness cost *in vivo*, because they lessen the thermodynamic and kinetic stability of a protein fold [117]. This deleterious effect of mutations appears to be a major confinement to the evolvability of proteins, i.e. their access to changes in sequence and function. In this respect, the chaperone machinery not only plays a central role in reducing the burden of destabilized and misfolded proteins, but it also acts a capacitor for protein evolution [118, 119]. Although proteins are accurate, proficient and specific functional agents, their dynamic nature provides access to evolutionary innovation. Conformational variability, multi-specificity,

functional promiscuity and moonlighting are considered to be a source of diversity immanent to proteins, which can be enriched by evolution [120].

1.5.2 The Concept of Homology

Two or more biological structures are homologous if they are alike because of common ancestry [121]. In contrast to similarity homology is an absolute concept; there is no such condition as ‘degrees of homology’. Homology has to be distinguished from analogy, which is the evolution of similar structures (and functions) through different pathways [122]. Convergent evolution has been observed at various levels, for instance the wings of bats and birds or the catalytic triad of Subtilisin and Chymotrypsin [123].

Descent from a common ancestor can be hypothesized on the basis of similar properties detected in biological objects. For proteins, similarities could be reflected in sequence, structure, and function. Comprehensive analysis of all three aspects offers the best way to support homology [53]. Among these criteria statistically significant sequence similarity is the most robust marker for homology because the combinatorial complexity of polypeptide sequences results in a basically infinite sequence space [124]. However, protein structure is more conserved than its sequence, and therefore structure comparisons can reveal remote homologies beyond the detection limit of sequence-based methods [125, 126]. On the other hand structural similarities are more likely to be of analogous origin due to the relatively minor size of structure space [127].

For many purposes and with more and more genomes being sequenced it turns out that homology is not a sufficiently well defined term to characterize the relationship of two genes [128]. Orthologous genes derive from an ancestral gene in the last common ancestor of a considered group and are separated by speciation events. Paralogs stem from an ancestral gene that has been duplicated within a genome. Whereas orthologs often perform the same or a similar function [129], paralogs frequently experience neo/sub-functionalizations because the duplicated copy enjoys a greater freedom to evolve in the presence of the ancestor that is still fulfilling the original function. Genes that are acquired by lateral gene transfer escape these definitions, and are sometimes referred to as xenologs.

The BLAST heuristics facilitates the retrieval of homologous sequences and allows for subsequent gathering of information about the best-studied homolog of the gene under investigation [130]. The development of position specific substitution matrices and usage in PSI-BLAST increased sensitivity of sequence searches [131].

Currently, the generation of Hidden Markov Models (HMMs) from the query sequence and comparing it to databases of pre-calculated HMMs is the most sensitive way of sequence comparison enabling the discovery and substantiation of remote homologies that were previously only suspected on the basis of structural similarities [132, 133]. The HMM-HMM comparison method HHpred was frequently used in the course of this work in order to identify distant evolutionary relationships.

1.5.3 Evolution of Protein Diversity

The tremendous diversity of proteins has been and continuous to be generated through a variety of mechanisms, such as substitution, insertion and deletion [134], gene duplication [47, 135, 136], (unequal) recombination [137], domain shuffling [138], circular permutation [139], and other events of homologous fold change [53]. Modern proteins with their intricate domain composition are highly complex. Many proteins can be decomposed in domains, which define islands of autonomously folding sequences within the vast sequence space [8], resulting in a distinct topological arrangement of secondary structure elements. Systematic sequence comparisons detected about 10^5 domain families, which can be assigned to approximately 10^3 folds [140] indicating restrictions in structure space for polypeptide chains. These islands of structural stability are largely separate in an evolutionary sense; in a geometrical sense they can be connected through intermediates and localized regions of structural similarity [141].

Even autonomously folding domains as the basic modules of proteins represent elaborate structures. The sequences of protein domains have an average length of about 185 amino acids (in the SCOP database) [142], which implicates a combinatorial complexity whose random sampling excludes *de novo* evolution. However, domains are frequently constructed from smaller supersecondary-structure elements, αα-/ββ-hairpins, βαβ-elements, that have a length of about 30 amino acids [143] reducing the combinatorial complexity to a level that could have been sampled by biological systems [8].

The theory of an ancient peptide world describes a scenario for the evolution of folding, according to which such supersecondary structure elements gave rise to the modern complement of domains by modular repetition and recombination [144-146]. These ancient peptide modules would originate in the context of an ‘RNA world’ when RNA was used for both information storage and catalysis [147]. Short polypeptide chains would have acted as cofactors and extended the catalytic repertoire of ancient

ribozymes [148]. A gradual process of emancipation, would lead to autonomously folding, domain-sized polypeptide chains. Internal symmetry and repetitive patterns [149-151] provide evidence for such modules and suggest that this emancipation process unfolded via oligomeric intermediates, at least in case of globular proteins [146]. Evolutionary 'improvement' of central information processing machineries (i.e. RNA polymerase and ribosome) would have eventually empowered the accumulation of ancient peptides on one chain removing the necessity of oligomerization and enabling more efficient folding.

All cellular organisms descend from the last universal common ancestor. In contrast, the protein universe is polyphyletic and has its roots in a number of primordial forms. Growth of sequence and structure databases and the improvements in remote homology detection methods enabled the delineation of such ancient peptide modules, revealing homologous relationships across folds. Their number is smaller than the number of known folds, which suggests that the protein universe is less polyphyletic than previously thought [140].

1.5.3 Protein Classification

Protein classification efforts generate order among the diversity of proteins. Structural classification systems, such as SCOP (Structural classification of proteins) [142] or CATH (Class-Architecture-Topology-Homology) [152], group high-resolution protein structures deposited at the Protein Data Bank [153] based on structural similarity. They define different levels of hierarchy and account for varying degrees of similarity by employing homologous criteria at lower levels of hierarchy and analogous criteria at higher levels of hierarchy. The SCOP database, which we frequently used as a reference, implies homology on the lower four levels of hierarchy: Species, Protein, and Family contain closely related sequences, and the Superfamily level groups together homologous families with common structural and functional features. The relationships at the two top levels of hierarchy are intended to be rather geometrical. The fold brings together superfamilies sharing a general structural similarity. At the root the Class level orders proteins based on their secondary structure content and organization.

Advancements in sequence comparison methods revealed that many of the superfamilies grouped in a fold indeed share a common ancestry [140]. Additionally, more and more evolutionary relationships across folds become apparent [154], which are based on events of homologous fold change [53]. The core regions of homology

shared by one or more folds may encompass such ancient peptide modules mentioned above. In order to further group such related folds, the term metafold was coined [155]. Accordingly, the metafold would contain homologous proteins with a similar architecture but not necessarily identical topology contributing to the goal of a protein classification system that is based on natural descent [52].

Sequence based classification schemes like the Pfam database offer information about protein domains of sufficient prevalence regardless whether the structure or function is known or not [156]. Domains of unknown function (DUFs) are general targets of structural genomics initiatives with the goal to explore uncharted territory in the protein universe [157, 158]. In this work, we present the initial characterizations of two such domains of unknown function, (DUF120 – 4.1; DUF2121 – 4.3).

The evolutionary history of a homologous group of genes is traditionally inferred by phylogenetic analysis [159]. Although phylogenetic trees can be highly informative they are problematic in certain respects (dependence on the quality of the multiple sequence alignment; error-prone if unrelated sequences or laterally transferred genes are included). By contrast, the cluster analysis of proteins, uses the power of BLAST for an all-against-all comparison of the sequences under consideration and places them in a force field where they experience attraction proportional to the BLAST P-value [160]. This allows for the classification of large protein families whose phylogeny would have been hard to compute. Furthermore, it is insensitive to the inclusion of non-related sequences. On the other hand clustering is not able to delineate the evolutionary history of a gene family. This work, however, frequently relies on CLANS (**Cl**uster **An**alysis of **S**equences) in order to visualize relationships of subgroups in protein families, and to identify uncharted regions or novel domain co-occurrences, applied in combination with remote homology detection methods.

1.6 Archaea

Archaea are single-celled microorganisms that do not contain a nucleus or any other membrane-bound organelle. The three-domain system of biological classification treats archaea as a third basic domain of life in addition to bacteria and eukaryotes [161]. It has replaced the five-kingdom taxonomy or the prokaryote-eukaryote dichotomy that were based on classical inspections of phenotypes, The tripartite division of the living

world was put forward by Carl Woese and uses molecular sequence comparison, originally the sequences of 16S ribosomal RNA [162].

1.6.1 Archaeal Phylogeny and Taxonomy

Molecular phylogeny subdivides the domain of archaea into two main taxa, the crenarchaeota and the euryarchaeota. Furthermore, a separate phylum is formed by *Nanoarchaeon equitans*, a symbiont of the crenarchaeon *Ignicoccus hospitalis* [163]. The availability of more completely sequenced genomes will provide clarity regarding yet unassigned groups like thaumarchaeota [164] and korarchaeota [165]. Currently, about 80 genome sequences of cultured archaea are available. Archaea contain some of the most extremophilic organisms, among them the most halophilic, thermophilic (*Methanopyrus kandleri* T=122°C), and acidophilic (*Picrophilus torridus*: pH=0) organisms. However, psychrophily is found among archaea as well as mesophiliy. Archaea contain two groups of methanogenic organisms [166] and contribute significantly to the earth's biomass [167].

1.6.2 Archaea and the Origin of Eukaryotes

Although the defining feature of eukaryotes, as implied by the name, is a nucleus, the endosymbiosis of an α-proteobacterium, which gave rise to the mitochondrion, is of crucial importance for the origin of the eukaryotic cell [168]. The mitochondrion is an ancestral trait of the eukaryotic lineage, which is supported by evidence that hydrogenosomes are degenerate mitochondria [169]. In general mitochondria are considered to provide an energetic advantage important for multicellularity [170]. The nature of the host of this endosymbiosis, however, is debated. Two alternative scenarios based on various gene phylogenies are offered, relying on different genes, multiple alignment techniques, and methods of tree reconstruction [171]. The three primary domains (3D) scenario states that host of was a primitive proto-eukaryote of largely unknown character implying that eukaryotes and archaea are two distinct sister lineages. In contrast, the two primary domains (2D) scenario postulates a unique endosymbiotic event according to which an archaeon ingested a proteobacterium, which proposes that eukaryotes derived from within archaea. The diversity of the scenarios illustrates inherent problems of using phylogenetic trees in order to resolve the deepest evolutionary relationships. This is due to the high frequency of horizontal gene transfer

at the earlier stages of life and to the fact that tree reconstruction methods are prone to artefacts regarding the deeper branches. The reconciliatory synthetic view of the ring of life hypothesis recognizes multiple prokaryotic sources of the eukaryotic cell and thereby refuses bifurcating trees for the deepest nodes on the basis of proposed genome fusion events [172].

Apart from evolutionary inventions specific to eukaryotes, e.g. endomembrane systems, spliceosomal introns, meiotic sex, sterol synthesis etc., the genome of the last common ancestor of all eukaryotes is inferred to have a chimeric character with various features originating from both, the archaeal and the bacterial lineage. The most general and oversimplifying distinction is that the majority of eukaryotic operational and metabolic systems can be traced back to bacteria whereas the information processing machinery (replication, transcription, translation, and protein quality control) shares greater similarity with the archaeal counterpart [173]. Additionally, there are important operational systems that are closely related between archaea and eukaryotes, including a cell division machinery in archaea homologous to the eukaryotic endosomal sorting complex. Although the number of bacteria-derived genes is higher than the number of archaea-derived genes (about 4x in yeast) in reconstructed genomes of the eukaryotic ancestor. Archaea-derived genes turned out to be significantly more important in terms of essentiality, expression level, and network connectedness [174]. Therefore, bacterial genes are considered to arrive relatively late in the process that defined the nucleus and the eukaryotic cell, favouring the 2D scenario according to which eukaryotes originated from within an already established archaeal lineage [175]. Recently, the 2D and 3D scenarios are challenged by the identification of bacterial planctomycetes that have an endocytosis-like ability [176], because of which they have been proposed as an alternative route to the eukaryotic cell [177].

Archaeal proteins play a central role in this work. We describe the characterization of three hypothetical archaeal proteins, which have retained ancestral, highly derived, or intermediate features and are therefore informative from an evolutionary perspective. In addition to their sometimes remote connection to the proteasome, their archaeal origin constitutes a common denominator of the proteins studied in the following projects: A CTP-specific archaeal riboflavin kinase that forms an evolutionary bridge between basal DNA-binding cradle-loop barrels and ATP-specific bacterial/eukaryotic riboflavin kinases – present in all archaea except for *Nanoarchaeum equitans* – (4.1); A β-clam protein with a homooligomeric β-propeller, which illustrates the evolution of

monomeric β-propellers via oligomeric precursors – present in pyrococci and thermococci – (4.2); A symmetric, six-stranded OB-fold with an internal sequence repeat, being potentially informative for the evolution of five-stranded OB-folds – present in both groups of methanogenic archaea – (4.3). Furthermore, we predict a regulatory network for proteasomal protein degradation in archaea (5.1), contrasting the fully differentiated 26S proteasome of eukaryotes.

2 AIMS AND CONTRIBUTIONS

The goal of the first part of this work was to investigate the divergent evolution of N-terminal substrate recognition domains of AAA proteins studzing three selected target proteins. Various N-domains have been recruited to the homologous core of AAA proteins, the AAA+ module, by independent evolutionary events conferring specificity to reactions catalyzed by sub-groups of this protein family. However, N-domains do not only provide input to AAA proteins; their homologs also occur in the context of other, unrelated domains or have acquired functions differing from those performed within AAA proteins. In order to explore their structural and functional diversity we selected three targets, hypothetical proteins, based on sequence analysis and characterized them experimentally: Mj0056, a homolog of double-ψ β-barrels with CTP-dependent riboflavin kinase activity; PH1500, a protein containing a β-clam domain and a homohexameric, twelve-bladed β-propeller involved in DNA-repair; Mj0548 (DUF2121), a protein combining a PAN-like OB-fold with a proteasome-like Ntn-hydrolase that lost the ability to self-compartmentalize.

The second part of this work aims at tracing the origins of proteasomal protein degradation. Proteasomal ATPases regulate access of substrates to the proteasome and stimulate gate opening through their C-terminal HbYX interaction motif. In the first project of this part we collect *in silico* evidence for the hypothesis that a network of AAA ATPases regulates the archaeal proteasome. Therefore, we conducted a kingdom-wide analysis of the C-termini of AAA proteins suggesting that, in addition to PAN, CDC48 and AMA proteins function as proteasomal ATPases in archaea. The second project analyses the genetic environment of Anbu, an uncharacterized proteasome homolog, predicting that the Anbu operon constitutes a tagging system for targeted protein degradation. Finally, we investigate the global distribution of proteasome-like Ntn-hydrolases and (putative) proteasomal ATPases on the tree of life.

2.1 Contributions

This PhD Thesis was conducted in the department of Prof Andrei Lupas, which provided a framework of experimental and evolutionary knowledge. The projects described in this work were developed on the basis of numerous discussions between Andrei Lupas and myself. It may be difficult to disentangle the individual ideas each of

us contributed in these discussions. However, Andrei Lupas' continuous scientific advice was substantial for this work. Furthermore, contributions of collaboration partners are of importance for projects of this thesis. These contributions will be stated in the following.

Mj0056 – A CTP-specific Archaeal Riboflavin Kinase (4.1):

The homology of Mj0056 and the double-ψ barrel domain of CDC48-like AAA proteins was identified by Andrei Lupas [49]. Sergej Djuranovic cloned the gene and purified the protein resulting in the NMR-structure solved by Murray Coles. In contrast to the initial prediction that Mj0056 functions as a transcription factor, the hypothesis that Mj0056 is an archaeal riboflavin kinase is my contribution. I purified Mj0056 using the expression construct provided by Sergej Djuranovic and I performed the biochemical characterization including the determination of its CTP-specificity. The latter was aided by the deposition of a crystal structure of Mj0056 by a structural genomics consortium (MCSG), because it contained a CDP-moiety. The MS-based riboflavin kinase assay was done in collaboration with Guido Sauer. I produced the protein for co-crystallization with CDP and FMN, which was done in collaboration with Marcus Hartmann who also solved the structures. In this respect the existence of the NMR-structure by Murray Coles and the crystal structure by the MCSG were of importance. Andrei Lupas recognized the evolutionary significance and embedded archaeal riboflavin kinases in his broad scenario for the evolution of the cradle-loop barrel metafold. This work resulted in a publication, on which I am the first author sharing equal contribution with Marcus Hartmann [4].

PH1500 – A β-Clam in a Homohexameric Twelve-bladed β-Propeller (4.2):

Andrei Lupas identified the presence of a β-clam domain in PH1500. Murray Coles and Ilka Varnay conducted the structural characterization of the isolated N- and C-terminal domains by NMR-spectroscopy using purified protein provided by Sergej Djuranovic. Inspecting the large pore of the β-propeller formed by the C-terminal domain of PH1500, Andrei Lupas coined the hypothesis that it could function as a PCNA-analog. Together with Johannes Schiff, who collaborated with me as a HiWi, we cloned full-length PH1500 – including the correct start codon – and the genetically coupled EndoIII, purified the proteins, and performed the pull down and DNA binding assays. Our preparation of the full-length protein led to the crystal structure of dodecameric

PH1500 solved by Marcus Hartmann using the NMR structure by Murray Coles *et al.* [178] as a model for molecular replacement. Marcus Hartmann provided the unpublished coordinates of AMA-N for the structure based-sequence alignment of β-clam domains and model of PH1500 tentatively placing DNA in the central pore. I performed sequence and structure analyses, which were guided by Andrei Lupas.

DUF2121 – A PAN-like OB-fold in a Monomeric Proteasome Homolog (4.3):

I performed a compehensive classification of proteasome-like Ntn- hydrolases and identified the relationship to the DUF2121 protein family including the presence of a putative OB-fold at the C-terminal end, which is similar to the OB-fold of PAN-like AAA proteins. I cloned, purified and characterized the ortholog from *M. jannaschii*. The crystallography was done in collaboration with Marcus Hartmann who solved the structure by MAD phasing. I conducted bioinformatics and structure analysis frequently consulting Andrei Lupas.

Regulation of the Archaeal Proteasome by a Network of AAA ATPases (5.1):

Andrei Lupas put forward the hypothesis that a network of AAA ATPases regulates the archaeal proteasome. I performed the systematic analysis of archaeal AAA proteins and their C-terminal interaction motifs. Together with Dara Forouzan, in charge, and we are currently collecting experimental evidence confirming the network hypothesis.

The Anbu Operon – A Tagging System for Targeted Degradation? (5.2):

I performed the analysis of the genetic context of the Anbu protease, which was previously identified as a homolog of the proteasome [179]. I proposed that the Anbu operon is a tagging system for targeted protein degradation. This hypothesis was further developed in discussions with Andrei Lupas who also initiated a collaboration (Medizinische Mikrobiologie, Universität Tübingen – Monika Schütz, Sebastian Klein, Ingo Autenrieth), which aims at providing support for the hypothesis through in vivo experiments in *Y. enterocolitica*. After this analysis had been performed, Iyer et al. [180] suggested that the Anbu operon rather functions as a peptide synthesis system. However, preliminary characterization of an Anbu protease points to a self-compartmentalizing phenotype.

Implications for the Evolution of Proteasomal Protein Degradation (5.3):
I investigated the global distribution of proteasome-like Ntn-hydrolases and putative proteasomal ATPases as a summary and extension of analyses presented in previous chapters, discussing the results with Andrei Lupas.

All analyses of sequences, structures and, their relationships, which are presented in the following, are my own work including all graphical representations.

3 GENERAL PROCEDURES

This chapter covers general aspects of gene cloning, protein production, purification and quality control that were constantly applied as part of sample preparations for biochemical and structural characterizations throughout the projects. Furthermore, the typical workflow of bioinformatic analyses is outlined. Descriptions of particular experiments, including functional assessment of target proteins with tailored assays, are detailed within individual sections of chapter 4.

3.1 Cloning and Expression of Target Genes

The coding DNA of target genes was amplified by polymerase chain reaction (PCR) with gene specific oligonucleotides [181]. Plasmid DNA or genomic DNA of organisms hosting the target gene served as templates for the reaction. Restriction sites engineered into end-specific primers were digested with type II restriction endonucleases from the PCR product, which was subsequently ligated into the multiple cloning site of an expression vector cleaved with appropriate restriction enzymes. Circular plasmids were brought into *E. coli* competent cells by heat-shock or chemical transformation. Positive clones were selected based on resistance against antibiotics and confirmed by DNA sequencing of the target gene region.

For each target gene, expression conditions were optimized in small-scale test expressions with respect to growth temperature, concentration of inducer, and *E. coli* expression strain. Standard conditions for large-scale expression included induction with 0.5 mM IPTG at an optical density of 1.0, and an expression temperature of 20°C for 16 hours. Harvested cells were lysed using a French press.

Labelling of proteins with selenomethionine for MAD phasing of X-ray diffraction data [182] was achieved by expression of the target gene in the methionine-auxtrophic strain *E. coli* B834 (DE3) grown in M9 minimal Medium supplemented with 4 mg/mL selenomethionine.

3.2 Protein Production, Purification and Quality Control

Protein purification strategies differed depending on the presence of an affinity tag, solubility and isoelectric point of the target protein. Parameters of proteins of interest were calculated with the ProtParam tool (http://expasy.org/tools/protparam.html). The standard purification procedure included the application of soluble fractions of cellular extracts to a Nickel affinity column followed by ion exchange and size-exclusion chromatography using Äkta FPLC devices (GE Healthcare). Buffer conditions and chromatography columns were chosen with respect to the target protein. At each step purity was monitored by SDS-PAGE analysis. Protein identity was confirmed by electrospray ionization mass-spectrometry on a Bruker HCTulra ion trap coupled to a nano liquid chromatography system. Guido Sauer, Department of Biochemistry, MPI for Developmental Biology, preformed mass spectrometry analyses.

The folding state of purified proteins was tested using circular dichroism (CD) spectroscopy (Jasco CD-Spectropolarimeter J-810) and fluorescence spectroscopy (Jasco FP-6500 Spektrofluorometer). CD-spectroscopy was used to estimate the secondary structure content of target proteins and compared to the results of sequence-based secondary structure prediction programs such as PSI-PRED [183]. Additionally heat denaturation was followed by CD-spectroscopy to determine the melting temperature of target proteins. Sigmoidal melting curves, indicative of cooperative unfolding, provided evidence for folding.

Hydrophobic burial of tryptophan residues was assayed by fluorescence spectroscopy and compared to chemically denatured samples of the same protein. Tryptophan fluorescence was excited at a wavelength of 293 nm. A significant red shift of the emission maximum (depending on the target protein) to 357 nm in samples treated with chaotropic agents indicated the formation of a hydrophobic core. Target proteins that passed the quality control were concentrated to final concentrations of 2 to 15 mg/mL according to the solubility, shock frozen and stored at -80°C.

3.3 Structural Analysis of Target Proteins

Analytical size exclusion chromatography was performed in order to determine the quaternary structure of target proteins. The choice of the gel-sizing column was governed by the expected size of the target. The molecular mass was determined with

calibration curves calculated from runs with appropriate standard proteins conducted at preferably similar conditions.

Proteins forming sufficiently large assemblies were subjected to negative staining electron microscopy intended to serve as a means of quality control. Dilution series of protein preparations were adsorbed to glow-discharged carbon-coated support films, washed with water and stained with 1-2% uranyl acetate. Samples were inspected under a Philips CM10 transmission electron microscope. Heinz Schwarz, Microscopy Unit MPI for Developmental Biology, performed electron microscopy.

In order to obtain high-resolution structural information target proteins were crystallized and subjected to X-ray crystallographic analysis. Crystallization trials were performed at 295 K with ca. 1000 conditions by mixing 400 nL of protein solution with 400 nL of reservoir solution in 96-well Corning 3550 plates using the honeybee 961 crystallization robot (Genomic Solutions). Drop images were obtained with the RockImager 54 device (Formulatrix) and visually inspected. Crystallization conditions were optimized using different protein concentrations and orthogonal screens detailing successful conditions. Diffraction measurements were conducted at the Swiss Light Source (Paul-Scherrer-Institut, Villigen Switzerland), Beamline X10SA equipped with a PILATUS 6M hybrid detector. Kerstin Bär and Reinhard Albrecht operated the crystallization robot. Marcus Hartmann solved the crystal structures.

3.4 Bioinformatics

Homologs of proteins of interest were gathered with BLAST and PSI-BLAST [184] searching the non-redundant protein database (nr) or selected taxa as available at the National Centre for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/). Remote homologs of known structure were detected with HHpred searching the Protein Data Bank clustered to a maximum pairwise sequence identity of 70% (http://toolkit.tuebingen.mpg.de/hhpred) [132, 133]. Representative sequences of homologous groups were collected with HHsenser [185] searching the non-redundant database of NCBI or subsets thereof like the non-redundant database of archaeal proteins (nr_arc).

The relationships in groups of (homologous) proteins were analyzed and visualized by clustering at variable P-value cut-offs using CLANS [160]. All-against-all comparisons of sequences for clustering were performed with BLAST or PSI-BLAST

(three iterations) using a BLOSUM80 amino acid substitution matrix (for the first iteration) [186].

Multiple sequence alignments were produced with a variety of methods depending on the degree of similarity among the proteins of interest. Highly similar sequences were aligned with MUSCLE or CLUSTALW [187, 188]. The alignment provided by HHpred was used for alignment of remote homologs enabling re-alignment with the maximum accuracy algorithm. Additionally, structure-based sequence alignments generated by DALI [189] or by manual superimposition with the Swiss-PDB Viewer [190] were taken into account. Multiple alignments containing highly diverse sequences were interactively generated relying on the methods outlined above. Aligned sequences were subjected to phylogenetic inference with the neighbour-joining method as implemented in the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) [159].

Structural comparisons were performed at the DALI server (http://ekhidna.biocenter.helsinki.fi/dali_server/) [189] and by interactive superimposition using the Swiss-PDB Viewer [190]. Modeller was used for homology modelling with template alignments provided by HHpred [133, 191]. Representations of protein structures were conducted with PyMol (http://www.pymol.org/).

Comparative genetic context analysis exploits gene order and distance, gene fusions, and co-occurrences in order to provide information about potential functional associations of genes of interest. Context analyses were performed using STRING (http://string-db.org/) [192], BioCyC (http://biocyc.org/) [193], and interactive genome browsing at the KEGG database (http://www.genome.jp/kegg/) [194] in order to obtain additional hints regarding the function of hypothetical proteins. The iToL platform was used to visualize gene content and phlogentic trees.

Furthermore, the Bioinformatics Toolkit of the MPI for Developmental Biology (http://toolkit.tuebingen.mpg.de/) [195] served as a platform for a variety of sequence analysis methods, including repeat detection (HHrepID) [196], alignment comparison (HHalign), coiled coil detection (Pcoils) [197], secondary structure/transmembrane/disorder prediction (Quick2D).

4 N-Domains of AAA Proteins in the Light of Evolution

4.1 The Double-Ψ Barrel is homologous to CTP-specific Riboflavin Kinases

The cradle-loop barrel metafold comprises three main folds of different topology sharing a duplicated ββαβ-element in their common core. The basal RIFT barrel fold gave rise to the swapped-hairpin barrel by strand invasion at the C-terminal ends of the two symmetry related halves [198, 199]. The double-ψ barrel is connected to the RIFT barrel by a swap of the second β-strand of the ββαβ-element enabled by a substantial elongation of the cradle-loops connecting the first two strands (Figure 6) [50, 54, 198].

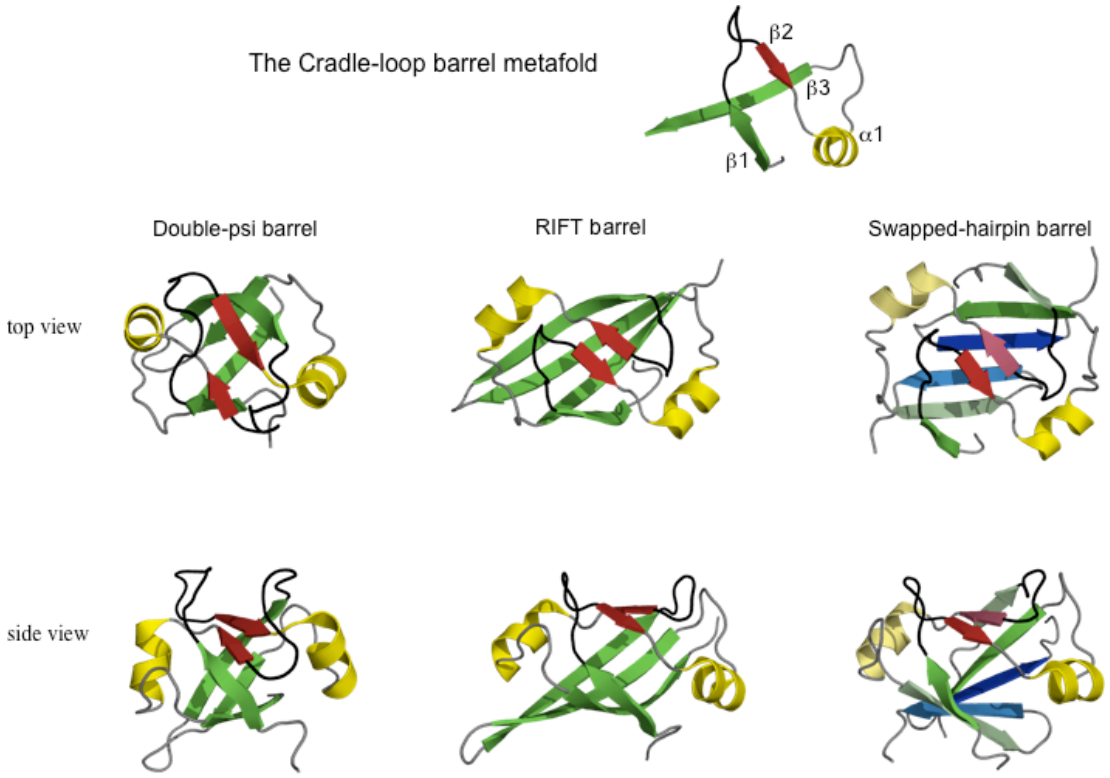


Figure 6. Gallery of the Cradle-loop Barrel Metafold

The gallery depicts the three basic topologies of cradle-loop barrels. The left panel shows the double-ψ barrel domain of VAT from *T. acidopilum* (1CZ4), the middle panel shows the RIFT barrel of the hypothetical protein Phs018 from *P. horikoshii* (2GLW), and the right panel shows the transition state regulator AbrB from *E. coli* (1YFB). All three folds display (pseudo)-twofold internal symmetry, indicating their evolution by duplication of an ancestral ββαβ-element (top). The strands β2 and β2' (coloured in red) of the double-ψ barrel are swapped with respect to the homologous elements in the RIFT barrel. The additional β-strands of the dimeric swapped hairpin barrel (blue) invade the paired strands β3 and β3` of the RIFT barrel. The second monomer of dimeric AbrB is distinguished by light colors. The cradle-loops are in black.

In addition to these three topologies, the cradle-loop barrel metafold accommodates various other topologies that are also related by events of evolutionary fold change

illustrating the evolution of a diverse group of proteins from a simple precursor [52]. This group is not only structurally diverse but also functionally as it contains transcription factors, protein interaction modules and various enzymes.

The name RIFT barrel indicates its presence in proteins like **ri**boflavin synthases, **F**1 ATPase, and **t**ranslation factors [198]. Because of the occurrence in these ancient proteins and because of its simpler topology the RIFT barrel is considered to be the ancestral form of cradle-loop barrels. Furthermore, it is adopted by eukaryotic and bacterial riboflavin kinases as well as the archaeal transcription factor PhS018. Despite the pronounced similarities residing in the structural core their sequences are only weakly similar. A group of archaeal proteins exemplified by Open Reading Frame (ORF) number 56 from *Methanocaldococcus jannaschii* (Mj0056) serves as a sequence intermediate linking eukaryotic and bacterial riboflavin kinases to basal cradle-loop barrels.

The Mj0056 group was initially predicted to function as transcription factors, because they are more similar in sequence to PhS018 and other transcription factors than to known **r**ibo**f**lavin **k**inases (RFKs) [54]. Here, we describe the characterization of Mj0056 revealing that it is a CTP-specific archaeal riboflavin kinase [4].

4.1.1 Experimental Procedures

Bioinformatics

Homologs of Mj0056 were gathered by searching the non-redundant protein sequence database at NCBI with HHsenser, a method for exhaustive transitive profile searches based on Hidden Markov Model (HMM) comparisons [185]. Six different starting points were used as queries for HHsenser runs with default settings: Mj0056, the Double-ψ barrel domain of the archaeal CDC48 homolog from *T. acidophilum* (VAT-Nn, PDB-ID 1CZ4, residues 1-94), the swapped-hairpin barrel domain of the transcription factor AbrB from *B. subtili*s (1YFB, 7-58), the RIFT barrel PhS018 from *P. horikoshii* (2GLW), riboflavin kinase from *S. pombe* (1N07), and the N-terminal RIFT barrel of riboflavin synthase from *S. pombe* (1KZL, 1-88). From each search the strict alignment was obtained and subsequently filtered with HHfilter for the approximately 300 most dissimilar sequences [195]. For Mj0056 and PhS018 the complete strict alignment were used, because of the smaller number of sequences

contained. This procedure yielded an array of 1452 unique sequences restricted to the domain of interest.

Two cluster maps were generated with CLANS employing BLOSUM80 as a substitution matrix [160]. For the first one BLAST [130] was applied as a tool for an all-against-all comparison, and sequences were clustered using P-values < 1. Usage of a more stringent P-value abolished most connections across the groups. Therefore a second NxN comparison was calculated with three iterations of PSI-BLAST [131] using an inclusion P-value of 10^{-3}. Clustering was performed at various P-values including $P<10^{-3}$ and $P<10^{-5}$. The increased sensitivity of PSI-BLAST and the usage of significant P-values clarify the patterns observed in the map based on BLAST.

Selection of sequences of basal-cradle-loop barrels and bacterial/eukaryotic RFKs for the multiple alignment was based on the presence of a high-resolution structure. Sequences of archaeal RFKs were selected such that all major phyla of the archaeal kingdom are covered. The multiple sequence alignment was interactively generated relying on a variety of methods. Within the groups - basal cradle-loop barrels, bacterial/eukaryotic RFKs, archaeal RFKs - the alignment was guided by alignments obtained from HHpred searches [132, 133]. For the alignment of critical positions across the groups a structure based-sequence alignment was taken into account. Therefore, all proteins of known structure contained within the multiple alignment were interactively superimposed using Swiss-PDB Viewer [190]. The repeats within basal cradle-loop barrels were identified with HHrepID [151, 196]. Analysis of the genetic context of Mj0056 and its orthologs was preformed using STRING and the KEGG database [192, 200].

Protein Production and Purification

The expression construct, the Mj0056 gene cloned into pET-30b was obtained from the in-house stock collection. It was previously generated by Sergej Djuranovic and used for the determination of the NMR structure by Murray Coles. The target protein was expressed in *E. coli* C41 (DE3) RIL which were grown at 37°C up to an OD of 1.0, induced with 0.5 mM IPTG and harvested after 4 hrs. Soluble fractions of cellular extracts were subjected to an anion exchange (MonoQ, Amersham) followed by a cation exchange chromatography (SP Sepharose FF, Amersham). Bound protein was eluted by a linear sodium chloride gradient from 50 mM to 1 M in Tris buffer (pH 6.8). Monitoring by SDS-PAGE indicated the presence of the target protein in a yellow and an uncoloured fraction, which were pooled separately. Both pools were heated to 80° C

for 20 min to precipitate thermolabile *E. coli* proteins, cooled to 4° C, and centrifuged. The yellow fraction was concentrated by ultrafiltration by using Vivaspin 10 kDa membranes and used directly for crystallization trials without additives. The uncoloured fraction was applied to a Superdex G-75 preparative column that had been equilibrated in 25 mM HEPES buffer (pH 7.4) containing 150 mM NaCl. Eluted fractions were tested by SDS-PAGE, combined, and concentrated with Vivaspin 10 kDa concentrators. The resulting solution was used for both enzymatic assays and crystallizations trials with various additives. The oligomeric state of pure uncoloured Mj0056 was analyzed on a calibrated analytical gel-sizing column (Superose 12, Amersham).

Riboflavin Kinase Assay

Riboflavin kinase activity was assayed in reaction mixtures containing 40 mM Tris/HCl (pH 8) buffer, 50 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 50 μM riboflavin, 3 mM nucleotide (ATP, CTP, GTP, or UTP), and 1 μM Mj0056. Reaction mixtures were incubated at various temperatures (25, 37, 50, 70, and 85°C) for 60 min and subsequently cooled to 4 C. Controls were processed identically but in the absence of enzyme. FMN controls contained 50 μM FMN instead of riboflavin. Riboflavin, FMN, ATP, CTP, and GTP were obtained from Sigma, UTP from Roth. 100 μL of reaction mixtures were desalted prior to MS analysis with C18 extraction tips and eluted in 50 ml 50% acetonitrile/0.1% formic acid. MS data was acquired on an HCT Ultra ion trap (Bruker Daltonics, Bremen) by electrospray ionization in alternating positive and negative ion mode. MS measurements were performed in collaboration with Guido Sauer.

X-Ray Crystallography

In all crystallization trials, 400 nL of protein solution were mixed with 400 nL of reservoir solution in 96-well Corning 3550 plates with 75 ml reservoir volume by using the honeybee 961 crystallization robot (Genomic Solutions). Drop images were obtained with the RockImager 54 device (Formulatrix) and visually inspected. From crystallization trials with various additives three structures were obtained. The Mj0056-MgCDP complex crystallized by mixing a reservoir solution containing 35% v/v MPD and 0.1 M imidazole with a protein solution additionally containing 10 mM MgCDP, 10 mM MgADP, and a saturated amount of riboflavin. The Mj0056-MgCDP-FMN complex crystallized by mixing a reservoir solution containing 20% w/v PEG8000 and

0.2 M NaCl with a protein solution additionally containing 10 mM MgCDP and 10 mM FMN. The colourless protein solution concentrated to 10 mg/mL in 25 mM Hepes pH 7.4 and 150 mM NaCl yielded these two structures. The Mj0056-NaCDP-PO_4 complex crystallized by mixing a reservoir solution containing 40% v/v ethylene glycol and 0.1 M phosphate-citrate (pH 4.2), 0.2 M NH_4SO_4 with the yellow fraction concentrated to 20 mg/mL. The crystal structures were solved and deposited by Marcus Hartmann (Mj0056-MgCDP: PDB-ID 2VBU, Mj0056-MgCDP-FMN: 2VBV, Mj0056-NaCDP-PO_4, 2VBT). Solving the structures was supported by a solution structure of Mj0056 (2P3M) that served as a search model for molecular replacement. Murray Coles determined the NMR structure.

4.1.2 Results

Bioinformatics

Searches with HHpred indicated sequence similarity of Mj0056 to transcription factors of the AbrB superfamily, to double-ψ barrels found in the N-terminal substrate recognition domains of AAA proteins like CDC48 and NSF, and to eukaryotic RFKs (Table 3). In order obtain a comprehensive overview of the relationships among these proteins, they were classified with two CLANS cluster maps (Figure 7). For the first one an all-against-all comparison was calculated with BLAST, for the second map comparisons were calculated with three iterations of PSI-BLAST. Both maps show five major groups representing the starting points for searches with HHsenser. Most tightly connected are the double-ψ barrels of AAA proteins and the swapped hairpin barrels of the AbrB superfamily. Simpler RIFT barrels like Phs018 are immersed within the latter. The Mj0056 group, however, displays stronger connections (note the darker lines) to these basal cradle-loop barrels than to eukaryotic and bacterial RFKs providing an evolutionary bridge.

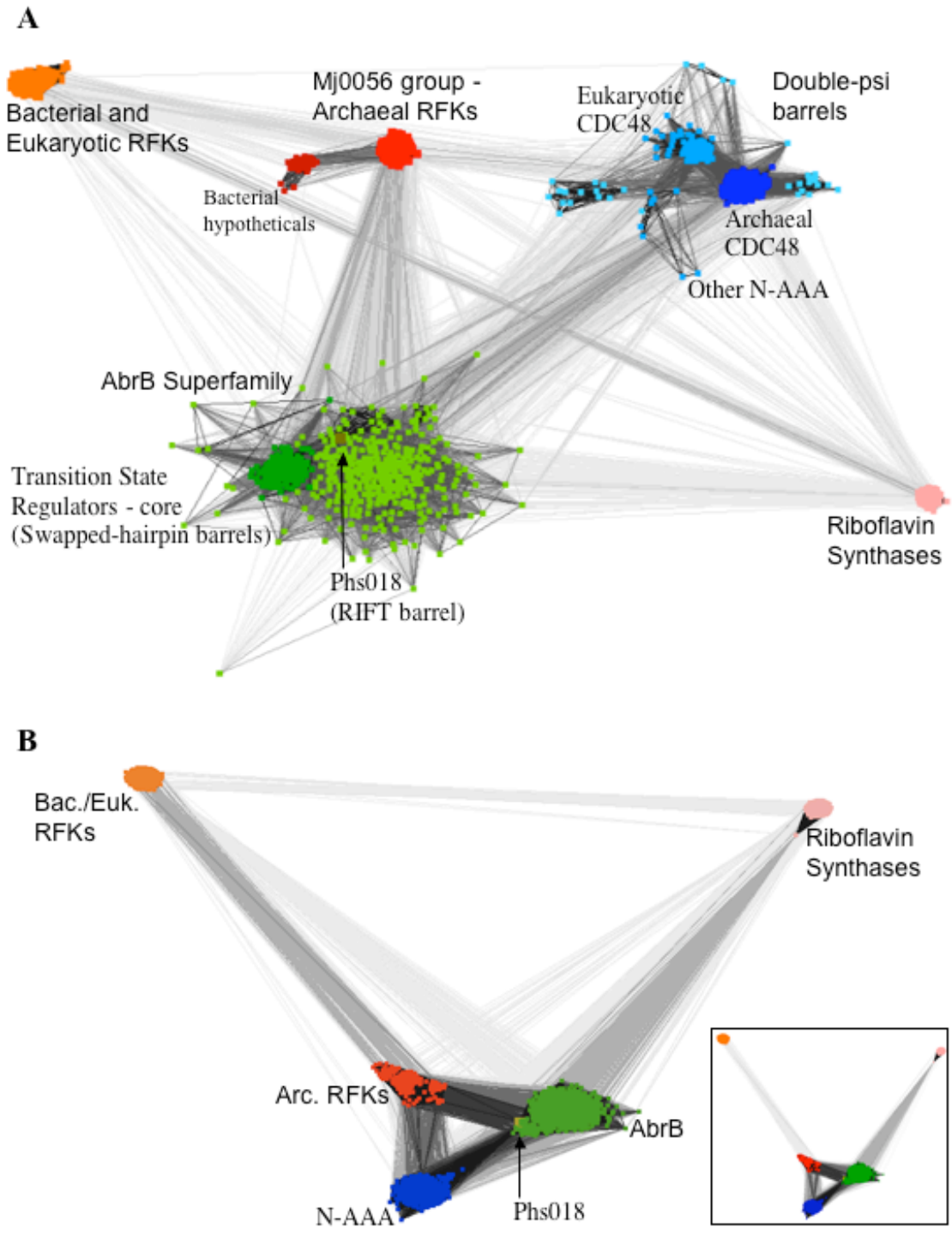


Figure 7. Cluster Map of Mj0056 Homologs

The CLANS cluster maps illustrate the relationships of the Mj0056 group to other groups of the cradle-loop barrel metafold. (A) Pairwise similarities of all sequences were computed with BLAST. Sequences were clustered at P-values < 1 indicated by grey lines between two dots, representing a pair of sequences that share a P-value better than the cutoff. Darker lines signify stronger similarities. (B) Similarities for the identical set of sequences were computed with three iterations of Psi-Blast at an inclusion P-value of 0.001 that was also used for clustering. Relying on the greater sensitivity of PSI-BLAST panel (B) underlines that archaeal RFKs are strongly related to basal cradle-loop barrels, and that they link them to eukaryotic and bacterial RFKs. At more restrictive P-values eukaryotic RFKs are exclusively connected via archaeal RFKs (see the inlet that only shows connections better than $P<10^{-5}$). Both maps contain 1452 sequences.

