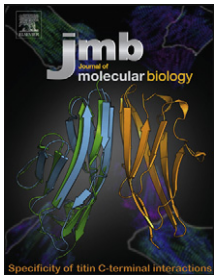
Article



Structure and Evolution of N-domains in AAA Metalloproteases

Franka Scharfenberg, Justyna Serek-Heuberger, Murray Coles, Marcus D. Hartmann, Michael Habeck[1], Jörg Martin, Andrei N. Lupas and Vikram Alva

Department of Protein Evolution, *Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany*

Correspondence to Andrei N. Lupas: *andrei.lupas@tuebingen.mpg.de*
http://dx.doi.org/10.1016/j.jmb.2014.12.024
Edited by A. Panchenko

Abstract

Metalloproteases of the AAA (*A*TPases *a*ssociated with various cellular *a*ctivities) family play a crucial role in protein quality control within the cytoplasmic membrane of bacteria and the inner membrane of eukaryotic organelles. These membrane-anchored hexameric enzymes are composed of an N-terminal domain with one or two transmembrane helices, a central AAA ATPase module, and a C-terminal Zn^{2+}-dependent protease. While the latter two domains have been well studied, so far, little is known about the N-terminal regions. Here, in an extensive bioinformatic and structural analysis, we identified three major, non-homologous groups of N-domains in AAA metalloproteases. By far, the largest one is the FtsH-like group of bacteria and eukaryotic organelles. The other two groups are specific to Yme1: one found in plants, fungi, and basal metazoans and the other one found exclusively in animals. Using NMR and crystallography, we determined the subunit structure and hexameric assembly of *Escherichia coli* FtsH-N, exhibiting an unusual α + β fold, and the conserved part of fungal Yme1-N from *Saccharomyces cerevisiae*, revealing a tetratricopeptide repeat fold. Our bioinformatic analysis showed that, uniquely among these proteins, the N-domain of Yme1 from the cnidarian *Hydra vulgaris* contains both the tetratricopeptide repeat region seen in basal metazoans and a region of homology to the N-domains of animals. Thus, it is a modern-day representative of an intermediate in the evolution of animal Yme1 from basal eukaryotic precursors.

Introduction

The AAA+ (*A*TPases *a*ssociated with various cellular *a*ctivities) superfamily of proteins represents one of the largest and most diverse clades of ring-shaped P-loop NTPases [1]. They are ubiquitous to all domains of life and are involved in the energy-dependent unfolding and disaggregation of macromolecules. AAA+ proteins are characterized by the presence of a non-ATPase N-terminal domain, one or two central copies of an extended P-loop ATPase harboring the conserved Walker A and B motifs, and a C-terminal α-helical subdomain (the C-domain; see Ref. [2]).

AAA proteins form a family within the AAA+ superfamily that is distinguished by the "second region of homology" found in their ATPase domain [3]. They assemble into hexameric complexes and play a significant role in many cellular processes, including protein degradation and maturation, gene expression, membrane fusion, membrane complex formation, and microtubule regulation. We have previously classified AAA proteins into six clades: D1 domains, D2 domains, proteasome subunits, metalloproteases, the "meiotic" group, and BCS1 [4].

The AAA metalloprotease subfamily has been identified so far in bacteria and eukaryotes but not in archaea. All members of this subfamily are membrane anchored through their N-terminal domain and are followed by one AAA ATPase module and a C-terminal metallopeptidase domain of the M41 family, which harbors the conserved Zn^{2+}-binding motif HEXXH. Hexameric complexes of these proteins are located in the cytoplasmic membrane of