These observations are supported by the usage of various P-value cutoffs at which clustering was performed. Whereas the BLAST-based map (Figure 7 A) uses and shows all connections with a P-value <1, the PSI-BLAST based map (Figure 7 B) only uses significant P-values better than 10^{-3} or 10^{-5} (Figure 7 B inlet). The increased sensitivity of PSI-BLAST allows for a tighter packing of certain groups if the members share a similarity better than the inclusion P-value of 10^{-3}. This leads to a more focused appearance of the groups and to a stronger connection of the Mj0056 group to basal cradle-loop barrels (Figure 7 B). At a P-value of 10^{-5} eukaryotic RFKs are only bound to the core clusters of cradle-loop barrels via the Mj0056 group. Interestingly, the only remaining attachment of Riboflavin synthases to the core group at this P-value is to the AbrB superfamily, but their similarity requires further corroboration in order to answer the question whether riboflavin synthases are indeed homologous to these groups of cradle-loop barrels.

A multiple alignment was generated for a more detailed investigation of the sequence similarity between the Mj0056 group, eukaryotic RFKs and basal cradle-loop barrels. The alignment revealed that the C-terminal ββαβ-element of the Mj0056 group shows a more pronounced sequence similarity to both repeats of the ββαβ-element of basal cradle-loop barrels including a highly conserved ariginine and the GD-box [201] preceding β3 (Figure 8). In eukaryotic RFKs this region has diverged substantially partly explaining the clustering behaviour. Despite the presence of two large insertions in the N-terminal ββαβ-element of Mj0056, there are two conserved patterns shared with eukaryotic RFKs. Both patterns, four glycine residues involved in the formation of the pocket binding the isoalloxazine moiety and the TxN motif responsible for coordination of the γ-phosphate of ATP in eukaryotic RFKs, are also present in the Mj0056 group. The conservation of these functional residues supports the hypothesis that Mj0056 functions as an RFK.

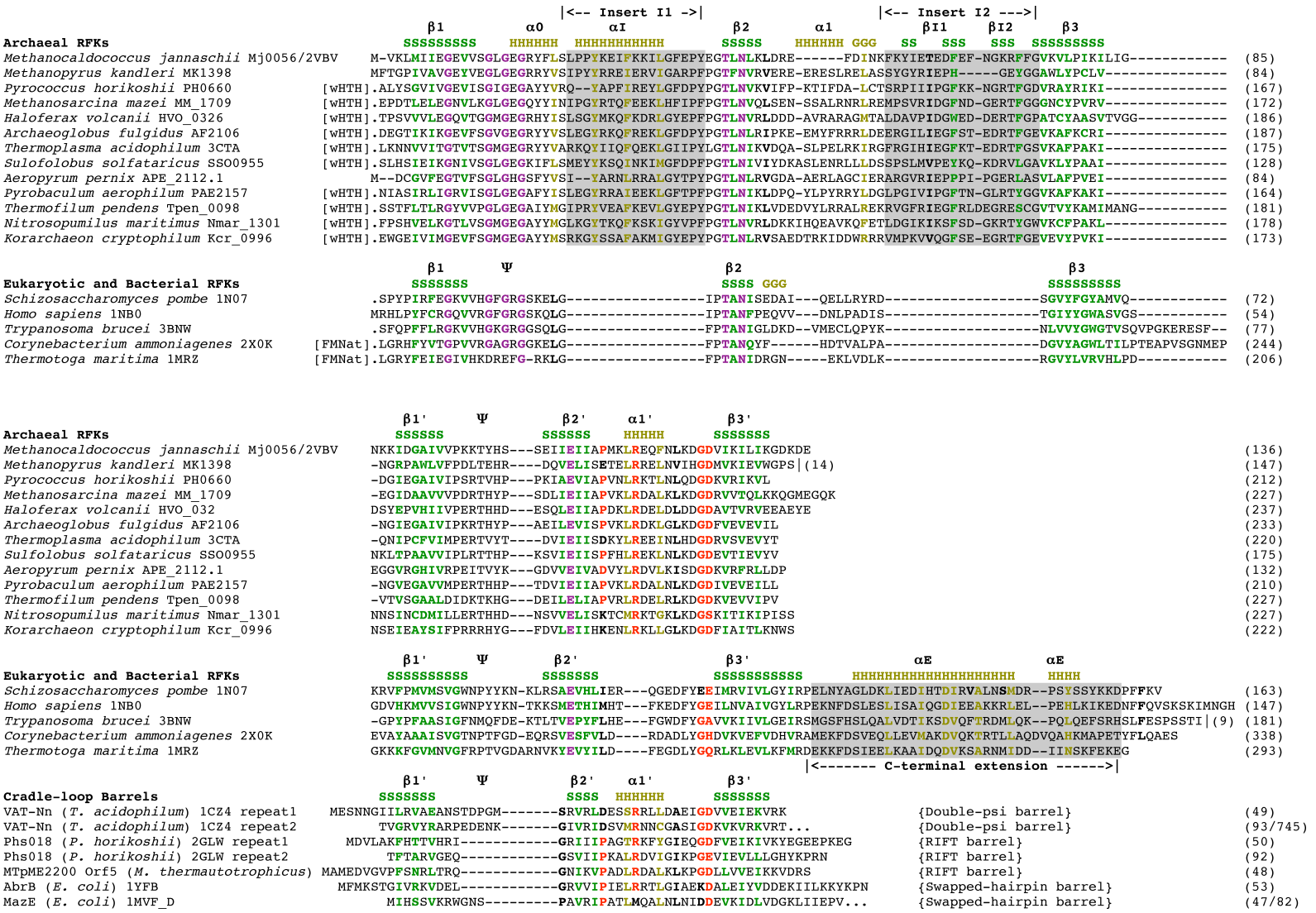

|<-- Insert I1 ->|
|<-- Insert I2 --->|
β1 α0 αI β2 α1 βI1 βI2 β3
Archaeal RFKs
Methanocaldococcus jannaschii Mj0056/2VBV
Methanopyrus kandleri MK1398
Pyrococcus horikoshii PH0660
Methanosarcina mazei MM_1709
Haloferax volcanii HVO_0326
Archaeoglobus fulgidus AF2106
Thermoplasma acidophilum 3CTA
Sulofolobus solfataricus SSO0955
Aeropyrum pernix APE_2112.1
Pyrobaculum aerophilum PAE2157
Thermofilum pendens Tpen_0098
Nitrosopumilus maritimus Nmar_1301
Korarchaeon cryptophilum Kcr_0996
β1 Ψ β2 β3
Eukaryotic and Bacterial RFKs
Schizosaccharomyces pombe 1N07
Homo sapiens 1NB0
Trypanosoma brucei 3BNW
Corynebacterium ammoniagenes 2X0K
Thermotoga maritima 1MRZ
β1' Ψ β2' α1' β3'
Archaeal RFKs
Methanocaldococcus jannaschii Mj0056/2VBV
Methanopyrus kandleri MK1398
Pyrococcus horikoshii PH0660
Methanosarcina mazei MM_1709
Haloferax volcanii HVO_032
Archaeoglobus fulgidus AF2106
Thermoplasma acidophilum 3CTA
Sulfolobus solfataricus SSO0955
Aeropyrum pernix APE_2112.1
Pyrobaculum aerophilum PAE2157
Thermofilum pendens Tpen_0098
Nitrosopumilus maritimus Nmar_1301
Korarchaeon cryptophilum Kcr_0996
β1' Ψ β2' β3' αE αE
Eukaryotic and Bacterial RFKs
Schizosaccharomyces pombe 1N07
Homo sapiens 1NB0
Trypanosoma brucei 3BNW
Corynebacterium ammoniagenes 2X0K
Thermotoga maritima 1MRZ
|<------- C-terminal extension ------>|
β1' Ψ β2' α1' β3'
Cradle-loop Barrels
VAT-Nn (T. acidophilum) 1CZ4 repeat1
VAT-Nn (T. acidophilum) 1CZ4 repeat2
Phs018 (P. horikoshii) 2GLW repeat1
Phs018 (P. horikoshii) 2GLW repeat2
MTpME2200 Orf5 (M. thermautotrophicus)
AbrB (E. coli) 1YFB
MazE (E. coli) 1MVF_D
{Double-psi barrel}
{Double-psi barrel}
{RIFT barrel}
{RIFT barrel}
{RIFT barrel}
{Swapped-hairpin barrel}
{Swapped-hairpin barrel}

Figure 8. Multiple Alignment of Archaeal Riboflavin Kinases and their Homologs

The alignment contains riboflavin kinases from all major archaeal phyla, eukaryotic and bacterial riboflavin kinases of known structure, and five basal cradle-loop barrels. The latter are aligned to the C-terminal halves of RFKs to which they show significant sequence similarity. Residues contributing to the hydrophobic core are shown in bold and coloured according to their occurrence in secondary structure elements (green for strands, yellow for helices). Secondary structure assignment is based on DSSP of known structures (H, helix; S, strand; G, 3_{10} helix). Residues conserved across major groups are in red. Grey boxes mark two insertions in archaeal RFKs and the C-terminal extension of eukaryotic and bacterial RFKs, which are structurally equivalent. Locus tags and PDB-identifiers are denoted next to the name of the respective organism. Co-occurring N-terminal domains are abbreviated in angular brackets: wHTH, winged Helix-Turn-Helix; FMNat, FMN acetyl transferase.

Evidence from the genetic context of Mj0056 and its orthologs reinforced this hypothesis. In many archaeal organisms the Mj0056 ortholog is found in an operon with 3,4-dihydroxy-2-butanone-4-phosphate synthase (DHBP synthase, gene *ribB*), an enzyme that catalyzes the reaction two steps before the riboflavin kinase reaction in FAD biosynthesis [202]. This is the case for example in *M. jannaschii* (ORFs Mj0056, Mj0055), *Methanopyrus kandleri* (ORFs MK1396 and MK1398), *Archaeoglobus fulgidus* (Af2106, Af2107) and *Methanosarcina mazei* (Mm_1709, Mm_1710). Furthermore, the riboflavin kinase was elusive within the archaeal kingdom, and several members of the Mj0056 group contain a winged helix-turn-helix domain involved in sequence-specific DNA-binding additionally favouring the idea that Mj0056 functions rather as an RFK than a transcription factor (Figure 8).

Mj0056 is a CTP-specific Riboflavin Kinase

A riboflavin kinase (RFK) is a phosphotransferase catalyzing the production of flavin mononucleotid (FMN) from riboflavin and a donor nucleotide [203] (Figure 9 A). The hypothesis that Mj0056 is an archaeal RFK was tested in a Mass Spectrometry (MS)-based assay.

During sample preparation a fraction containing the target protein exhibited a bright yellow colour. Its analysis by UV/VIS Absorption spectroscopy revealed two bands at 350 and 420 nm in addition to the most intense peak at 280 nm, which is caused by aromatic amino acids and generally observed in absorption spectra of proteins. The presence of these bands at lower energies corresponds to those seen in absorption spectra of free riboflavin and (FMN), in which they are caused by the shared isoalloxazin moiety. This was taken as initial evidence that Mj0056 catalyzes the riboflavin kinase reaction. Nevertheless, in the MS-based assay the formation of FMN could not be detected.

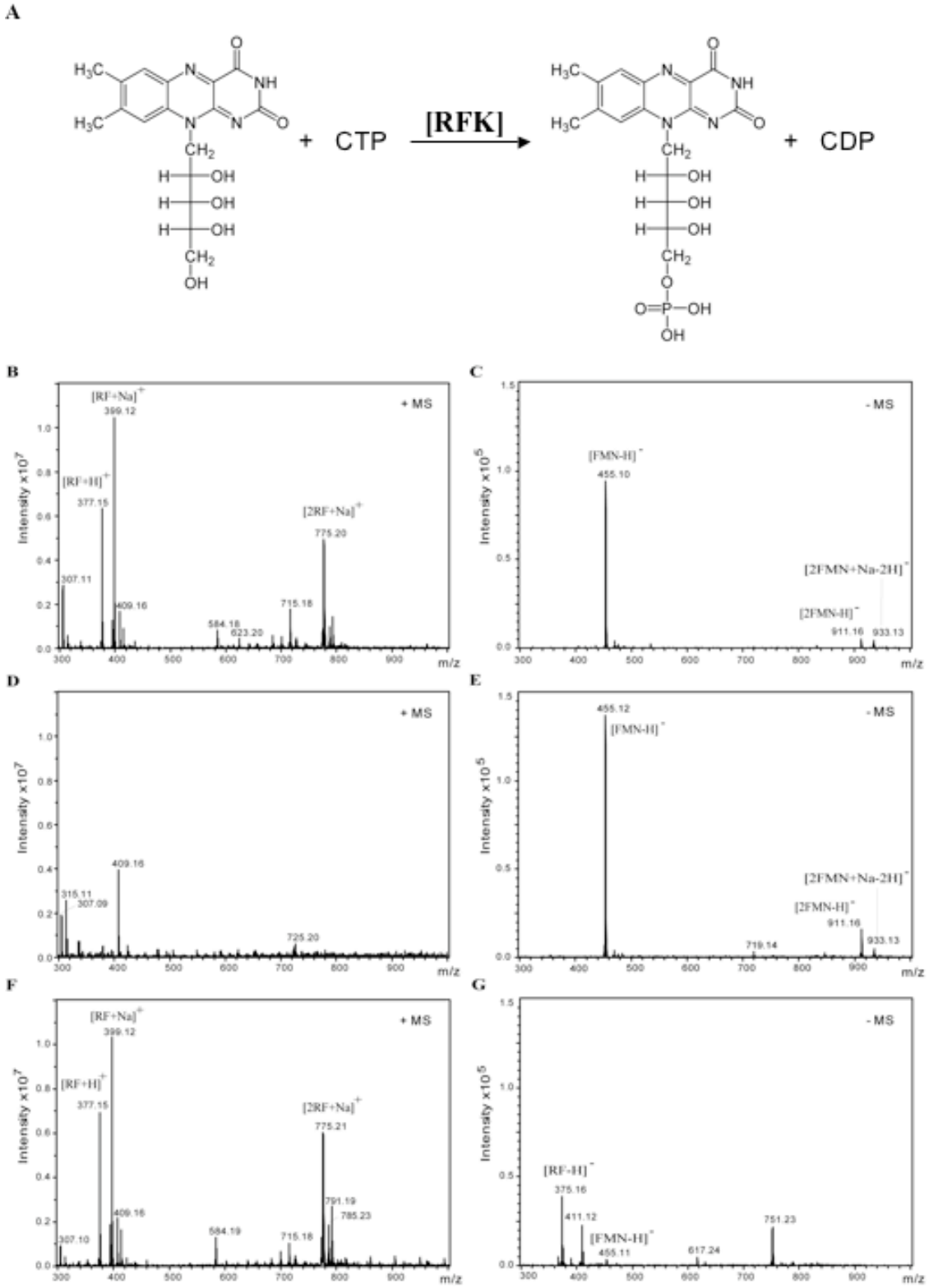


Figure 9. Mj0056 is a CTP-specific Riboflavin Kinase

Panel (A) shows the reaction catalyzed by archaeal riboflavin kinases. Panels (B) and (C) show ESI-MS spectra of riboflavin and FMN controls detected in positive and negative ion mode, respectively. Adducts of Riboflavin and FMN are annotated in brackets and assigned to the corresponding peaks. Panels (D) and (E) show the positive and negative ion mode MS spectra of samples containing Mj0056, riboflavin and CTP. Complete conversion of riboflavin to FMN is observed after 60 min at 50°C. Panels (F) and (G) show the spectra of samples containing ATP as donor nucleotide instead of CTP. No product formation is detected indicating the specificity of Mj0056 for CTP.

A crucial hint was obtained from a crystal structure of Mj0056 deposited by a structural genomics consortium. In this structure (PDB-ID 2OYN) Mj0056 was in complex with CDP. Therefore the functional assay was henceforth conducted with a range of donor nucleotides. In electrospray MS, adducts of Riboflavin (MW=376 g/mol) are detected at mass to charge ratios of 377, 399, and 775, adducts of FMN (MW=456 g/mol) at 455, 911 and 933 (Figure 9 B and C). Whereas riboflavin is identifiable in standard positive ion mode, FMN shows a response only in negative ion mode due the addition of the negatively charged phosphate moiety [204]. Formation of FMN was observed with CTP (Figure 9 D and E) and UTP. The latter, however, supports product formation approximately one order of magnitude less efficiently. Usage of ATP (Figure 9 F and G) and GTP did not yield significant amounts of FMN at any reaction temperature up to 85°C, the temperature of the natural habitat of *M. jannaschii*. In contrast, excess of CTP leads to a complete conversion of riboflavin to FMN after 60 min reaction time at all tested temperatures underlining the specificity of Mj0056 for this nucleotide.

Structure of Mj0056

After the activity of Mj0056 had been established, crystallographic experiments were conducted in order to study the binding modes of the reaction partners. Therefore Mj0056 crystals were grown with various additives among them ADP, CDP, Riboflavin, and FMN. Crystal structures were obtained for Mj0056 in complex with CDP and in complex with CDP and FMN. Furthermore, the structure of the yellow fraction, crystallized without additives, contained CDP and inorganic phosphate but no flavin moiety as expected from the absorption spectrum.

Mj0056 forms a six-stranded β-barrel adopting the topology of a RIFT barrel (Figure 10). The fold is constructed on the basis of two ββαβ-elements that shape the core of the cradle-loop barrel metafold. Whereas the C-terminal ββαβ-element is quite conserved in sequence and structure, the N-terminal half has substantially diverged by acquisition of two large insertions (see the multiple alignment, Figure 8).

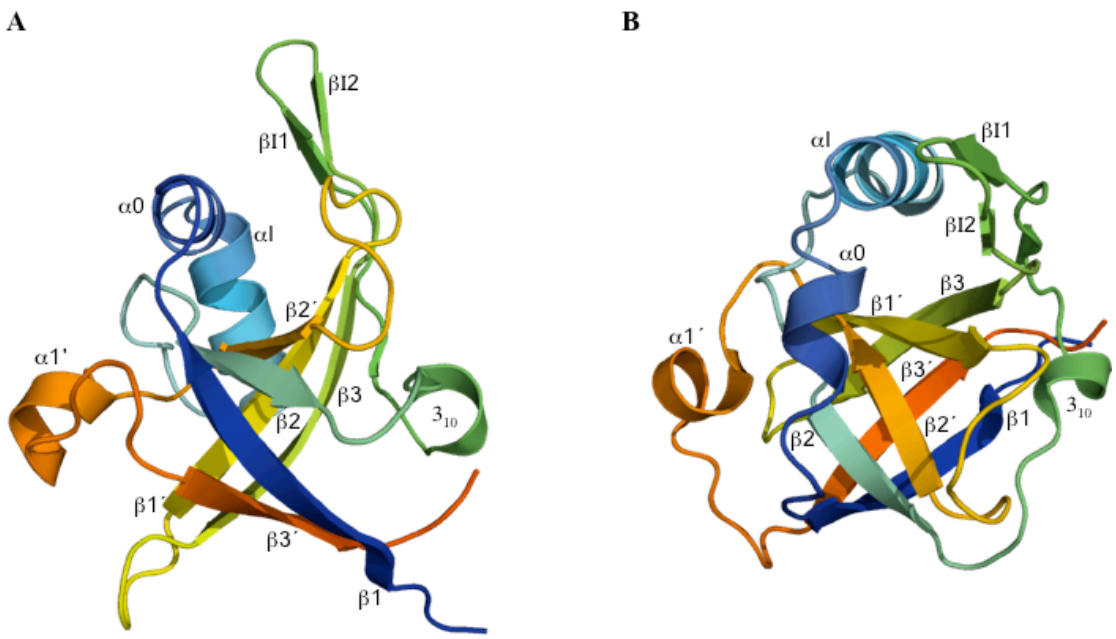


Figure 10. Overall Structure of Mj0056

The structure of Mj0056 is represented as a secondary structure cartoon in rainbow colouring starting with blue at the N-terminus and ending with red at the C-terminus. (A) and (B) show a side view and a top view, respectively. Mj0056 adopts a RIFT barrel fold, according to which the secondary structure elements are annotated. Secondary structure elements that are inserted into the basal RIFT barrel fold are denoted with "I". In contrast to most archaeal RFKs, Helix α1 of Mj0056 is degenerate to a 3_{10} helix.

The first insertion to the N-terminal half is located within the cradle-loop following β1 and leading into β2. It consists of two helices, α0 and αI, the first one of which is formed upon ligand binding as revealed by a comparison to the solution structure in the apo state (see below). The second insertion comprises an extended β-hairpin (βI1-βI2) leading directly into β3, which accompanied the loss of a type II β-turn that is usually found in cradle-loop barrels at this position. The beginning of the second insertion is marked by a 3_{10} helix. In Mj0056 this region is shortened in contrast to most of the archaeal RFKs in which a helix is found that is elongated with respect to α1 of basal cradle-loop barrel. The structure of the ortholog from *T. acidophilum* (PDB-ID 3CTA, deposited, to be published by Bonanno *et al.*) illustrates that this helix participates in the formation of the interface to the N-terminal wHTH domain. Nevertheless, among archaeal RFKs the short 3_{10} helix is a derived feature of methanococci – organisms closely related to *M. jannaschii* - because most archaeal RFKs contain a proper α-helix and not only members that contain the N-terminal wHTH domain (Figure 8).

The C-terminal half of Mj0056 is built around an extended GD-box that links β2' to β3' via helix α1'. The GD-box is a supersecondary structure element connecting two unpaired β-strands via an orthogonal type II β-turn [201]. It is

frequently found in cradle-loop barrels and forms the basis for the significant sequence and structure similarity between archaeal RFKs and basal cradle-loop barrels. A second conserved string of amino acids is represented by the PxxxR motif. Despite the similarity in sequence this motif adopts a different conformation in archaeal RFKs playing an important role in binding the cytosine moiety.

Table 3. Sequence and Structure Comparisons of Cradle-loop Barrels

		HHPRED-SCORES[1]			DALI-SCORES[2]			
Protein	**PDB-ID**	**Prob. [%]**	**E-Value**	**P-Value**	**Z-Score**	**RMSD [A]**	**Lali[2]**	**Fold**
Archaeal RFK	**2VBV**	**100**	**0**	**0**	**29.5**	**0.0**	**134/134**	**RIFT-barrel**
Archaeal RFK	3CTA	100	0	0	8.3	2.2	94	RIFT-barrel
AbrB-N	1YFB	95.1	0.03	5.8e-07	-	-	-	Swapped-hairpin barrel
PhS018	2GLW	93.6	0.09	1.9e-06	4.5	3.2	77	RIFT-barrel
MazE	1MVF_D	92.0	0.25	4.9e-06	-	-	-	Swapped-hairpin barrel
CDC48-Nn	1CZ4	56.2	11	0.0002	3.5	3.2	72	Double-Psi barrel
Eukaryotic RFK	1N07[3]	37.7	0.32	1.3e-05	5.2	3.3	91	RIFT-barrel
Bacterial RFK	1MRZ[3]	36.8	0.57	2.3e-05	5.1	2.7	75	RIFT-barrel
Eukaryotic RFK	3BNW[3]	22.1	0.49	2.0e-05	3.6	3.9	85	RIFT-barrel
Eukaryotic RFK	1NB0	-	-	-	4.8	3.1	84	RIFT-barrel
Bacterial RFK	2X0K	-	-	-	4.1	2.7	78	RIFT-barrel

The targets were taken from the HHpred and Dali hit lists and represent a highly diverse selection of proteins, which are included in the multiple alignment (Figures 8), rather than the ranking by the servers. Despite low scores for certain targets the degree of sequence and structure similarity is indicative of homology.
1: HHpred searches were performed in default settings against the Protein Data Bank, release of May 14 2011, filtered for a maximum of 70% pairwise sequence identity at http://toolkit.tuebingen.mpg.de/HHpred.
2: Dali searches were done at http://ekhidna.biocenter.helsinki.fi/dali_server/. Lali denotes the number of residues included in the superimposition.
3: These HHpred hits are only obtained when HHpred is run in Global Alignment Mode.

Conformational changes upon ligand binding

The structure of Mj0056 in complex with MgCDP reveals the architecture of the Nucleotide binding site and conformational changes involved in binding. The Mg^{2+} ion is mainly bound by the conserved TLN motif (T43-N45) in strand β2 and coordinates the α- and β-phosphates of CDP, which also interact with the glycine-rich motif (G14-G18) at the beginning of the first cradle-loop. In this region the formation of helix α0 (E17-S23) is induced representing a difference to the solution structure in the apo state

where this region forms an extended loop (Figure 11 A and B). Residue R115 in α1' is of importance for the interaction with the ribose providing a rationale for its invariance in all archaeal RFKs. The cytosine ring packs on one side against Y40 in the first cradle-loop where all archaeal RFKs contain an aromatic amino acid. On the other side of the cytosine ring, the PxxxR motif comprising the loop between β2' and α1' of the symmetry related half determines the specificity for CTP (see below).

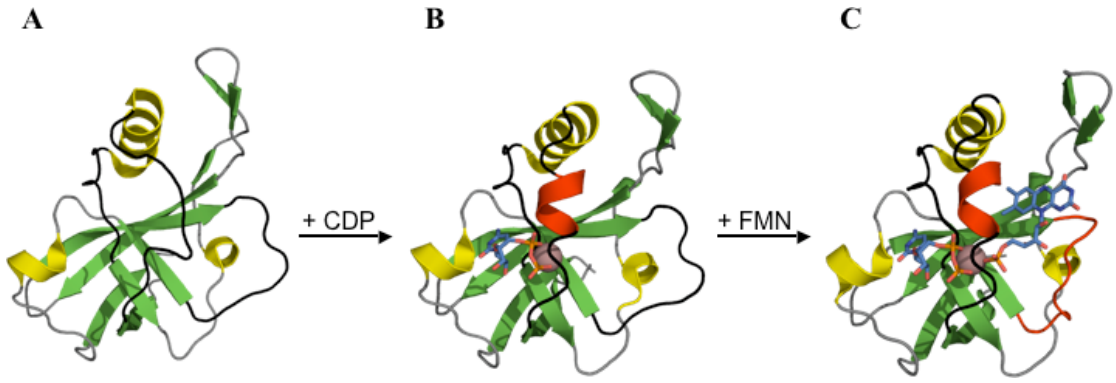


Figure 11. Conformational Changes of Mj0056 upon Ligand Binding

The apo-strucure of Mj0056 solved by NMR-spectroscopy (2P3M) is shown in (A). The complex of Mj0056 with Mg-CDP (2VBU) is shown in (B). Nucleotide binding induces formation of the transient helix a0 in the first cradle loop highlighted in red (B). Additional binding of FMN leads to the closure of the second cradle-loop demonstrated by the complex of Mj0056 with Mg-CDP and FMN (2VBV, C). The change with respect to (B) is indicated by the red colour of the second cradle-loop in (C). Because FMN-binding experiments were only successful in the presence of CDP, the sequential substrate binding cycle of Mj0056 from (A) to (C) is proposed.

The ternary complex of Mj0056, MgCDP, and FMN elucidates the binding mode of the reaction product FMN (Figure 11 C). Both insertions into the N-terminal ββαβ-element contribute to the confinement of the isoalloxazine moiety. The terminal phosphate group coordinates the Mg^{2+} ion and is positioned by E107 that interacts with the 4' and 5' oxygens of FMN. The crucial position of this residue and its invariance in archaeal RFKs suggests that it plays an important role in the kinase reaction, perhaps by functioning as a base that activates the terminal hydroxyl group of Riboflavin for nucleophilic attack. Furthermore, the second cradle-loop contacts all moieties of FMN. It adopts a different conformation in the ternary complex in comparison to the binary complex providing closure to the FMN binding site (Figure 11 B and C). The closed conformation of the second cradle-loop is also observed in structure of Mj0056 in complex with CDP and inorganic phosphate that was obtained from the yellow fraction mentioned above.

The fact that co-crystallization efforts with Mj0056 and the flavin alone -neither with FMN nor with the poorly soluble riboflavin - were not successful suggests a sequential substrate binding cycle. First, the binding of CTP induces the formation of helix α0 in the first cradle-loop, which appears to provide the basis for flavin binding. Furthermore, only the nucleotide confers the ability to appropriately mount the Mg^{2+} ion within the active site, a crucial factor in this kinase reaction. Subsequent binding of riboflavin involves the closure of the second cradle-loop, which may trigger the transfer of the phosphate from CTP to riboflavin.

Analytical size-exclusion chromatography indicated that Mj0056 has a slight tendency to form dimers. This tendency is reflected in the crystal structures of the ternary complex and of the NaCDP-PO_4-complex (yellow fraction). The dimerization occurs via antiparallel pairing of the β1 strands and does not include functionally important regions. Furthermore, the structure of the binary complex is monomeric. Therefore enzymatic activity should not be dependent on dimerization. In orthologs containing a wHTH domain at the N-terminus, dimerization might be of functional relevance. Winged-HTH domains are known to dimerize upon sequence-specific DNA-binding, and the crystal structure of the RFK from *T. acidophilum* shows dimerization through the wHTH domain. Whether there is a co-regulation of RFK and transcription factor activity in these proteins - perhaps in form of an up-regulation of the expression of genes involved in FAD biosynthesis (including the operon encoding itself and the *ribB* gene) - remains to be studied.

4.1.3 Discussion

The structural characterization of Mj0056 bound to the products of the riboflavin kinase reaction, its specificity for CTP as a donor nucleotide, and its intermediate position between ATP-dependent RFKs and basal cradle-loop barrels provide mechanistic and evolutionary implications.

Comparison of CTP- and ATP-dependent Riboflavin Kinases

Archaeal RFKs on the one hand and bacterial and eukaryotic RFKs on the other hand do not share a high degree of overall sequence similarity, which is reflected by the low scores retrieved from the most sensitive sequence comparison method. HHpred returns

a probability of 60% and an E-value of 10 for the RFK from Fission Yeast (1N08) using Mj0056 as query, and fails to properly detect the second insertion (Table 3). Nevertheless, residues of crucial importance for the kinase reaction are highly conserved and placed in equivalent positions. Among them is the glycine-rich (G14 - G18) motif in the first cradle-loop, the TxN (T43-N45) motif that coordinates the Mg^{2+} ion leading into β2, and the invariant E107 interacting with the 4' and 5' oxygens of FMN. These motifs are also placed in equivalent positions in a superposition of structures from both groups. The RFK from fission yeast (1N07, [205]) can be superimposed onto Mj0056 with an RMSD of 1.27 Å over 53 residues (Figure 12 A) catching the conserved core of the RIFT barrel fold and the phosphate transfer site. Taken together this evidence argues for the homology of both types of RFKs.

The architectures of the binding sites of both reaction partners show considerable differences outside of the centre involved in phosphate transfer. In Mj0056 the cytosine moiety is sandwiched between Y40 and L44, and encompassed by the PxxxR motif in the region around helix α1'. In contrast, eukaryotic and bacterial RFKs do not contain tyrosine 40, which is part of the first insertion, and have an alanine in position of L44. Most importantly the deletion of three residues between β2' and β3' leads to the resolution of helix α1' into an extended loop allowing the accommodation of the larger adenine moiety by eukaryotic and bacterial RFKs (Figure 12 B). The degree of conservation around the PxxxR motif suggests that all archaeal RFKs are specific for CTP as the donor nucleotide.

For the formation of the flavin-binding site of both groups, elaborations to the RIFT barrel fold play a crucial role. In archaeal RFKs the inserts I1 and I2 embed the flavin, in common RFKs the helical extension at the C-terminus provides enclosure of the flavin. Although these elaborations are of independent origin, they are found in strikingly similar locations (Figure 12 C and Figure 8). Nevertheless, their analogous origin also results in slightly differing positions of FMN.

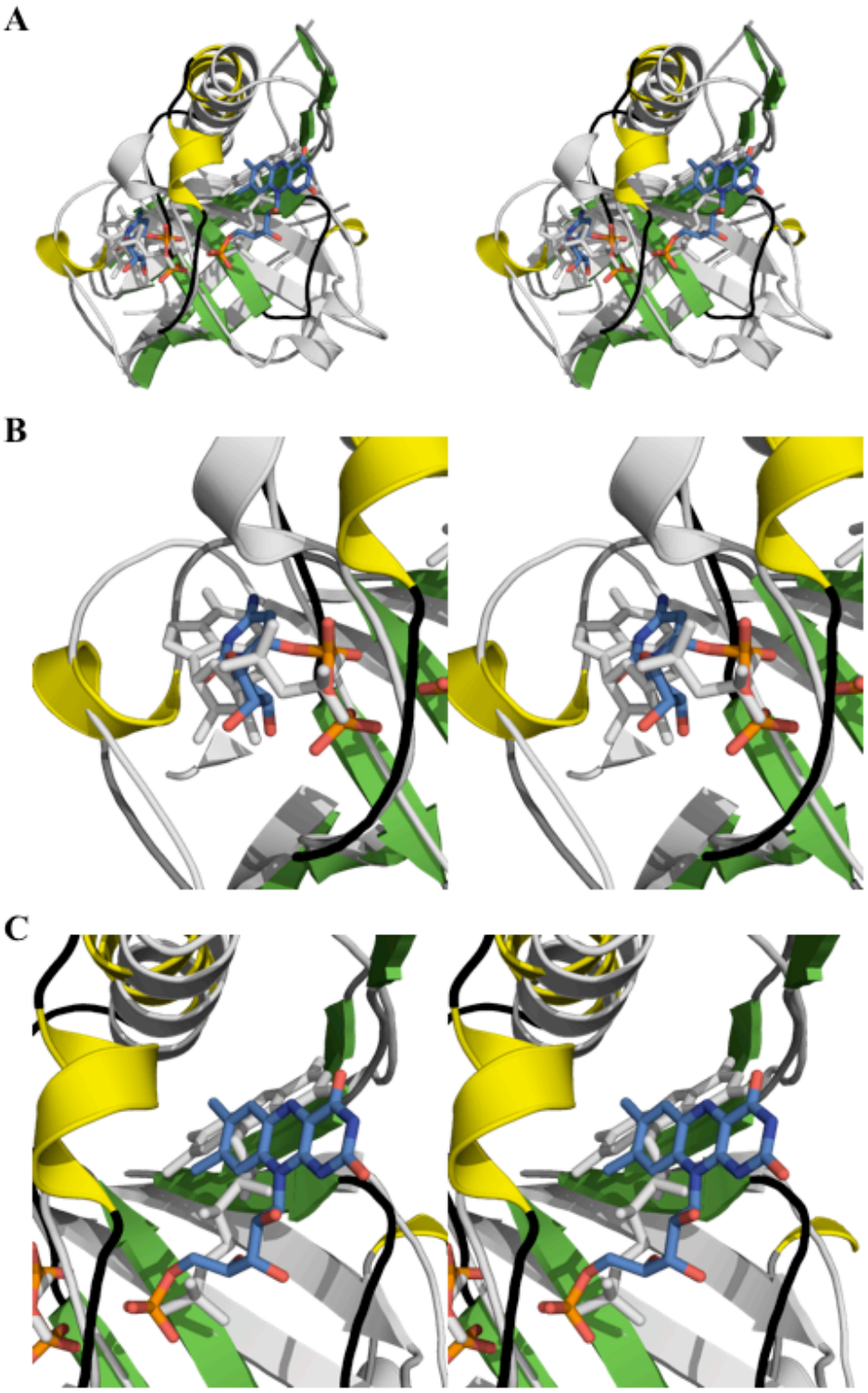


Figure 12. Structural Comparison of Archaeal and Bacterial/Eukaryotic RFKs

Mj0056 in complex with MgCDP and FMN (2VBV) is compared to the riboflavin kinase of *S. pombe* bound to MgADP and FMN (1N07, coloured in light grey including ligands). The overall stereo view of a superposition of the two structures (A) illustrates their similarity. 53 residues contributing to the core of the RIFT barrel fold are superimposed with an RMSD of 1.27 A. Panel (B) shows a close up stereo view of the nucleotide binding sites. Decomposition of helix α1` into a widened loop allows for the accommodation of the larger adenine moiety in SpRFK. Presence of α1` in Mj0056 provides the basis for CTP-specificity. The close up stereo view of the FMN binding sites (C) indicates the structural equivalence of the inserted helix αI in Mj0056 with the helix αE of the C-terminal extension in the eukaryotic homolog.

Evolution of Enzymatic from DNA-binding Activity

The core of the cradle loop barrel metafold is formed by the ββαβ-element. The unifying activity of basal members of the metafold is DNA-binding, which is the case for dimeric swapped-hairpin barrels and simple RIFT barrels like Phs018 [198, 206]. The latter arose from a dimeric RIFT barrel ancestor by duplication and fusion of the ββαβ-element because the internal repeat is significantly detected on the sequence level and reflected by a clear two-fold pseudo-symmetry in structure. Subsequent divergence led to a reduction in internal symmetry and acquisition of enzymatic activity on the evolutionary trajectory to riboflavin kinase activity (Figure 13).

Characterization of Mj0056 provides insights into the evolutionary changes necessary to convert a transcription factor into an enzyme, because it forms an evolutionary bridge between basal cradle-loop barrels and bacterial and eukaryotic RFKs (Figures 7, 8). Most of the changes necessary to bring about nucleotide binding capability in Mj0056 occurred in the N-terminal half. This included the acquisition of insert I1 resulting in a considerable elongation of the first cradle-loop. Parts of the insertion contribute to the formation of the cytosine-binding pocket, especially Y40. Together with the attainment of the TLN motif it might have enabled the binding of CTP, because T43 and N45 coordinate the Mg^{2+} ion together with the phosphate groups while L44 and Y40 sandwich the base. At this point, no considerable changes were required in the C-terminal half in order to establish CTP binding. This is reflected by the similarity of the second ββαβ-element of archaeal RFKs to both repeats of basal cradle-loop barrels. Therefore, we conclude that CTP-specificity used to be ancestral to ATP-specificity (Figure 13). Continuing this scenario, the appearance of a base in β2' -in extant RFKs most likely E107- might have formed a phosphotransfer centre yielding a primordial CTP-dependent kinase with the ability to phosphorylate a variety of small molecules.

Increase in specificity towards riboflavin was achieved by changes in the second cradle-loop, growth of insert I1, and emergence of insert I2. At this stage, archaeal and bacterial RFKs must have already diverged from their common ancestor, because specificity towards riboflavin is achieved on different routes in both lineages. Whereas the second cradle-loop still retains a certain sequence similarity in both, the elongated insert I1 and insert I2 of archaeal RFKs clearly represent an analogous development to the helical extension αE of the bacterial RFKs. Nevertheless, both elaborations occupy similar positions in structure (Figure 12 C) and converge on conferring specificity to

riboflavin. ATP-specificity, however, is a derived property of bacterial RFKs. The increased steric requirements of the adenine moiety were matched by a deletion that led to the loss of helix α1' resulting in an extended loop just wide enough to accommodate adenine (Figure 12 B and alignment Figure 8).

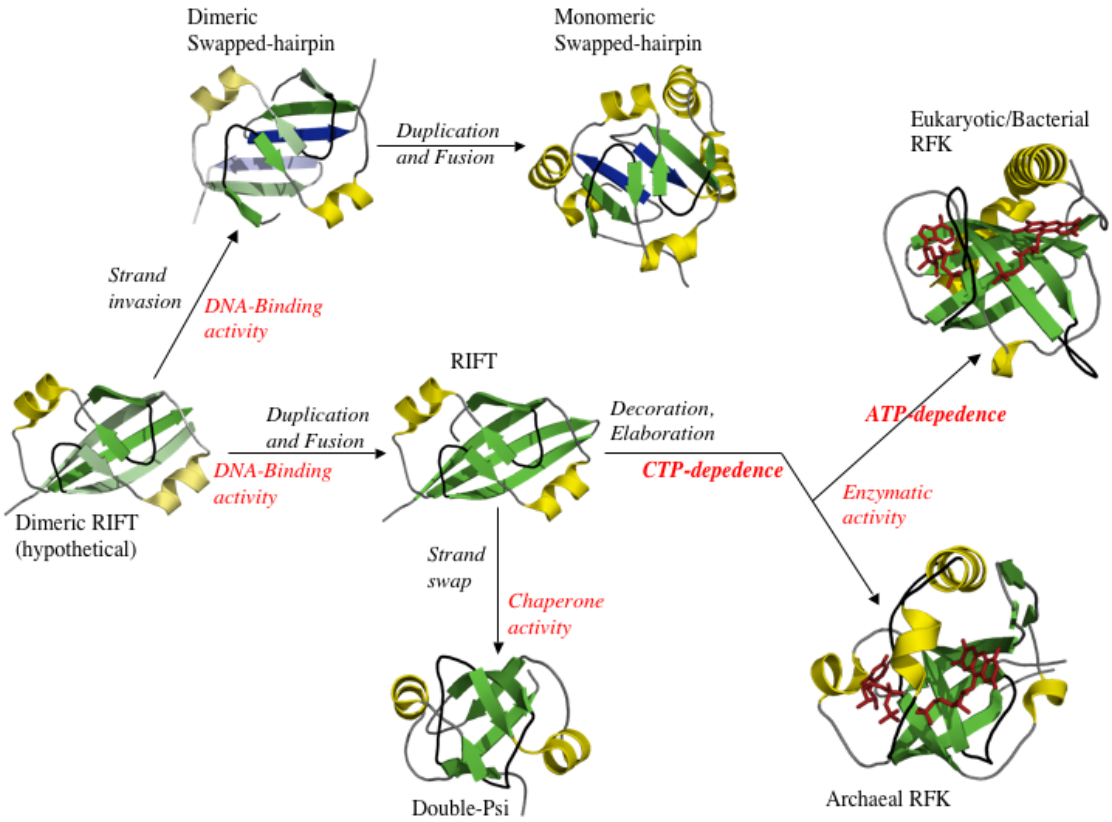


Figure 13. A scenario for the evolution of RFKs within the Cradle-loop barrel metafold
The scenario illustrates the topologies of cradle-loop barrels and the evolution of RFK enzymes from simpler DNA-binding ancestors. A hypothetical dimeric RIFT barrel ancestor gave rise to dimeric swapped hairpin barrels (AbrB, 1YFB) by strand invasion (invading strands are in blue), and to the monomeric RIFT barrel (PhS018, 2GLW) by duplication and fusion. The protein Mtpme2200 Orf5 consisting of only one ββαβ-element is a candidate for such an ancestor (model of PhS018, second monomer in light colours). Another gene duplication event led to the monomeric swapped hairpin barrels (MraZ, 1N0E). Double-ψ barrels (VAT-Nn, 1CZ4) evolved from monomeric RIFT barrels by a swap of the strands β2 and β2` enabled by elongation of the cradle loops (black in all structures). This was accompanied by a gain of chaperone activity. The RFKs describe an evolutionary path from DNA-binding RIFT barrels to elaborated RIFT barrels with enzymatic activity, both catalyzing a similar reaction, but with specificity for different donor nucleotides (ATP in euRFK, 1N07; CTP in arcRFK, 2VBV; ligands in red), of which CTP may represent the ancestral property. Double-ψ barrels encountered another route to enzymatic activity, which is omitted for clarity.

These ancient events have supposedly taken place well before the emergence of the eukaryotic cell. In the light of this scenario, the tight grouping of bacterial and eukaryotic RFKs suggests that the bacterial endosymbiont, which evolved into the mitochondrion, contributed its RFK to the genetic repertoire of the eukaryotic ancestor.

Regardless of the three-domain or two-domain view of the tree of life [171, 175], the gene encoding for RFKs adds another example to the concept, according to which most of the operational genes of the eukaryotic ancestor share greater similarity with bacterial homologs, whereas the most important informational genes (e.g. RNA polymerase, Ribosome) are more closely related to their archaeal homologs [173]. Because RFKs catalyze the production of redox cofactors (FMN and subsequently FAD) playing important roles in energy metabolism, they are generally assigned to the operational group of genes.

Specificity for CTP is a rather unusual property among kinases. A comprehensive classification of kinases describes 25 families, only one of which is specific for CTP [207, 208]. Dolichol kinase is an α-helical transmembrane protein and should therefore constitute an analogous development [209].

4.1.4 Conclusions

In conclusion, this is the first characterization of an archaeal riboflavin kinase providing an evolutionary link between highly symmetric DNA-binding cradle-loop barrels and bacterial/eukaryotic riboflavin kinases. Comparison of the molecular architectures of RFKs illustrates divergent and convergent elements in the evolution of these enzymes. Whereas both main lineages of RFKs independently evolved specificity for a flavin, CTP-specificity was an ancestral property that gave divergently rise to ATP-dependence. More general, these results underline the relevance of evolutionary intermediates for understanding structural and functional diversity of proteins.

4.2 A β-Clam in a Homohexameric Twelve-bladed β-Propeller

The β-clam domain is found in AAA proteins like CDC48 (cell division cycle 48), PEX1 (peroxisome biogenesis factor 1), and NSF (N-ethylmaleimide sensitive fusion protein) [49, 50]. In these proteins the β-clam occurs in tandem with a double-ψ barrel domain. In proteins of the AMA group, encoded by archaeoglobales and methanogenic archaea, the double-ψ barrel domain is not present, and they contain only one AAA+ module in contrast to CDC48, PEX1, or NSF [39]. Eukaryotic UFD1 (**u**biquitin **f**usion and **d**egradation) proteins, however, have the tandem of double-ψ barrel and β-clam, but the AAA+ modules are replaced by a disordered region.

A previously unobserved combination of the β-clam domain is detected in hyperthermophilic archaea of the clade of pyrococci. The N-terminal domain of this protein family shows clear sequence similarity to other members of the β-clam fold, but the C-terminal part does not display significant similarity to other proteins in the database. Only its predicted secondary structure indicated a resemblance to the OB (oligonucleotide and oligosaccharide binding)-fold [60]. In contrast to the β-clam domain, the OB-fold domain is widely distributed and also found in the N-terminal part of proteasomal ATPases, in which the OB-fold is involved in binding to proteins destined for degradation [3]. Therefore, the this family was considered to form an interesting co-occurrence of two domains so far known to operate as substrate recognition domains in separate subgroups of AAA proteins in absence of an ATPase.

The characterization of a member of this family, Open Reading Frame (ORF) number 1500 from *Pyrococcus horikoshii* (PH1500, PH1498.1n after reannotation), shows that the N-domain indeed forms a β-clam domain. The C-terminal part folds as a homohexameric β-propeller with the unusual number of twelve blades [178], shedding light on the evolution of the vast group of β-propellers via oligomeric intermediates (in the following we refer to PH1500 as **HP12** – **h**omohexameric **p**ropeller with **12** baldes). Furthermore, interaction with a genetically coupled type III endonuclease (EndoIII) implicates HP12 in DNA repair.

4.2.1 Experimental Procedures

Bioinformatics

Homologs of HP12 were gathered by searching the non-redundant protein sequence database at NCBI with HHsenser, a method for exhaustive transitive profile searches based on Hidden Markov Model (HMM) comparisons [185]. Five different starting points, identified by HHpred searches and the SCOP database (entry d.31.1) were used as queries for HHsenser runs with default settings: the b-clam domain of HP12, of the archaeal CDC48 homolog from *T. acidophilum* (VAT-Nc, PDB-ID 1CZ4, residues 95-185), of NSF from *C. griseus* (NSF-Nc, 1QCS, 86-201), of PEX-1 from *M. musculus* (1WLF, 100-179), and of UFD1 from *S. cerevisiae* (1ZC1, 121-208). From each search the strict alignment was obtained. The pool of sequences was filtered for duplicates resulting in an array of 1630 sequences. Sequences were clustered with CLANS using BLAST with a BLOSUM80 substitution matrix as a comparison tool. Clustering was performed at default parameters using P-values < 1 [160].

Selection of sequences for the multiple alignment was based on the presence of a high-resolution structure for CDC48, NSF/SEC18, and UFD1. Furthermore, the two paralogs of CDC48 in *Pyrococcus horikoshii*, two CDC48 from actinobacterial organisms, and four diverse sequences representing the AMA group were selected. For the HP12 group all sequences were used with exception of the ortholog from *Thermococcus sp.* AM4. The multiple sequence alignment was interactively generated relying on a variety of methods. Alignments obtained from HHpred searches served as starting points [133]. Closely related sequences were aligned with MUSCLE. For the alignment of critical positions across the groups a structure based-sequence alignment was taken into account. Therefore, the proteins of known structure contained within the multiple alignment were superimposed using Swiss-Pdb Viewer [190].

Analysis of the genetic context of HP12 orthologs was done with STRING and the KEGG database [192, 194]. The DALI server was used for searches for similar structures [189]. Structural alignments were interactively conducted with the Swiss-Pdb Viewer guided by HHpred and DALI searches. HHrepID was used for the identification of repeats within proteins [196]. Representations of protein structures were done with PyMol (www.pymol.org).

Protein Production and Purification

The DNA sequences encoding HP12 (GI: 3257925) and EndoIII (GI: 14591284) were amplified from genomic DNA of *P. horikoshii* by polymerase chain reaction (PCR) and cloned into pET30b and pET28b expression vectors (Novagen) with NdeI and XhoI restriction sites using the following primers: HP12-for 5'-GGAATCCATATGTCGGAGCTGAAGTTAAAGCCG-3′ HP12-rev 5'-CGCCTCGAGTAACGTTCTGATAAGGGTAAGTTTTTG-3′, EndoIII-for 5′-GGCGGCCATATGAACAAAAACTTACCCTTATCAG-3′, EndoIII-rev 5′-CGCCTCGAGTTATTGGCTAGAGGTATCCTGAACGCC-3′ (restriction sites underscored). In case of HP12 an alternative start codon was used six residues downstream of the sequence initially deposited at NCBI (PH1500). The target proteins were expressed in *E. coli* BL21 (DE3) RIL cells, which were grown in LB-medium at 37°C up to an OD of 1.0 (with the addition of 0.1% glucose in case of EndoIII), induced with 0.5 mM IPTG, and harvested after over night expression at 20°C (HP12), or 1 hour expression at 37°C (EndoIII). Soluble fractions of cellular extracts were subjected to a Ni2+ affinity column. Bound protein was eluted with a linear imidazole gradient. Fractions containing the target protein (as monitored by SDS-PAGE analysis) were heated to 80°C for 20 min to precipitate thermolabile E. coli, cooled to 4°C and centrifuged. The supernatant was subjected to size-exclusion chromatography (Superdex 200, Amersham). Using Amicon ultrafiltration devices pure protein was concentrated to 10 mg/mL (HP12) and 0.5 mg/mL (EndoIII), respectively.

DNA binding assays

Electrophoretic mobility shift assays were performed in agarose gels stained with ethidium bromide. Prior to gel electrophresis reaction mixtures containing 20 nM EndoIII, 20 nM HP12 (hexamer), 20 mM Tris (pH7.8), 150 mM NaCl, 5% (v/v) glycerol, 5% β–Mercaptoethanol, and different DNA constructs (2:1 molar excess protein:DNA) were incubated for 20 min at 40°C. Reductive trapping of EndoIII on DNA constructs containing 5,6-Dihydrodeoxyuridine was achieved by addition of 50 mM sodium borohydride ($NaBH_4$) to the buffered solution described above. The reaction proceeded until completion at 4°C as monitored by SDS-PAGE analysis.

Co-Immunoprecipitation

Purified His_6-tagged EndoIII (10 μg) and un-tagged HP12 (30 μg) were incubated in binding buffer containing 20 mM Tris (pH7.8), 150 mM NaCl, 5% (v/v) glycerol, 5% β–Mercaptoethanol for 1h at 42°C. 2.5 μg of Anti-His antibody (Invitrogen) were added to the mixture, incubated for 1h at 4°C, followed by the addition of 20 μL of Protein A Sepharose beads (GE Healthcare). Complexes were washed three times, eluted with 4x SDS sample buffer and analyzed by SDS-PAGE.

X-Ray Crystallography

For crystallization 400 nL of HP12, which was concentrated to 5 mg/mL in a buffered solution containing 20 mM Tris (pH7.8) and 150 mM NaCl, were mixed with 400 nL of reservoir solution in 96-well plates with 75 mL reservoir volume by using a honeybee crystallization robot (Genomic Solutions). Drop images were obtained with the RockImager device (Formulatrix) and visually inspected. Full-length HP12 crystallized by mixing the protein solution with a reservoir solution containing 20% w/v polyethylene glycol 3350 and 0.2 M sodium chloride. The crystal structure of dodecameric HP12 was solved by Marcus Hartmann using the NMR-structure of hexameric HP12-C (determined by Ilka Varnay and Murray Coles) as a search model for molecular replacement.

4.2.2 Results

Bioinformatics

In order to clarify the relationships between the various proteins containing a β-clam domain, we performed a comprehensive clustering of all detectable β-clam homologs using CLANS [160]. The analysis, based just on the sequences of β-clams, returns separate groups for all clades of AAA proteins that have a β-clam domain (Figure 14). The position of UFD1 proteins in the map and their presence being restricted to eukaryotes suggests a secondary loss of both AAA+ modules contained by NSF, PEX1, AFG2, CDC48 and spermatogenesis factors. Of these, CDC48-like proteins form the central and most tightly connected clusters in the map separated by the lineage in which they are found. Archaeal CDC48 proteins are the dominant group, because all archaeal organisms contain at least one gene encoding for CDC48, some possess up to four

closely related CDC48 paralogs. In bacteria, CDC48 is only present in a few organisms, most of which are actinobacteria (see Figure 41). Although AMA proteins represent the ATPase with the simplest domain architecture within the map, their restricted occurrence suggests that they are a derived feature, which evolved in the common ancestor of archaeoglobales and methanogenic archaea (although we cannot exclude that AMA represents an ancestral form of CDC48). A similar pattern most likely applies to HP12 proteins that are solely found in all sequenced pyrococci. The HP12 group forms a small satellite cluster to archaeal CDC48 proteins. The strongest connections are observed to CDC48 proteins of pyroccoci and other closely related archaeal organisms pointing to a recent origin of this family within the lineage of pyrococci, which involved a loss of the AAA+ modules and the acquisition of a different C-terminal domain.

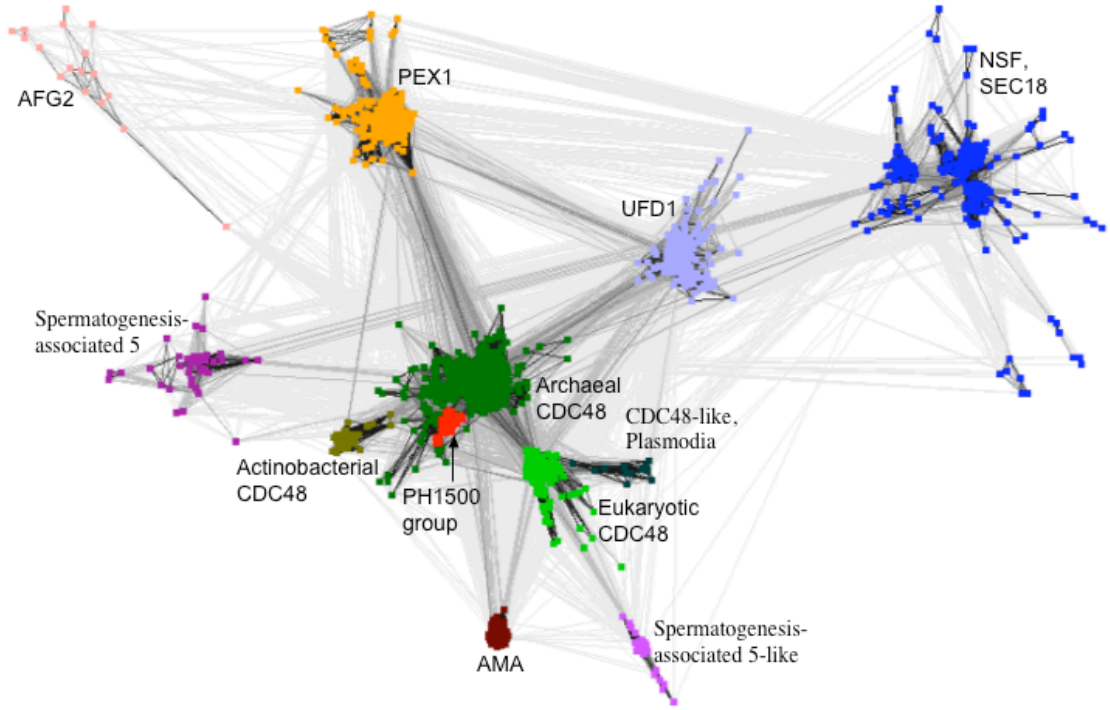


Figure 14. Cluster Map of β-Clam Domains

The cluster analysis illustrates the relationships between homologous β-clam domains. They group according to the protein family, in which they occur. The HP12 group (red) clusters tightly with archaeal CDC48 proteins (green). Darker lines indicate lower BLAST P-values. The clustering uses all P-values < 1. The map contains 1608 sequences.

The multiple alignment of the N-domains of all members of the HP12 group and a diverse set of β-clam domains representing the major clusters elucidates a number of conserved motifs including a GD-box between β2 and β3 and patterns of hydrophobic residues that match across the groups (Figure 15 A). The loops connecting the

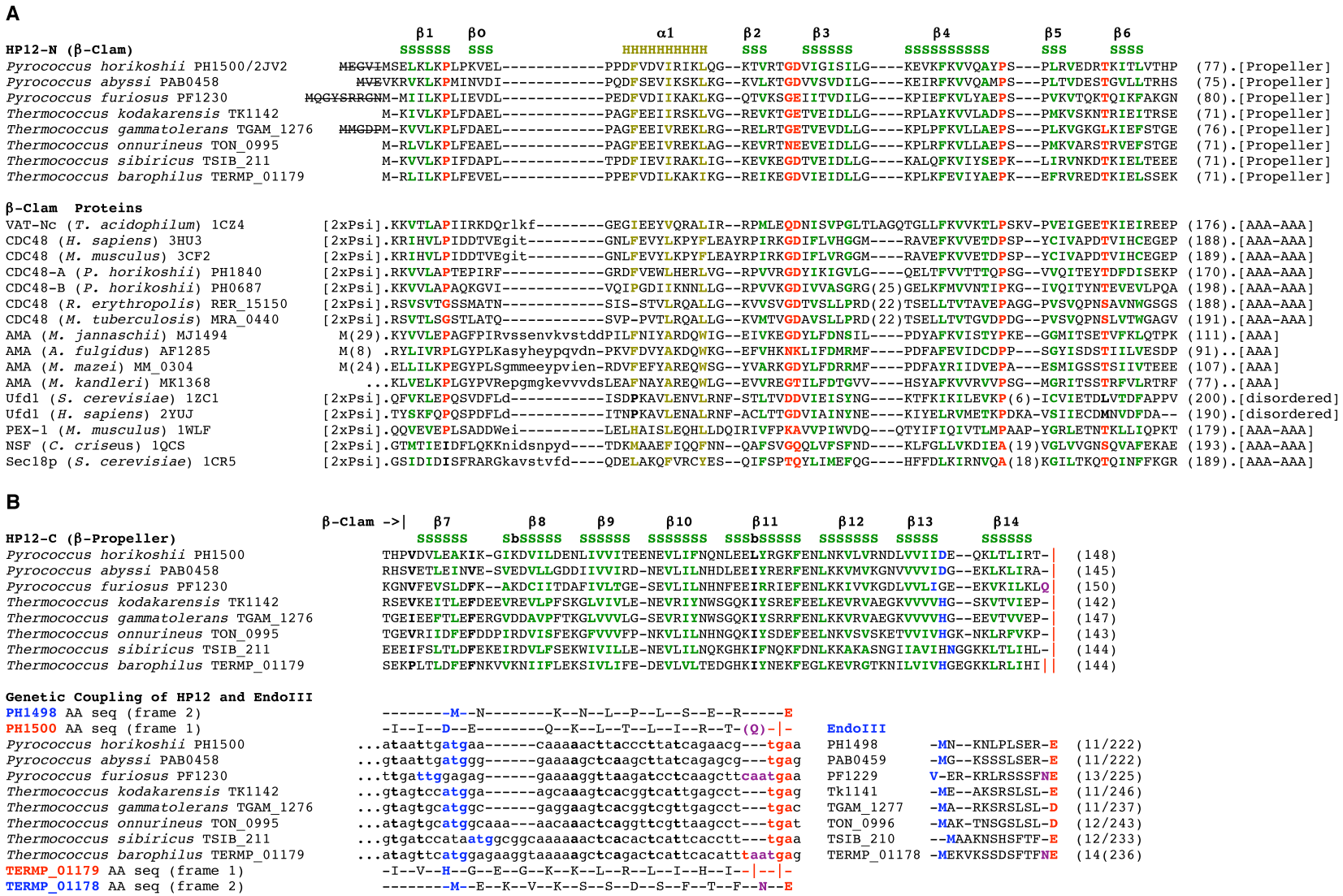
conserved secondary structure elements frequently accommodate sub-group specific insertions with the exception of the conserved connectors between β2-β3 and β5-β6. The β-clam domains of the HP12 group however do not contain larger insertions and have rather short loop regions.

Analysis of the genetic environment of the HP12 locus revealed a coupling to a type III endonuclease (EndoIII). In all pyroccoci, the Endo III start codon is placed about 30 nucleotides upstream of the HP12 stop codon (Figure 15 B). The region of entanglement shows a high degree of conservation on the nucleotide level underpinning the strong operon structure. Furthermore, these EndoIII proteins are N-terminally extended by approximately 10 residues in comparison to typical bacterial orthologs. EndoIII proteins are bifunctional DNA glycosylases that catalyze both the excision of damaged nucleobases and subsequent strand nicking 3' to the lesion [210]. With the recognition of certain DNA lesions EndoIII initiates the base-excision DNA-repair (BER) pathway that proceeds by complete removal of the nucleoside by A/P nucleases, repolymerization, and resealing of DNA. EndoIII belongs to the large superfamily of HhH-GPD repair enzymes and displays highest activity towards oxidized pyrimidine substrates [211]. Because homologs of most enzymes constituting the BER pathway are readily detected in pyrococci, a role for HP12 in this pathway was not evident during sequence analysis. However, the fact that HP12 and EndoIII are strictly found in an operon with a rather long overlapping region suggested a physical interaction.

Figure 15. Sequence Analysis of HP12-like Proteins

(A) The multiple alignment of β-Clam domains shows the complete HP12 group (except for *Thermococcus sp.* AM4) in the upper half. The lower half contains, sorted by protein family, all β-Clam domains of known structure, the two CDC48 paralogs from *P. horikoshii*, two actinobacterial CDC48 proteins, and a phylogenetically representative selection of AMA proteins. Hydrophobic core residues are in bold (green in β-strands, yellow in α1). Residues that are conserved across groups are in red. Upstream or downstream domains are denoted in angular brackets. Locus tags or, if available, PDB-identifiers are given next to the name of the organism. The consensus secondary structure is shown above the sequences (H, helix; S, strand; b, bulge).

(B) Mutliple alignment of HP12-C and genetic coupling to EndoIII. The upper half shows the C-terminal domain of the HP12 group. The lower half shows the aligned overlapping nucleotide sequences on the left, including two translation frames that correspond to the HP12 group and the downstream Endonuclease. On the right, the N-terminal residues of endonucleases are aligned and positioned with respect to the upper half. The stop codons of HP12-C and the corresponding position in the EndoIII reading frame are in red. The start codons of endonucleases and the corresponding position inHP12-C are in blue. Two insertions at the C-terminus of PF1230 and TERMP_01179 are in magenta.


A
β1 β0 α1 β2 β3 β4 β5 β6
HP12-N (β-Clam)
Pyrococcus horikoshii PH1500/2JV2 (77).[Propeller]
Pyrococcus abyssi PAB0458 (75).[Propeller]
Pyrococcus furiosus PF1230 (80).[Propeller]
Thermococcus kodakarensis TK1142 (71).[Propeller]
Thermococcus gammatolerans TGAM_1276 (76).[Propeller]
Thermococcus onnurineus TON_0995 (71).[Propeller]
Thermococcus sibiricus TSIB_211 (71).[Propeller]
Thermococcus barophilus TERMP_01179 (71).[Propeller]
β-Clam Proteins
VAT-Nc (T. acidophilum) 1CZ4 [2xPsi].[AAA-AAA]
CDC48 (H. sapiens) 3HU3 [2xPsi].[AAA-AAA]
CDC48 (M. musculus) 3CF2 [2xPsi].[AAA-AAA]
CDC48-A (P. horikoshii) PH1840 [2xPsi].[AAA-AAA]
CDC48-B (P. horikoshii) PH0687 [2xPsi].[AAA-AAA]
CDC48 (R. erythropolis) RER_15150 [2xPsi].[AAA-AAA]
CDC48 (M. tuberculosis) MRA_0440 [2xPsi].[AAA-AAA]
AMA (M. jannaschii) MJ1494 M(29).[AAA]
AMA (A. fulgidus) AF1285 M(8) .[AAA]
AMA (M. mazei) MM_0304 M(24).[AAA]
AMA (M. kandleri) MK1368 ...[AAA]
Ufd1 (S. cerevisiae) 1ZC1 [2xPsi].(6)[disordered]
Ufd1 (H. sapiens) 2YUJ [2xPsi].[disordered]
PEX-1 (M. musculus) 1WLF [2xPsi].[AAA-AAA]
NSF (C. crise us) 1QCS [2xPsi].[AAA-AAA]
Sec18p (S. cerevisiae) 1CR5 [2xPsi].[AAA-AAA]
B
β-Clam ->| β7 β8 β9 β10 β11 β12 β13 β14
HP12-C (β-Propeller)
Pyrococcus horikoshii PH1500 (148)
Pyrococcus abyssi PAB0458 (145)
Pyrococcus furiosus PF1230 (150)
Thermococcus kodakarensis TK1142 (142)
Thermococcus gammatolerans TGAM_1276 (147)
Thermococcus onnurineus TON_0995 (143)
Thermococcus sibiricus TSIB_211 (144)
Thermococcus barophilus TERMP_01179 (144)
Genetic Coupling of HP12 and EndoIII
PH1498 AA seq (frame 2)
PH1500 AA seq (frame 1)
Pyrococcus horikoshii PH1500
Pyrococcus abyssi PAB0458
Pyrococcus furiosus PF1230
Thermococcus kodakarensis TK1142
Thermococcus gammatolerans TGAM_1276
Thermococcus onnurineus TON_0995
Thermococcus sibiricus TSIB_211
Thermococcus barophilus TERMP_01179
TERMP_01179 AA seq (frame 1)
TERMP_01178 AA seq (frame 2)
EndoIII
PH1498 -MN--KNLPLSER-E (11/222)
PAB0459 -MG--KSSSLSER-E (11/222)
PF1229 V-ER-KRLRSSSFNE (13/225)
Tk1141 -ME--AKSRSLSL-E (11/246)
TGAM_1277 -MA--RKSRSLSL-D (11/237)
TON_0996 -MAK-TNSGSLSL-D (12/243)
TSIB_210 --MAAKNSHSFTF-E (12/233)
TERMP_01178 -MEKVKSSDSFTFNE (14/236)

HP12 Forms a Ternary Complex with EndoIII and DNA

HP12 and the genetically coupled EndoIII from *P. horikoshii* were purified to homogeneity and subjected to various interaction assays. Prior characterization showed that both proteins exhibit the thermal stability expected for proteins from a hyperthermophilic organism. Analytical size-exclusion chromatography indicated that HP12 forms a large oligomer in solution, most likely a hexamer (not shown). Despite an obvious toxicity of EndoIII during recombinant expression in *E. coli*, sufficient amounts of soluble protein were obtained. Proper functionality of EndoIII was ascertained by "borohydride trapping" on a DNA construct containing 5,6-dihydro-deoxyuridine (DHU) [210, 212]. The reaction product showed a characteristic shift in SDS-PAGE analysis. However, tests for co-migration during size-exclusion chromatography could not detect an interaction of HP12 and EndoIII.

Co-immunoprecipitation experiments demonstrated the ability of His-tagged EndoIII, to pull down HP12 confirming a physical interaction of both proteins (Figure 16 B), which was suggested by genetic association (Figure 15 B). Subsequently, the formation of a ternary complex with DNA was investigated by electrophoretic mobility shift assays. EMSA with linearized plasmid DNA showed that addition of HP12 alone does not alter the migration behaviour of DNA substantially, whereas EndoIII causes a considerable band shift. In samples that additionally contained HP12 the band shift was even more pronounced indicating the recruitment of HP12 to an EndoIII-DNA complex (Figure 16 A). A comparable pattern was observed with short 100 base pair DNA constructs containing a missing-T site. These results showed that HP12 indeed forms a ternary complex with EndoIII and DNA.

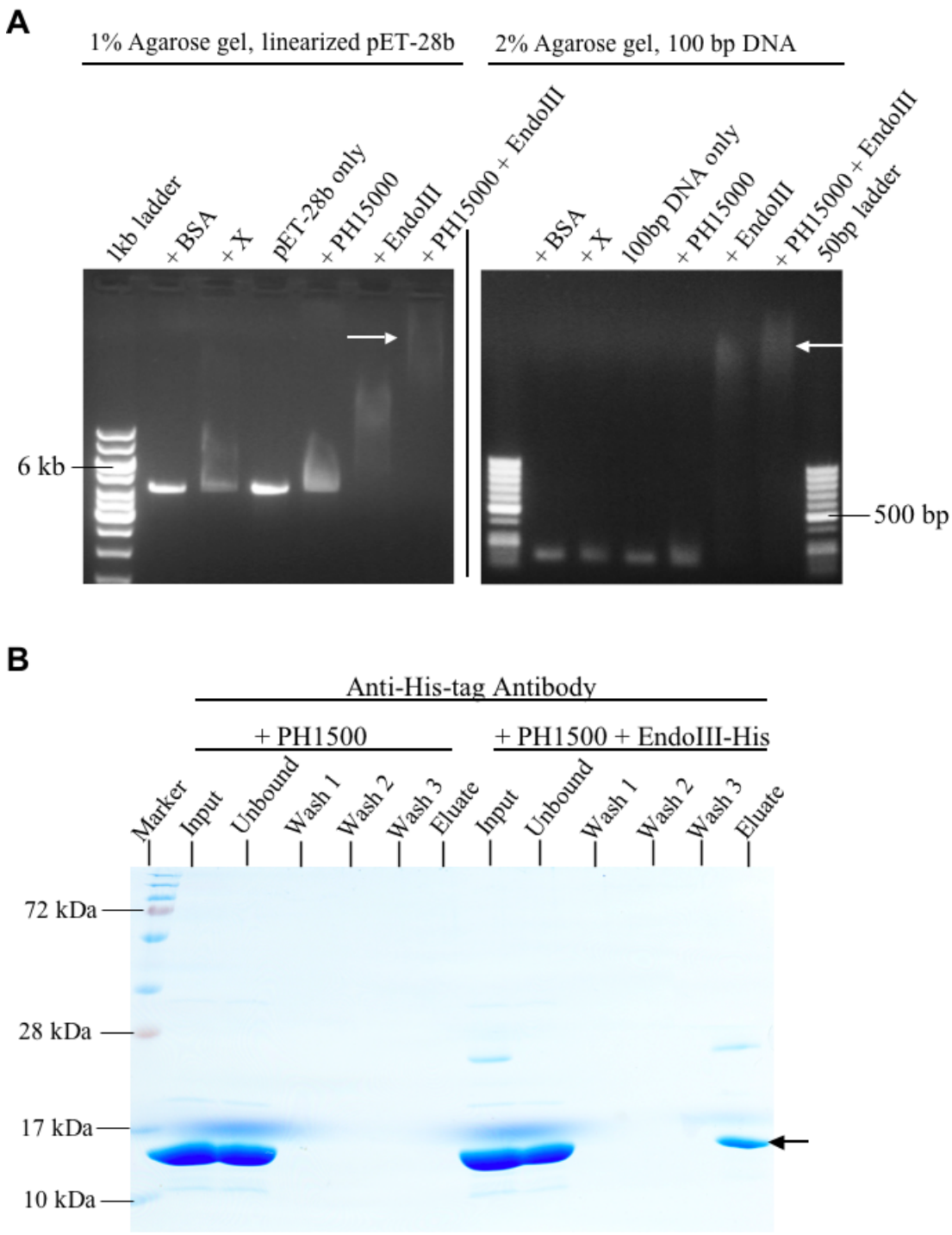


Figure 16. HP12 Forms a Ternary Complex with EndoIII and DNA

(A) HP12 binds to an EndoIII-DNA complex. In electrophoretic mobility shift assays using EtBr-stained agarose gels, HP12 alone is not influencing the mobility of short 100 bp DNA constructs, and the mobility of linearized plasmid DNA only to a minor degree. In presence of EndoIII, which causes a substantial shift by itself, HP12 contributes to a further decrease of DNA mobility indicating complex formation (white arrows).

(B) HP12 physically interacts with EndoIII. The Coomassie blue stained SDS-PAGE (15%) analysis shows that HP12 binds to (black arrow) his-tagged EndoIII using an anti-his antibody. HP12 alone is not precipitated by the anti-his antibody. The molecular weight of HP12 is 16.3 kDa (using the alternative start codon); the molecular weight of EndoIII-His is 26.9 kDa.

The N-domain Adopts a β-Clam Fold

As predicted by sequence analysis, the N-terminal domain is highly similar to members of the SCOP fold CDC48 domain 2-like [213]. For this fold, the term β-clam has been coined because its six β-strands do not form a closed barrel [49]. Instead, the central β-sheet consisting of β1, β3, β4, and β6 forms a clam-like structure that embeds a long α-helix, which provides closure and contributes three conserved hydrophobic residues to the core of the fold (F23, I27, L31) (Figure 17 A and B).

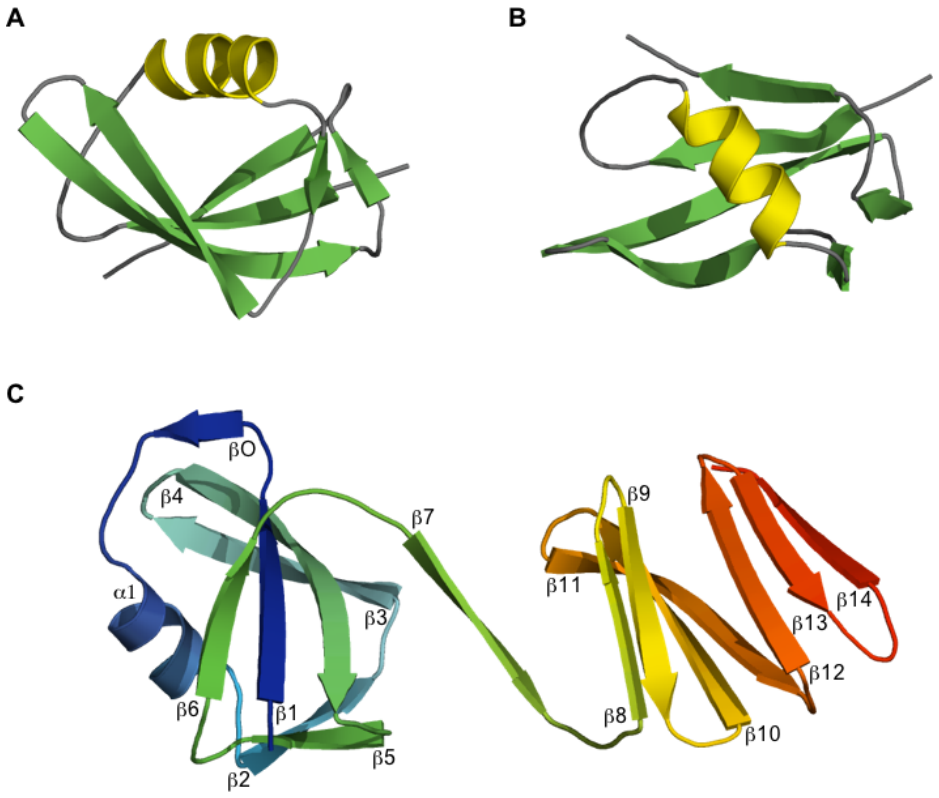


Figure 17. The Structure of the HP12 Monomer

(A and B) Cartoon representation of the solution structure of isolated HP12-N in side and top view. Strands are coloured in green helix α1 is in yellow.

(C) Cartoon representation of the full-length HP12 monomer coloured in a rainbow-like succession from blue at the N-terminus to red at the C-terminus. Secondary structure elements are annotated according to the multiple sequence alignment (Figure 15 A).

The presence of a GD-box (G39-D40) between β2 and β3 prevents the formation of a β-barrel by inducing a hydrogen-bonding pattern that leaves the strands β2 and β3 unpaired. The GD-box is a super-secondary structure element connecting two unpaired β-strands via an orthogonal type II β-turn [201]. This β-turn forms additional hydrogen bonds with the last third of β4 (V55 and A58). The region comprising the extended GD-box (K34-I44) is the most conserved region, in sequence (Figure 15 A) and structure, across the different groups of β-clam domains.

Superpositions of various β-Clam domains onto HP12-N yield root mean square deviations (RMSD) of about 1.2 Å over approximately 50 residues located in the core of the fold. Additionally, DALI searches underpin the structural similarity of β-clam domains (Table 4). In contrast to other β-clam domains, the N-domain of HP12 has shorter loop regions whereas the conserved secondary structure elements are slightly extended, which results in an increased compactness.

Structure of Full-length HP12

Although crystallization trials aiming at the ternary complex were not successful, we obtained a crystal structure of HP12 in full length with a resolution of 2.5 Å. Therefore, the usage of an alternative start codon turned out to be of importance. Comparison to other members of the HP12 group as well as the location of the ribosome-binding site suggested a methionine six residues downstream of the sequence deposited at NCBI as the proper start codon. (A re-annotation of the *P. horikoshii* genome placed the start codon twelve residues further upstream of the originally deposited sequence, and HP12 was renamed to PH1498.1n). Usage of the alternative start codon removed a rather unstructured tail from the N-terminus as being judged from the solution structure of the isolated β-clam domain (PDB-ID 2JV2). This led to improved solubility and crystallization behaviour of the full-length protein.

Determination of the isolated C-terminal domain by NMR spectroscopy [178] enabled the solution of the crystal structure of full-length HP12 by molecular replacement. It reveals a dodecamer consisting of two stacked hexamers held together by the equatorially positioned N-terminal domains (Figure 5). They provide all contacts between the hexamers without any contribution from the C-terminal domains. Most relevant for dodecamerization is the formation of an additional β-strand, βO (K16-V18, Figure 17 C), between β1 and α1, which pairs with the last β-strand (β6) of the neighbouring β-clam domain. Two different conformations of the loop that connects the β-clam with the C-terminal domain (T75-V78) are observed in adjacent monomers within each hexamer resulting in two different relative positions of the N-domain with respect to its C-domain. Tight pairing of the N-domains of two such alternative conformers, each of which is contributed from a different hexamer, leads to their interdigitation at the equatorial surface of the hexamers giving the dodecamer a disc-shaped appearance (Figure 18 A and C). Nevertheless, dodecamerization appears to be

accentuated by crystallization because size-exclusion chromatography and DOSY-NMR indicated that the hexamer is the main oligomerization state in solution.

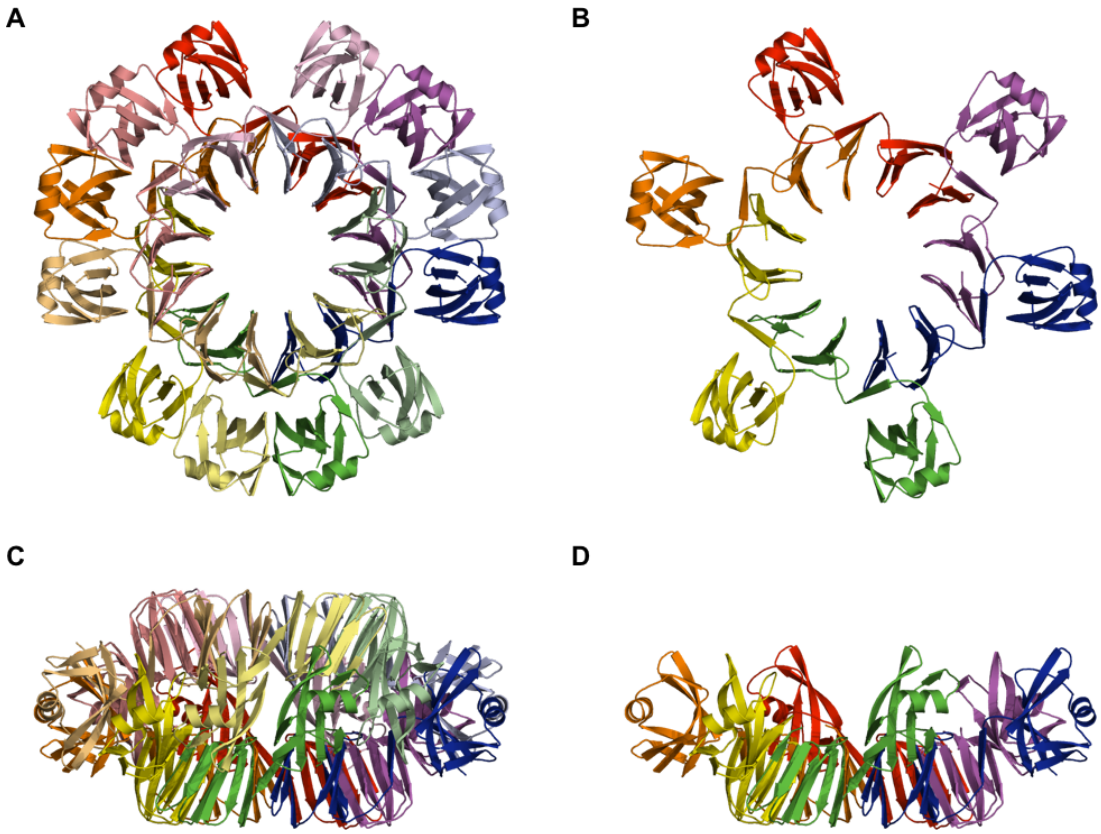


Figure 18. The Structure of the HP12 Oligomer

(A and C) Cartoon representation of the HP12 dodecamer in top and side view. Colouring is by chain using pairs of bright and pale colours for interacting monomers from the different hexamers. The interaction of the hexamers occurs via the N-terminal β-clam domains, resulting in an additional β-strand, βO, between β1 and α1 (K16-E18).

(B and D) Cartoon representation of the HP12 hexamer in top view, facing the narrower opening of the funnel-shaped pore, and side view. The hexamer in bright colour was retained from the dodecamer shown in A and C. Size-exclusion chromatography and NMR-spectroscopy (not shown) indicate that the hexamer is the oligomerization state of HP12 in solution.