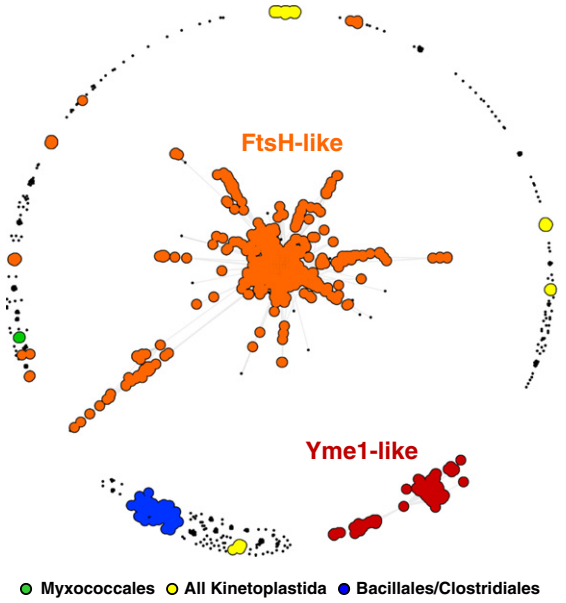


Fig. 1. Cluster map of AAA metalloprotease N-domains. Sequences were clustered in CLANS at a BLAST *P*-value cutoff of 1e-10. Each dot represents one protein sequence. BLAST connections are shown as gray lines; the darker a line, the higher the similarity. Sequences that matched the N-domain of *E. coli* FtsH with an HHpred probability of ≥50% were assigned to the FtsH-like group (see Materials and Methods) and are colored orange. Yme1-like proteins are colored red, paralogs of canonical FtsH from Bacillales and Clostridiales are in blue, paralogs of canonical FtsH from Myxococcales are in green, and the four highly divergent non-canonical groups from Kinetoplastida are in yellow. Sequences that could not be assigned to a particular group are shown as black dots. Further details are provided in the text.

bacteria and in the inner membranes of mitochondria and chloroplasts, where they are involved both as proteases and as chaperones in processing soluble and membrane-associated proteins. The first reported protein of the AAA metalloprotease subfamily was FtsH from *Escherichia coli* (occasionally also referred to as HflB; see Refs. [5] and [6]). It forms a homohexamer where each polypeptide chain spans the cytoplasmic membrane twice, thereby localizing the AAA and protease domains to the cytoplasm and the N-domain to the periplasm. The N-domain, including the two transmembrane (TM) helices, is involved in oligomerization and can regulate the activity of the hexamer in conjunction with the membrane proteins HflK and HflC [7,8].

Most bacteria contain only one FtsH homolog, whereas varying numbers of AAA metalloproteases have been identified in eukaryotic cells. One of the best studied cases is that of the three orthologs in yeast, which form two complexes with opposite topology within the inner membrane of mitochondria, termed i-AAA and m-AAA [9]. The homooligomeric i-AAA complex, named for the location of the catalytically active parts within the intermembrane space, is formed by Yme1 (Yta11), a protein with a single TM helix. The heterooligomeric m-AAA complex, where the C-terminal domains face the matrix of the organelle, is formed by the orthologs Yta10 and Yta12, each containing two TM helices. Notably, in photosynthetic organisms such as cyanobacteria and plants, the number of AAA metalloprotease genes is significantly increased in comparison to mitochondria and to bacteria that produce energy only by cellular respiration. For instance, in the genome of *Synechocystis sp.* PCC 6804, four FtsH homologs have been identified [10], and in *Arabidopsis thaliana*, as many as 17 genes encode such proteases [11].

Our previous work on the classification of AAA proteins showed that the N-domains of AAA metalloproteases are less conserved by comparison to their AAA and catalytic domains [4]. Yme1-like N-domains exhibit no homology to the N-domains of other metalloproteases, and additionally, they form two distinct groups that share no apparent sequence similarity. Spurred by these results and by the tremendous growth of sequence data in recent years, we decided to revisit the N-domains of AAA metalloproteases in order to gain further insight into their evolution.

Results and Discussion

To gather the N-domains of AAA metalloproteases, we searched the non-redundant protein sequence database at the National Center for Biotechnology Information (NCBI) using HMMER3 [12], with the profile of the M41 metallopeptidase from the Pfam database as seed. This yielded 11,816 sequences that were subsequently filtered to remove partial and redundant sequences and sequences without an AAA domain. In the resulting 10,352 sequences, the AAA domain and amino acids following it were masked out to obtain the set of N-domains. We employed cluster analysis for inferring the evolution of these sequences as, unlike phylogenetic methods that require well-curated multiple alignments and only allow calculation of trees with at most a few thousand sequences, clustering allows handling of large datasets comprising highly diverse, unaligned sequences. For cluster analysis, we used CLANS [13], an implementation of the Fruchterman–Reingold graph drawing algorithm, which treats sequences as point masses in a virtual multidimensional space, wherein they attract or repel each other depending on their pairwise sequence similarities. Sequences find their equilibrium position based on the force vectors resulting from all pairwise interactions. In the equilibrated map, groups of sequences with statistically significant pairwise similarities form tightly connected clusters, whereas dissimilar sequences tend to drift to the periphery.

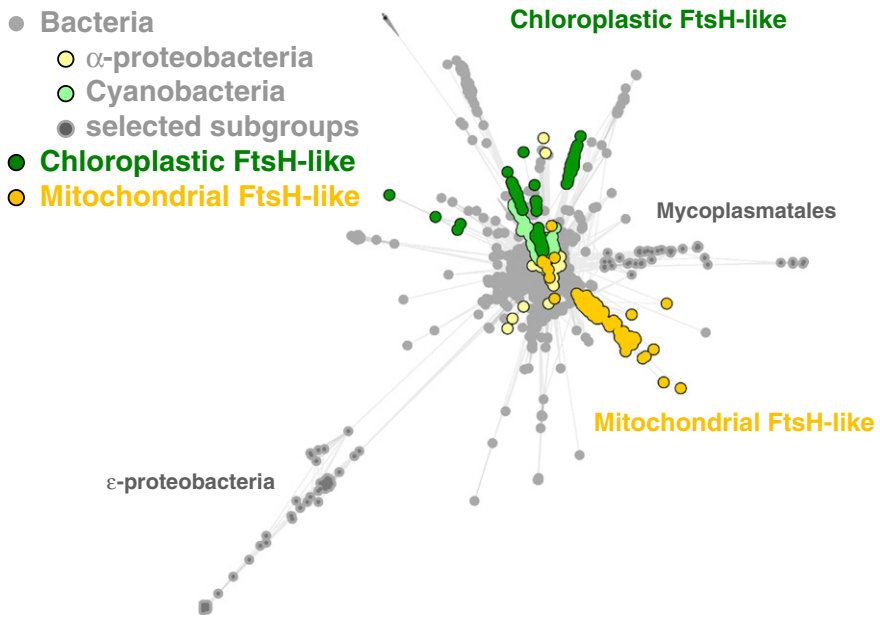


Fig. 2. Cluster map of FtsH-like N-domains. Sequences were clustered at a BLAST *P*-value cutoff of 1e-10. Each dot represents one protein sequence. BLAST connections are shown as gray lines; the darker a line, the higher the similarity.

The N-domains were clustered in CLANS by their pairwise BLAST *P*-values and the resulting map shows many distinct clusters and a number of singletons scattered in the periphery (Fig. 1). The largest cluster is formed by several tightly connected subclusters comprising FtsH-like N-domains from different bacterial and eukaryotic phyla. A divergent projection from this cluster comprises sequences from ε-proteobacteria. Two other clusters, loosely connected to each other, contain Yme1-like proteins, which are exclusive to eukaryotes. Additionally, the periphery of the map contains distinct clusters of N-domains from Kinetoplastida, Bacillales/Clostridiales, and Myxococcales that show no sequence similarity to each other or to other clusters in the map. Our map indicates that AAA metalloproteases have recruited non-homologous N-domains several times in the course of their evolution.

FtsH-like N-domains

The largest cluster within the map (Fig. 1) is the central FtsH-like cluster, comprising 9,069 of the 10,352 sequences, which are to 87% of bacterial origin. In most of these sequences, the FtsH-like N-domains can be readily detected even with the least sensitive sequence comparison methods such as BLAST and PSI-BLAST. Radiating from this cluster are a few divergent branches containing orthologs from species with special lifestyles, such as the Mycoplasmatales as intracellular pathogens (Fig. 2). The most divergent branches, however, contain paralogous sequences from organisms that also have a canonical FtsH copy in the central cluster. Particularly conspicuous is a distant branch of paralogs from ε-proteobacteria (Fig. 2), which have all lost their proteolytic activity, as judged by the absence of the HEXXH motif in the protease domain. Outside this branch, proteolytically inactive paralogs are almost exclusively found in cyanobacteria and organelles.

Our map shows a clear separation of chloroplastic and mitochondrial FtsH paralogs. The grouping of chloroplast FtsH-like N-domains with the ones from cyanobacteria and the close proximity of N-domains of mitochondrial proteins to those from α-proteobacteria reflect the endosymbiotic origin of the organelles (Fig. 2). Unlike bacteria, in plastids, multiple FtsH paralogs allow the formation of both homooligomeric and heterooligomeric complexes (for review, see Ref. [14]). In plants, the number of paralogs in chloroplasts is much higher than that in mitochondria. Out of the 17 paralogs described for *A. thaliana*, 13 are known to be targeted to the chloroplast [15], 3 to the mitochondria [16], and 1 to both organelles [17]. Even though we only clustered the N-domains, our results are consistent with those of previous studies, which were based on the phylogenetic analysis of full-length sequences [18], and we could reproduce the subgrouping of the closely related copies. As exemplified for *A. thaliana*, the FtsH paralogs formed five groups: AtFtsH1/5, AtFtsH2/6/8, AtFtsH7/9, AtFtsH3/10, and AtFtsH12 (Supplementary Fig. S1). In two of these groups, AtFtsH1/5 and AtFtsH2/8, the N-terminal TM sequence appears to be interpreted as a signal