The Hexameric C-domain is a Twelve-bladed β-Propeller

The C-terminal domain constitutes the core of each hexamer and adopts the fold of a twelve-bladed β-propeller (Figure 18 B and D). Although significant sequence similarity is currently not detectable, HP12-C is structurally similar to β-propellers independent of the number of blades. This is underlined by DALI searches superimposing all eight strands of the HP12-C monomer with RMSD of 2.5 A (Figure 19, Table 4).

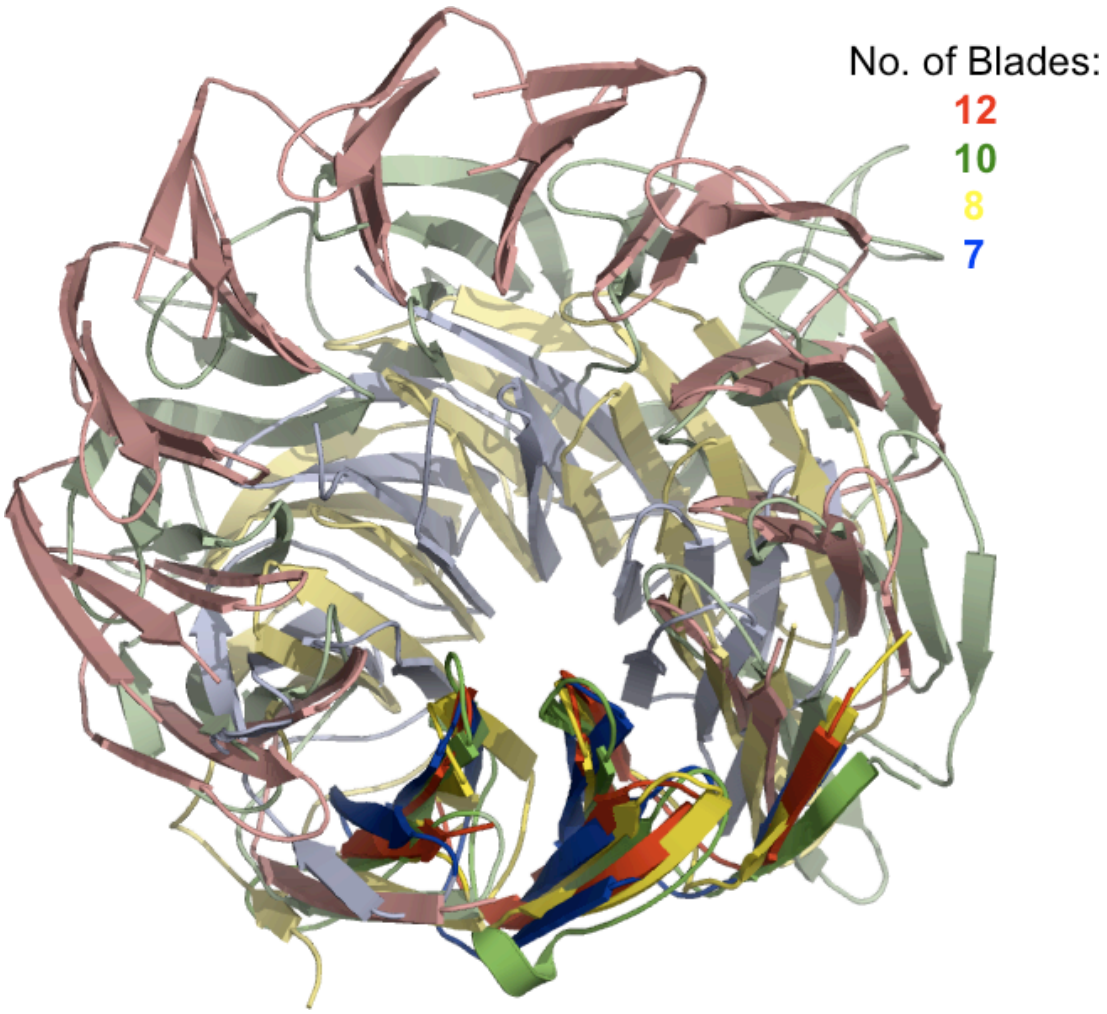


Figure 19. HP12-C is Structurally Similar to β-Propellers

(A) Superimposition of HP12-C onto structurally similar β-Propellers. The superimposition includes all eight strands of the HP12-C monomer (red) and the first eight strands of a seven- (blue), an eight- (yellow), and a ten-bladed (green) β-propeller. An increasing number of blades results in a larger volume of the pore. HP12-C is a homohexamer, whereas the other β-propellers are monomers. Superimposed regions are shown in bright colours. Insertions to the core propellers are omitted. The structures shown are 3F6K (10), 1NEX (8), and 1PGU (7). The superimposition was interactively conducted with Swiss-Pdb Viewer. The structural representation was done with PyMol. Table 4 contains details of the superimposition.

Each monomer consists of two four-stranded β-meanders that form the blades of the propeller (Figure 17 C). The first blade is formed by β8 to β11, of which β8 lines the central hole and β11 faces the outer surface of the propeller. The second blade is composed by β12 to β14 plus β7, which occupies the position of the external strand in the second meander. This points to a circular permutation of the last strand of the second blade to the N-terminus of the C-terminal portion resulting in a velcro closure through β7 [214]. In this arrangement, a swap of the very N-terminal strand into the position of the last strand of the second blade of the adjacent monomer leads to an interlocking of the chains and presumably provides increased stability.

Table 4. Summary of Sequence and Structure Comparisons for HP12

		HHPRED-SCORES[1]			DALI-SCORES[2]			
Protein	**PDB-ID**	**Prob. [%]**	**E-Value**	**P-Value**	**Z-Score**	**RMSD [A]**	**LALI[2]**	**Fold**
HP12-N (6-77)	**2JV2[3]**	**100**	**7.3e-37**	**1.4E-39**	**9.9**	**1.9**	**71/71**	**β-Clam**
CDC48	1CZ4	99.8	6.6e-19	1.3e-23	7.0	2.3	66	β-Clam
UFD1	2YUJ	98.9	3.4e-09	6.7e-14	8.2	2.1	67	β-Clam
CDC48	3CF2	98.6	4.2e-08	8.3e-13	7.5	2.6	69	β-Clam
CDC48	3HU3	98.6	1.0e-07	2.0e-12	7.3	2.1	69	β-Clam
UFD1	1ZC1	98.4	7.2e-07	1.4e-11	7.7	2.2	71	β-Clam
PEX1	1WLF	85.3	3.3	6.4e-05	8.5	2.0	67	β-Clam
SEC18P	1CR5	58.4	43	0.00085	8.2	2.3	68	β-Clam
NSF	1QCS	55.0	48	0.00095	8.1	2.3	71	β-Clam
HP12-C[4] (78-148)	**–**	**Query (structure not yet deposited)**						**12-bladed Propeller**
Actin Interacting	1PGU_B				7.5	2.5	71	7-bladed Propeller
Centromere Binding P.	1NEX_B	-	-	-	5.8	3.1	70	8-bladed Propeller
Lectin	2BT9	-	-	-	4.5	2.9	61	6-bladed Propeller
Sortilin	3F6K	-	-	-	-	1.8[5]	70[5]	10-bladed Propeller
HP12-C *Blade2*[6]	–	-	-	-	-	0.96[5]	30[5]	β–Meander

The targets taken from the HHpred and Dali hit lists represent a highly diverse selection of proteins, which are included in the multiple alignment of β-clams (Figure 15A, HP12-N) and the superposition of propellers with differing numbers of blades (Figure 21), rather than the ranking by the servers.

1: HHpred searches were performed in default settings against the Protein Data Bank, release of May 7 2011, filtered for a maximum of 70% pairwise sequence identity at http://toolkit.tuebingen.mpg.de/HHpred.

2: Dali searches were done at http://ekhidna.biocenter.helsinki.fi/dali_server/. Lali denotes the number of residues included in the superimposition.

3: This target is the self-hit. The Dali-scores reflect the difference between HP12-N in the crystal structure of the full-length protein and the NMR-structure of the isolated N-terminal domain

4: For the C-terminal domain the HHpred hit list does not contain targets with scores considered significant.

5: Superimposition was interactively conducted using Swiss-PDB Viewer.

6: Superimposition of strands β11- β14 onto β7- β10 revealing the internal repeat of HP-12-C (Figure 20).

Both blades are structurally similar suggesting that an internal repeat resides within the C-terminal domain. A superposition of four contiguous strands, including the last strand of the neighbouring blade, yields an RMSD of 0.96 Å over 30 residues indicating that

also the connections between blades adopt a very similar conformation. This procedure practically revises the circular permutation and produces a structure-based alignment, which uncovers the repetitive nature of the C-terminal domain on the sequence level (Figure 20). The sequence similarity between the blades rests on matching hydrophobic patterns and polar motifs providing evidence that a duplication event followed by a circular permutation gave rise to the C-terminal domain.

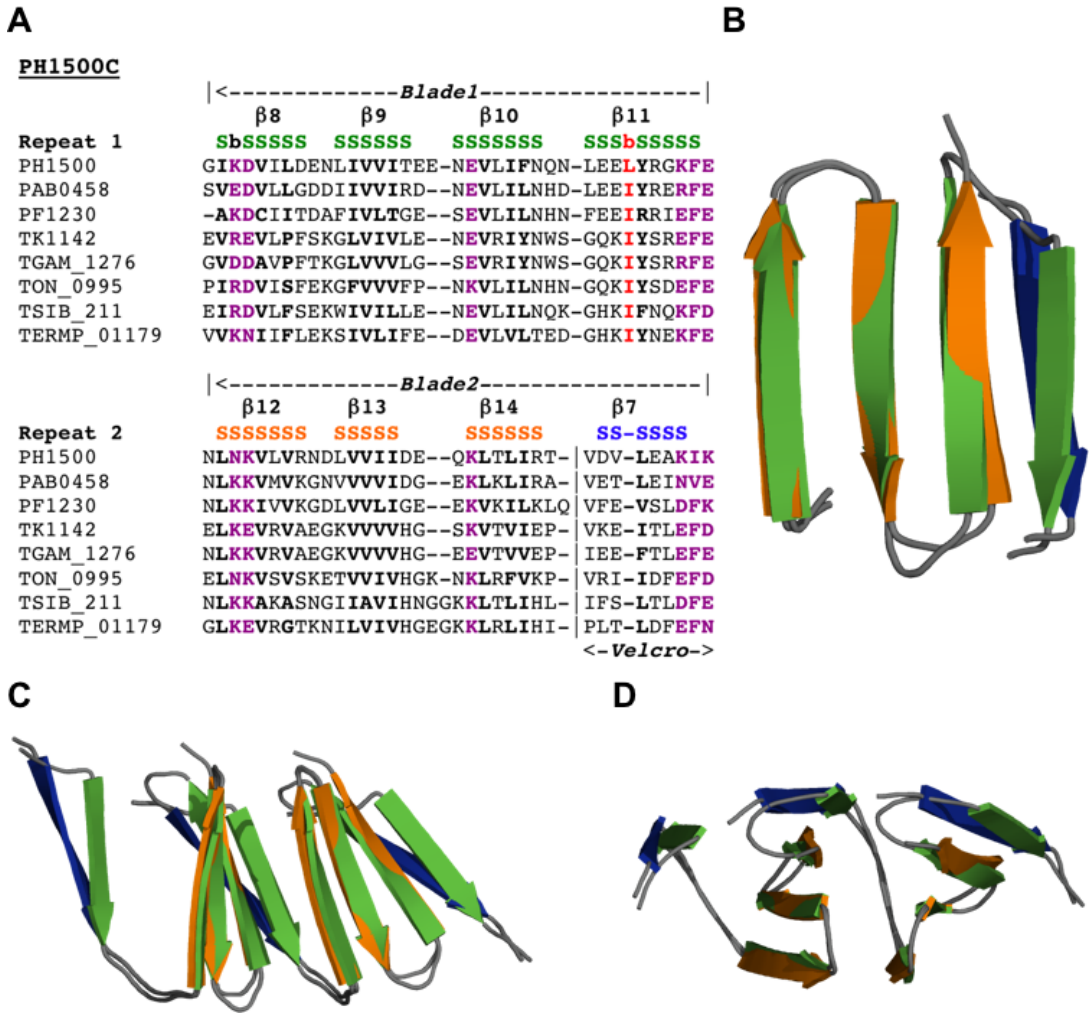


Figure 20. HP12-C Contains an Internal Repeat

(A and B) Sequence alignment and superposition of the two propeller-blades of HP12-C. The circular permutation has been revised so that the velcro-strand is placed in the position of the fourth strand of the second blade (blue). The remaining strands of the second blade are in orange, the strands of the first blade in green. In the alignment, the hydrophobic patterns of both blades (bold) match without exception; conserved polar patterns are in magenta, the β-bulge in the external strand of blade1 is in red.

(C and D) Consecutive superposition of full-length HP12-C to its internal repeats. The superposition includes four strands and illustrates that the inter-blade loops adopt a very similar conformation. Panel D is rotated by 90° about the x-axis with respect to panel C. Colouring corresponds to (B).

The blades of the propeller are arranged in a rather steep fashion and radially tilted with respect to the pseudo-twelve-fold symmetry axis yielding a funnel-shaped pore with a

diameter of 27 nm at the narrow opening and 38 nm at the wide opening (measured from the Cα atoms of opposite residues, R131-R131 and D90-D90, respectively). Despite descent from a common ancestor and remarkable similarities in sequence and structure the two types of blades occupy sectors of different size within the toroid. Blade 1 contains a β-bulge in the external strand β11 (L117) (Figure 20 A), which leads to an increased twist angle of the β-meander. This, in turn, results in a larger space requirement of blade 1 as compared to blade 2. Nevertheless, the twist angle of both blades of HP12 is rather small in comparison to the dominant group of seven-bladed propellers, which necessitates the previously unobserved number of twelve blades to close the toroid generating a propeller with an exceptionally wide central pore.

4.2.3 Discussion

In the light of the initial prediction that the C-terminal domain might form an OB-fold and that HP12 may constitute a novel co-occurrence of two major N-domains of AAA proteins, these results were unexpected. HP12 interacts with a type III endonuclease implying a role in DNA repair, and the C-terminal domain folds as a homohexameric β-propeller. One the one hand function and appearance provoke the notion that HP12 shares certain features with DNA sliding clamps. On the other hand the symmetric nature of the C-terminal domain provides clues about the evolution of β-propellers.

HP12 - a PCNA Analog?

Despite absence of homology, HP12 and the Proliferating Cell Nuclear Antigen (PCNA) have functional and structural characteristics in common. PCNA encircles DNA and orchestrates numerous processes such as DNA replication and DNA repair including mismatch repair, nucleotide excision repair, and base excision repair (BER) [215]. In the latter, PCNA operates as an assembly platform that binds to early DNA glycosylases like EndoIII [216] and recruits downstream A/P nucleases. For one of two PCNA paralogs from *Pyrococcus furiosus* it has been shown that it interacts with an A/P nuclease and stimulates its 3' to 5' exonuclease activity; not its A/P nuclease activity though [217]. The ability of PCNA to interact with various DNA-glycosylases, responsible for the recognition of certain DNA lesions, resembles the interaction of HP12 with its genetically coupled EndoIII. Apart from the HP12 operon *P. horikoshii* encodes two PCNA paralogs (one of which is truncated and most likely inactive), three

early DNA-glycosylases of the HhH-GPD superfamily, and two A/P nucleases, one of which is found in the widened genetic neighbourhood of HP12. Therefore, HP12 could function as an assembly platform specialized in coordinating the removal of oxidized pyrimidines by EndoIII and cooperate with PCNA in the diverse branches of the BER pathway.

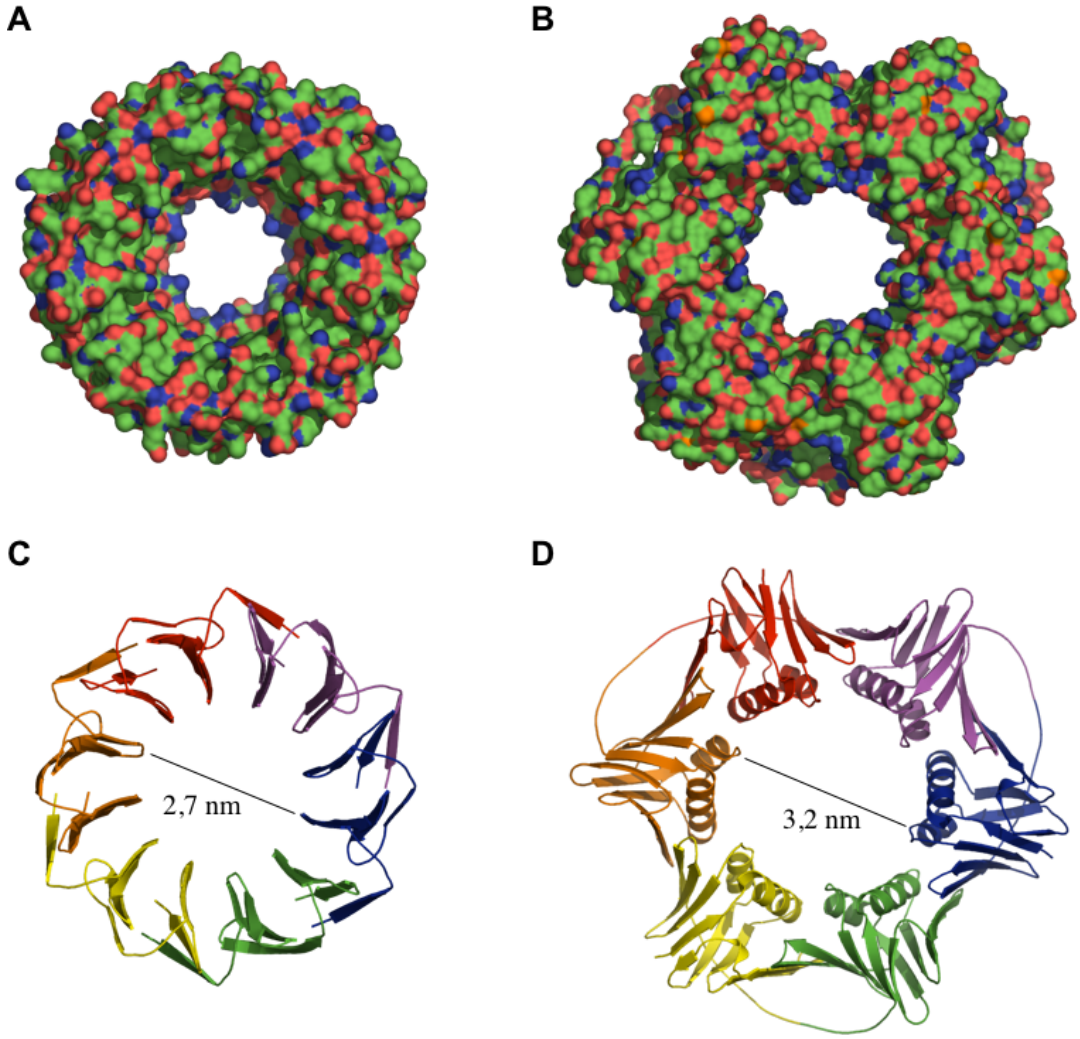


Figure 21. Comparison of HP12 and PCNA

(A and B) Surface representation of HP12 and Tk-PCNA. Surfaces are coloured by charge; blue denotes positively charged residues, red negatively charged residues. In both molecules, positively charged residues line the pore facilitating an interaction with DNA. The depicted PCNA is from closely related *T. kodakarensis* (3LX2).
(C and D) Cartoon representation of HP12 and PCNA. Colouring of homohexameric HP12 is by chain. Colouring of homotrimeric Tk-PCNA distinguishes the two domains of each monomer, illustrating the pseudo-six-fold symmetry. The diameter is calculated as the distance between two opposite Cα atoms lining the pore (R131-R131 in case of HP12 and K81-R210 in case of Tk-PCNA). The diameter of HP12 is somewhat smaller at the narrow opening of the funnel, but still large enough to encircle B-DNA. The diameter at the wide opening of the pore is 38 nm (D90-D90, not shown).

On the structural level, the large central pore formed by HP12 and PCNA is most conspicuous. In case of PCNA the pore is 3.3 nm wide and in case of HP12 it is 2.7 nm

wide at the narrow opening of the funnel and 3.8 nm at the wide opening. Although somewhat smaller on the narrow side, the pore of HP12 is wide enough to accommodate B-DNA with a diameter of 2 nm. Positively charged residues line the inner surface of the pore of PCNA as well as of HP12 allowing an interaction with the backbone of DNA (Figure 21 A and B). The similarity also includes the internal symmetry. The monomers of homotrimeric PCNA proteins are constructed by two similar domains (three in case of homodimeric sliding clamps of bacteria), which are connected by a long so-called inter-domain loop leading to a pseudo-six-fold symmetry [215]. Analysis of PCNA sequences with sensitive repeat detection algorithms reveals that each PCNA monomer indeed contains four internal repeats (detected by HHrepID, P-value 10^{-7}; [196]) resulting in a pseudo-twelve-fold symmetry reminiscent of the β-propeller of HP12 (Figure 21 C and D). These analogous features suggest that encircling DNA is the functional reason for the evolution of HP12. Nevertheless, further experimental evidence is required to show that HP12 is indeed able to thread DNA through its pore.

Twelve-fold vs. the Preferred Seven-fold Symmetry of β-Propellers

The large central pore of HP12 is a consequence of the rather small twist angle of the β-meanders requiring twelve blades to close the toroid. A model by Murzin based on general packing principles of β-sheet proteins rationalized the structural parameters of β-propellers and returned a seven-fold symmetry as the preferred packing mode with six- and eight-fold symmetry being well possible [218]. According to this model the main parameters are the distance between adjacent strands within a sheet (4.5 Å), the mean perpendicular distance between adjacent non-intercalating sheets (10 Å), the number of packed β-sheets/blades, and the twist of the β-sheets measured as the dihedral angle between adjacent strands. Of these, the latter two variably govern the assembly of β-meanders into a β-propeller toroid. Whereas the twist decreases with an increasing number of blades, the size of the central pore increases with the number of blades (Figure 22). HP12 represents a protein, in which these two effects are counterbalanced in an unusual form. Twelve blades of small twist-angles (on average 20° in contrast to 23.8° of the Murzin model for a seven-bladed propeller) are packed to a propeller with an exceptionally large central hole, potentially large enough to encircle DNA.

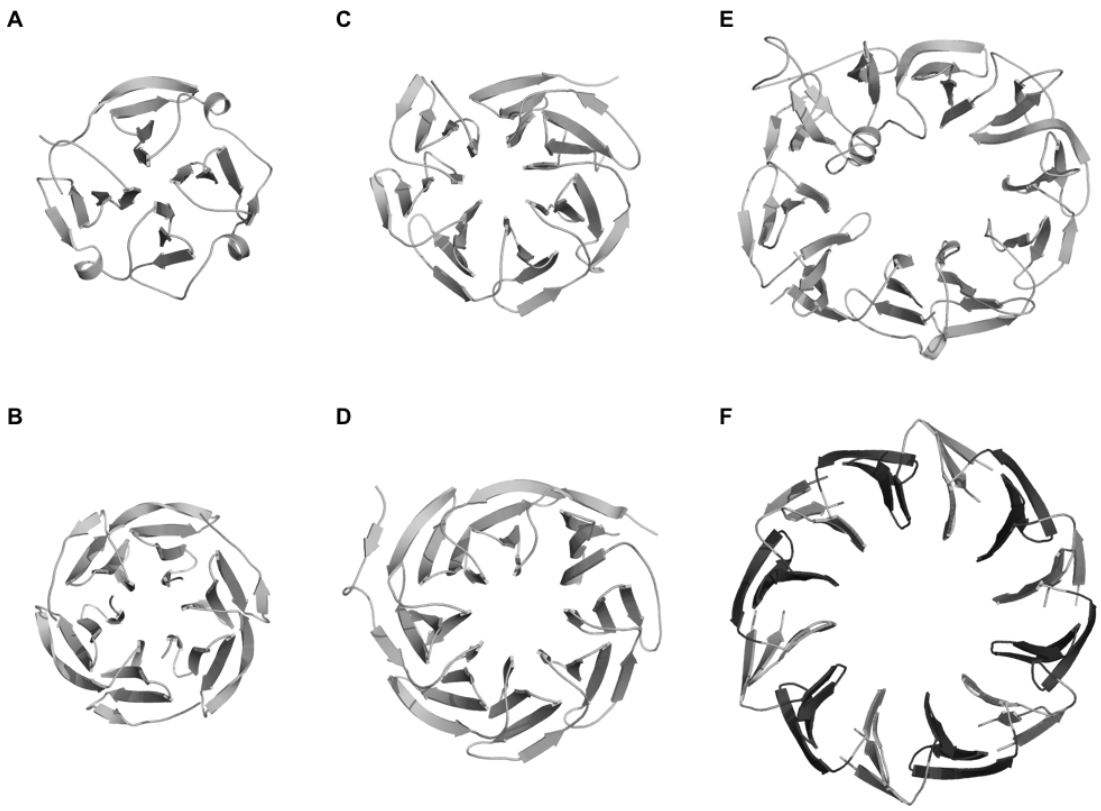


Figure 22. HP12-C is Structurally Similar to β-Propellers

The gallery shows a four- (A, 1FKL), a six- (B, 2BT9), a seven- (C, 1PGU_B), an eight- (D, 1NEX), a ten- (E, 3F6K) and a twelve-bladed (F) propeller. It illustrates the relationship between the twist angle of the β-meanders, the number of blades, and the size of the central pore. In the HP12 hexamer (F) two types of blades occupy sectors of different size. The less twisted β-meanders are in light grey, the blade containing a β-bulge in its rather twisted external strand is in dark grey. In the four-bladed propeller (A) small helices, which pack between the external strands of the β-meanders, are required for closure of the toroid. The six-bladed propeller (B) is a trimer.

Within the C-terminal domain of HP12 alternating blades occupy sectors of different size, which is caused by different twist-angles in neighbouring blades. The first of the two blades of one monomer contains a β-bulge in the external β-strand (β11, L117; see Figure 20 A) resulting in a larger twist angle than in blade 2 (Figure 22 F). This suggests that a removal of the one residue insertion might allow the accommodation of a larger number of less twisted blades in a propeller with an even larger pore. While HP12 exemplifies one extreme, the four-bladed propeller of collagenase [219], in which small helices are packed between the external β-strands in order to close the toroid, marks the other (Figure 22 A). Both extremes illustrate the plasticity inherent to the β-propeller architecture. The prevailing number of seven-bladed propellers, however, reflects the least restrictive sequence requirements and the favourable nature of the seven-fold symmetry. [218].

The Evolution of β-Propellers

Although the most sensitive homology detection programs do not provide evidence for significant sequence similarity of HP12-C and β-propellers, the structural similarity is striking and supported by superimpositions and DALI searches. DALI searches with the HP12-C monomer retrieve β-propellers with a velcro arrangement of one strand, which allows for alignment of the permuted strand β7, as top hits with Z-scores of ca. 7 [189]. Among them are β-propellers with varying numbers of blades. Superimpositions comprise up to seventy residues with an RMSD as low as 1.7 Å, and include all eight strands of the C-terminal domain. However, the structure-based sequence alignment does not reveal significantly similar patterns, which might have been missed by sequence similarity searches. Because sequence similarity is the primary marker of homology, currently robust evidence for homology is lacking, a situation that might change with improved sensitivity of homology detection programs and an increasing number of genomes being sequenced. This is in contrast to the vast majority of β-propellers that can be evolutionarily connected by sequence similarity searches independent of the number of blades [220]. On the other hand, structural similarity often substantiates evolutionary relationships when sequence similarity has eroded [125, 126]. Therefore, the pronounced structural similarity of HP12 to other β-propellers has implications for the evolution of this fold (which comprises all β-propellers independent of the number of blades [220] in contrast to the SCOP classification [62]).

Amplification of single blades and subsequent divergence are the major evolutionary mechanisms that shape β-propeller proteins. At one end of the spectrum are propellers that have diverged to an extent that allows recognition of the origin by repetition of single-blade units only on the structural level. At the other end are propellers consisting of basically identical blades that have hardly differentiated or drifted [220]. This is exemplified by a group of proteins found in certain cyanobacterial species like *Nostoc punctiforme* and *Anabaena variabilis* whose propeller blades are encoded by identical nucleotide sequences providing evidence that amplification continues to enrich the repertoire of propellers [220]. Interestingly, one of these proteins contains fourteen blades but folds into two seven-bladed propeller domains [221] underpinning the preference for seven-fold symmetry [218]. However, ongoing amplification based on blades as the evolutionary unit suggests that primordial four-stranded β-meanders could have formed propellers by oligomerization prior to repeated duplication and fusion events.

The Protein Data Bank (PDB) contains only three other structures of oligomeric propellers, two of which have been generated by directed evolution of fragments of five-bladed tachylectin-2. The fragments however did not correspond to blades as the evolutionary units yet they demonstrate the metamorphic character of propeller pieces [222-224]. The only other case of a naturally occurring oligomeric propeller in the PDB is the homotrimeric, six-bladed fucose-binding lectin from *Ralstonia solanacearum* (PDB-ID 2BT9) [225], which, in contrast to HP12, does not show a velcro arrangement. Its origin by duplication of an ancestral blade is comfortably detected in sequence (HHrepID score 10^{-13}). Furthermore, the structure of a closely related fungal lectin from *Aleuria arantia* (1OFZ, 31% sequence identity) displays the fully amplified phenotype in which all blades have been placed on one chain. This observation underlines what is implied by the scarcity of oligomeric propellers: They are readily replaced by their apparently more efficient, monomeric descendents. HP12 represents such a rare oligomeric intermediate that preserves traces of its single-bladed ancestor and assembles to a homohexameric propeller with twelve blades.

These results provide support for the more general hypothesis that modern protein domains originated by duplication and recombination of an ancestral set of peptides capable of forming super-secondary structure elements [144-146]. Their role at early stages of protein evolution is indicated by the pervasive phenomenon of repetition within domains and the significant sequence similarity detected beyond the boundaries of folds [140]. The repeated motifs as well as the regions of similarity shared by more than one fold correspond to such ancient peptides, some of which accommodate basic functions like metal or nucleic acid binding. The four-stranded β-meander, which gave rise to β-propellers, is one of these peptides. Other examples include the ββαβ-element shared by members of the cradle-loop barrel metafold [52] (see chapter 4.1), αβα-peptide of helix-turn-helix transcription factors, the β-hairpin of outer membrane β-barrels [226], or the αβα-peptide found in histones and the C-domain of the AAA+ module [28]. Oligomerization of such sub-domain-sized fragments might have been an intermediate step taken by short primordial polypeptide chains in order to gain a size sufficient for the emergent property of folding. Evolutionary ‘improvement’ of central information processing machineries (i.e. transcription and translation) would have eventually allowed the assembly of ancient peptides on one chain removing the necessity of oligomerization and enabling more efficient folding.

4.2.4 Conclusions

The structural and functional characterization of HP12 revealed a protein that contains the expected β-clam domain and a C-terminal domain, which is reminiscent of DNA sliding clamps, and folds as a homohexameric propeller thereby extending the upper limit of blades found in one propeller domain from ten to twelve. It provides a link in the evolutionary scenario that proposes the descent of fully amplified, monomeric propellers from ancestral blades. Because propellers do not show any other symmetry than that based on single blades [220], this scenario claims that, in principle, single blades should be capable of forming propellers by oligomerization. Interestingly, the two naturally occurring oligomeric intermediates characterized so far, HP12 and the Ralstonia-lectin, assemble from subunits containing two blades. For both proteins, the internal duplication signal testifies to the single bladed ancestor, at least on the genetic level. The next step in order to confirm the evolutionary scenario would be the identification or design of a single-bladed polypeptide chain forming an oligomeric propeller. An alternative explanation is that a first duplication and fusion event yielding a two-bladed monomer represents the acquisition of a 'critical mass', which could be required for folding purposes, at least under the conditions of extant organisms.

4.3 A PAN-like OB-fold in a Monomeric Proteasome Homolog

Within the proteasome-ATPase complex substrate recognition is performed by the N-domain of the regulatory ATPases, which includes a coiled coil domain and an OB-fold, whereas Ntn-hydrolase domains of the 20S proteasome are responsible for proteolysis.

Using sensitive HMM-HMM comparisons [133] we identified an archaeal protein family that accommodates on one polypeptide chain a proteasome-like Ntn-hydrolases domain and a C-terminal domain, which shares weak sequence similarity with the OB-fold of the proteasomal ATPase PAN. We present the structure of the hypothetical protein Mj0548 from *Methanocaldococcus jannaschii*, a member of this family, which is currently annotated as domain of unknown function 2121 (DUF2121) by the PFAM database [156]. We show that this protease lost the pro-peptide as well as the ability to self-compartmentalize, but acquired an OB-fold to the C-terminus, which may function as a substrate recognition domain. Sequence similarity and the distribution of this protein family suggest an origin from the proteasomal β-subunit in the last common ancestor of methanogenic archaea. Therefore, we refer to this protein as **m**onomeric **p**roteasome homolog of **m**ethanogens (MPM).

4.3.1 Experimental Procedures

Bioinformatics

The homology of MPM and proteasome-like Ntn-hydrolases was found through HHpred searches [133] with the β-subunit of archaeal proteasomes against the Pfam database [156]. Homologs of MPM were gathered by searching the non-redundant protein sequence database at NCBI with HHsenser, a method for exhaustive transitive profile searches based on Hidden Markov Model (HMM) comparisons [185]. Eight different starting points were used as queries for HHsenser runs with default settings: MPM from *M. jannaschii* the α- and β-subunits of the 20S proteasomes from *T. acidophilum* (PDB-ID 1PMA, chains A and B), *M. tuberculosis* (3MI0, chains A and C), and *S. cerevisiae* (1RYP, chains D and L) HslV from *T. maritima* (1M4Y, chain A). From each search the strict alignment was obtained. After removal of duplicates the pool of sequences was clustered to a maximum pairwise identity of 70% resulting in a set of 1711 sequences. These were clustered with CLANS using BLAST with a

BLOSUM80 substitution matrix as a comparison tool [131, 160]. Clustering was performed at default parameters using P-values < 1.

The Selection of sequences for the multiple alignment of proteasome-like Ntn-hydrolases included an archaeal (*T. acidophilum* – 1PMA), an actinobacterial (*M. tuberculosis* – 3MI0), and a eukaryotic (*S. cerevisiae* – 1RYP) proteasome as well as two HslV proteins (*E. coli* – 1G3I; *T. maritima* – 1M4Y). Furthermore, two members of the uncharacterized Anbu group from *Y. enterocolitica* and *T. elongatus* were selected. For the MPM group a phylogenetically representative selection was used including family members with and without the additional C-terminal domain. For the alignment of the C-terminal domain all MPM family members containing an OB-fold were selected and OB-folds from proteasome activating nucleotidases representing the major lineages of the archaeal kingdom including two proteins of known structure (*A. fulgidus* – 2WG5; *M. jannaschii* – 3H43)

Both alignments were interactively generated relying on a variety of methods. Alignments obtained from HHpred searches served as starting points. Closely related sequences were aligned with MUSCLE [187]. For the alignment of critical positions across the groups a structure based-sequence alignment was taken into account. Therefore, the proteins of known structure contained within the multiple alignment were superimposed. The DALI server was used for searches for similar structures [189]. Structural alignments were interactively conducted with the Swiss-Pdb Viewer [190] guided by HHpred and DALI searches. Representations of protein structures were done with PyMol (www.pymol.org).

For the phylogeny of MPM proteins, orthologs were extracted with PSI-BLAST [131], aligned with MUSCLE (the N-terminal Ntn-hydrolase domain only) [187], and subjected to neighbour-joining phylogenetic inference using the PHYLIP package (JTT matrix, 100 bootstrap replicates) [159]. The interactive tree of life project (iToL) was used to comparatively map the presence of MPM versus the 20S proteasome, as identified by BLAST searches, onto a phylogenetic tree [227, 228].

Protein production and purification

The DNA sequence encoding MjMPM (GI: 15668728, locus tag MJ_0548) was amplified from genomic DNA of *M. jannaschii* by polymerase chain reaction (PCR) and cloned into a pET28b expression vector (Novagen) with NdeI and XhoI restriction sites using the following primers: Mj2121-for 5'-GGCGGCCATATGAGTTTAATTATTTGCTACTATGG-3' and Mj2121-rev 5'-

CGCCTCGAGTTATTTATGAATTATGATATATTTGG-3′ (restriction sites underscored). The target protein was expressed in *E. coli* BL21 (DE3) RIL cells, which were grown in LB-medium at 37°C up to an OD of 1.0, induced with 0.5 mM IPTG, and harvested after over night expression at 20°C. Soluble fractions of cellular extracts were subjected to a Ni^{2+} affinity column. Bound protein was eluted with a linear imidazole gradient. Fractions containing the target protein (as monitored by SDS-PAGE analysis) were heated to 80°C for 20 min to precipitate thermolabile E. coli, cooled to 4°C and centrifuged. The supernatant was subjected to size-exclusion chromatography (Superdex 200, Amersham) after cleavage of the N-terminal His-tag with Thrombin. Using Amicon ultrafiltration devices pure protein was concentrated to 15 mg/mL. Analytical size-exclusion chromatography was performed with a calibrated 11/300 GL S200 column (Amersham). Labelling with selenomethionine was achieved by expression of the target protein in the methionine-auxotrophic strain E. coli B834 (DE3) grown in M9 minimal medium containing 4 mg/mL of selenomethionine.

Protease assays

Protease activity was assayed in a buffered solution containing 20 mM Hepes (pH 7.5), 150 mM NaCl, 0.5 mM TCEP, and 200 nM quenched TR-X BODIPY-Casein (Molecular Probes). Subsequently, different proteases (MjMPM, 20S proteasome from *M. mazei* and the ClpP homolog from *M. mazei* Orf Mm2878) were added to a final concentration of 25 nM. Increase of BODIPY fluorescence reflecting Casein cleavage (excitation wavelength λ_{ex}=580 nm, emission wavelength λ_{em}=620 nm) was followed on a Fluostar Optima (bmg) spectrometer for 2 hours at 37°C.

X-Ray Crystallography

For crystallization 400 nL of MjMPM, which was concentrated to 15 mg/mL in a buffered solution containing 20 mM Hepes (pH7.5), 150 mM NaCl, and 0.5 mM TCEP, were mixed with 400 nL of reservoir solution in 96-well plates with 75 ml reservoir volume by using a honeybee crystallization robot (Genomic Solutions). Drop images were obtained with the RockImager device (Formulatrix) and visually inspected. The reservoir solution contained 100 mM Hepes (pH 7.5) and 70% MPD. Full-length MPM crystallized by mixing the protein solution with a reservoir solution containing 100 mM Hepes (pH 7.5) and 70% MPD. Marcus Hartmann solved the crystal structure by

selenomethionine MAD phasing at a resolution of approximately 2.6 Å using 5 Se sites. The structure is not yet deposited at the Protein Data Bank.

4.3.2 Results and Discussion

Bioinformatics

The attempt to connect the various branches of the Ntn-hydrolase superfamily [79] on the sequence level using sensitive HMM-HMM comparisons was not successful. However, HHpred searches [133] with the β-subunit of the archaeal proteasome pointed to a weak but significant sequence similarity shared with Domain of unknown function 2121 (MPM), which is only found in methanogenic archaea. Closer inspection of the 26 members of this family revealed that all of them have a serine or a threonine residue following the N-terminal methionine suggesting a proteolytic function and the absence of a pro-peptide. Furthermore, the family can be divided into two sub-groups, the smaller of which contains just the Ntn-hydrolase domain whereas the other group contains an additional C-terminal domain. Secondary structure prediction indicated a long α-helix connecting the Ntn-hydrolase to a small all-β domain. HHpred searches retrieved the N-domain of proteasome activating nucleotidases (PAN) [3] as the only yet low scoring hit for MPM-C. In sum, this analysis pointed to an Ntn-hydrolase that could carry a separate substrate recognition domain on the same polypeptide chain.

In order to clarify the relationships among proteasome-like Ntn-hydrolases, we clustered all homologs detectable by sequence similarity alone (Figure 23). Sequences of the 20S proteasome occupy the centre of map separated by the lineage in which they are found. The main groups are archaea, eukaryotes, and actinobacteria, whose ancestor could have received the 20S proteasome including regulatory ATPase(s) by lateral gene transfer [87] (see chapter 5.3). Repeated gene duplication in the eukaryotic lineage yielded the seven different paralogs of α–subunits and up to ten different paralogs of β-subunits - seven plus three additional β-subunits of the immunoproteasome in mammals - forming the heteroheptameric rings of the eukaryotic proteasome [70]. The paralogy can be resolved by clustering at more stringent P-values (Figure 23, inlet). Within the proteasomal super-cluster, actinobacterial α-subunits form the most divergent group, which contains sequences of some verrucomicrobial organisms that, in turn, received the proteasome including the Pupylation-based tagging system from the actinobacterial lineage.

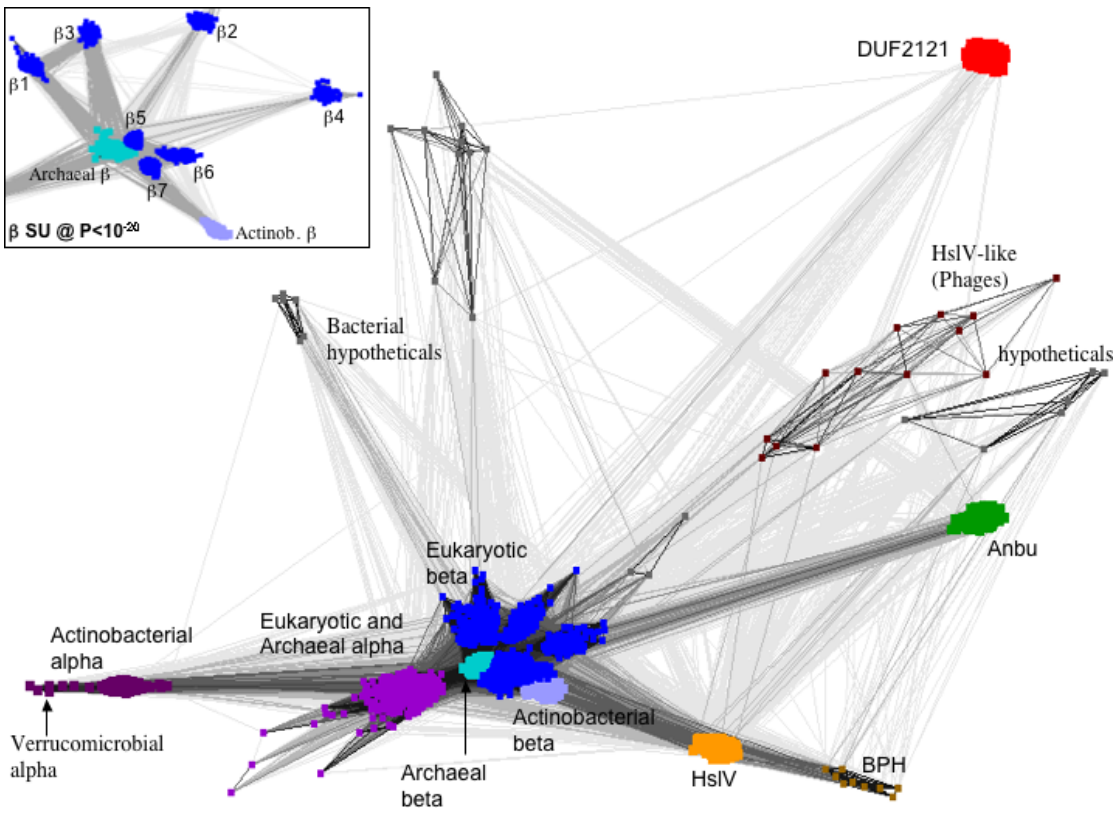


Figure 23. Cluster map of proteasome-like Ntn-hydrolases

The cluster analysis illustrates the relationships between Ntn-hydrolases whose similarity is detectable on the sequence level. Darker lines indicate lower BLAST P-values. The clustering uses all P-values < 1. The map contains 1711 sequences. Proteasomal sequences (α-subunits in magenta, β-subunits in blue) are in the centre of the map separated by the lineage, in which they occur. The paralogy of proteasomal subunits of the eukaryotic lineage are resolved by clustering at a more stringent P-value (see inlet, $P<10^{-20}$). Catalytically active subunits are more closely related to archaeal subunits. Sequences of the simpler HslV (yellow) and Anbu (green) proteases show a close relationship to Proteasomes reflected by rather low P-values ($P<10^{-15}$), but not to each other. The MPM group (MPM) forms the most divergent cluster (red) loosely connected ($P>10^{-4}$) especially to archaeal β-subunits.

The HslV (heat shock locus V) and the Anbu (ancestral β-subunit) cluster [179] are tightly connected to the proteasome cluster (P-values of about 10^{-15}). Both proteases are only found in bacterial organisms and consist of only one type of subunit ([75], see chapter 5.2). Whereas both genes frequently co-occur in one organism they are never found in organisms encoding a 20 S proteasome [229], with the exception of mitochondrially localized HslV of certain unicellular eukaryotes [230]. These aspects led to the proposal that HslV and Anbu are ancestral to the proteasome [179, 229]. However, the map illustrates that the DUF2121/MPM cluster forms a highly divergent group among proteasome-like Ntn-hydrolases, which is mainly connected to proteasomal β-subunits by relatively large P-values of about 10^{-4}.

MPM is a Protease

We selected the MPM family members from *Methanocaldococcus jannaschii* and *Methanosarcina mazei* for characterization, the latter of which does not contain the C-terminal all-β domain. It turned out to be insoluble under various expression conditions in *E. coli* and refolding attempts were not successful. The ortholog from *M. jannaschii*, MjMPM exhibits the thermal stability expected for a protein from a hyperthermophilic organism with a melting temperature of 91°C as determined by CD-spectroscopy. In Size-exclusion chromatography MjMPM eluted at a molecular mass corresponding to a monomer with a minor tendency to form a dimer. This is in contrast to all proteasome-like Ntn-hydrolases that form large self-compartmentalizing oligomers. In protease assays, however, MjMPM shows a caseinolytic activity comparable to the 20S proteasome and an archaeal ClpP homolog (Figure 24).

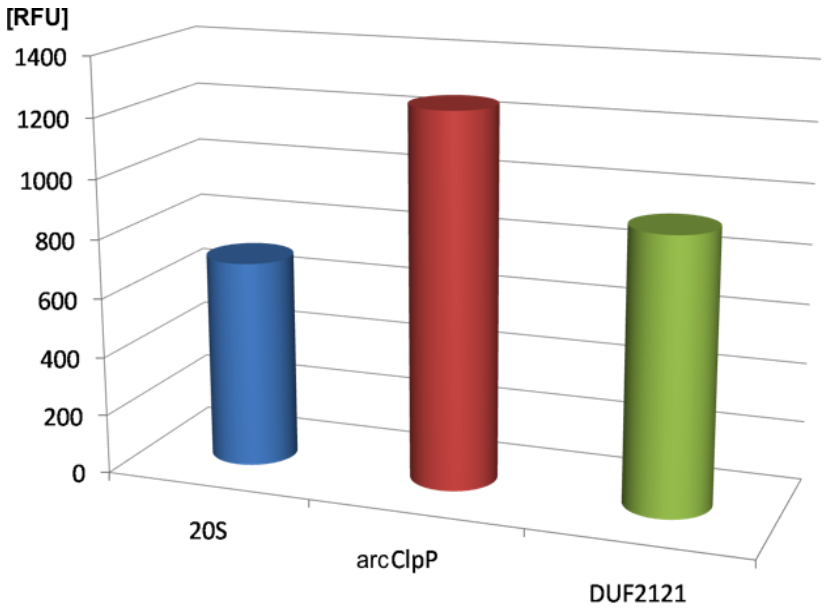


Figure 24. MjMPM (DUF2121) is a Caseinolytic Protease

The degradation assay with fluorescently labelled Casein shows that MjMPM has proteolytic activity comparable to the 20S proteasome from *Methanosarcina mazei* and an archaeal ClpP homolog (Mm2878, DUF114) from *Methanosarcina mazei*. Depicted is the relative fluorescence after two hours. The assay was conducted at 37°C, which is well below the temperature of the natural habitat of *M. jannaschii*, and shows the general proteolytic activity of MjMPM.

Casein is a model substrate for the characterization of endopeptidases. A covalently attached dye changes its fluorescence upon hydrolysis of the peptide chain. The assay confirms the protease activity of MjMPM expected from sequence similarity, especially because of the conservation of catalytically important residues (S1, D17, K67; Figure 27).

Structure of full-length MPM

The crystal structure of MPM from *M. jannaschii* was solved by selenomethionine MAD phasing at 2.5 Å using 5 Se sites. It contains a dimer whose interface is mainly formed by the extended β7 strands (not shown). Because the dimeric fraction corresponded to less than 5% of the monomer in size-exclusion chromatography, we conclude that crystallization accentuates dimerization yet it does not reflect the main functional state in solution. The monomeric full-length protein consists of an Ntn-hydrolase domain [76] connected to an OB-fold domain [60] via a long helical linker conforming the prediction (Figure 25). However, on the basis of the sequence analysis the formation of larger oligomers remained a possibility, because the most-similar Ntn-hydrolases are self-compartmentalizing and the remotely similar OB-fold of PAN hexamerizes (also in the absence of the coiled coil domain) [3], but this is not observed in the crystal structure of MjMPM

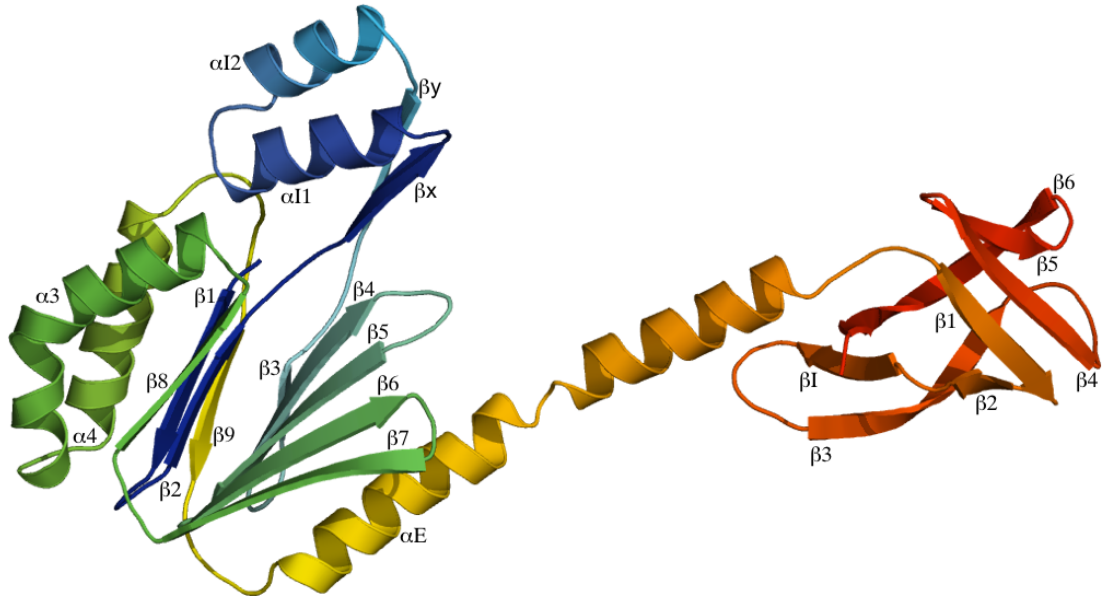


Figure 25. Structure of Full-length MPM from *M. jannaschii*

The structure of full-length MPM shows an N-terminal Ntn-hydrolase domain connected to a C-terminal OB-fold domain via a long helical linker. The cartoon representation is coloured in a rainbow-like succession starting with blue at the N-terminus and ending with red at the C-terminus. The annotation of the secondary structure elements of MPM-N follows the nomenclature for the archaeal proteasome. Helices α1 and α2 are not present in MPM-N. Their position is occupied by the C-terminal extension αE that leads to the OB-fold domain (see Figure 27). The short strands βx and βy accommodate a helical insertion (αI1 and αI2). The OB-fold contains an additional sixth strand and an insertion between β2 and β3. Their annotation is independent of the Ntn-hydrolase domain.

Whereas the proteasome and its regulatory ATPase form a highly complex assembly, MjMPM is a simpler protease whose substrates might be delivered by the C-terminal

OB-fold. Although the nature of the protein or peptide substrates of MjMPM is not known, the helical linker could assist in substrate delivery from the C-terminal binding domain to the N-terminal proteolytic domain. The presence of a heptad-repeat pattern within the linker initially suggested the formation of a coiled coil (not shown), which is not observed in the crystal structure. Instead, conserved hydrophobic residues located within αE remain exposed potentially allowing an interaction with partially unfolded substrates.

MPM-N is a Divergent Ntn-hydrolase

The N-terminal Ntn-hydrolase domain contains two central anti-parallel β-sheets (bI and bII) that pack against each other and are capped by a layer of α-helices forming an αββα-sandwich (Figure 25 and 26). Whereas sheet bII is constructed by the consecutive strands β3-β7, sheet bI consists of a central β-hairpin decorated by strands β8 and β9 on each side. The latter two strands are interrupted by the helices α3 and α4 providing the cap to sheet bI. Strand β9 departs into a loop that crosses sheet bII and leads into helix αE completing the αββα-architecture. The inner layers of the sandwich are joined by the short loop connecting β7 of sheet bII and β8 of sheet bI and by the extended region connecting β2 and β3. This region is not part of the αββα-core and comprises the short strands βx and βy, between which an α-helical hairpin (αI1, αI2) is accommodated.

Figure 26. Structural Comparison of MPM-N and Proteasomes (continued)

(E) Superimposition of proteasome-like Ntn-hydrolases. The superposition contains MPM-N (red), subunits of the 20S proteasome (*T. acidophilum*; α in green, β in blue), and HslV (*T. maritima*, yellow), and underlines the conservation the core of the αββα-sandwich. Proteasomal α-subunits use an additional helix at the N-terminus (α0) to mount the regulatory gates.

(F) Close up view of the active sites of proteins superimposed in (E). Side chains of the N-terminal nucleophile (T1/S1), a highly conserved acidic residue (E17/D17) and the invariant lysine (K33/K67) of active Ntn-hydrolases are shown in a sticks representation. The conservation includes the orientation of relevant side chains.

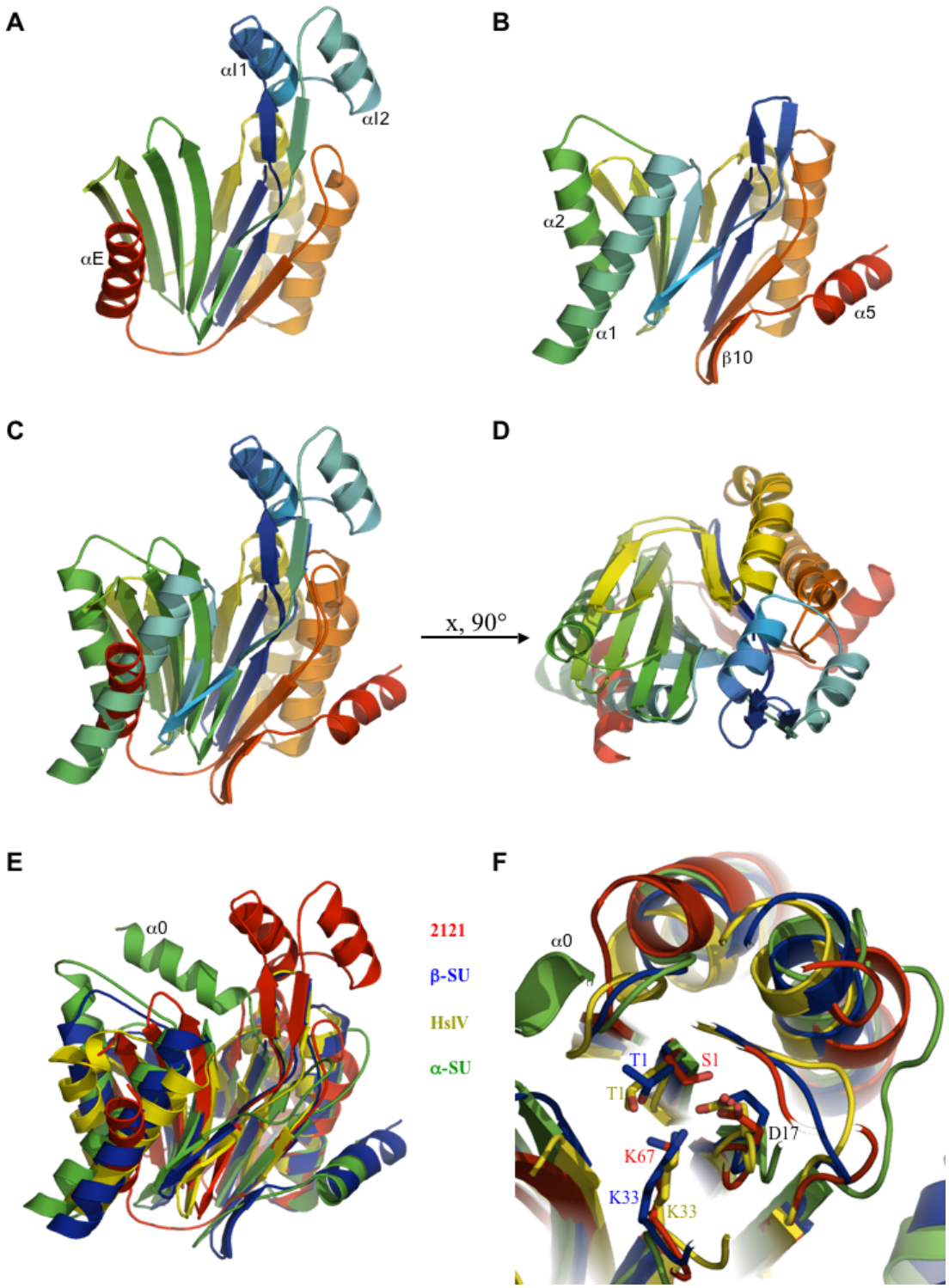


Figure 26. Structural Comparison of MPM-N and Proteasomes

(A and B) Cartoon representation of MPM-N (A) and a proteasomal β-subunit (*T. acidophilum*; B) coloured in rainbow succession. Differing secondary structure elements are annotated. Helix αE of MPM-N is found in a similar position as α1 and α2 in the β-subunit. The β-subunit contains additional strand (β10) and helix (α5) at the C-terminus. In MPM-N, the short hairpin outside of the core αββα-sandwich harbours an inserted α-helical hairpin (αI1, αI2).

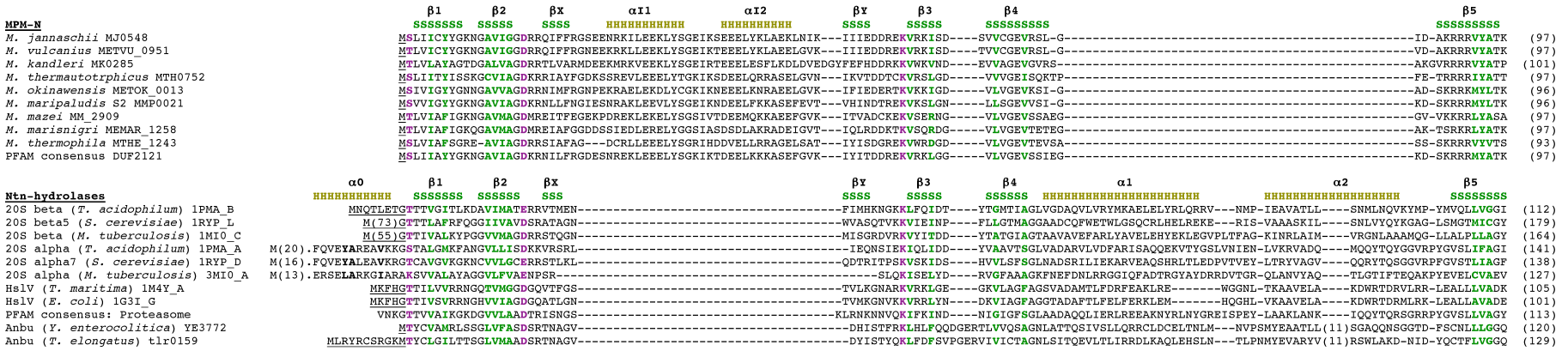
(C and D) Superimposition of MPM-N and a proteasomal β-subunit in side and top view. Colouring is as in A and B. The superimposition includes 98 Cα atoms with an RMSD of 1.48 Å.

The N-domain of MPM shares a conserved core with proteasome-like Ntn-hydrolases [1, 70, 231], which includes strands β1 to β9 of the two central β-sheets, and helices α3 and α4 that cap sheet bII (Figure 26). Helices α1 and α2, which cap sheet bI in proteasome-like Ntn-hydrolases, are not present in MPM-N. Instead, helix αE following strand β9 is placed in a structurally equivalent position. While strand β9 marks the C-terminus of the domain in HslV, in proteasome subunits the additional strand β10 decorates sheet bII and leads into helix α5 that joins α3 and α4 completing this layer of the αββα-sandwich. Furthermore, the hairpin βx-βy furnishes a large insertion in MPM (αI1, αI2). Superposition of proteasome-like Ntn-hydrolases onto MPM captures the main differences including the absence of helices α1 and α2 in MPM together with the structurally equivalent position of αE, and the insertion of αI1 and αI2 between βx and βy (Figure 26 C and D). The conservation of the core of the αββα-sandwich is reflected by root mean square (RMS) deviations of 1.5 Å over up to 100 residues located in the central β-sheets and the helices α3 and α4 (Figure 26 E). However, the highest degree of conservation resides in the region of the active site including the side chains of the N-terminal nucleophile (T1 in catalytically active proteasome-like Ntn-hydrolases, S1 in MjMPM), an acidic residue at the end of β2 (D17, D17) and an invariant lysine preceding β3 (K33, K67) (Figure 26 F, Figure 27). Interestingly, these residues are often still conserved in inactive proteasomal α-subunits.

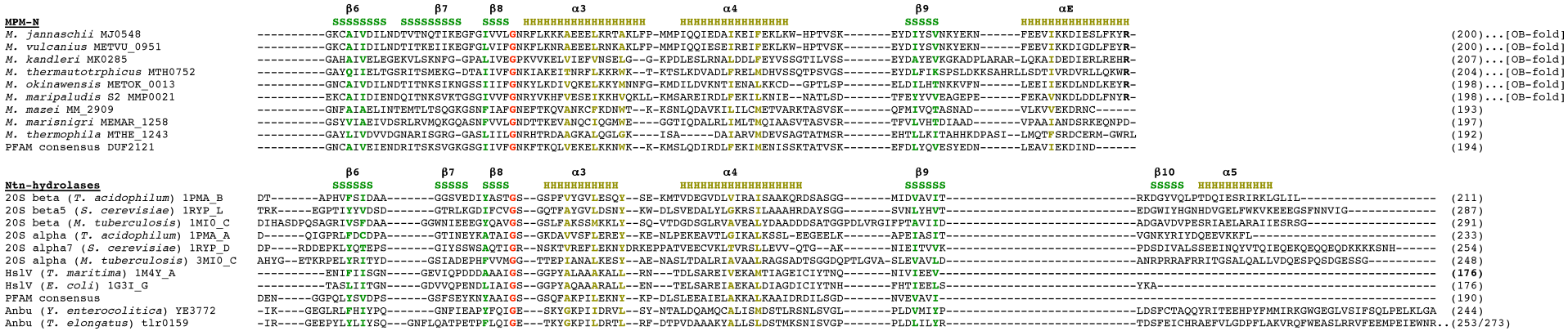
Figure 27. Multiple Ailgnment of MPM-N and Proteasomes

The alignment shows a representative selection of the MPM group in the upper half. The lower half contains an archaeal (*T. acidophilum*), an actinobacterial (*M. tuberculosis*), and a eukaryotic (*S. cerevisiae*; one α- one β-subunit) proteasome, two HslV proteins - all of known structure -, and two Anbu sequences. Additionally, both halves contain the corresponding Pfam consensus. Hydrophobic core residues are in bold (green in β-strands, yellow in helices). Residues important for catalysis are in magenta. The presence of a downstream OB-fold is denoted in angular brackets. Locus tags or, if available, PDB-identifiers are given next to the name of the organism. The consensus secondary structure is shown above the sequences (H, helix; S, strand).

A


MPM-N
M. jannaschii MJ0548
M. vulcanius METVU_0951
M. kandleri MK0285
M. thermautotrphicus MTH0752
M. okinawensis METOK_0013
M. maripaludis S2 MMP0021
M. mazei MM_2909
M. marisnigri MEMAR_1258
M. thermophila MTHE_1243
PFAM consensus DUF2121
Ntn-hydrolases
20S beta (T. acidophilum) 1PMA_B
20S beta5 (S. cerevisiae) 1RYP_L
20S beta (M. tuberculosis) 1MI0_C
20S alpha (T. acidophilum) 1PMA_A
20S alpha7 (S. cerevisiae) 1RYP_D
20S alpha (M. tuberculosis) 3MI0_A
HslV (T. maritima) 1M4Y_A
HslV (E. coli) 1G3I_G
PFAM consensus: Proteasome
Anbu (Y. enterocolitica) YE3772
Anbu (T. elongatus) tlr0159

B


MPM-N
M. jannaschii MJ0548
M. vulcanius METVU_0951
M. kandleri MK0285
M. thermautotrphicus MTH0752
M. okinawensis METOK_0013
M. maripaludis S2 MMP0021
M. mazei MM_2909
M. marisnigri MEMAR_1258
M. thermophila MTHE_1243
PFAM consensus DUF2121
Ntn-hydrolases
20S beta (T. acidophilum) 1PMA_B
20S beta5 (S. cerevisiae) 1RYP_L
20S beta (M. tuberculosis) 1MI0_C
20S alpha (T. acidophilum) 1PMA_A
20S alpha7 (S. cerevisiae) 1RYP_D
20S alpha (M. tuberculosis) 3MI0_C
HslV (T. maritima) 1M4Y_A
HslV (E. coli) 1G3I_G
PFAM consensus
Anbu (Y. enterocolitica) YE3772
Anbu (T. elongatus) tlr0159

The presence of the insertion comprising αI1 and αI2 and the loss of helices α1 and α2 provide an explanation for the loss the self-compartmentalizing phenotype in MPM proteins. The helices α1 and α2 are involved in the interaction of α- and β-rings in the 20S proteasome in a manner that is unlikely to be fulfilled by helix αE of MPM alone. Furthermore, the short βx-βy-hairpin that extends the αββα-sandwich in HslV and proteasomal β-subunits is important for interlocking of the hexameric rings of HslV and of the inner β-rings of the 20S proteasome. Although both strands are present in MPM, the inserted helical hairpin αI1-αI2 would collide with the assembly of subunits found in HslV and the 20S proteasome.

Table 5. Summary of Sequence and Structure Comparisons for MjMPM

		HHPRED-SCORES[1]			DALI-SCORES[2]			
Protein	**PDB-ID**	**Prob. [%]**	**E-Value**	**P-Value**	**Z-Score**	**RMSD [A]**	**lali[2]**	**Fold**
MPM-N (1-201)	-	**Query (Structure not yet deposited)**						**Ntn-hydrolase**
Archaeal β-Subunit	1PMA_B	96.3	0.13	2.6e-06	13.0	2.5	133	Ntn-hydrolase
Eukaryotic β-Subunit	1RYP_L	93.1	3.9	7.6e-05	11.4	2.4	133	Ntn-hydrolase
Actinob. β-Subunit	3MI0_C	80.4	45	0.00088	10.1	2.5	116	Ntn-hydrolase
Archaeal α-Subunit	1PMA_A	94.0	0.46	9.0e-06	10.2	3.8	127	Ntn-hydrolase
Eukaryotic α-Subunit	1RYP_D	52.3	11	0.00021	9.9	3.6	127	Ntn-hydrolase
Actinob. α-Subunit	3MI0_A	86.6	9.2	0.00018	11.5	2.8	136	Ntn-hydrolase
HslV	1M4Y_A	78.4	1.6	3.1e-05	10.6	2.6	123	Ntn-hydrolase
HslV	1G3I_A	71.8	3.0	5.8e-05	10.5	2.6	124	Ntn-hydrolase
MPM-C (224-293)	-	**Query**						**OB-fold**
PAN-N	2WG5_A[3]	63.1	69	0.0014	2.6	2.0	36	OB-fold
PAN-N	3H43_A	-	-	-	-	1.2*	28[4]	OB-fold
MPM-C 2nd half[5]	-	-	-	-	-	1.5*	15[4]	β-Meander

The targets are taken from the HHpred and Dali hit lists and represent a highly diverse selection of proteins, which are included in the Multiple Aignments (Figures 27 and 29), rather than the ranking by the servers. Despite low scores for certain targets the degree of sequence and structure similarity is indicative of homology.

1: HHpred searches were performed in default settings against the Protein Data Bank, release of May 14 2011, filtered for a maximum of 70% pairwise sequence identity at http://toolkit.tuebingen.mpg.de/HHpred.

2: Dali searches were done at http://ekhidna.biocenter.helsinki.fi/dali_server/. Lali denotes the number of residues included in the superimposition.

3: This HHpred hit is only obtained when full-length MjMPM (residues 1-293) is used as the query.
4: Superimposition was interactively conducted using Swiss-PDB Viewer.
5: Superposition of both three stranded β-meanders of the OB-fold pointing to an internal repeat (Figure 30).

Autocatalytic processing of an N-terminal pro-peptide is observed in various branches of the Ntn-hydrolase fold allowing the α-amino nitrogen of the N-terminal nucleophile to function as a general base [76]. For the proteasomal β-subunit of *T. acidophilum* it has been shown that only a threonine residue confers full functionality in proteolysis and autolysis, the latter of which is impaired by a serine residue in this position [67, 232]. Moreover, a highly conserved glycine residue, which adopts a γ-turn conformation in proximity to the nucleophile, facilitates autoprocessing [78]. MPM proteins do not contain a propeptide and mostly serine instead of threonine residues in position 2 (Figure 27) suggesting that, instead of autolysis, the universally conserved enzyme methionine aminopeptidase removes the N-terminal methionine from the nascent chain *in vivo*.

MPM-C Forms a Symmetrical OB-fold

The C-terminal domain forms a six-stranded β-barrel. The additional strand, βI, which is inserted between β2 and β3 and pairs with the N-terminal portion of β3 and the C-terminal of β6, does not participate in the barrel architecture. HHpred searches and DALI searches indicated a weak similarity to OB-fold proteins including the N-domain of PAN proteins [133, 189] (Table 5). Interactive superimpositions of MPM-C and PAN-N using Swiss-PDB Viewer [190] resulted in RMS deviations as low as 1.0 Å over 28 residues located in all five canonical strands of the OB-fold (Figure 28). This procedure illustrated the displacement of the β4-β5-hairpin and the invasion of strand β6, which closes the barrel between β3 and β5 in MPM-C. In contrast, direct pairing of β3 and β5 provides closure in regular five-stranded OB-fold barrels. Furthermore, the superimposition generated a structure-based sequence alignment that enabled the

recognition of conserved hydrophobic residues, a GD-box, and the large insertion on the sequence level (Figure 29), suggesting a homologous relationship.

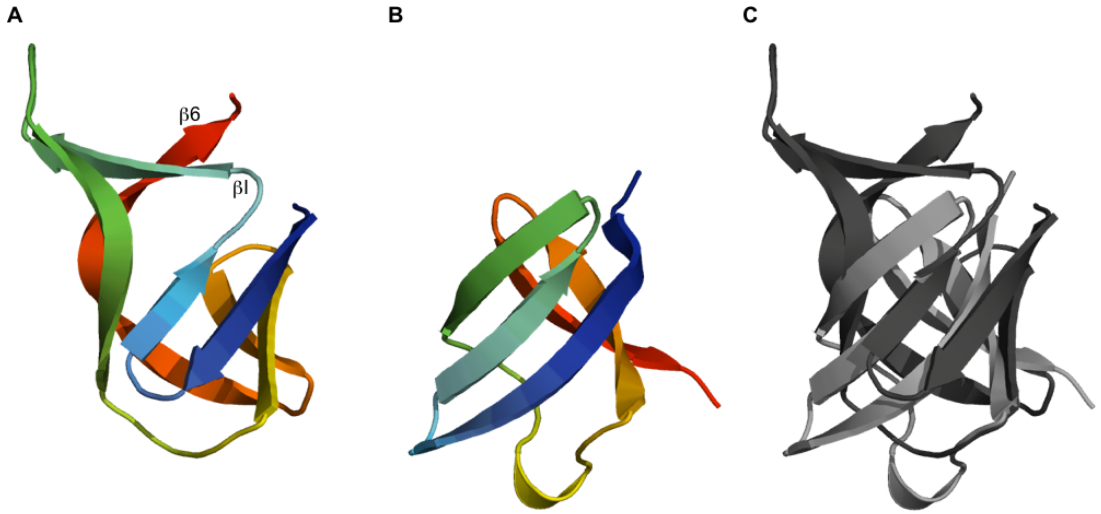
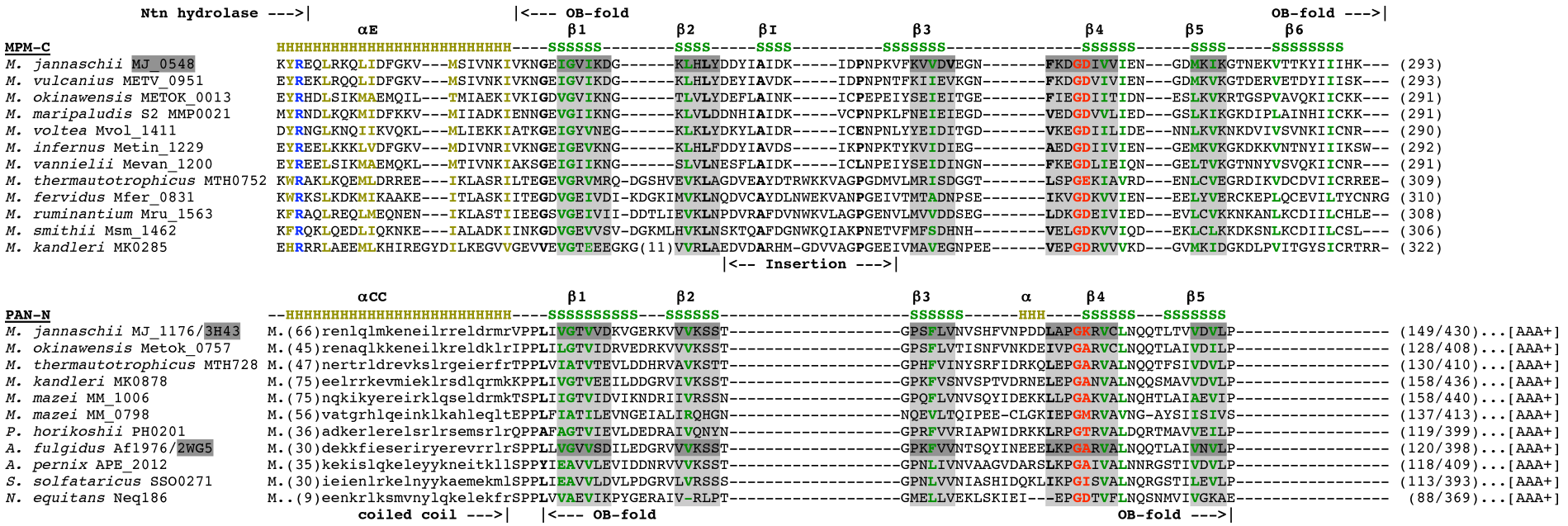


Figure 28. MPM-C Forms a Divergent OB-fold

(A and B) Cartoon representation of MPM-C (A) and PAN-N (*M. jannaschii*; 3H43, chain B) coloured in rainbow succession. The additional sixth strand and the inserted strand between β2 and β3 of MPM-N are annotated. These secondary structure elements constitute deviations from canonical five-stranded OB-folds.
(C) Superimposition of MPM-C (dark grey) and PAN-N (light grey). 28 residues located within the five conserved strands are superimposed with an RMSD of 0.98 Å.

Figure 29. Multiple Alignment of MPM-C and the OB-fold of PAN

The alignment shows a selection of OB-folds of the MPM group in the upper half, and a phylogenetically representative selection of the N-domains of archaeal proteasome activating nucleotidases (PAN-N) in the lower half. Hydrophobic core residues are in bold (green in β-strands, yellow in helices), the core residues of the GD-box are in red. These patterns suggest a homologous relationship. Grey boxes mark positions included in the superimposition (dark grey for available structures used for the generation of the structure-based sequence alignment). Locus tags or PDB-identifiers are given next to the name of the organism. Residues of the non-equivalent coiled-coil of PAN-N are in lower case letters.


Ntn hydrolase ---> |<--- OB-fold OB-fold ---> |
αE β1 β2 βI β3 β4 β5 β6
MPM-C
M. jannaschii MJ_0548 (293)
M. vulcanius METV_0951 (293)
M. okinawensis METOK_0013 (291)
M. maripaludis S2 MMP0021 (291)
M. voltea Mvol_1411 (290)
M. infernus Metin_1229 (292)
M. vannielii Mevan_1200 (291)
M. thermautotrophicus MTH0752 (309)
M. fervidus Mfer_0831 (310)
M. ruminantium Mru_1563 (308)
M. smithii Msm_1462 (306)
M. kandleri MK0285 (322)
|<-- Insertion --> |
αCC β1 β2 β3 α β4 β5
PAN-N
M. jannaschii MJ_1176/3H43 (149/430)...[AAA+]
M. okinawensis Metok_0757 (128/408)...[AAA+]
M. thermautotrophicus MTH728 (130/410)...[AAA+]
M. kandleri MK0878 (158/436)...[AAA+]
M. mazei MM_1006 (158/440)...[AAA+]
M. mazei MM_0798 (137/413)...[AAA+]
P. horikoshii PH0201 (119/399)...[AAA+]
A. fulgidus Af1976/2WG5 (120/398)...[AAA+]
A. pernix APE_2012 (118/409)...[AAA+]
S. solfataricus SSO0271 (113/393)...[AAA+]
N. equitans Neq186 (88/369)...[AAA+]
coiled coil ---> |<--- OB-fold OB-fold ---> |

The OB-fold is among the most widespread binding domains. The SCOP database assigns 16 superfamilies to the OB-fold [213], and only some of these can be connected by sequence similarity [140]. Therefore, the fold of MPM-C can be viewed as a non-canonical OB-fold, but significant similarity to other groups than the OB-fold of PAN is currently not detected.

The invading strand β6 gives the OB-fold a symmetrical appearance, which is broken by the insertion between β2 and β3. Both β-meanders, consisting of β1-β3 and β4-β6, can be reasonably superimposed with an RMSD of 1.4 Å comprising 15 residues in all three β-strands despite the distortion caused by the inserted strand βI in the first meander. This uncovers internal sequence similarity, most pronounced in the hairpins β1-β2 and β4-β5, and suggests that this OB-fold originated by duplication of a three-stranded β-meander (Figure 30).

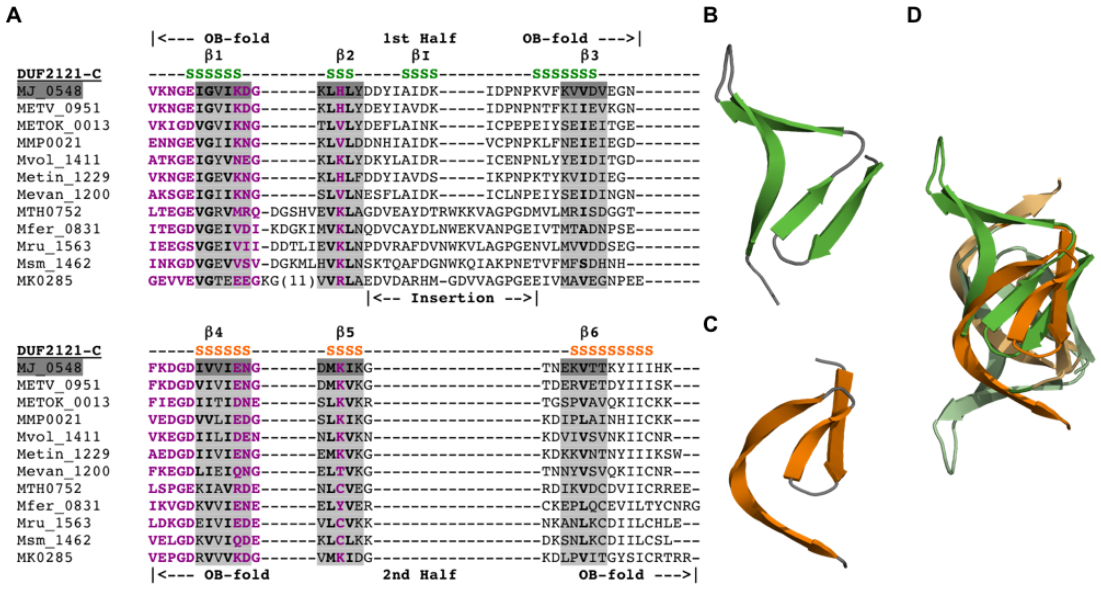


Figure 30. The Symmetrical OB-fold of MPM-C Contains an Internal Repeat

The closure of the β-barrel by the invading strand β6 results in an OB-fold with internal pseudo-two-fold symmetry broken by the inserted strand βI. Superimposition and structure –based sequence alignment were interactively conducted using Swiss-PDB Viewer.

(A) The Multiple alignment of both three-stranded β-meanders within the OB-fold reveal a duplication signal reflected by conserved hydrophobic residues (bold) and polar pattern (magenta). Grey boxes mark structurally similar positions based on the two halves of MjMPM-C.

(B and C) The b-meanders β1-β3 (green) and β4-β6 (orange) are structurally similar

(D) Superposition of the two halves β1-β3 and β4-β6 illustrates the internal symmetry within the C-terminal domain of MPM-proteins.

Currently it is unclear, whether this scenario applies to other OB-fold proteins, because MPM-C lacks significant sequence similarity to major groups of the OB-fold.

Furthermore, we could not detect an internal duplication signal in five-stranded OB-fold proteins. Nevertheless, it is tempting to speculate that the OB-fold originated by the duplication of a three-stranded β-meander followed by a loss of the sixth strand. The five-stranded OB-fold performs a plethora of binding functions and is highly populated, reflected by the call to one of ten superfolds [146, 233]. By contrast, the number of symmetrical six-stranded OB-folds is sparse. This argues for an extraordinary stability and plasticity of this type of a five-stranded β-barrel, which might have been the driving force for the loss of the sixth strand. In this context it is interesting to note that in a directed evolution experiment the N-terminal three stranded β-meander of the OB-fold of cold shock protein A was able to form a folded protein, 1b11, with the N-terminal meander of the OB-fold of S1 RNA-binding protein [234], yielding a six-stranded β-barrel reminiscent on an OB-fold [235].

MPM is a Derived but Simplified Proteasome Homolog

The distribution of the MPM family is restricted to closely related methanogenic archaea whereas all archaeal organisms encode the 20S proteasome, which was shown to be essential in *Haloferax volcanii* [73]. This pattern suggests that the MPM protease is a derived characteristic, which originated in an ancestor of methanogenic archaea (Figure 31). Because methanogens form a paraphyletic group [166], which excludes for instance halobacteria and archaeoglobales, secondary loss of MPM in certain clades is likely (evidence for lateral gene transfer is absent in the MPM family). Nevertheless, we cannot exclude that the MPM family represents an ancestral feature that experienced massive secondary loss. However, we could not detect significant sequence similarity to other members of the Ntn-hydrolase fold, comprising penicillin acylase, asparaginase, γ-glutamyltranspeptidase and others [62, 142].