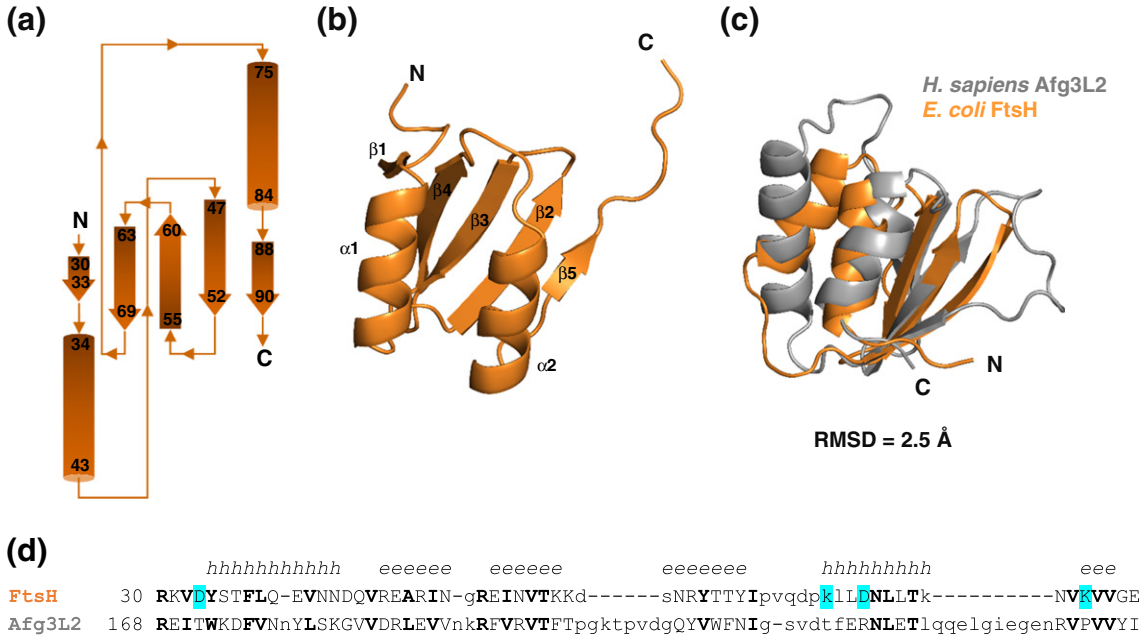


Fig. 3. NMR structure of the *E. coli* FtsH-N monomer. (a) Schematic topology diagram of the N-domain. β-Strands are shown as arrows and α-helices are shown as cylinders. (b) Cartoon representation of the FtsH-N monomer. (c) Superimposition of the *E. coli* FtsH-N (PDB code 2MUY) and the human Afg3L2-N structures (PDB code 2LNA). (d) Structure-based sequence alignment of *E. coli* FtsH-N and *Homo sapiens* Afg3L2-N. Structurally equivalent residues are shown in capital letters and conserved residues are in boldface. The secondary structure is shown above the sequences (h, helix; e, strand). Residues forming the two intersubunit salt bridges, Asp33-Lys87 and Asp79-Lys76, are marked in cyan.

sequence and cleaved off after membrane insertion, leaving the mature proteins with a single TM helix [19]. The cleavage site in these proteins appears to be conserved in cyanobacterial proteins, raising the possibility that this process predates the origin of chloroplasts. Certainly, cyanobacteria contain a separate, paralogous group that only contains a single TM helix, showing that proteins with this topology emerged several times in the evolution of the FtsH group.

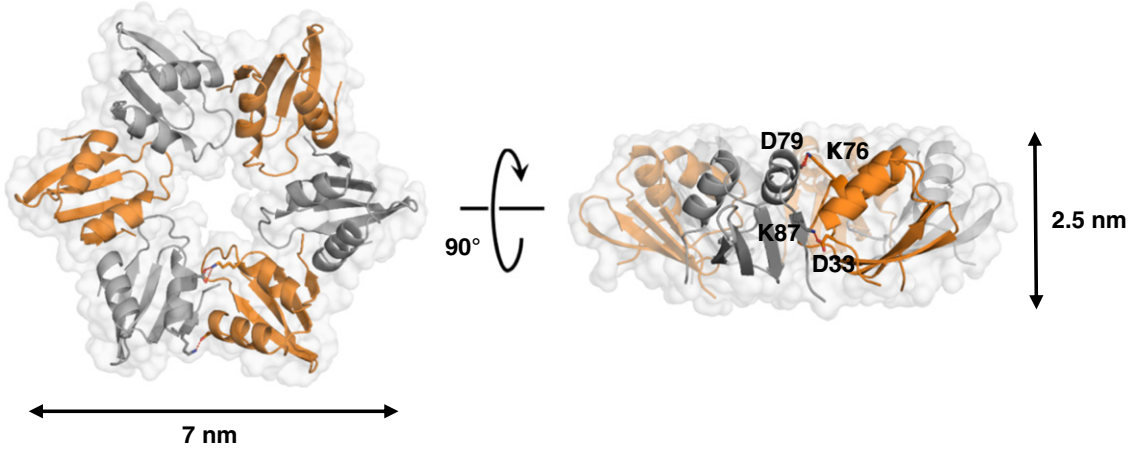


Fig. 4. Crystal structure of the *E. coli* FtsH-N hexamer in cartoon representation, embedded in a transparent surface model (top and side views). Individual subunits are colored orange and gray. The complex has a lateral dimension of approximately 5–7 nm and a height of approximately 2.5 nm. Subunit interfaces are stabilized by hydrophobic interactions and two salt bridges, Asp33-Lys87 and Asp79-Lys76.