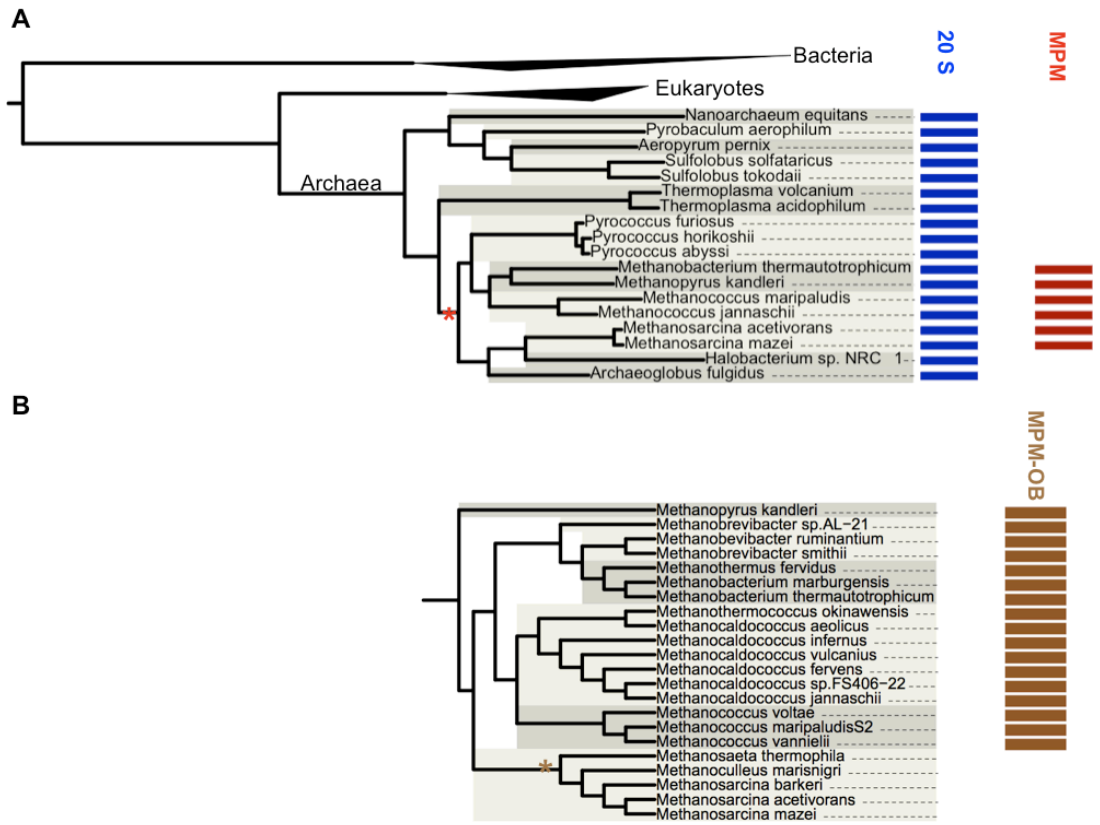


Figure 31. MPM is a Derived Character of Methanogenic Archaea

(A) MPM vs. the 20S proteasome in Archaea. The presence of MPM and the 20S proteasome is mapped onto the interactive tree of life. The bacterial and eukaryotic kingdoms are collapsed. Although methanogens are considered to form a paraphyletic group [166], the restricted occurrence of MPM proteins and their similarity to the universally conserved archaeal proteasome suggests a derived origin in the last common ancestor of methanogens (presumably at the node marked with a red asterisk).

(B) Phylogeny of MPM proteins. The congruence of the MPM phylogeny and general archaeal phylogenies (Figure 9 B) suggests the absence of lateral gene transfer of MPM proteins. The OB-fold is an ancestral trait of the MPM family and was lost in methanosarcinales (brown asterisk). Most MPM proteins use a serine residue instead of threonine as in proteaseomes for the nucleophile (Figure 27).

Although the organismal distribution suggests that MPM is derived, the molecular phenotype appears to be rather simplified. The ability to form elaborate micro-compartments has been lost as well as the pro-peptide removing the requirement for autoprocessing, which is reflected by the acceptance of serine as the nucleophile. Instead, MPM has acquired an OB-fold to the C-terminus, which might serve as a substrate recognition domain, through an unspecified recombination or shuffling event. In contrast, substrate recognition in self-compartmentalizing proteases is fulfilled by the N-domains of hexameric AAA (+) proteins like PAN or HslU. Therefore, we consider MPM as a secondary simplification.

4.3.3 Conclusions

The Pfam database is a widely used repository for annotating the function of sequenced genomes [156]. However, a growing fraction of families in the Pfam database is in search of a function. These domains of unknown function are considered to "represent biological functions that are specific for certain groups of organisms or environmental conditions rather than being part of the core machinery common to all life [157]". We presented the structure and function of a member of one of these families, DUF2121 (MPM), which is restricted to methanogenic archaea. Although our characterization does not provide details for the function of MPM proteins *in vivo*, it reveals a basal proteolytic activity and proposes an evolutionary scenario for their descent from proteasomal β-subunits within a subgroup of the archaeal lineage. The derived character of the MPM family suggests a secondary simplification in comparison to its proteasomal ancestor being substantiated in a loss of the self-compartmentalizing phenotype and an acquisition of a C-terminal substrate recognition domain, which homologous to the N-domain of proteasomal ATPases. Furthermore, the internal symmetry of the OB-fold could reflect its origin by the duplication of an ancestral three-stranded β-meander. Whether this scenario is informative for the evolution of canonical five-stranded OB-folds by a loss of the sixth strand remains to be substantiated. However, domains of unknown function are considered to represent uncharted regions of the protein universe [158]. Our characterization of a highly divergent niche family, MPM (DUF2121), explores uncharted areas of Ntn-hydrolases and OB-folds.

5 Origins of Proteasomal Protein Degradation

5.1 Prediction of a Network of AAA ATPases Regulating the Archaeal Proteasome

The AAA ATPases Rpt1-6 (in eukaryotes), PAN (in archaea), and ARC/MPA (in actinobacteria) play a crucial role in regulating the 20S proteasome [55, 56, 236]. Members of this orthologous group share a common domain composition consisting of an N-terminal substrate recognition part, which can be subdivided into a coiled coil and a OB-fold domain (two in case of ARC) [3, 59], and a C-terminal AAA+ unfoldase module followed by the HbYX interaction motif [83]. On the one hand, their N-domain is involved in substrate recognition [58, 59] and interaction with other subunits of the regulatory particle thereby binding degradation tags like Pup [100, 237]. On the other hand, their C-terminal tails penetrate the pockets between proteasomal α-subunits and stimulate gate opening via binding of the conserved HbYX motif [68, 83]. In the fully differentiated Rpt heterohexamer of the eukaryotic proteasome the presence of the HbYX motif is restricted to Rpt2 and Rpt5, which bind to specific pockets of the heteroheptameric α-rings [238, 239].

The dynamic N-terminal tails of the α-subunits form the gate of the 20S proteasome. These tails exist in two distinct conformations placing them either inside of the opening of the barrel or outside. The equilibrium between these conformations is influenced by the presence of regulatory particles shifting it towards the open gate conformation thereby stimulating the rate of hydrolysis [69]. Within the binding pockets of the α-subunits, binding of the C-terminal interaction motif causes a repositioning of the proline17 reverse turn, inducing the open gate conformation [68, 84]. The penultimate tyrosine of the HbYX motif contacts glycine19 of the α-subunits and stabilizes this confromation. The eukaryotic 11S/PA26 regulator stimulates gate-opening by the same mechanism [84]. This non-ATPase particle forms a heptamer and has been used as a model to obtain these insights, because absence of symmetry mismatchs fascilate tighter binding and co-crystalization [69, 84, 238].

Although the proteasome degrades intrinsically disordered proteins in absence of regulatory particles [240, 241], the presence of regulatory ATPases is crucial for viability of *S. cerevisiae* [242]. Whereas the 20S proteasome is essential (requiring a knock-out of both α-subunits) for the archaeon *Haloferax volcani*, it tolerates a double knock-out of both PAN proteins [73]. We observed that certain archaeal organisms like *Thermoplasma acidophilum* do not encode PAN [243]. Instead, we detected the HbYX motif at the C-terminus of closely related AAA ATPases of the CDC48 group in *T.*

acidophilum and *H. volcanii*. This prompted us to a systematic analysis of archaeal AAA ATPases and their C-termini.

5.1.1 Procedures

Homologs of archaeal AAA proteins were identified with HHsenser searching the unclustered non-redundant database of archaeal proteins (NCBI, nr_arc) with the AAA+ module of AMA from *Methanosarcina mazei* (GI: 21226406, Mm_0304, residues 119-372) [185]. Assignment to orthologous groups of full-length sequences was based on cluster analyses using CLANS [160]. P-values for clustering were selected interactively in order to achieve formation of orthologous groups. Groups of AAA proteins were distinguished from other members of the AAA+ superfamily using different P-value cutoffs and relying on our classification of AAA+ proteins [27]. Members of orthologous groups were verified by testing for concordant domain composition using HHpred and MUSCLE multiple alignments [133, 187].

Proteins of archaeal organisms included in the interactive tree of life project were pooled from the purified sets of sequences of each orthologous group. Subsequently, the presence or absence of proteins was assigned to the respective organisms and visualized using iToL [227]. For putative proteasomal ATPases of the PAN, CDC48, and AMA group the presence of group members was mapped onto the archaeal taxonomy available at NCBI containing 81 fully sequenced genomes (deposited when this analysis was performed) and visualized with iToL.

C-terminal peptides comprising the last seven residues of AAA proteins were extracted from full-length sequences. A multiple alignment of archaeal α-subunits, generated with MUSCLE [187], and especially the conservation of residues forming the binding pocket for the HbYX interaction motif (proline17, glycine19) indicated that archaeal proteasomes are most likely activated by the same mechanism.

5.1.2 Results and Discussion

AAA proteins cluster as a distinct group within the AAA+ superfamily and are classified based on the presence of the so-called second region of homology (SRH). Sequence variations within the ATPase domain, especially in the SRH, and the domain

architecture in the entire proteins enabled the identification of six major clades [30, 38]. Whereas the clades of metalloproteases and BCS1-like proteins are not found in archaea, the clades of meiotic proteins, proteasome subunits, and the D1 and D2 ATPase domains are represented by archaeal Vps4 [244, 245], PAN and CDC48 proteins, forming the main groups of AAA proteins in archaea. Additionally, the three subgroups of MBA, a group exemplified by open reading frame 854 from *M. mazei* (Mm0854), and AMA represent lineage-specific inventions that also experienced secondary loss in certain taxa. MBA proteins consisting of a transmembrane helix and a duplicated AAA+ module are restricted to sulfolobales [246]. AMA proteins are found in archaeoglobales and methanogens [39] in contrast to Mm854-like proteins, which are also present in archaeoglobales, and additionally in halobacteria. Only certain methanogens, the methanosarcinales, contain a Mm0854-like protein (Figure 32). Inspection of the full-length sequences revealed the presence of the HbYX motif in PAN, many CDC48 proteins, and the AMA group.

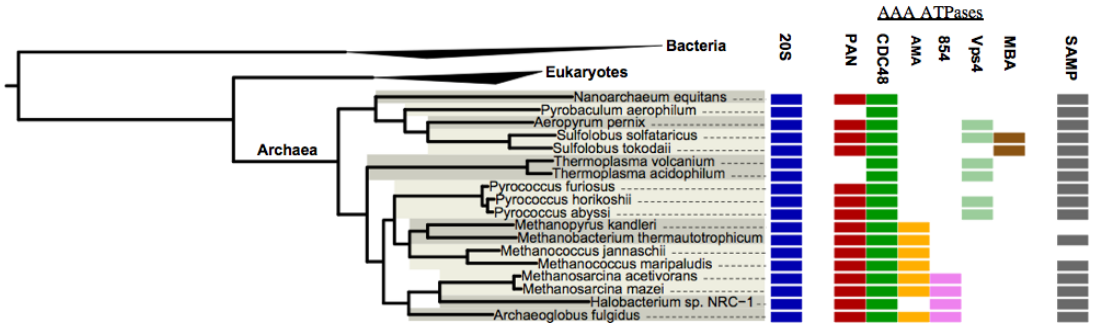
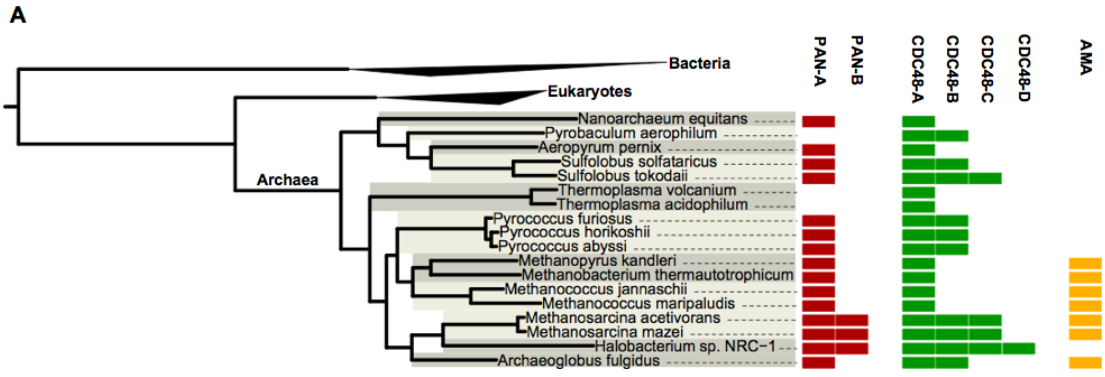


Figure 32. Distribution of AAA ATPases in Archaea

The figure illustrates the distribution of all subgroups of AAA proteins found in the archaeal kingdom including CDC48, PAN, AMA, Mm0854-like, the archaeal homolog of the vacuolar sorting protein Vps4 and the membrane-bound ATPase (MBA) of sulfolobales. Whereas CDC48 is found in all sequenced archaea, the canonical proteasome activating nucleotidase PAN is not present in thermoplasmata and thermoproteales like *Pyrobaculum aerophilum*. The pattern of occurrence suggests that PAN has been lost in these lineages. In contrast, the 20S proteasome is universally conserved in archaea. Because the presence of a regulatory ATPase is important for proteasome function, CDC48 could serve as a proteasomal ATPase at least in species lacking PAN. Despite a high degree of conservation the degradation tag SAMP seems to be dispensable indicated by the loss in certain methanogens. The figure was generated with iToL [227].

CDC48 is the only AAA protein, which is contained by all archaeal organisms sequenced to date, while major archaeal groups, among them thermoplasmata, thermoproteales, and the deep-branching thaumarchaeota and korarchaeota, lack the canonical proteasome activating nucleotidase. CDC48 proteins show a larger number of paralogs than PAN, which experienced duplication in the ancestor of halobacteria and

methanosarcinales. Up to four genes encoding CDC48-like proteins are present in some halobacteria (e.g. *Halobacterium sp.* NRC 1) and methanosarcinales (*Methanosarcina barkeri*); most crenarchaeota contain two CDC48 paralogs (Figure 33).



B

C-terminal Peptide

Organism	PAN-A	PAN-B	CDC48-A	CDC48-B	CDC48-C	CDC48-D	AMA
Nanoarchaeum_equitans	KGLSYFG		KNNPYLG				
Pyrobaculum_aerophilum			ATPPLYL	RRAIAGL			
Sulfolobus_solfataricus	ERREKYS		VKPSTYA	IKRSVIG			
Sulfolobus_tokodaii	DRTEKYV		LKPSTFT	RRASGYA	ELKRSVL		
Aeropyrum_pernix	DTIATVI		EAPTITL	LKKLTLG			
Thermoplasma_volcanium			QDQGLYL				
Thermoplasma_acidophilum			QDQGLYL				
Pyrococcus_furiosus	SHEVIYG		EPDYFTG	KFKQARG			
Pyrococcus_horikoshii	SHEVIYG		EPDYFTG	QFKQARG			
Pyrococcus_abyssi	SHEIIYG		EPDYFTG	QFKQARG			
Methanopyrus_kandleri	EFKRAYH		EYQRLAG				AVRIYHR
Methanobacter_thermautotrophicus	ETGVMFG		EEQVQYH				EPKHMFA
Methanocaldococcus_jannaschii	HLDVLYR		EEGKEVI				APQHLYL
Methanococcus_maripaludis	QLTVMYG		ENTGYHG				EPLHLYA
Methanosarcina_acetivorans	PETTMFV	IPQAMFV	AGYVGYR	IENPYYV	VPNFPKV		DPTHLYV
Methanosarcina_mazei	LPETMFV	IPQAMFV	AGLVGYR	IESPYYI	KDQETRV		DPTHLYV
Halobacterium_sp._NRC-1	DVSRTFA	RYPSYIQ	SEHIGFQ	PASGAFR	EKPAEAV	PSVDGAG	
Archaeoglobus_fulgidus	LKGVMFV		DFHRMYA	KPFMGYG			PPKQMFV

Figure 33. Putative Proteasomal ATPases and their C-terminal Peptides

Analysis of the C-termini of archaeal AAA proteins indicated the presence of the HbYX motif not only in PAN but also CDC48 and AMA providing support for the hypothesis that a network of AAA proteins regulates the archaeal proteasome.

(A) The number of putative proteasomal ATPases varies throughout the archaeal kingdom ranging from one CDC48 ortholog in thermoplasmata to five PAN, CDC48 and AMA proteins in certain methanosarcinales and halobacteria. CDC48 is not only more conserved than PAN, but also shows a higher degree of paralogy.

(B) The sequences of the last seven residues of the ATPases shown in panel A (same order) illustrate the presence of the HbYX motif (Hb in blue, Y/F in magenta) in AAA proteins other than PAN. Most notably, species lacking PAN contain at least one CDC48 ortholog that shows the HbYX motif in canonical form (grey boxes). At least in these species, CDC48 is likely to function as the proteasomal ATPase, because the conservation of the binding pocket in proteasomal α-subunits suggests that all archaeal proteasomes are regulated by the same mechanism. Furthermore, all AMA proteins contain the HbYX motif in form suggesting that it represents an invention, which increases the repertoire of proteasomal protein degradation in archaeoglobales and methanogens. Although Halobacteria presumably lost AMA, the number of PAN and CDC48 paralogs also indicates a network of putative proteasomal ATPases in these particularly complex archaea.

Systematic analysis of their C-termini revealed the presence of the HbYX-motif in all PAN proteins, and in certain CDC48 proteins. The pattern emerged that organisms

lacking PAN encode at least one CDC48 protein, which contains the HbYX motif in canonical form. Furthermore, the HbYX-motif is detected in all members of the AMA group increasing the repertoire of putative proteasomal ATPases in these organisms.

In all archaea, there is at least one AAA protein with the HbYX-motif, supporting the notion that regulation of the proteasome through an ATPase is a conserved functionality, despite the fact that a double knock-out of both PAN proteins is not lethal for *Haloferax volcanii* [73]. However, *H. volcanii* encodes three CDC48 proteins, two of which have the HbYX motif. We propose that these CDC48 proteins compensate for the deletion of PAN. The number of putative proteasomal ATPases varies between organisms from one in thermoplasmata to five in certain methanosarcinales (Figure 34) consistent with the idea that archaea employ a network of AAA proteins in proteasomal protein degradation. Although the importance of the HbYX motif within the last seven residues of PAN and Rpt proteins has been shown [83], the association of AMA and CDC48 proteins with the proteasome awaits experimental verification. Nevertheless, the observed patterns are robust in 81 archaeal organisms (Figure 34).

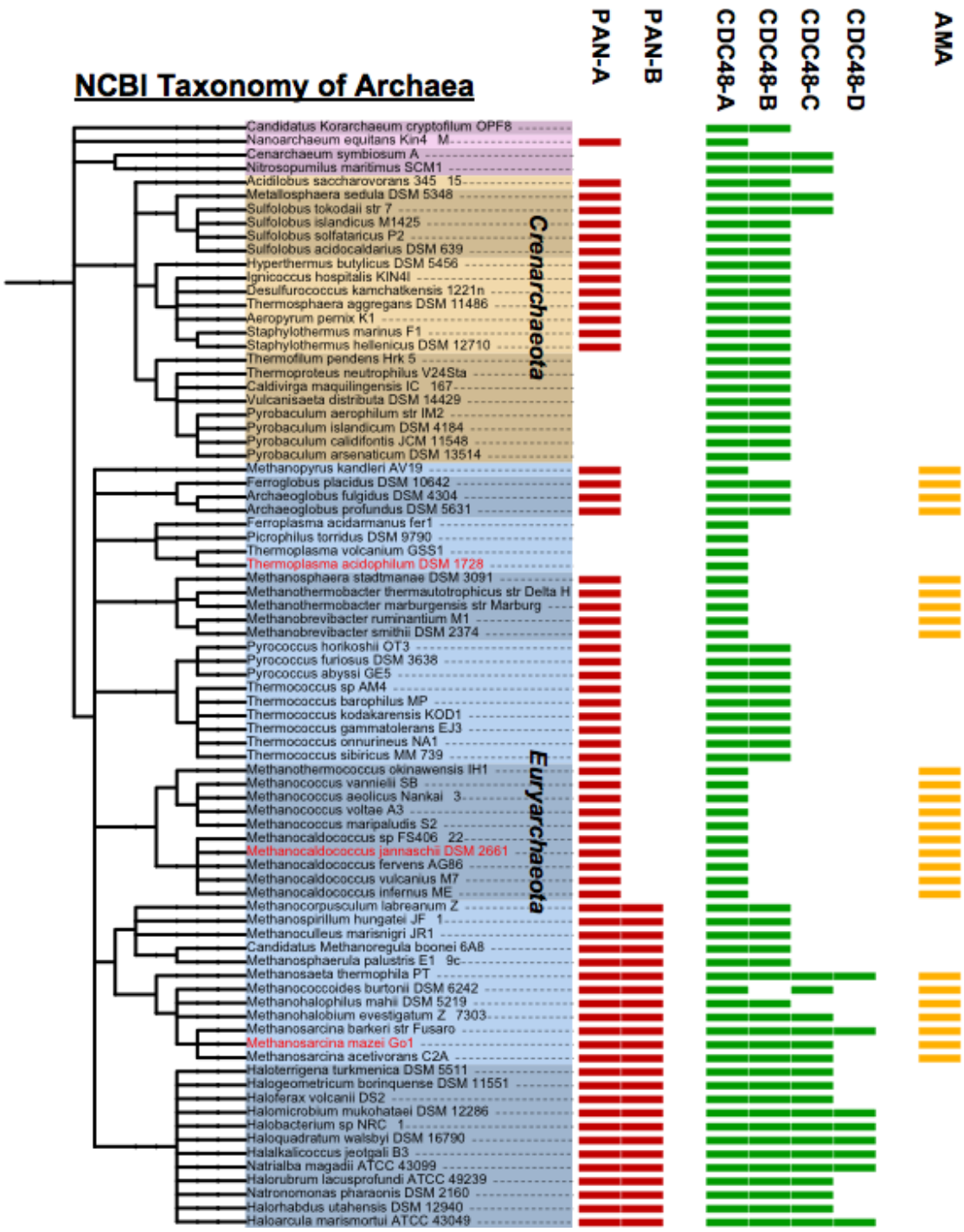


Figure 34. PAN, CDC48, and AMA in all classified Archaea

Extending the analysis shown in Figure 33 provides further support for the network hypothesis (shown are 81 organisms included in the NCBI taxonomy). Notably deep-branching thaumarchaeota and korarchaeota (clades in violett) do not contain PAN, but at least one CDC48 that has the HbYX motif. The figure was generated with iToL.

The analysis points to CDC48 and AMA proteins as putative proteasomal ATPases in addition to the known proteasome activating nucleotidase PAN. Moreover, conservation and degree of paralogy suggest that CDC48 is the primary proteasomal ATPase in the archaeal kingdom, whereas PAN has been lost in certain sub-groups. The degree of

conservation of PAN suggests secondary loss of PAN in species like *Thermoplasma acidophilum* rather than frequent lateral gene transfer [243]. Our analysis favours a complex ancestral state with respect to PAN and CDC48 proteins, which most likely included one PAN and two CDC48 homologs. Nevertheless, the orthologs of PAN prevailed as proteasomal ATPases in the eukaryotic lineage. By contrast AMA proteins are only found in methanogens and archaeoglobales. The conservation of the HbYX motif in AMA proteins suggests that they further diversify the regulation of the proteasome in these species.

The variable numbers of putative proteasomal ATPases in archaea indicate organism-specific solutions in search for adequate regulation of the proteasome. A network of AAA ATPases could increase the capabilities of a system targeting proteins for degradation by the proteasome. Their different N-domains would allow for an increase of potentially recognizable substrates in combination with or addition to the tagging system of sampylation. Sampylation appears to be restricted to a rather small set of substrates [73] and is dispensable in certain methanogens like *Methanopyrus kandleri* [247]. This marks a difference to the fully differentiated 26S proteasome of eukaryotes, which is basically fixed with respect to the base of the 19S regulatory particle including the heterohexameric Rpt1-6 ATPase [248]. Variability within the eukaryotic ubiquitin-proteasome pathway rather takes place at the level of E3 ubiquitin ligases whose number varies between ca. 80 in yeast and ca. 600 in human [249]. CDC48 (p97) is also involved in the ubiquitin-proteasome pathway, but there is no evidence for a direct interaction with the 20S proteasome. It is considered to function as a ‘gearbox’ in this pathway that segregates ubiquitylated substrates from unmodified partners [41]. Interestingly, the penultimate tyrosine of the HbYX motif, which is highly conserved in eukaryotic CDC48 proteins, is a phosphorylation site modulating ER-associated degradation [250].

5.1.3 Conclusions

The kingdom-wide analysis of the C-termini of archaeal AAA ATPases reveals the presence of the HbYX motif, which is crucial for the interaction with the 20S proteasome, in CDC48 and AMA proteins. Therefore, we predict that all AMA proteins and CDC48 proteins, which contain the interaction motif, function as proteasomal ATPases. Furthermore, we show that not only *Thermoplasma acidophilum* [243] but also major archaeal lineages including thermoproteales, thaumarchaeota and

korarchaeota do not encode canonical proteasome activating nucleotidases. These organisms, however, have one or more CDC48 paralog that accommodate the HbYX motif. This provides additional support for the prediction that these CDC48 proteins serve as proteasomal ATPases, given that regulation of the proteasome by gate-keeping ATPases constitutes an important functionality. In this context we note that a double knock-out of PAN-A and PAN-B has little impact on standard growth of *Haloferax volcanii* [73], which encodes two CDC48 proteins with the HbYX motif that may compensate for the elimination of PAN. Because some archaeal species, e.g. *Methanosarcina mazei*, contain five putative proteasomal ATPases, we propose that a regulatory network of ATPases - potentially in cooperation with sampylation - increases the capabilities of proteasomal protein degradation in archaea.

5.2 The Anbu Operon - A Tagging System for Targeted Degradation?

Conjugation of single amino acids or polypeptides to target proteins (peptide tagging) is a widespread phenomenon. However, peptide tagging is not only used as a signal for targeted protein degradation (Table 2), but also as a general post-translational modification regulating activity and interactions of the substrate protein. Examples include tyrosinylation and polyglutamylation of tubulin in eukaryotes [251, 252], glutamylation of ribosomal protein S6 in bacteria [253], or S-glutathionylation of cysteins as a response to oxidative stress [254]. Furthermore, ubiquitin chains function in signal transduction, endocytosis, and DNA-repair when the linkage between ubiquitins occurs via lysine63 [91].

Peptide tagging is achieved through the formation of peptide or isopeptide bonds, which is catalyzed by amidoligases. These enzymes are broadly classified into four groups [180]: Glutamine synthetase-like (Pup, [101]) and ATP-grasp fold proteins (glutamylation, tyrosinylation) hydrolyze ATP in order to directly condense carboxylate- and amino-groups of their substrates; E1 ubiquitin-activating enzymes (Ubi) drive ubiquitylation by adenylating the C-terminus of the tag; GNAT acetyltransferases (N-end rule) make use of charged and already activated tRNAs (a special tRNA is also used for SsrA-tagging directly at the ribosome). In addition to peptide tagging, members of these folds are involved in cofactor biosynthesis (E1), amine utilization (glutamine-synthetase), glutathione biosynthesis (ATP-grasp), or synthesis of secondary metabolites (GNAT).

Within the proteasomal sequence landscape the Anbu cluster forms a group of proteasome homologs whose function is still elusive [179] (Figure 23). Analyzing the genetic environment of the Anbu proteasome homolog, we identified a robust operon structure in approximately 250 bacterial organisms, which comprises three other proteins including an ATP-grasp fold peptide ligase and a transglutaminase. Based on sequence analysis of the components of the Anbu operon, we propose that the Anbu operon constitutes a novel tagging system for targeted protein degradation in which Anbu acts as the proteolytic component. This is in contrast to a previous prediction suggesting that the Anbu operon is involved in synthesis and removal of an unknown peptide [180].

5.2.1 Procedures

Orthologs of the Anbu protease were gathered using PSI-BLAST [131] and subsequently clustered with CLANS [160]. Clustering of 265 sequences was performed using P-values < 1.0e-70. HHpred was used to detect homologs known structure for the proteins of the Anbu operon from *Yersinia enterocolitca* (Open reading frames YE3769, YE3770, YE3771 and YE3772). Homology models for YE3769, YE3771 and YE3772 were generated using Modeller with alignments provided by HHpred [133, 191]. The selected templates were the protein structures 3N6X for YE3769, 3ISR for YE3771 and multiple proteasome β-subunits for Anbu (YE3772). Analysis of the genetic context was conducted with STRING [192], BIOCYC [193], and by interactive genome browsing at the KEGG database [194].

5.2.2 Results and Discussion

Distribution of the Anbu Operon

Sequence analysis shows that Anbu belongs to the family of proteasome-like Ntn-hydrolases (Figure 23). Although it is yet experimentally uncharacterized, the conservation of residues involved in catalysis in proteasomal β-subunits, HslV and MPM strongly suggest that Anbu functions as a protease using an N-terminal threonine residue as the nucleophile (Figure 27). Anbu is found in 265 bacterial organisms with a broad phylogenetic distribution. Therefore, it has been referred to as **an**cestral **β**-s**u**bunit [179]. Organisms containing Anbu mainly comprise cyanobacteria, including deep-branching *Gloeobacter violaceus*, and proteobacteria but also a few instances of nitrospira (*Leptospirillum ferrooxidans*) and bacteroidetes (*Cytophaga hutchinsonii*). Cluster analysis indicates that Anbu sequences group according to the lineage in which they occur with rare instances of lateral gene transfer arguing for frequent secondary loss, especially in proteobacteria (Figure 35).

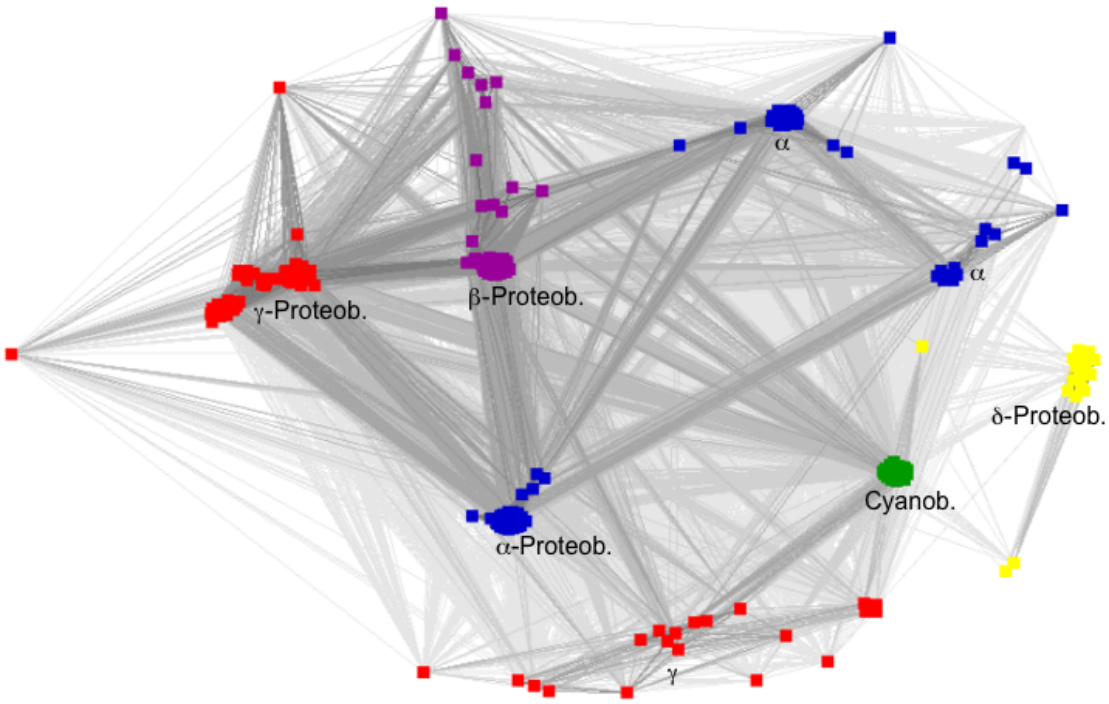


Figure 35. Cluster Map of Anbu Proteases

The cluster analysis illustrates the relationships between Anbu proteases. Anbu is mainly found in proteobacteria and cyanobacteria but sparsely distributed. The cluster of δ-proteobacteria contains sequences from bacteroidetes like *Cytophaga hutchinsonii*, which might have acquired Anbu by lateral gene transfer. Darker lines indicate lower BLAST P-values. The clustering uses all P-values < 1.0e-70. The map contains 265 sequences.

Analysis of the genetic environment of prokaryotic target genes is frequently used for function prediction. In case of proteasome like Ntn-hydrolases, context analysis has been successfully applied revealing the strong genetic coupling of HslV and its gate-keeping ATPase HslU [255] or the Pup-proteasome operon in actinobacteria (more precisely actinomycetes) [102], which often contains all main components of this system for targeted protein degradation. Furthermore, in archaea proteasome loci are embedded in the proteasomal-exosomal superoperon implicated in a general co-regulation of mRNA and protein abundance [256]. However, the genetic context of the Anbu locus points to a rather conserved environment of three most frequently co-occurring genes [180] (Figure 36).

A

B

YE3772: Anbu Protease
YE3771: Transglutaminase-like
YE3770: alpha-E (DUF403)
YE3669: Glutathione synthase-like Ligase

Figure 36. The Anbu Operon

(A) The figure shows the genetic context of the Anbu protease in selected organisms. Anbu is frequently embedded in an operon structure that contains four proteins: A glutathione synthase-like ligase, the alpha-E repeat protein, a transglutaminase-like enzyme, and Anbu. The operon is shown for two α-, β-, three γ-proteobacterial, and one cyanobacterial species. The Anbu proteasome homologue is in the centre of the graph. The synthase is in yellow, the alpha-E repeat protein is in green, and the transglutaminase in violet. The representation was generated with BioCyc [193].
(B) The four proteins of the Anbu operon in *Yersinia enterocolitica* are annotated based on remote homology detection (Table 6).

Components of the Anbu Operon

The first member of the operon, exemplified by Orf 3769 of *Yersinia enterocolitica*, is predicted to adopt an ATP-grasp fold found in proteins catalyzing peptide ligation reactions (Figure 37 A, Table 6). The closest homolog of known function with a sparse sequence identity of 19% is glutathionylspermidine synthase, which couples ATP hydrolysis to the formation of an amide bond between the polyamide spermidine and the glycine carboxylate of glutathione (Figure 38 A). In most organisms the synthase domain is part of a bifunctional enzyme fused to an N-terminal amidase domain regulating the concentration of the antioxidant glutathionylspermidine based on the concentration of these metabolites. Because the amidase domain is not present in YE3669 it retains the capability to catalyze a peptide ligation reaction.

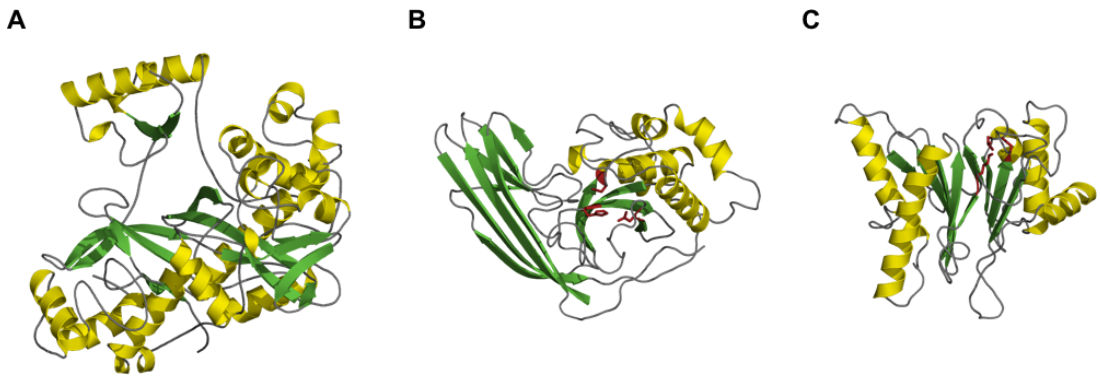


Figure 37. Homology Models of Components of the Anbu Operon

(A) The homology model of YE3769 shows an ATP-grasp fold similar to glutathionylspermidine synthase. The hypothetical protein Mfla_0391 from *Methylobacillus flagellatus* was used as a template (PDB-ID 3N6X).
(B) The homology model of YE3771 shows a cysteine proteinase fold capped by an N-terminal β-sheet. Residues of the conserved putative catalytic triad are shown in red including side chains (Cys166-His202-Asp217). The transglutaminase-like enzyme from *Cytophaga hutchinsonii* was used as a template (PDB-ID 3ISR). The templates in panel (A) (3N6X) and panel (B) (3ISR) are orthologs of YE3769 and YE3771, respectively. Structural genomics consortia deposited both structures without any functional information.
(C) The homology model of the Anbu protease shows an Ntn-hydrolase fold. Catalytically important residues including an N-terminal threonine found in HslV and the β-subunit of the 20S proteasome are in red (see Figure 27 for an alignment of Anbu and other proteasome-like Ntn-hydrolases). Multiple templates were used for the homology model.
Homology models of monomers were generated with HHpred and Modeller [133, 191].

For the second member of the operon, YE3770, there are no homologs of known structure or function being detected. It is a member of domain of unknown function 403 (DUF403) in the Pfam database [156], and contains two copies of an α-helical internal repeat, which is referred to as the alpha-E domain [180]. The presence of coiled coils or TPR-repeats is not predicted suggesting that the protein rather forms globular helical bundles. In addition to several conserved arginine residues, there are two glutamate residues in the N-terminal portion of each repeat, which are invariant in all members of this family. Their γ-carboxylate group could play a role in a potential peptide tagging system via participating in the formation of an isopeptide bond, for instance as an intermediate carrier of the tag, which could then be transferred to substrate proteins by the transglutaminase.

Table 6. Summary of Sequence Comparisons of the Anbu Operon

PROTEIN	PDB-ID	PROB. [%]	E-VAL.	P-VAL	QUERY HMM	TEMPLATE HMM	FOLD	NOTE
Hypothetical protein YE3769	**-**	-	-	-	**478**	-	-	**DUF404, DUF407**
Hypothetical protein Mfla_0391	3N6X	100	0	0	4-475	2-474 (474)	ATP-grasp	deposited by JCSG*
Bifunctional glutathionylspermidine synthase/amidase	2IO8	99.5	2.9e-13	5.8e-18	79-452	253-601 (619)	ATP-grasp	
Trypanothione synthase	2VOB	98.7	3.8e-07	7.5e-12	70-438	257-606 (652)	ATP-grasp	
Glutathione synthase (*E.coli*)	1GSA	98.2	8.6e-06	1.7e-10	234-469	22-244 (316)	ATP-grasp	
Glutathione synthase (*H. sapiens*)	2HGS	97.6	0.76	1.5e-06	37-452	11-453 (474)	ATP-grasp	
Hypothetical protein YE3770	**-**	-	-	-	**309**	-	-	**DUF403, alpha-E**
Hypothetical protein YE3771	**-**	-	-	-	**273**	-		**Bacterial TG-like N-domain**
Transglutaminase-like enzyme	3ISR	100	0	0	1-265	12-276 (293)	Cysteine proteinase	deposited by MCSG*
Peptide N-glycanase	2F4M	99.7	1.3e-16	2.5e-21	154-226	131-196 (287)	Cysteine protease	
DNA repair protein RAD4	2QSF	98.3	6.0e.06	1.2e-10	127-223	100-238 (533)	Cysteine protease	
Protein-glutamine γ-glutamyltransferase	1G0D	98.3	2,4e-06	4.7e-11	163-223	269-361 (695)	Cysteine protease	
Coagulation factor XIII	1EX0	98.3	1.9e-06	3.8e-11	103-227	227-405 (731)	Cysteine protease	
Arylamine N-acetyltransferase	1W4T	96.1	0.31	6.2e-06	108-231	27-162 (299)	Cysteine protease	
Hypothetical protein YE3772	**-**	-	-	-	**244**	-		**Anbu**
Eukaryotic β-subunit	1RYP_L	100	1.8e-33	3.6e-38	2-224	1-202 (212)	Ntn-h	
Archaeal β-subunit	1PMA_B	100	1.0e-29	2.0e-34	2-202	9-189 (217)	Ntn-h	
Actinobac. β-subunit	3MI0_C	100	1.2e-28	2.5e-33	2-204	58-259 (291)	Ntn-h	
Eukaryotic α-subunit	1RYP_D	100	7.2e-28	1.4e-32	2-224	29-236 (241)	Ntn-h	
Archaeal α-subunit	1PMA_A	100	2.9e-31	2.0E-33	2-203	35-219 (233)	Ntn-h	
Actinobac. α-subunit	3MI0_A	100	6.0e-28	1.2e-32	2-214	28-227 (248)	Ntn-h	
HslV	1G3I_G	99.6	1.7e-15	3.4e-20	2-200	1-174 (174)	Ntn-h	
HslV	1M4Y_A	99.6	2.3e-14	4.6e-19	2-201	1-170 (171)	Ntn-h.	

The targets taken from the HHpred output represent the diversity of significantly similar matches rather than the ranking by the servers. HHpred searches were performed in default settings against the Protein Data Bank, release of May 14 2011, filtered for a maximum of 70% pairwise sequence identity at http://toolkit.tuebingen.mpg.de/HHpred.

The Anbu protease, YE3772, shares significant sequence similarity with proteasome-like Ntn-hydrolases [179] (Figure 37 C, Table 6). However, it is considerably more similar to proteasome subunits than to HslV (Figure 23). Interestingly, Anbu and the 20S proteasome never co-occur in one organism (including Leptospirilla that contain either Anbu or the proteasome) whereas Anbu and HslV frequently co-occur (Figure 39). However, the degree of similarity between Anbu and proteasome-like Ntn-hydrolases suggests that it is capable of self-compartmentalization. In contrast to the monomeric proteasome-homolog of methanogens Anbu does not contain insertions that would collide with the general type of assembly formed by the proteasome or HslV (Figure 27 – Alignment of proteasome-like Ntn-hydrolases).

A Tagging System for Targeted Degradation?

The sequence analysis of the Anbu operon points to a system, which has the basic biochemical capabilities for peptide synthesis (ATP grasp), tagging/removal (transglutaminase) plus a potential effector protease (Anbu) for the degradation of targeted substrates (Figure 38, Table 6). The broader genetic vicinity of the operon suggests that it is generally involved in the protection against oxidative stress and transformation of xenobiotics. Among the proteins more loosely associated with the Anbu operon there are a thioredoxin, other components of glutathion metabolism, and the DNA repair protein MutS. The synthesized and potentially tagged molecule, which was not identified in this context analysis, could be a small, not-genetically encoded peptide like glutathione or one of its derivatives like glutathionylspermidine. Nevertheless, this prediction misses the crucial component of a AAA+ ATPase, which would serve as a regulator of the proteasome-like protease. Although there is no AAA+ protein detected in the genetic environment of the Anbu operon, most bacterial genomes encode several AAA+ proteins, whose function is not assigned yet.

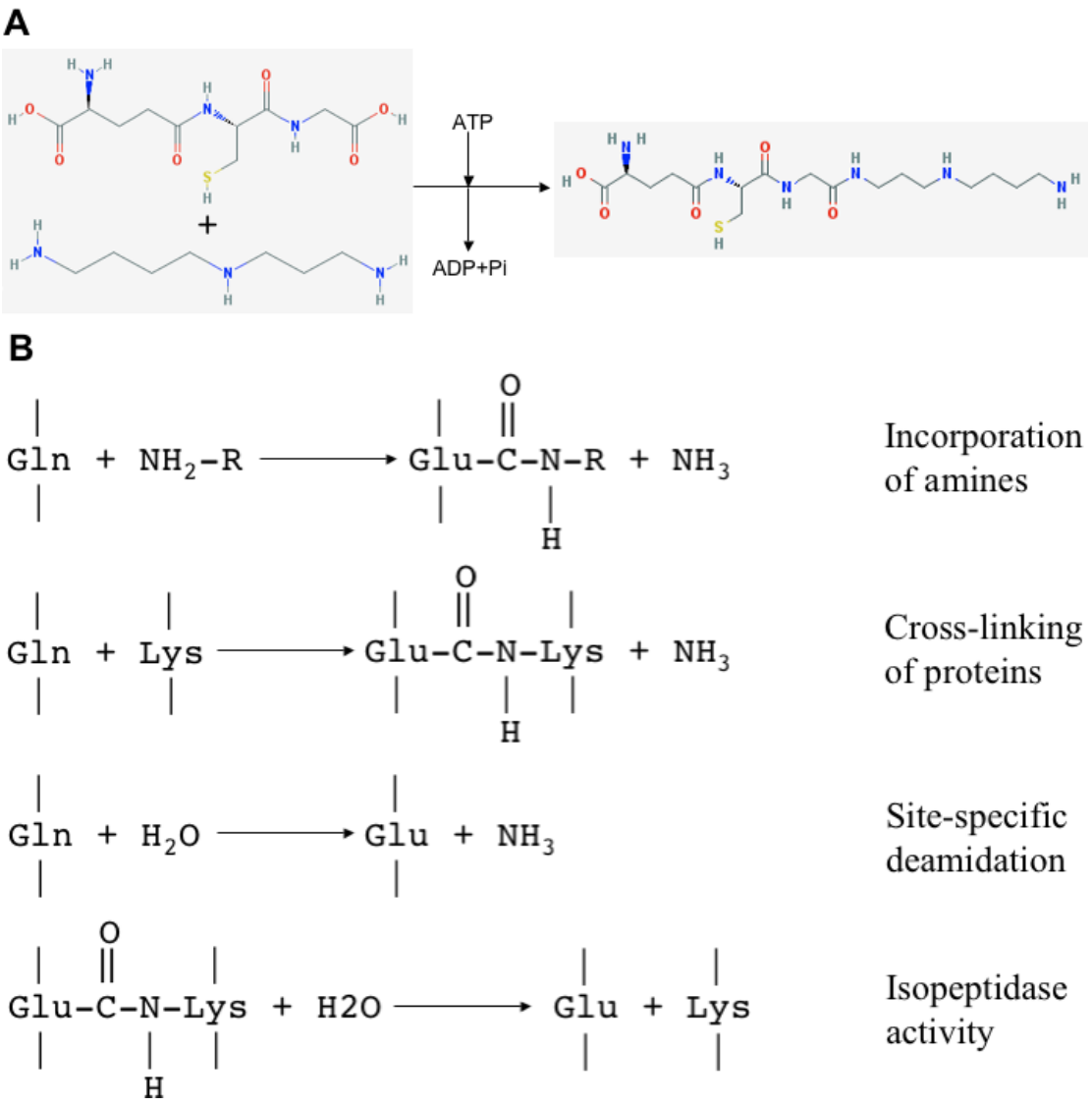


Figure 38. Reactions of Enzymatic Components of the Anbu Operon

(A) Gluthionylspermidine synthase, the closest homolog of known function of YE3769 catalyzes the formation of an amide bond between the glycine carboxylate of glutathione and spermidine using the energy of ATP-hydrolysis. Compounds are drawn with Pubchem (http://pubchem.ncbi.nlm.nih.gov/).

(B) Transglutaminases, the homologs of YE3771, catalyze a variety of reactions including cross-linking of proteins and cleavage of isopeptide bonds. These reactions are compatible with the transfer of a peptide tag to a substrate or the cleavage of the ligation product formed through catalysis by YE3769. TG-Reactions are adapted from [257].

Although the nature of the substrates of YE3769 and YE3771 are unknown, the reactions shown in panels A and B illustrate basic biochemical abilities displayed by tagging systems.

Alternatively, it has been proposed that this operon may be involved in the synthesis of a small peptide metabolite by catalysis of the ATP-grasp fold protein [180]. The peptide would be tagged to the alpha-E protein, which would serve "as a substrate for elongation of a peptide via the γ-carboxylate [180]" of the side chain of the invariant glutamate residues. Because of the absence of a regulatory ATPase in the operon Iyer and colleagues exclude tagging of the peptide to protein substrates with the purpose of

degradation. Instead Iyer and colleagues postulate that the peptide product is removed by two distinct peptidase reactions catalyzed by the transglutaminase and Anbu. Although HslU and HslV are genetically coupled in most cases and ARC is frequently found in the proteasomal operon of actinobacteria, we note that components of the archaeal proteasome system are distributed over several loci in the proteasomal-exosomal superoperon [256]. Genetic association of proteasome subunits and putative regulatory ATPases is not observed in archaea. However, experimental evidence will be required to determine whether the Anbu operon is a system just for the synthesis and removal of a peptide as proposed by Iyer and colleagues [180] or whether it is, as we suggest, a tagging system for targeted protein degradation.

Prelimenary characterization of the Anbu protease from *Thermosynechococcus elongatus* indicated the formation of a large oligomer compatible with the size of a self-compartmentalizing protease. In collaboration with the Institute for Medical Microbiology of the University of Tübingen a knock out mutant for the Anbu operon of the pathogen *Yersinia enterocolitica* has been designed. Characterization of the Anbu knock out strain showed a growth phenotype in comparison to the wild type, especially at elevated temperatures. Proteomics experiments revealed a significant upregulation of the HslUV protease in the Anbu knock out strain (Sebastian Klein). This suggests a compensatory mechanism employed to deal with the increased burden of non-native protein species in the absence of the Anbu operon, which favours protein degradation as a function of the operon. Future *in vitro* characterization of the Anbu protease and further *in vivo* characterization of *Y. enterocolitica* mutants lacking the Anbu operon will be our next efforts.

5.2.3 Conclusions

Analyzing the genetic environment of the uncharacterized Ntn-hydrolase Anbu [179], we detected an operon structure wide-spread in cyanobacteria and proteobacteria. The conserved core of this operon comprises a peptide ligase of the ATP-grasp fold, an a-helical repeat protein, alpha-E, and a transglutaminase. In contrast to a previous analysis that suggests a function of this operon in the synthesis of an unknown peptide [180], we propose that it is a tagging system for targeted protein degradation.

5.3 Evolutionary Implications for Proteasome-like Ntn-hydrolases and Regulatory ATPases

The analyses of the putative network of proteasomal AAA ATPases in archaea, the Anbu operon, and the characterization of the monomeric proteasome-homolog of methanogens (MPM) have implications for the evolution of targeted protein degradation by proteasome-like Ntn-hydrolases. In order to obtain a global overview of the organismal distribution of relevant genes we have mapped their presence or absence onto a highly resolved tree of life. This analysis of 191 species includes on the one hand the Ntn-hydrolases Anbu, HslV, 20S proteasome, and MPM and on the other hand the AAA ATPases HslU, PAN/ARC, CDC48 and AMA.

The analysis confirms and extends observations that have been made for Anbu [179, 180], HslUV and the 20S proteasome [87, 229, 258] in the light of newly sequenced genomes. Furthermore, our analysis provides additional aspects to current evolutionary scenarios because of the differential consideration of (putative) proteasomal ATPases.

5.3.1 Procedures

Homologs of Anbu, HslU and HslV, 20S proteasome, MPM, PAN, CDC48, and AMA were identified with BLAST [130]. HslU and HslV are generally treated as a unit, because we did not detect the presence of one of these in absence of the other. Assignment to orthologous groups of full-length sequences was based on cluster analyses using CLANS [160]. P-values were selected interactively for each protein family in order to achieve grouping by orthology. Members of orthologous groups were verified by testing for concordant domain composition using HHpred and MUSCLE multiple alignments [133, 187]. Proteins of 191 organisms included in the interactive tree of life project (iToL) were pooled from the purified sets of sequences of each protein family [227]. Subsequently, the presence or absence of proteins was assigned to the respective organisms. In order to exclude artefacts based on the selective inclusion of organisms into the iToL project emerging patterns were confirmed by BLAST searches against selected taxa containing the full set of organisms currently included in a particular taxon at NCBI. Potential lateral gene transfer events were analyzed by testing the congruence of protein family trees with respective species trees.

Phylogenetic trees were generated with Phylip-neighbour (JTT matrix, 100 bootstrap replicates) [159]. Species trees were obtained from http://www.bacterialphylogeny.info.

5.3.2 Results and Discussion

Distribution of Proteasome-like Ntn-hydrolases

The distribution of proteasome-like Ntn-hydrolases on the tree of life shows that the 20S proteasome is universally conserved in all sequenced archaeal and eukaryotic organisms, whereas HslV is mainly found in certain unicellular eukaryotes like trypanosomatids (Figure 39). Although encoded in the nuclear genome, the HslUV complex is targeted to the mitochondria of these organisms [258]. Phylogenetic analysis suggests that the eukaryotic ancestor acquired HslUV through the engulfed α-proteobactrium, which evolved into the mitochondrion [259]. However, in most eukaryotic lineages HslUV was not retained. In contrast, HslUV seems to be an ancestral character of bacteria indicated by its high degree of conservation in gram-positive and gram-negative bacteria including deep-branching aquificae and thermotogales. Two major bacterial taxa that do not contain HslUV are cyanobacteria and actinobacteria. The latter is the only bacterial taxon in which the 20S proteasome is found, namely the actinobacterial sub-group of actinomycetales [260, 261]. An exception is marked by a few verrucomicrobial species that have most like acquired the 20S proteasome through lateral gene transfer (LGT) from an actinomycete [262].

Whereas the proteasome is conserved and essential in eukaryotes and *Haloferax volcanii* [73], a proteasome knock out mutant of the actinomycete *Mycobacterium tuberculosis* is viable. However it is required for resistance to nitric oxide and for persistence of the pathogen in its host [263, 264]. The lack of essentiality mirrors the degree of conservation, which shows the absence of the proteasome in several actinomycetales, most notably in corynebacteria [265] and tropherymna. The conservation pattern is consistent with a secondary loss of the proteasome in corynebacteria, along with the diversification of actinomycetes. Furthermore, conservation and gene importance provide support for an evolutionary scenario according to which an ancestor of actinomycetales received the proteasome via LGT from an archaeal-eukaryotic ancestor [87, 260].

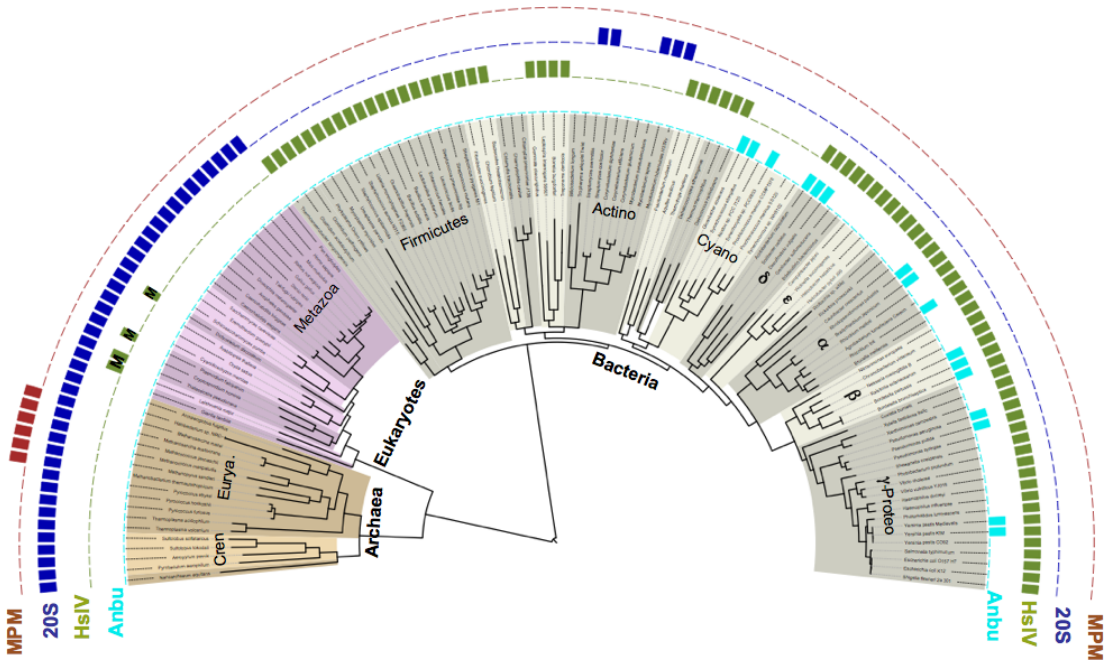


Figure 39. Distribution of Proteasome-like Ntn-hydrolases

MPM is a highly divergent member of the proteasome-like Ntn-hydrolase family. We proposed that it is a derived feature of methanogenic archaea. Whereas Anbu is only sparsely distributed among cyanobacteria and proteobacteria, HslV is highly conseved in the bacterial kingdom. Notably, HslV has been lost in chlamydia, cyanobaceria, and actinobacteria, the latter of which is the only bacterial taxon containing a 20S proteasome. It is unclear whether actinobacteria lost HslV independent of the acquisition of the proteasome or whether the proteasome displaced HslV in this taxon. In contrast, the proteasome is universally conserved in archaea and eukaryotes. Both kingdoms, however, do not retain Anbu or HslV with exception of HslV found in the mitochondria of certain unicellular eukaryotes (M). The proteasome arose by gene duplication from an ancestral Ntn-hydrolase. Whether the ancestor was more similar to HslV or to Anbu is currently unknown. The fact that several bacteria do not encode proteasome-like Ntn-hydrolases points to compensation by other systems for targeted protein degradation like ClpXP.

The loss of HslUV in cyanobacteria and actinobacteria indicates its dispensable nature, which matches the viability of knock out mutants of HslUV, supporting the notion that other AAA+ proteases like ClpA/X-P or Lon are able to compensate for the absence of HslUV [266]. Whereas cyanobacteria contain neither HslV nor the proteasome, the conservation pattern in actinobacteria poses the question whether the acquisition of the proteasome lead to a displacement of HslV [229]. Within the class of actinobacteria not only proteasome-encoding actinomycetales lack HslV but also deep-branching bifidobacteria and rubrobacteridae do not contain the proteasome. This opens two possible scenarios. Either an early actinobacterial ancestor lost HslV independent of the acquisition of the proteasome, or the proteasome was received earlier then previously thought, displaced HslV and was again lost in certain lineages like bifidobacteria. The high conservation of HslV in other gram-positive bacteria and the presence of ARC in

bifidobacteria despite absence of the proteasome support the latter scenario (Figure 39 and 40).

We have already described the sub-group of MPM as a derived development in methanogenic archaea, which is underlined by the global distribution. For the Anbu protease, however, the situation appears to be rather complicated (Figure 39). Despite the absence of evidence for frequent lateral gene transfer the degree of conservation is sparse, restricted mainly to cyanobacteria and proteobacteria, and cannot be matched to certain metabolic or environmental preferences. Despite the name Anbu (ancestral β-subunit) it is not found in deep-branching aquificae and thermotogales and generally sparsely distributed, implying massive secondary loss, which is in contrast to the high degree of conservation of HslV. Whereas Anbu and HslV frequently co-occur, Anbu is not present in any organism that encodes the 20S proteasome resembling the situation for HslV. Nevertheless, the distribution argues for a bacterial origin of Anbu. Interestingly, Anbu and HslV are more similar in sequence to proteasome subunits than to each other.

The complexity of the architecture and the consistence of two different types of subunits suggest the evolution of the proteasome by gene duplication of an ancestral Ntn-hydrolase. Previously, duplication and subsequent divergence of HslV was considered as the origin of the proteasome. Because phylogenetic inference of proteasome-like Ntn-hydrolases does not provide a conclusive answer, it remains unclear whether the proteasomal ancestor was more similar to HslV or to Anbu. Determination of the molecular architecture of the Anbu protease will contribute further information.

Distribution of Proteasomal ATPases

The distribution of putative proteasomal ATPases shows a universal conservation of CDC48 in archaea and eukaryotes (Figure 41), whereas PAN is not found in a number of archaeal clades including korarchaeota, thaumarchaeota, thermoproteales and thermoplasmata (Figure 34). Furthermore, the presence of the C-terminal HbYX interaction motif in archaeal CDC48 proteins, their conservation and degree of paralogy suggested that CDC48 was the primordial proteasomal ATPase in archaea (see 5.1). Therefore, we proposed a network of ATPases, including CDC48, PAN and AMA if present, regulating the archaeal proteasome. In eukaryotes however, the PAN orthologs have prevailed as the regulatory ATPase in the fully differentiated 26S proteasome. The spreading of CDC48 in bacteria is sporadic with a few instances found in cyanobacteria,

α-proteobacteria and actinobacteria. In actinobacteria, CDC48 does not display the C-terminal interaction motif. Therefore, CDC48 does not function as a proteasomal ATPase in this taxon. In contrast, PAN/ARC is only found in actinobacteria throughout the bacterial kingdom. Because an actinobacterial ancestor acquired the 20S proteasome most likely via a single event of LGT, PAN was part of the genes included. The sparse distribution of CDC48 in bacteria also suggests an acquisition through LGT from the archaeal-eukaryotic lineage, but frequent secondary loss cannot be excluded. Although CDC48 seems to be enriched in actinonbacteria in comparison to other bacteria, it is unclear whether it was included in the proteasomal LGT event.

Comparison with the presence of the proteasome in actinobacteria reveals that ARC is also found in organisms that do not contain a proteasome, among them corynebacteria and bifidobacteria. However, proteasome containing actinomycetales show a highly focused C-terminal interaction motif, whereas the C-termini of ARC in actinobacteria lacking the proteasome show no apparent conservation (Figure 40). In these species the selective pressure on the C-terminal tails of ARC has obviously disappeared. Nevertheless, the fact that ARC was retained points to valuable functionality, likely as an unfoldase or disaggregase. Interestingly, ARC is found in all bifidobacteriales (11 genomes), which are deep branching within the class of actinobacteria and do not belong to the actinomycetales. Furthermore, the only acidimicrobium (*A. ferrooxidans*) contains ARC and proteasome. This suggests that the recipient of the LGT was not the ancestor of actinomycetales but an even deeper ancestor within the class of actinobacteria. Therefore, the LGT event took place earlier than previously thought. Alternatively, PAN was transferred independently of the proteasome.

	C-terminal Peptide		
	ARC	ARC-B	CDC48
ACTINOBACTERIA			
Bifidobacterium longum	RRIR**TA**E		-
Bifidobacterium dentium	RIRP**TA**Q		-
Bifidobacterium catenulatum	TSIR**PV**A		-
Bifidobacterium breve	AIRA**VG**C		-
Tropheryma whipplei_TW08/27	-		-
Tropheryma whipplei_Twist	-		-
Streptomyces coelicolor	ANTG**QY**L		-
Streptomyces sp. AA4	TNTG**QY**L		AQAQQER
Streptomyces avermitilis	ANTG**QY**L		-
Streptomyces lividans	ANTG**QY**L		-
Corynebacterium efficiens	VAEV**EV**V		-
Corynebacterium glutamicum	THAE**VV**I		-
Corynebacterium jeikeium	DIDV**HK**V		-
Corynebacterium accolens	SCTW**PA**C		-
Mycobacterium gilvum	SNLG**QY**L		RQFADTR
Mycobacterium avium	SNLG**QY**L		RAFAEAP
Mycobacterium leprae	SNLG**QY**L		-
Mycobacterium paratuberculosis	SNLG**QY**L		RAFAEAP
Mycobacterium ulcerans	SNLG**QY**L		RAFGEEL
Mycobacterium tuberculosis	SNLG**QY**L		TKGDLRS
Rhodococcus erythropolis	SNTG**QY**L		QAYADNR
Rhodococcus equi	SNTG**QY**L		EAYAENR
Frankia alni	ANTG**QY**L		-
Frankia sp. EAN1pec	ANTG**QY**L		-
NITROSPIRA			
Leptospirillum rubarum	VPAG**HY**L	LERRAIE	-
Leptospirillum ferrodiazotrophum	LPVG**HY**L	RIPRAIE	-
VERRUCOMICROBIA			
Methylacidiphilum infernorum	KVVS**GV**V		-
Chthoniobacter flavus	PKPS**AI**V		-

Figure 40. The HbYX motif in proteasomal ATPases of Actinobacteria

The C-terminal HbYX motif is highly conserved in actinobacterial species encoding a 20S proteasome, in the form of a QYL motif. In species lacking the proteasome (in blue), the C-terminal peptide is degenerate. The presence of ARC in deep-branching bifidobacteria suggests that acquisition of ARC and the 20S proteasome occurred before the divergence of actinomycetales (given that there was only a single LGT event comprising the proteasome including the regulatory ATPase). Corynebacteria, which are part of the taxon actinomycetales, and bifidobacteria have lost the proteasome but retain ARC, whereas Tropheryma whippei lost both. In turn, some verrucomicrobia and leptospirilla most likely received the proteasome including ARC and the Pup-based tagging system, which are frequently encoded in an operon, via lateral gene transfer from an actinomycete.

The Pup-based targeting system was an original development within the actinobacterial lineage, because neither the tag nor the peptide ligases have counterparts in archaea or eukaryotes. Interestingly, Pup as well as the Pup ligase PafA are found in bifidobacteria [265]. That these proteins are present in species lacking the proteasome point to an evolutionary origin of Pup-based tagging, which is independent of targeted protein degradation. The genetic context of Pup in bifidobacteria does not suggest a particular

function, but the GGQ/E motif at the C-terminus and the ligase PafA are strictly conserved. Therefore, Pup may fulfil a carrier function, remotely resembling the evolutionary roots of ubiquitin in β-Grasp fold proteins like ThiS, MoaD or Urm1, which function as sulphur carriers in cofactor biosynthesis and tRNA-sulphuration [91]. However, the entire Pup-proteasome based tagging system experienced lateral transfer to certain leptospirillum and verrucomicrobia species [262] (Figure 40) underlined by the fact that crucial components of this system are frequently encoded in one operon in both actinomycetes and the few verrucomicrobia. According to the selfish operon concept [267] the presence of similar operons in prokaryotes is often a result of LGT rather than selection for co-regulation. The chances for fixation are supposed to be greatly increased, if the transferred DNA encodes a functional system or complete pathway. This provides circumstantial support for the idea that an actinobacterial ancestor received proteasome and regulatory ATPase via LGT. However, investigation of the genetic neighbourhood of proteasome subunits and regulatory ATPases in archaea does not contribute further information regarding the donor of this LGT package, because they are not genetically coupled. Instead, archaeal proteasome subunits are part of the proteasomal-exosomal superoperon, a system for the co-regulation of protein and mRNA levels [256].

Evolutionary Scenario

In the light of these analyses and in line with previous studies we propose an evolutionary origin of the proteasome by a gene duplication event in the last common ancestor of archaea and eukaryotes [87]. It has been suggested that the proteasome directly evolved from the HslV gene, because both proteins are essentially never found together in one species [229]. In the eukaryotic lineage, HslV has been reacquired with the endosymbiosis of the proteobacterium that gave rise to the mitochondrion. During the gene flow from the engulfed bacterium to the nucleus HslUV was only retained by certain unicellular species, in which it is targeted to this organelle [259]. The evolutionary trajectory led from HslV to the proteasome because of the substantially increased complexity of the proteasome. Although increase in complexity does not necessarily provide a polarizing argument for the direction of an evolutionary path, the number of subunits, the higher order symmetry, and the size of the particle support the view that the proteasome evolved out of HslV [229].

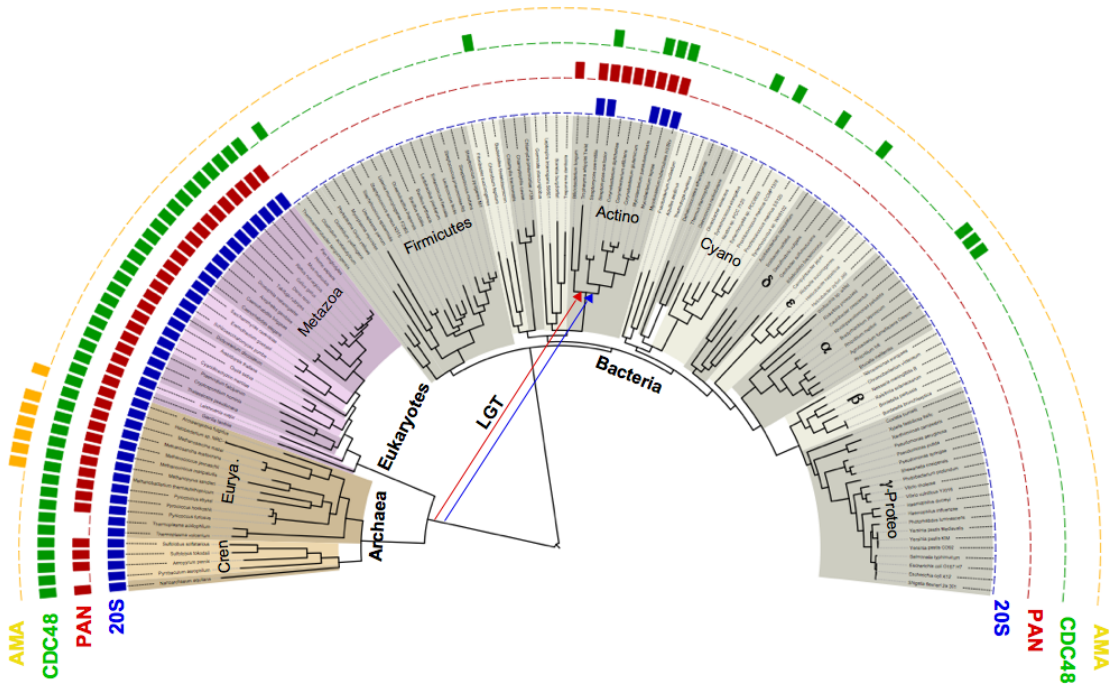


Figure 41. Distribution of (Putative) Proteasomal ATPases

Despite its simple architecture, AMA (yellow) has most likely evolved from a CDC48-like protein in the common ancestor of methanogens and archaeoglobales. The more complex CDC48 protein (green) is universally conserved in archaea and eukaryotes. In bacteria it is only sparsely distributed and either lost frequently or acquired by LGT. Actinobacteria are the only bacterial taxon encoding PAN/ARC (red). Conservation and gene importance suggest that PAN and the 20S proteasome (blue) were acquired most likely by a single LGT event from an unknown ancestor of the archaeal-eukaryotic lineage (red and blue arrows). ARC is not only found in actinomycetales but also in early-branching bifidobacteria, advancing the receipt of the PAN-proteasome transfer to an earlier point in actinobacterial evolution. The proteasome has been lost frequently in the course of actinobacterial evolution, for instance in corynebacteria, whereas ARC is more conserved. Notably, the HbYX motif of ARC is conserved only in species containing the proteasome (Figure 40).

The identification of the Anbu protease, however, adds complexity to the evolutionary scenario [268], and HslV is not the only candidate for a hypothetical direct ancestor of the proteasome through gene duplication. Like HslV, Anbu does not co-occur with the proteasome, also consists out of one subunit, and shares a very similar degree of sequence similarity with the proteasome subunits. However, Anbu is less conserved than HslV, and therefore its ancestry would require massive secondary loss. Our analysis underlines that the presence of a regulatory ATPase is a necessity for proteasome function illustrated by the fact that ARC is often retained in the absence of the proteasome, but never the other way around (lending support for the hypothesis that CDC48 functions as the proteasomal ATPase in archaea lacking PAN). The regulatory ATPase of Anbu is still elusive, whereas the tight co-regulation of HslU and HslV might contribute to the conservation in bacteria. The determination of the molecular architecture of Anbu including size, symmetry and assembly-type will provide

additional information in this case. A third possibility is that the proteasome evolved from an ancestral Ntn-hydrolase whose traces have completely vanished after the gene duplication event.

A rather different scenario by Cavalier-Smith places the origin of the proteasome within actinobacteria [229]. HslV or Anbu are also considered as direct ancestors of the proteasome via gene duplication and subsequent divergence [268]. However, this scenario excludes lateral gene transfer of the proteasome. Instead, it assumes that eukaryotes and archaea, jointly referred to as neomura, originated from an actinomycete ancestor. This is part of a reconstruction of the tree of life inferred by transition analysis, which is based on polarizing arguments derived from molecular cladistics. The evolution of the proteasome is one of the key characters considered in this approach that places the root of the tree of life within bacteria. Furthermore, the bacterial taxon of actinomycetes would divergently give rise to archaea and eukaryotes explaining the presence of the proteasome in these lineages.

Regardless of the reconstruction of the deep roots of the tree of life, this view of proteasome evolution is not supported by our analysis. Essentiality and conservation of the proteasome in eukaryotes and archaea, and the operon structure in actinomycetes favour lateral gene transfer of proteasome and regulatory ATPase from an ancestor of archaea and eukaryotes to the ancestor of bifidobacteria, acidimicrobia and actinomycetales.

5.3.3 Conclusions

This analysis provides a comprehensive view of the distribution of proteasome-like Ntn-hydrolases and putative regulatory ATPases. It supports the hypothesis that actinomycetes acquired PAN/ARC and proteasome through LGT from an unspecified ancestor of the archaeal-eukaryotic lineage. The presence of ARC in deep-branching bifidobacteria suggests that the LGT event(s) occurred earlier in actinobacterial evolution than previously thought, implying that bifidobacteria lost the proteasome, unless both proteins were independently transferred. Secondary loss of the proteasome is also observed in other actinobacteria including corynebacteria. We show that the C-terminal interaction motif, in actinobacteria as a QYL motif, is highly conserved, in proteasome-containing species, whereas it is degenerate when the proteasome is absent.

6 CONCLUSIONS

In the first part, we have characterized homologs of N-domains of AAA proteins whose properties draw attention to the pervasive phenomenon of duplication [47, 135, 269, 270], in particular the emergence of autonomously folding polypeptide chains by amplification of sub-domain-sized supersecondary structure elements [145, 146]. Our characterization of archaeal RFKs illustrates how an erosion of internal symmetry was accompanied by a gain of enzymatic activity from an ancestral DNA-binding activity [4]. The homohexameric twelve-bladed β-propeller of HP12 provides evidence for the evolution of fully amplified monomeric propellers from single blades via oligomeric intermediates. The C-terminal domain of MPM proteases is a symmetric, six-stranded example of an OB-fold that most likely originated by the duplication of a three-stranded β-meander. In all three cases, traces of internal symmetry and repetitive patterns on the sequence level, often only recognizable in the light of structural information, support the hypothesis of ancient peptide precursors, whose oligomerization might have been an intermediate step in the evolution of modern domains.

In the second part we traced the origins of proteasomal protein degradation. Based on sequence analysis of C-terminal interaction motifs of AAA proteins, we predict that CDC48 and AMA proteins function as proteasomal ATPases in archaea supported by the absence of canonical proteasome activating nucleotidases in major archaeal taxa. Up to five putative proteasomal ATPases may form a network that increases the spectrum of recognizable substrates through the participation of different N-domains. In collaboration with Dara Forouzan we are currently collecting experimental evidence in support of the network hypothesis.

Analysis of the genetic neighbourhood of the yet uncharacterized proteasome homolog Anbu identified an operon [179, 180], which we predict to constitute a tagging system for targeted protein degradation. Preliminary characterization of an Anbu protease suggests that it forms a large oligomer of self-compartmentalized architecture. Therefore, the highly derived MPM family is an exception among proteasome-like Ntn-hydrolases. Loss of the ability to self-compartmentalize provides an example for divergence towards a more simplified molecular phenotype. This process was accompanied by the incorporation of a potential substrate recognition domain similar to the one found in the N-domain of ATPases (PAN) that regulate the highly complex proteasome machinery.

7 BIBLIOGRAPHY

1. Löwe, J., et al., *Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution.* Science, 1995. **268**(5210): p. 533-9.
2. Zhang, X., et al., *Structure of the AAA ATPase p97.* Mol Cell, 2000. **6**(6): p. 1473-84.
3. Djuranovic, S., et al., *Structure and activity of the N-terminal substrate recognition domains in proteasomal ATPases.* Mol Cell, 2009. **34**(5): p. 580-90.
4. Ammelburg, M., et al., *A CTP-Dependent Archaeal Riboflavin Kinase Forms a Bridge in the Evolution of Cradle-Loop Barrels.* Structure, 2007. **15**(12): p. 1577-90.
5. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.
6. Nicholls, A., K.A. Sharp, and B. Honig, *Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons.* Proteins, 1991. **11**(4): p. 281-96.
7. Dobson, C.M., *Protein folding and misfolding.* Nature, 2003. **426**(6968): p. 884-90.
8. Lupas, A.N., Koretke, K., K., *Evolution of Protein Folds*, in *Computational Structural Biology*, T. Schwede, Editor. 2008, World Scientific Publishing Company. p. 131-152.
9. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.* Cell, 2008. **134**(2): p. 341-52.
10. Koonin, E.V. and Y.I. Wolf, *Constraints and plasticity in genome and molecular-phenome evolution.* Nat Rev Genet, 2010. **11**(7): p. 487-498.
11. Frauenfelder, H., et al., *A unified model of protein dynamics.* Proc Natl Acad Sci USA, 2009. **106**(13): p. 5129-34.
12. Bucciantini, M., et al., *Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases.* Nature, 2002. **416**(6880): p. 507-11.
13. Ellis, R.J. and S.M. Hemmingsen, *Molecular chaperones: proteins essential for the biogenesis of some macromolecular structures.* Trends Biochem Sci, 1989. **14**(8): p. 339-42.
14. Hartl, F.U., *Molecular chaperones in cellular protein folding.* Nature, 1996. **381**(6583): p. 571-9.
15. Goldberg, A.L., *The mechanism and functions of ATP-dependent proteases in bacterial and animal cells.* Eur J Biochem, 1992. **203**(1-2): p. 9-23.
16. Gottesman, S., S. Wickner, and M.R. Maurizi, *Protein quality control: triage by chaperones and proteases.* Genes Dev, 1997. **11**(7): p. 815-23.
17. Wickner, S., M. Maurizi, and S. Gottesman, *Posttranslational quality control: folding, refolding, and degrading proteins.* Science, 1999. **286**(5446): p. 1888-93.
18. Houry, W.A., et al., *Identification of in vivo substrates of the chaperonin GroEL.* Nature, 1999. **402**(6758): p. 147-54.
19. Palombella, V.J., et al., *The ubiquitin-proteasome pathway is required for processing the NF-kappa B1 precursor protein and the activation of NF-kappa B.* Cell, 1994. **78**(5): p. 773-85.
20. Neuwald, A., et al., *AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes.* Genome Res, 1999. **9**(1): p. 27-43.

21. Erdmann, R., et al., *PAS1, a yeast gene required for peroxisome biogenesis, encodes a member of a novel family of putative ATPases.* Cell, 1991. **64**(3): p. 499-510.
22. Kunau, W.H., et al., *Two complementary approaches to study peroxisome biogenesis in Saccharomyces cerevisiae: forward and reversed genetics.* Biochimie, 1993. **75**(3-4): p. 209-24.
23. Saraste, M., P.R. Sibbald, and A. Wittinghofer, *The P-loop--a common motif in ATP- and GTP-binding proteins.* Trends Biochem Sci, 1990. **15**(11): p. 430-4.
24. Koonin, E.V., Y.I. Wolf, and L. Aravind, *Protein fold recognition using sequence profiles and its application in structural genomics.* Adv Protein Chem, 2000. **54**: p. 245-75.
25. Erzberger, J. and J.M. Berger, *Evolutionary relationships and structural mechanisms of AAA+ proteins.* Annu Rev Biophys Biomol Struct, 2006. **35**(1056-8700 (Print)): p. 93-114.
26. Walker, J., et al., *Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold.* EMBO J, 1982. **1**(8): p. 945-51.
27. Ammelburg, M., T. Frickey, and A.N. Lupas, *Classification of AAA+ proteins.* J Struct Biol, 2006. **156**(1): p. 2-11.
28. Alva, V., et al., *On the origin of the histone fold.* BMC Struct Biol, 2007. **7**: p. 17.
29. Hanson, P.I. and S.W. Whiteheart, *AAA+ proteins: have engine, will work.* Nat Rev Mol Cell Biol, 2005. **6**(7): p. 519-29.
30. Lupas, A.N. and J. Martin, *AAA proteins.* Curr Opin Struct Biol, 2002. **12**(6): p. 746-53.
31. Jeruzalmi, D., M. O'Donnell, and J. Kuriyan, *Crystal structure of the processivity clamp loader gamma (gamma) complex of E. coli DNA polymerase III.* Cell, 2001. **106**(4): p. 429-41.
32. Qi, S., et al., *Crystal Structure of the Caenorhabditis elegans Apoptosome Reveals an Octameric Assembly of CED-4.* Cell, 2010. **141**(3): p. 446-457.
33. Iyer, L.M., et al., *Evolutionary history and higher order classification of AAA+ ATPases.* J Struct Biol, 2004. **146**(1-2): p. 11-31.
34. Sauer, R.T., et al., *Sculpting the proteome with AAA(+) proteases and disassembly machines.* Cell, 2004. **119**(1): p. 9-18.
35. Lupas, A.N., et al., *Self-compartmentalizing proteases.* Trends Biochem Sci, 1997. **22**(10): p. 399-404.
36. Rawlings, N.D., A.J. Barrett, and A. Bateman, *MEROPS: the peptidase database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D227-33.
37. Djuranovic, S., *Evolution of substrate recognition domains of AAA proteins*, in *Fakultät für Chemie und Pharmazie*. 2007, Eberhard-Karls-Universität: Tübingen.
38. Frickey, T. and A.N. Lupas, *Phylogenetic analysis of AAA proteins.* J Struct Biol, 2004. **146**(1-2): p. 2-10.
39. Djuranovic, S., et al., *Characterization of AMA, a new AAA protein from Archaeoglobus and methanogenic archaea.* J Struct Biol, 2006. **156**(1): p. 130-8.
40. Gerega, A., et al., *VAT, the thermoplasma homolog of mammalian p97/VCP, is an N domain-regulated protein unfoldase.* J Biol Chem, 2005. **280**(52): p. 42856-62.
41. Jentsch, S. and S. Rumpf, *Cdc48 (p97): a "molecular gearbox" in the ubiquitin pathway?* Trends Biochem Sci, 2007. **32**(1): p. 6-11.

42. Ye, Y., H.H. Meyer, and T.A. Rapoport, *The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol.* Nature, 2001. **414**(6864): p. 652-6.
43. Koller, K.J. and M.J. Brownstein, *Use of a cDNA clone to identify a supposed precursor protein containing valosin.* Nature, 1987. **325**(6104): p. 542-5.
44. Fröhlich, K.U., et al., *Yeast cell cycle protein CDC48p shows full-length homology to the mammalian protein VCP and is a member of a protein family involved in secretion, peroxisome formation, and gene expression.* J Cell Biol, 1991. **114**(3): p. 443-53.
45. Pamnani, V., et al., *Cloning, sequencing and expression of VAT, a CDC48/p97 ATPase homologue from the archaeon Thermoplasma acidophilum.* FEBS Lett, 1997. **404**(2-3): p. 263-8.
46. Golbik, R., et al., *The Janus face of the archaeal Cdc48/p97 homologue VAT: protein folding versus unfolding.* Biol Chem, 1999. **380**(9): p. 1049-62.
47. McLachlan, A.D., *Gene duplications in the structural evolution of chymotrypsin.* J Mol Biol, 1979. **128**(1): p. 49-79.
48. Murzin, A.G., A.M. Lesk, and C. Chothia, *Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis.* J Mol Biol, 1994. **236**(5): p. 1369-81.
49. Coles, M., et al., *The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple betaalphabetabeta element.* Curr Biol, 1999. **9**(20): p. 1158-68.
50. Castillo, R., et al., *A six-stranded double-psi beta barrel is shared by several protein superfamilies.* Structure, 1999. **7**(2): p. 227-236.
51. Park, S., et al., *Ufd1 exhibits the AAA-ATPase fold with two distinct ubiquitin interaction sites.* Structure, 2005. **13**(7): p. 995-1005.
52. Alva, V., et al., *Cradle-loop barrels and the concept of metafolds in protein classification by natural descent.* Curr Opin Struct Biol, 2008. **18**(3): p. 358-65.
53. Grishin, N.V., *Fold change in evolution of protein structures.* J Struct Biol, 2001. **134**(2-3): p. 167-85.
54. Coles, M., et al., *The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple betaalphabetabeta element.* Curr Biol, 1999. **9**(20): p. 1158-68.
55. Zwickl, P., et al., *An archaebacterial ATPase, homologous to ATPases in the eukaryotic 26 S proteasome, activates protein breakdown by 20 S proteasomes.* J Biol Chem, 1999. **274**(37): p. 26008-14.
56. Wolf, S., et al., *Characterization of ARC, a divergent member of the AAA ATPase family from Rhodococcus erythropolis.* J Mol Biol, 1998. **277**(1): p. 13-25.
57. Glickman, M.H., et al., *The regulatory particle of the Saccharomyces cerevisiae proteasome.* Mol Cell Biol, 1998. **18**(6): p. 3149-62.
58. Zhang, F., et al., *Mechanism of substrate unfolding and translocation by the regulatory particle of the proteasome from Methanocaldococcus jannaschii.* Mol Cell, 2009. **34**(4): p. 485-96.
59. Zhang, F., et al., *Structural insights into the regulatory particle of the proteasome from Methanocaldococcus jannaschii.* Mol Cell, 2009. **34**(4): p. 473-84.
60. Murzin, A.G., *OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences.* EMBO J, 1993. **12**(3): p. 861-7.
61. Arcus, V., *OB-fold domains: a snapshot of the evolution of sequence, structure and function.* Curr Opin Struct Biol, 2002. **12**(6): p. 794-801.