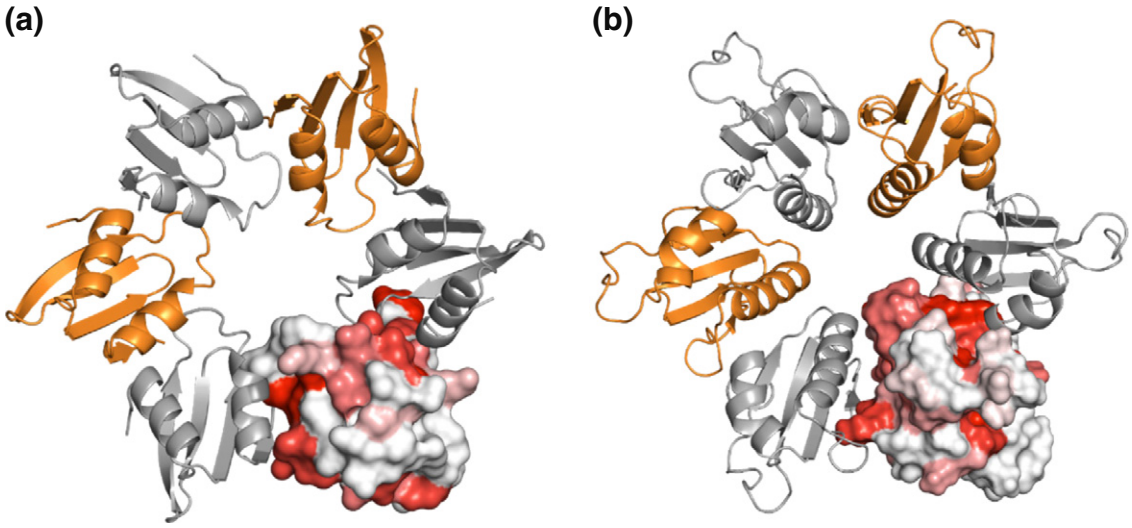


Fig. 5. Amino acid conservation of *E. coli* FtsH and human Afg3L2 N-domains. Surface mapping of evolutionarily conserved residues onto the hexameric structure of *E. coli* FtsH-N is shown in (a) and onto a homology model of human Afg3L2-N is shown in (b). Amino acids are colored by residue conservation: red (high conservation) to white (no conservation). Conservation is especially high at the subunit interfaces.

The NMR structure of *E. coli* FtsH-N

To obtain structural information for the FtsH-like group of N-domains, we expressed and purified the periplasmic region (amino acid residues 25–96) of the FtsH protein from *E. coli*. The structure of this region, solved using NMR, revealed a compact monomeric α + β fold in solution, comprising two α-helices and five β-strands, with a topology of β1-α1-β2-β3-β4-α2-β5 (Fig. 3a). The hydrophobic core of the molecule is formed by the packing of the two helices against the β-strands β2-β5 (Fig. 3b). Soon after we solved our bacterial structure, a representative structure of a eukaryotic FtsH homolog, that of the inner membrane space domain of human mitochondrial protein Afg3L2, was reported {Protein Data Bank (PDB) code 2LNA [20]}. Despite exhibiting a low level of sequence identity, the two domains are structurally similar, with a root-mean-square deviation (RMSD) of 2.5 Å over 51 C^{α} positions (Fig. 3c and d). By comparison to the *E. coli* structure, the loop regions are in general longer in Afg3L2. Nonetheless, these two structures establish the conservation of this fold across bacteria and mitochondria.

A hexameric assembly of *E. coli* FtsH-N

In addition to the solution structure, we obtained a 2.55-Å crystal structure, which was solved by using the NMR structure as a molecular replacement model. The monomer is very similar to the solution structure, with an RMSD of 0.78 Å over 62 C^{α} positions. The asymmetric unit contains three monomers belonging to two hexameric rings. These rings can be constructed by crystallographic symmetry, one from a single monomer by 6-fold symmetry and the other one as a trimer of the other two monomers. The two hexamers are virtually identical, with an RMSD of 0.5 Å over all C^{α} positions, and have a disk-like shape of approximately 5–7 nm diameter and 2.5 nm height (Fig. 4). The interfaces between adjacent monomers are mainly hydrophobic and are further stabilized by the two salt bridges Asp33-Lys87 and Asp79-Lys76.