62. Andreeva, A., et al., *Data growth and its impact on the SCOP database: new developments.* Nucleic Acids Res, 2008. **36**(Database issue): p. D419-25.
63. Tanaka, K. and A. Ichihara, *Involvement of proteasomes (multicatalytic proteinase) in ATP-dependent proteolysis in rat reticulocyte extracts.* FEBS Lett, 1988. **236**(1): p. 159-62.
64. McGuire, M.J. and G.N. DeMartino, *Purification and characterization of a high molecular weight proteinase (macropain) from human erythrocytes.* Biochim Biophys Acta, 1986. **873**(2): p. 279-89.
65. Baumeister, W., et al., *The proteasome: paradigm of a self-compartmentalizing protease.* Cell, 1998. **92**(3): p. 367-80.
66. Seemüller, E., et al., *Proteasome from Thermoplasma acidophilum: A Threonine Protease.* Science, 1995. **268**(5210): p. 579-582.
67. Seemuller, E., A.N. Lupas, and W. Baumeister, *Autocatalytic processing of the 20S proteasome.* Nature, 1996. **382**(6590): p. 468-71.
68. Rabl, J., et al., *Mechanism of Gate Opening in the 20S Proteasome by the Proteasomal ATPases.* Mol Cell, 2008. **30**(3): p. 360-8.
69. Religa, T.L., R. Sprangers, and L.E. Kay, *Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR.* Science, 2010. **328**(5974): p. 98-102.
70. Groll, M., et al., *Structure of 20S proteasome from yeast at 2.4 A resolution.* Nature, 1997. **386**(6624): p. 463-71.
71. Rock, K.L., et al., *Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules.* Cell, 1994. **78**(5): p. 761-71.
72. Morel, S., et al., *Processing of some antigens by the standard proteasome but not by the immunoproteasome results in poor presentation by dendritic cells.* Immunity, 2000. **12**(1): p. 107-17.
73. Humbard, M.A., et al., *Ubiquitin-like small archaeal modifier proteins (SAMPs) in Haloferax volcanii.* Nature, 2010. **463**(7277): p. 54-60.
74. Tamura, T., et al., *The first characterization of a eubacterial proteasome: the 20S complex of Rhodococcus.* Curr Biol, 1995. **5**(7): p. 766-74.
75. Sousa, M.C., et al., *Crystal and solution structures of an HslUV protease-chaperone complex.* Cell, 2000. **103**(4): p. 633-43.
76. Brannigan, J.A., et al., *A protein catalytic framework with an N-terminal nucleophile is capable of self-activation.* Nature, 1995. **378**(6555): p. 416-9.
77. Duggleby, H.J., et al., *Penicillin acylase has a single-amino-acid catalytic centre.* Nature, 1995. **373**(6511): p. 264-8.
78. Ditzel, L., et al., *Conformational constraints for protein self-cleavage in the proteasome.* J Mol Biol, 1998. **279**(5): p. 1187-91.
79. Oinonen, C. and J. Rouvinen, *Structural comparison of Ntn-hydrolases.* Protein Sci, 2000. **9**(12): p. 2329-37.
80. Dodson, G. and A. Wlodawer, *Catalytic triads and their relatives.* Trends Biochem Sci, 1998. **23**(9): p. 347-52.
81. Smith, D.M., et al., *ATP binding to PAN or the 26S ATPases causes association with the 20S proteasome, gate opening, and translocation of unfolded proteins.* Mol Cell, 2005. **20**(5): p. 687-98.
82. Striebel, F., et al., *The mycobacterial Mpa-proteasome unfolds and degrades pupylated substrates by engaging Pup's N-terminus.* EMBO J, 2010. **29**(7): p. 1262-71.
83. Smith, D.M., et al., *Docking of the Proteasomal ATPases' Carboxyl Termini in the 20S Proteasome's alpha Ring Opens the Gate for Substrate Entry.* Mol Cell, 2007. **27**(5): p. 731-44.

84. Stadtmueller, B.M., et al., *Structural models for interactions between the 20S proteasome and its PAN/19S activators.* J Biol Chem, 2010. **285**(1): p. 13-7.
85. Kessel, M., et al., *Six-fold rotational symmetry of ClpQ, the E. coli homolog of the 20S proteasome, and its ATP-dependent activator, ClpY.* FEBS Lett, 1996. **398**(2-3): p. 274-8.
86. Nickell, S., et al., *Insights into the molecular architecture of the 26S proteasome.* Proc Natl Acad Sci U S A, 2009. **106**(29): p. 11943-7.
87. Gille, C., et al., *A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome.* J Mol Biol, 2003. **326**(5): p. 1437-48.
88. Förster, A., et al., *The 1.9 A structure of a proteasome-11S activator complex and implications for proteasome-PAN/PA700 interactions.* Mol Cell, 2005. **18**(5): p. 589-99.
89. Sadre-Bazzaz, K., et al., *Structure of a Blm10 complex reveals common mechanisms for proteasome binding and gate opening.* Mol Cell, 2010. **37**(5): p. 728-35.
90. Hershko, A. and A. Ciechanover, *The ubiquitin system.* Annu Rev Biochem, 1998. **67**: p. 425-79.
91. Hochstrasser, M., *Origin and function of ubiquitin-like proteins.* Nature, 2009. **458**(7237): p. 422-9.
92. Kerscher, O., R. Felberbaum, and M. Hochstrasser, *Modification of proteins by ubiquitin and ubiquitin-like proteins.* Annu Rev Cell Dev Biol, 2006. **22**: p. 159-80.
93. Schrader, E.K., K.G. Harstad, and A. Matouschek, *Targeting proteins for degradation.* Nat Chem Biol, 2009. **5**(11): p. 815-22.
94. Inobe, T., et al., *Defining the geometry of the two-component proteasome degron.* Nat Chem Biol, 2011. **7**(3): p. 161-7.
95. Peters, J.M., *The anaphase-promoting complex: proteolysis in mitosis and beyond.* Mol Cell, 2002. **9**(5): p. 931-43.
96. Iyer, L.M., A.M. Burroughs, and L. Aravind, *The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains.* Genome Biol, 2006. **7**(7): p. R60.
97. Striebel, F., W. Kress, and E. Weber-Ban, *Controlled destruction: AAA+ ATPases in protein degradation from bacteria to eukaryotes.* Curr Opin Struct Biol, 2009. **19**(2): p. 209-17.
98. Darwin, K.H. and K. Hofmann, *SAMPyling proteins in archaea.* Trends Biochem Sci, 2010.
99. Pearce, M.J., et al., *Ubiquitin-like protein involved in the proteasome pathway of Mycobacterium tuberculosis.* Science, 2008. **322**(5904): p. 1104-7.
100. Wang, T., K.H. Darwin, and H. Li, *Binding-induced folding of prokaryotic ubiquitin-like protein on the Mycobacterium proteasomal ATPase targets substrates for degradation.* Nat Struct Mol Biol, 2010. **17**(11): p. 1352-7.
101. Iyer, L.M., A.M. Burroughs, and L. Aravind, *Unraveling the biochemistry and provenance of pupylation: a prokaryotic analog of ubiquitination.* Biol Direct, 2008. **3**: p. 45.
102. Striebel, F., et al., *Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes.* Nat Struct Biol, 2009.
103. Sutter, M., et al., *Prokaryotic ubiquitin-like protein (Pup) is coupled to substrates via the side chain of its C-terminal glutamate.* J Am Chem Soc, 2010. **132**(16): p. 5610-2.

104. Wang, J., J.A. Hartling, and J.M. Flanagan, *The structure of ClpP at 2.3 A resolution suggests a model for ATP-dependent proteolysis.* Cell, 1997. **91**(4): p. 447-56.
105. Keiler, K.C., P.R. Waller, and R.T. Sauer, *Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA.* Science, 1996. **271**(5251): p. 990-3.
106. Gottesman, S., et al., *The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system.* Genes Dev, 1998. **12**(9): p. 1338-47.
107. Varshavsky, A., *The N-end rule: functions, mysteries, uses.* Proc Natl Acad Sci U S A, 1996. **93**(22): p. 12142-9.
108. Dohmen, R.J., et al., *The N-end rule is mediated by the UBC2(RAD6) ubiquitin-conjugating enzyme.* Proc Natl Acad Sci U S A, 1991. **88**(16): p. 7351-5.
109. Gavin, A.-C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.
110. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.
111. Darwin, C., *On the Origin of Species*. 1859, London: Murray.
112. Koonin, E.V., *Darwinian evolution in the light of genomics.* Nucleic Acids Res, 2009. **37**(4): p. 1011-34.
113. Kimura, M., *Evolutionary rate at the molecular level.* Nature, 1968. **217**(5129): p. 624-6.
114. Zuckerkandl, E. and L. PAULING, *Evolutionary divergence and convergence in proteins.* Evolving genes and proteins, 1965.
115. Drummond, D.A., et al., *Why highly expressed proteins evolve slowly.* Proc Natl Acad Sci USA, 2005. **102**(40): p. 14338-43.
116. Wolf, M.Y., Y.I. Wolf, and E.V. Koonin, *Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution.* Biol Direct, 2008. **3**: p. 40.
117. Soskine, M. and D.S. Tawfik, *Mutational effects and the evolution of new protein functions.* Nat Rev Genet, 2010. **11**(8): p. 572-82.
118. Rutherford, S.L. and S. Lindquist, *Hsp90 as a capacitor for morphological evolution.* Nature, 1998. **396**(6709): p. 336-42.
119. Tokuriki, N. and D.S. Tawfik, *Chaperonin overexpression promotes genetic variation and enzyme evolution.* Nature, 2009. **459**(7247): p. 668-73.
120. Tokuriki, N. and D.S. Tawfik, *Protein dynamism and evolvability.* Science, 2009. **324**(5924): p. 203-7.
121. Sommer, R.J., *Homology and the hierarchy of biological systems.* Bioessays, 2008. **30**(7): p. 653-8.
122. Doolittle, R.F., *Convergent evolution: the need to be explicit.* Trends Biochem Sci, 1994. **19**(1): p. 15-8.
123. Krem, M.M. and E. Di Cera, *Molecular markers of serine protease evolution.* EMBO J, 2001. **20**(12): p. 3036-45.
124. Aravind, L. and E.V. Koonin, *Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches.* J Mol Biol, 1999. **287**(5): p. 1023-40.
125. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* EMBO J, 1986. **5**(4): p. 823-6.
126. Murzin, A.G., *How far divergent evolution goes in proteins.* Curr Opin Struct Biol, 1998. **8**(3): p. 380-7.
127. Krishna, S.S. and N.V. Grishin, *Structurally analogous proteins do exist!* Structure, 2004. **12**(7): p. 1125-7.

128. Sonnhammer, E.L.L. and E.V. Koonin, *Orthology, paralogy and proposed classification for paralog subtypes.* Trends Genet, 2002. **18**(12): p. 619-20.
129. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
130. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
131. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
132. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
133. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
134. Benner, S.A., M.A. Cohen, and G.H. Gonnet, *Empirical and structural models for insertions and deletions in the divergent evolution of proteins.* J Mol Biol, 1993. **229**(4): p. 1065-82.
135. Ohno, S., U. Wolf, and N.B. Atkin, *Evolution from fish to mammals by gene duplication.* Hereditas, 1968. **59**(1): p. 169-87.
136. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.
137. Marcotte, E., et al., *A census of protein repeats.* J Mol Biol, 1999. **293**(1): p. 151-160.
138. Doolittle, R.F., *The multiplicity of domains in proteins.* Annu Rev Biochem, 1995. **64**: p. 287-314.
139. Lindqvist, Y. and G. Schneider, *Circular permutations of natural protein sequences: structural evidence.* Curr Opin Struct Biol, 1997. **7**(3): p. 422-7.
140. Alva, V., et al., *A galaxy of folds.* Protein Sci, 2010. **19**(1): p. 124-30.
141. Sadreyev, R., B. Kim, and N.V. Grishin, *Discrete-continuous duality of protein structure space.* Curr Opin Struct Biol, 2009.
142. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
143. Rao, S.T. and M.G. Rossmann, *Comparison of super-secondary structures in proteins.* J Mol Biol, 1973. **76**(2): p. 241-56.
144. Fetrow, J. and A. Godzik, *Function driven protein evolution. A possible proto-protein for the RNA-binding proteins.* Pac Symp Biocomput, 1998: p. 485-96.
145. Lupas, A.N., C. Ponting, and R.B. Russell, *On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world?* J Struct Biol, 2001. **134**(2-3): p. 191-203.
146. Soding, J. and A.N. Lupas, *More than the sum of their parts: on the evolution of proteins from peptides.* Bioessays, 2003. **25**(9): p. 837-46.
147. Gilbert, W., *The RNA world.* Nature, 1986.
148. Cech, T.R., *Crawling out of the RNA world.* Cell, 2009. **136**(4): p. 599-602.
149. McLachlan, A.D., *Repeating sequences and gene duplication in proteins.* J Mol Biol, 1972. **64**(2): p. 417-37.
150. Hocker, B., et al., *Dissection of a (betaalpha)8-barrel enzyme into two folded halves.* Nat Struct Biol, 2001. **8**(1): p. 32-6.
151. Soding, J., M. Remmert, and A. Biegert, *HHrep: de novo protein repeat detection and the origin of TIM barrels.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W137-42.

152. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. **5**(8): p. 1093-108.
153. Bernstein, F.C., et al., *The Protein Data Bank: a computer-based archival file for macromolecular structures.* J Mol Biol, 1977. **112**(3): p. 535-42.
154. Hocker, B., S. Schmidt, and R. Sterner, *A common evolutionary origin of two elementary enzyme folds.* FEBS Lett, 2002. **510**(3): p. 133-5.
155. Day, R., et al., *A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary.* Protein Sci, 2003. **12**(10): p. 2150-60.
156. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
157. Bateman, A., P. Coggill, and R.D. Finn, *DUFs: families in search of function.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2010. **66**(Pt 10): p. 1148-52.
158. Jaroszewski, L., et al., *Exploration of uncharted regions of the protein universe.* PLoS Biol, 2009. **7**(9): p. e1000205.
159. Felsenstein, J., *Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.* Methods Enzymol, 1996. **266**: p. 418-27.
160. Frickey, T. and A.N. Lupas, *CLANS: a Java application for visualizing protein families based on pairwise similarity.* Bioinformatics, 2004. **20**(18): p. 3702-4.
161. Woese, C.R., O. Kandler, and M.L. Wheelis, *Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.* Proc Natl Acad Sci USA, 1990. **87**(12): p. 4576-9.
162. Woese, C.R. and G.E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms.* Proc Natl Acad Sci U S A, 1977. **74**(11): p. 5088-90.
163. Huber, H., et al., *A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont.* Nature, 2002. **417**(6884): p. 63-7.
164. Brochier-Armanet, C., et al., *Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota.* Nat Rev Microbiol, 2008. **6**(3): p. 245-52.
165. Brunk, C.F. and N. Eis, *Quantitative measure of small-subunit rRNA gene sequences of the kingdom korarchaeota.* Appl Environ Microbiol, 1998. **64**(12): p. 5064-6.
166. Bapteste, E., C. Brochier, and Y. Boucher, *Higher-level classification of the Archaea: evolution of methanogenesis and methanogens.* Archaea, 2005. **1**(5): p. 353-63.
167. Lipp, J.S., et al., *Significant contribution of Archaea to extant biomass in marine subsurface sediments.* Nature, 2008. **454**(7207): p. 991-4.
168. Koonin, E.V., *The origin and early evolution of eukaryotes in the light of phylogenomics.* Genome Biol, 2010. **11**(5): p. 209.
169. Embley, T.M. and W. Martin, *Eukaryotic evolution, changes and challenges.* Nature, 2006. **440**(7084): p. 623-30.
170. Lane, N. and W. Martin, *The energetics of genome complexity.* Nature, 2010. **467**(7318): p. 929-34.
171. Gribaldo, S., et al., *The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse?* Nat Rev Microbiol, 2010. **8**(10): p. 743-52.
172. Rivera, M.C. and J.A. Lake, *The ring of life provides evidence for a genome fusion origin of eukaryotes.* Nature, 2004. **431**(7005): p. 152-5.
173. Rivera, M.C., et al., *Genomic evidence for two functionally distinct gene classes.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 6239-44.
174. Cotton, J.A. and J.O. McInerney, *Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function.* Proc Natl Acad Sci USA, 2010. **107**(40): p. 17252-5.

175. Logsdon, J.M., *Eukaryotic evolution: the importance of being archaebacterial.* Curr Biol, 2010. **20**(24): p. R1078-9.
176. Santarella-Mellwig, R., et al., *The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins.* PLoS Biol, 2010. **8**(1): p. e1000281.
177. Fuerst, J.A. and E. Sagulenko, *Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function.* Nat Rev Microbiol, 2011. **9**(6): p. 403-13.
178. Varnay, I., et al., *Optimized Measurement Temperature Gives Access to the Solution Structure of a 49 kDa Homohexameric β-Propeller.* J Am Chem Soc, 2010.
179. Valas, R.E. and P.E. Bourne, *Rethinking proteasome evolution: two novel bacterial proteasomes.* J Mol Evol, 2008. **66**(5): p. 494-504.
180. Iyer, L.M., et al., *Amidoligases with ATP-grasp, glutamine synthetase-like and acetyltransferase-like domains: synthesis of novel metabolites and peptide modifications of proteins.* Mol. BioSyst., 2009. **5**(12): p. 1636-60.
181. Saiki, R.K., et al., *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.* Science, 1988. **239**(4839): p. 487-91.
182. Hendrickson, W.A., *Determination of macromolecular structures from anomalous diffraction of synchrotron radiation.* Science, 1991. **254**(5028): p. 51-8.
183. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
184. Altschul, S., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
185. Soding, J., et al., *HHsenser: exhaustive transitive profile search using HMM-HMM comparison.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W374-8.
186. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci USA, 1992. **89**(22): p. 10915-9.
187. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
188. Chenna, R., et al., *Multiple sequence alignment with the Clustal series of programs.* Nucleic Acids Res, 2003. **31**(13): p. 3497-500.
189. Holm, L. and P. Rosenström, *Dali server: conservation mapping in 3D.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W545-9.
190. Schwede, T., et al., *SWISS-MODEL: An automated protein homology-modeling server.* Nucleic Acids Res, 2003. **31**(13): p. 3381-5.
191. Sali, A. and T.L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints.* J Mol Biol, 1993. **234**(3): p. 779-815.
192. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
193. Karp, P.D., et al., *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.* Nucleic Acids Res, 2005. **33**(19): p. 6083-9.
194. Kanehisa, M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2010. **38**(Database issue): p. D355-60.
195. Biegert, A., et al., *The MPI Bioinformatics Toolkit for protein sequence analysis.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W335-9.
196. Biegert, A. and J. Soding, *De novo identification of highly diverged protein repeats by probabilistic consistency.* Bioinformatics, 2008. **24**(6): p. 807.

197. Gruber, M., J. Soding, and A.N. Lupas, *Comparative analysis of coiled-coil prediction methods.* J Struct Biol, 2006. **155**(2): p. 140-5.
198. Coles, M., et al., *Common evolutionary origin of swapped-hairpin and double-psi Beta barrels.* Structure, 2006. **14**(10): p. 1489-98.
199. Coles, M., et al., *AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels.* Structure, 2005. **13**(6): p. 919-28.
200. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.
201. Alva, V., et al., *The GD box: a widespread noncontiguous supersecondary structural element.* Protein Sci, 2009. **18**(9): p. 1961-6.
202. Bacher, A., et al., *Biosynthesis of vitamin b2 (riboflavin).* Annu Rev Nutr, 2000. **20**: p. 153-67.
203. Hasegawa, E., O. Ando, and Y. Nose, *Studies on riboflavin kinase and FAD pyrophosphorylase.* J Vitaminol (Kyoto), 1968. **14**(4): p. 303-11.
204. Susin, S., et al., *Riboflavin 3'- and 5'-sulfate, two novel flavins accumulating in the roots of iron-deficient sugar beet (Beta vulgaris).* J Biol Chem, 1993. **268**(28): p. 20958-20965.
205. Bauer, S., et al., *Crystal structure of Schizosaccharomyces pombe riboflavin kinase reveals a novel ATP and riboflavin-binding fold.* J Mol Biol, 2003. **326**(5): p. 1463-73.
206. Coles, M., et al., *AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels.* Structure, 2005. **13**(6): p. 919-28.
207. Cheek, S., H. Zhang, and N.V. Grishin, *Sequence and structure classification of kinases.* J Mol Biol, 2002. **320**(4): p. 855-881.
208. Cheek, S., et al., *A comprehensive update of the sequence and structure classification of kinases.* BMC Struct Biol, 2005. **5**: p. 6.
209. Horst, M., *Dolichol phosphorylation occurs via a CTP-dependent reaction in Artemia larvae.* J Exp Zool, 1989. **252**(1): p. 16-24.
210. Dodson, M.L., M.L. Michaels, and R.S. Lloyd, *Unified catalytic mechanism for DNA glycosylases.* J Biol Chem, 1994. **269**(52): p. 32709-12.
211. Doherty, A.J., L.C. Serpell, and C.P. Ponting, *The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA.* Nucleic Acids Res, 1996. **24**(13): p. 2488-97.
212. Fromme, J.C. and G.L. Verdine, *Structure of a trapped endonuclease III-DNA covalent intermediate.* EMBO J, 2003. **22**(13): p. 3461-71.
213. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D226-9.
214. Neer, E.J. and T.F. Smith, *G protein heterodimers: new structures propel new questions.* Cell, 1996. **84**(2): p. 175-8.
215. Moldovan, G.-L., B. Pfander, and S. Jentsch, *PCNA, the maestro of the replication fork.* Cell, 2007. **129**(4): p. 665-79.
216. Oyama, M., et al., *Human NTH1 physically interacts with p53 and proliferating cell nuclear antigen.* Biochem Biophys Res Commun, 2004. **321**(1): p. 183-91.
217. Kiyonari, S., et al., *Studies on the base excision repair (BER) complex in Pyrococcus furiosus.* Biochem Soc Trans, 2009. **37**(Pt 1): p. 79-82.
218. Murzin, A.G., *Structural principles for the propeller assembly of beta-sheets: the preference for seven-fold symmetry.* Proteins, 1992. **14**(2): p. 191-201.

219. Li, J., et al., *Structure of full-length porcine synovial collagenase reveals a C-terminal domain containing a calcium-linked, four-bladed beta-propeller.* Structure, 1995. **3**(6): p. 541-9.
220. Chaudhuri, I., J. Soding, and A.N. Lupas, *Evolution of the beta-propeller fold.* Proteins, 2007. **71**(2): p. 795-803.
221. Chaudhuri, I., *Evolution der beta-Propeller Proteine*, in *Fakultät für Chemie*. 2007, Technische Universität: München. p. 122.
222. Yadid, I. and D.S. Tawfik, *Reconstruction of Functional beta-Propeller Lectins via Homo-oligomeric Assembly of Shorter Fragments.* J Mol Biol, 2007. **365**(1): p. 10-7.
223. Yadid, I., et al., *Metamorphic proteins mediate evolutionary transitions of structure.* Proc Natl Acad Sci USA, 2010. **107**(16): p. 7287-92.
224. Murzin, A.G., *Biochemistry. Metamorphic proteins.* Science, 2008. **320**(5884): p. 1725-6.
225. Kostlánová, N., et al., *The fucose-binding lectin from Ralstonia solanacearum. A new type of beta-propeller architecture formed by oligomerization and interacting with fucoside, fucosyllactose, and plant xyloglucan.* J Biol Chem, 2005. **280**(30): p. 27839-49.
226. Remmert, M., et al., *Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin.* Mol Biol Evol, 2010. **27**(6): p. 1348-58.
227. Letunic, I. and P. Bork, *Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.* Nucleic Acids Res, 2011.
228. Ciccarelli, F.D., et al., *Toward automatic reconstruction of a highly resolved tree of life.* Science, 2006. **311**(5765): p. 1283-7.
229. Cavalier-Smith, T., *Rooting the tree of life by transition analyses.* Biol Direct, 2006. **1**: p. 19.
230. Couvreur, B., et al., *Eubacterial HslV and HslU subunits homologs in primordial eukaryotes.* Mol Biol Evol, 2002. **19**(12): p. 2110-7.
231. Li, D., et al., *Structural basis for the assembly and gate closure mechanisms of the Mycobacterium tuberculosis 20S proteasome.* EMBO J, 2010. **29**(12): p. 2037-47.
232. Seemuller, E., et al., *Proteasome from Thermoplasma acidophilum: A Threonine Protease.* Science, 1995. **268**(5210): p. 579-582.
233. Orengo, C.A., D.T. Jones, and J.M. Thornton, *Protein superfamilies and domain superfolds.* Nature, 1994. **372**(6507): p. 631-4.
234. Riechmann, L. and G. Winter, *Novel folded protein domains generated by combinatorial shuffling of polypeptide segments.* Proc Natl Acad Sci USA, 2000. **97**(18): p. 10068-73.
235. de Bono, S., et al., *A segment of cold shock protein directs the folding of a combinatorial protein.* Proc Natl Acad Sci USA, 2005. **102**(5): p. 1396-1401.
236. Rubin, D.M., et al., *Active site mutants in the six regulatory particle ATPases reveal multiple roles for ATP in the proteasome.* EMBO J, 1998. **17**(17): p. 4909-19.
237. Imkamp, F., et al., *Dop functions as a depupylase in the prokaryotic ubiquitin-like modification pathway.* EMBO Rep, 2010. **11**(10): p. 791-7.
238. Yu, Y., et al., *Interactions of PAN's C-termini with archaeal 20S proteasome and implications for the eukaryotic proteasome-ATPase interactions.* EMBO J, 2010. **29**(3): p. 692-702.
239. Nickell, S., et al., *Insights into the molecular architecture of the 26S proteasome.* Proc Natl Acad Sci USA, 2009.

240. Kisselev, A.F., T.N. Akopian, and A.L. Goldberg, *Range of sizes of peptide products generated during degradation of different proteins by archaeal proteasomes.* J Biol Chem, 1998. **273**(4): p. 1982-9.
241. Liu, C.W., et al., *Endoproteolytic activity of the proteasome.* Science, 2003. **299**(5605): p. 408-11.
242. Bajorek, M., D. Finley, and M.H. Glickman, *Proteasome disassembly and downregulation is correlated with viability during stationary phase.* Curr Biol, 2003. **13**(13): p. 1140-4.
243. Ruepp, A., et al., *The Chaperones of the archaeon Thermoplasma acidophilum.* J Struct Biol, 2001. **135**(2): p. 126-38.
244. Hobel, C.F., et al., *The Sulfolobus solfataricus AAA protein Sso0909, a homologue of the eukaryotic ESCRT Vps4 ATPase.* Biochem Soc Trans, 2008. **36**(Pt 1): p. 94-8.
245. Makarova, K.S., et al., *Evolution of diverse cell division and vesicle formation systems in Archaea.* Nat Rev Microbiol, 2010.
246. Serek-Heuberger, J., et al., *Two unique membrane-bound AAA proteins from Sulfolobus solfataricus.* Biochem Soc Trans, 2009. **37**(Pt 1): p. 118-22.
247. Makarova, K.S. and E.V. Koonin, *Archaeal Ubiquitin-Like Proteins: Functional Versatility and Putative Ancestral Involvement in tRNA Modification Revealed by Comparative Genomic Analysis.* Archaea, 2010. **2010**.
248. Robinson, C.V., A. Sali, and W. Baumeister, *The molecular sociology of the cell.* Nature, 2007. **450**(7172): p. 973-82.
249. Li, W., et al., *Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling.* PLoS One, 2008. **3**(1): p. e1487.
250. Zhao, G., et al., *Studies on peptide:N-glycanase-p97 interaction suggest that p97 phosphorylation modulates endoplasmic reticulum-associated degradation.* Proc Natl Acad Sci USA, 2007. **104**(21): p. 8785-90.
251. Raybin, D. and M. Flavin, *Modification of tubulin by tyrosylation in cells and extracts and its effect on assembly in vitro.* J Cell Biol, 1977. **73**(2): p. 492-504.
252. Janke, C., et al., *Tubulin polyglutamylase enzymes are members of the TTL domain protein family.* Science, 2005. **308**(5729): p. 1758-62.
253. Kang, W.K., et al., *Characterization of the gene rimK responsible for the addition of glutamic acid residues to the C-terminus of ribosomal protein S6 in Escherichia coli K12.* Mol Gen Genet, 1989. **217**(2-3): p. 281-8.
254. Dalle-Donne, I., et al., *Protein S-glutathionylation: a regulatory device from bacteria to humans.* Trends Biochem Sci, 2009. **34**(2): p. 85-96.
255. Chuang, S.E., et al., *Sequence analysis of four new heat-shock genes constituting the hslTS/ibpAB and hslVU operons in Escherichia coli.* Gene, 1993. **134**(1): p. 1-6.
256. Koonin, E.V., Y.I. Wolf, and L. Aravind, *Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach.* Genome Res, 2001. **11**(2): p. 240-52.
257. Fesus, L. and M. Piacentini, *Transglutaminase 2: an enigmatic enzyme with diverse functions.* Trends Biochem Sci, 2002. **27**(10): p. 534-9.
258. Couvreur, B., et al., *Eubacterial HslV and HslU subunits homologs in primordial eukaryotes.* Mol Biol Evol, 2002. **19**(12): p. 2110-7.
259. Ruiz-Gonzalez, M.X. and I. Marin, *Proteasome-related HslU and HslV genes typical of eubacteria are widespread in eukaryotes.* J Mol Evol, 2006. **63**(4): p. 504-12.

260. Lupas, A., P. Zwickl, and W. Baumeister, *Proteasome sequences in eubacteria.* Trends Biochem Sci, 1994. **19**(12): p. 533-4.
261. Lupas, A.N., et al., *Eubacterial proteasomes.* Mol Biol Rep, 1997. **24**(1-2): p. 125-31.
262. De Mot, R., *Actinomycete-like proteasomes in a Gram-negative bacterium.* Trends Microbiol, 2007. **15**(8): p. 335-8.
263. Darwin, K.H., et al., *The proteasome of Mycobacterium tuberculosis is required for resistance to nitric oxide.* Science, 2003. **302**(5652): p. 1963-6.
264. Lin, G., et al., *Inhibitors selective for mycobacterial versus human proteasomes.* Nature, 2009. **461**(7264): p. 621-6.
265. Darwin, K.H., *Prokaryotic ubiquitin-like protein (Pup), proteasomes and pathogenesis.* Nat Rev Microbiol, 2009. **7**(7): p. 485-91.
266. Gottesman, S., *Proteolysis in bacterial regulatory circuits.* Annu Rev Cell Dev Biol, 2003. **19**: p. 565-87.
267. Lawrence, J.G. and J.R. Roth, *Selfish operons: horizontal transfer may drive the evolution of gene clusters.* Genetics, 1996. **143**(4): p. 1843-60.
268. Cavalier-Smith, T., *Predation and eukaryote cell origins: a coevolutionary perspective.* Int J Biochem Cell Biol, 2009. **41**(2): p. 307-22.
269. Wolfe, K.H. and D.C. Shields, *Molecular evidence for an ancient duplication of the entire yeast genome.* Nature, 1997. **387**(6634): p. 708-13.
270. Buck, L. and R. Axel, *A novel multigene family may encode odorant receptors: a molecular basis for odor recognition.* Cell, 1991. **65**(1): p. 175-87.

ZUSAMMENFASSUNG

AAA (+) Proteine sind ATPasen, die die Energie aus der Hydrolyse von ATP an die Remodellierung, Disaggregation und Entfaltung einer Vielfalt von Substraten koppeln. Die zentrale ATPase Domäne funktioniert als molekularer Schalter, der über die Bindung von Substraten an N-terminalen Erkennungsdomänen betätigt wird, und der die remodellierten Substrate an interagierende Partnerproteine weiterreicht. Im Fall von AAA+ Proteasen werden fehlgefaltete Proteine durch die N-domäne erkannt, durch die Pore des hexameren ATPase-Rings gefädelt und schließlich von Proteasen wie dem Proteasome zu kleineren Peptiden hydrolysiert.

Wir haben die divergente Evolution der N-Domänen von ATPasen, die als Regulator für das Proteasom fungieren (PAN), bzw. fungieren könnten (CDC48 und AMA), untersucht, sowie die C-terminalen Interaktionsmotive dieser ATPasen, die eine wichtige Rolle für die Bindung an das Proteasoms spielen.

Im ersten Teil dieser Arbeit beschreiben wir drei Fallstudien von hypothetischen Proteinen unbekannter Funtion, die sich mit der Evolution von drei dieser N-Domänen, der Double-ψ-Barrel-, der β-Clam- und der OB-Fold-Domäne, beschäftigen. Wir präsentieren die erste Charakterisierung einer CTP-spezifischen archaealen Riboflavinkinase, die mit dem Double-ψ-Barrel von AAA-Proteinen der CDC48-Gruppe durch ein dupliziertes ββαβ-Peptid evolutionär verwandt ist. Diese Riboflavinkinasen bilden eine evolutionäre Brücke zwischen ebenfalls verwandten Transkriptionsfaktoren und ATP-spezifischen Riboflavinkinasen, wie sie in Eukaryonten und Bakterien zu finden sind. Wir beschreiben die evolutionären Veränderungen, die nötig waren, um ein DNA-bindendes Protein in ein Enzym zu verwandeln.

Eine β-Clam Domäne, die in AAA-Proteinen wie CDC48 und AMA enthalten ist, wurde im Kontext einer C-terminalen Domäne detektiert, die keine Sequenzähnlichkeit zu bekannten Proteinen aufweist. Wir zeigen die Volllängenstruktur eines Mitglieds dieser Proteinfamilie, deren C-terminale Domäne einen als homohexamerer β-Propeller mit 12 Blättern faltet (HP12). Ein Monomer des Propellers besteht aus zwei Blättern, die auf ein Duplikationsereignis zurückgeführt werden können, was darauf hinweist, dass dieses Protein ein Intermediat in der Evolution von monomeren β-Propellern darstellen könnte. Wir zeigen außerdem, dass HP12 einen ternären Komplex mit einer genetisch gekoppelten Endonuclease vom Typ III und DNA bildet. Daher ist dieses Protein in die Reparatur von DNA involviert.

Wir haben eine OB-Fold Domäne, ähnlich derjenigen, die sich am N-terminus von proteasomalen PAN ATPasen befindet, in einer Proteinfamilie detektiert, die zusätzlich eine proteolytische Ntn-hydrolase Domäne enthält, welche wiederum Kernbestandteil des Proteasoms ist. Die Kristallstruktur eines Mitglieds dieser Familie zeigt ein monomers Proteasom-Homolog (MPM), das unserer Vorhersage entsprechend eine OB-fold Domäne am C-terminus enthält. Die interne Symmetrie dieser Domäne und die Repetition eines Sequenzmotivs verweisen auf eine Entstehung durch Duplikation eines dreisträngigen Supersekundärstruktureelements hin.

Duplizierte Supersekundärstrukturelemente wie das ββαβ-Peptid der Double-ψ-Barrel Domäne, das duplizierte Propellerblatt von HP12, sowie der dreisträngige β-Meander der OB-fold Domäne von MPM unterstützen die Hypothese, dass autonom faltende Domänen durch die Vervielfältigung, Fusion und Rekombination solcher ancestraler Peptide entstanden sein könnten.

Im zweiten Teil dieser Arbeit untersuchen wir die Ursprünge proteasomaler Proteindegradation. Wir präsentieren eine systematische Sequenzanalyse der C-termini archaealer AAA-Proteine, die das Vorhandensein des Motivs für die Interaktion mit dem Proteasom in ATPasen der CDC48- und der AMA-Gruppe enthüllt. Da die bekannte proteasom-aktivierende ATPase PAN in einer Vielzahl archaealer Organismen abwesend ist, schlagen wir vor, dass CDC48 und AMA ebenfalls als Regulatoren des Proteasoms fungieren können. Manche Archaeen besitzen bis zu fünf ATPasen mit dem Interaktionsmotif, weshalb wir die Existenz eines Netzwerks von ATPasen vorhersagen, welches das archaeale Proteasom reguliert. Dieses Netzwerk könnte die Leistungsfähigkeit proteasomaler Proteindegradation in Archaeen erhöhen, da verschiedene N-terminale Substraterkennungsdomänen involviert sind.

Die Analyse des genetischen Kontexts des bislang uncharakterisierten Proteasom-Homologen Anbu deckte eine Operonstruktur auf, die in Cyanobakterien und Proteobakterien weit verbreitet ist. Die Komponenten des Anbu-Operons deuten auf ein Markierungssystem zum gezielten Abbau von Proteinen hin, welches entfernte Ähnlichkeit mit der Ubiquitylierung besitzt, die Proteine zwecks Abbau zum Proteasom leitet. Experimentelle Evidenz für diese Hypothese sowie für die Netzwerk-Hypothese wird zur Zeit gesammelt.

Zuletzt haben wir die Verteilung von proteasom-ähnlichen Ntn-hydrolasen und vermeintlichen proteasomalen ATPasen auf dem Baum des Lebens untersucht. Diese Analyse unterstützt das Szenario, dass Actinobakterien, die einzige Gruppe bakterieller Organismen mit Proteasom, dieses durch lateralen Gentransfer erhalten haben. Die von uns charakterisierte MPM Protease ist ein vom Proteasom abgeleitetes Charakteristikum methanogener Archaeen und stellt eine Ausnahme unter den Proteasom-Homologen dar. Während das Proteasom und sein Homologes HslV große zylinderförmige Architekturen formieren, die einen Komplex mit hexameren ATPasen bilden, ist MPM ein monomer, das eine vermeintliche Substraterkennungsdomäne (den OB-fold) auf der selben Peptidkette besitzt. Daher dürfte die MPM Protease eine evolutionäre Entwicklung von einem komplizierten zu einem vereinfachten molekularen Phänotyp darstellen.

LEBENSLAUF

Persönliche Daten:

Name:	Moritz Ammelburg
Adresse:	Vogtshaldenstr. 5, 72074 Tübingen
Geburtsdatum und -ort:	17.01.1980; Marburg (Deutschland)
Staatsangehörigkeit:	deutsch

Schulbildung:

1986-1990	Grundschule Homberg
1990-1999	Abitur, Stiftsschule St. Johann Amöneburg
09/1999-08/2000	Zivildienst, Bettina-von-Arnim-Schule Marburg

Studium:

10/2000-08/2003	Bachelor of Science (Biochemie), Technische Universität München Bachelor Thesis: Interaktionsstudien mit dem Tumorsuppressor p53 und der Prolylisomerase Pin1; Betreuer: Prof. Johannes Buchner
10/2003-02/2006	Master of Science (Biochemie), Technische Universität München und ETH Zürich Master Thesis: The C-domain: A marker for AAA+ Proteins and a Homolog of the Histones; Betreuer: Prof. Andrei Lupas und Prof. Horst Kessler

Promotion:

09/2006-	Doktorarbeit, Max-Planck-Institut für Entwicklungsbiologie, Abteilung Proteinevolution und Eberhard-Karls-Universität Tübingen Dissertation: AAA Proteins and the Origins of Proteasomal Protein Degradation; Betreuer: Prof. Andrei Lupas und Prof. Thilo Stehle (Eingereicht am 31.5.11, verteidigt am 27.10.11)

Publikationen:

Ammelburg M, Frickey T, Lupas AN: Classification of AAA+ proteins, *J Struct Biol* 2006, 156: 2-11

Alva V*, Ammelburg M*, Söding J, Lupas AN: On the origin of the histone fold. *BMC Struct Biol* 2007, 7: 17

Ammelburg M*, Hartmann MD*, Djuranovic S, Alva V, Koretke KK, Martin J, Sauer G, Truffault V, Zeth K, Lupas AN, Coles M: A CTP-dependent riboflavin kinase forms a bridge in the evolution of cradle-loop barrels, *Structure* 2007, 15: 1577-90

Mir-Montazeri B, Ammelburg M, Forouzan D, Lupas AN, Hartmann MD: Crystal structure of a dimeric archaeal cleavage and polyadenylation specificity factor, *J Struct Biol* 2011, 173: 191-5

Heichlinger A, Ammelburg M, Kleinschnitz EM, Latus A, Maldener I, Flärdh K, Wohlleben W, Muth G: The MreB-like protein Mbl of Streptomyces coelicolor A3(2) depends on MreB for proper localization and contributes to spore wall synthesis, *J Bacteriol* 2011, 193: 1533-42

Vogelmann J, Ammelburg M, Finger C, Guezguez J, Linke D, Flötenmeyer M, Stierhof YD, Wohlleben W, Muth G: Conjugal plasmid transfer in Streptomyces resembles bacterial chromosome segregation by FtsK/SpoIIIE, *EMBO J* 2011 Apr 19 (Epub)