Using ConSurf [21], we found that regions of high sequence conservation are mainly located at the subunit interfaces (Fig. 5a), supporting the physiological nature of the observed hexamer. However, the two salt bridges that stabilize adjacent monomers in *E. coli* are not conserved. Since hexamer formation was not detectable in solution, it is likely that oligomerization is mainly driven by hydrophobic interactions and that this may need high local protein concentrations, which can be realized *in vitro* in protein crystals and *in vivo* by membrane anchoring.

Like our FtsH-N NMR structure, the solution structure of the human Afg3L2 N-domain was determined as a monomer. To evaluate whether it might also be hexameric *in vivo*, we built a homology model based on the hexameric FtsH-N structure and mapped the sequence conservation of Afg3 proteins onto the model. Again, the conservation is highest at the subunit interfaces (Fig. 5b), suggesting that eukaryotic FtsH-like N-domains also form hexameric rings with the same architecture as their bacterial homologs.

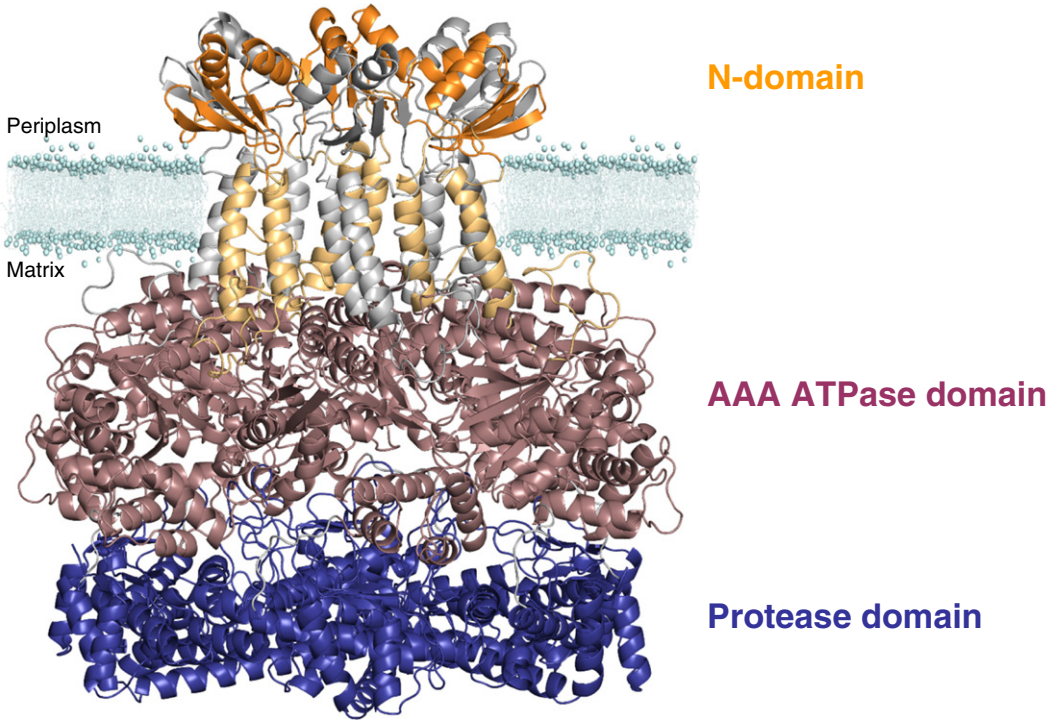


Fig. 6. Structural model of *E. coli* FtsH. The TM helices (shown in gray and orange) were built on the basis of the cryo-EM density map of the yeast m-AAA complex (EMD-1712; see Ref. [22]). The model shows that the α-helices face the periplasm and the β-sheets face the cytoplasmic membrane.

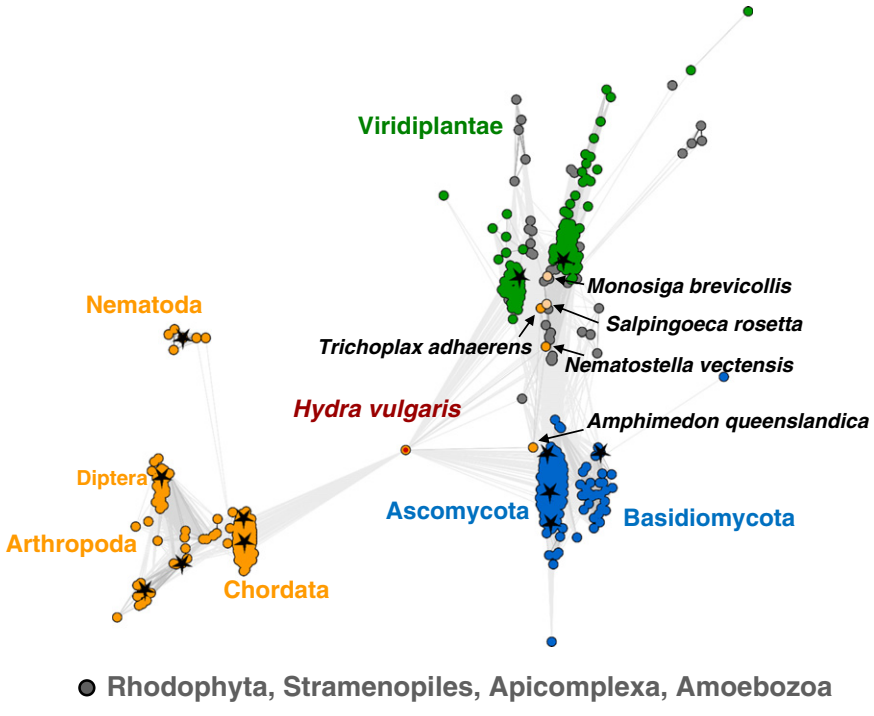


Fig. 7. Cluster map of Yme1-like N-domains. Sequences were clustered at a BLAST *P*-value cutoff of 1e-10. BLAST connections are shown as gray lines; the darker a line, the higher the similarity. Each dot represents one protein; sequences within one group are shown in the same color. Protein sequences included in sequence alignments of Figs. 8 and 9 are denoted by a star.

Fig. 8. Multiple sequence alignment of Yme1 N-domains (TPR fold) of plants, fungi, and basal metazoans. The alignment comprises representatives (Table S2) of the groups marked in the cluster map in Fig. 7. Numbers in parentheses represent the length of the omitted segments. Highly conserved residue columns are shown in boldface. Residues that could not be aligned are shown in lower case. The secondary structure is shown above the sequences (h, helix; e, strand) and TM helices are marked in red.

In order to illustrate the position of the N-domain ring within the AAA metalloprotease complex, we generated a model of *E. coli* FtsH. For this, we used our hexamer, the crystal structure of the cytoplasmic region of *Thermotoga maritima* FtsH (PDB codes 2CE7 and 3KDS), and the cryo-electron microscopy (EM) density map of the yeast m-AAA complex (EMD-1712; see Ref. [22]).

Fig. 9. Multiple sequence alignment of animal Yme1 N-domains. (a) The alignment of Yme1-NN domains of Chordata and *H. vulgaris*. This region appears to have been replaced by new forms in Nematoda and Diptera. (b) The alignment of Yme1-NC domains of Chordata. This domain is found in all animals. In both alignments, the secondary structure is shown above the sequences (h, helix; e, strand) and residues that could not be aligned are shown in lower case.

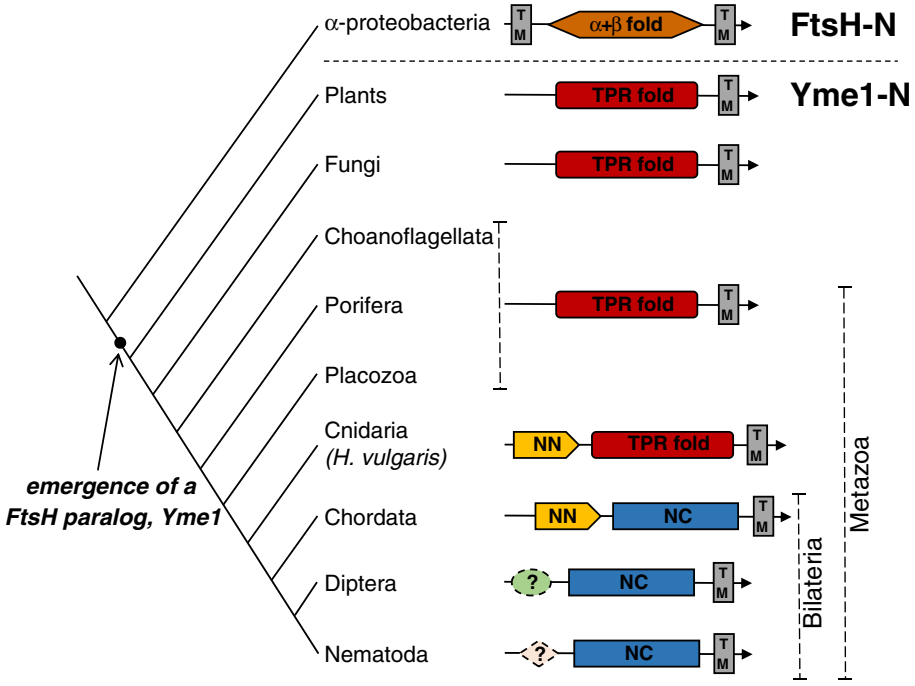


Fig. 10. Evolutionary scenario for the origin of diverse N-domains in AAA metalloproteases. Yme1-like proteins arose in organelles as a paralog of the FtsH-like proteins by replacing the N-terminal TM helix and α + β fold with the TPR fold. Details are described in the text.

As can be seen in Fig. 6, the N-domain ring is oriented such that the β-sheets of the subunits face the cytoplasmic membrane and the α-helices face the periplasm. Overall, the dimensions of this ring are comparable to those of the 12-helical TM domain and considerably smaller than those of the ATPase and protease rings.

Yme1-like N-domains

In the map, the second largest cluster, comprising 619 sequences, is formed by the N-domains of eukaryotic Yme1-like proteins. This cluster consists of two main subgroups, connected by a single sequence. One subgroup is formed by the Yme1 N-domains of plants, fungi, and basal metazoans and the other by those of all other metazoans. The bridging sequence belongs to the cnidarian *Hydra vulgaris* (Fig. 7). This map topology is in general agreement with our previous results, where, however, due to an oversight, we unfortunately misannotated the Yme1-like subgroups [4]. Sequence searches, even with the most sensitive methods such as HHsearch, detect no similarity between the two subgroups, indicating that they are evolutionarily unrelated. While the N-domains of plants, fungi, and basal metazoans make matches to proteins with tetratricopeptide repeat (TPR) folds, the N-domains of the other animals have no homologs of known structure. The N-domain of *H. vulgaris* bridges the two clusters as it comprises both a TPR domain and a region of homology to the animal N-domains (Fig. 7).

In an alignment of representative sequences from the plant/fungal group with basal metazoans, that is, the choanoflagellate *Monosiga brevicollis*, the sponge *Amphimedon queenslandica*, the placozoan *Trichoplax adhaerens*, and the cnidarians *Nematostella vectensis* and *H. vulgaris*, the conservation of the TPR domain is clearly apparent (Fig. 8). We conclude that Yme1-like proteins arose from FtsH-like AAA metalloproteases by the loss of the first TM helix and substitution of the α + β fold with the TPR fold (Fig. 10). The TPR fold represents the basal form of Yme1-like N-domains. At one point prior to the split between Cnidaria and Bilateria, Yme1 proteins appear to have acquired an additional domain preceding the TPR part (Yme1-NN), as still seen today in the N-domain of *Hydra* (Figs. 9a and 10). After separation from the Cnidaria, the Bilateria appear to have lost the TPR part by displacement with yet another domain (Yme1-NC) and this new two-domain structure became the canonical form of N-domains in animal Yme1 (Figs. 9b and 10). Nematoda and Diptera seem to have replaced the Yme1-NN domain with phylum-specific regions (Fig. 10), suggesting that

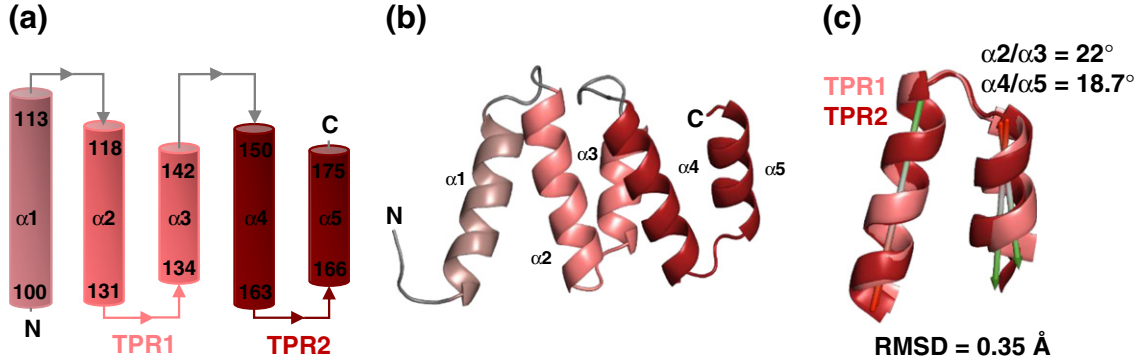


Fig. 11. NMR structure of *S. cerevisiae* Yme1-N. (a) Schematic topology diagram of the N-domain structure. α-Helices are shown as cylinders. (b) Cartoon representation of the NMR structure. α-Helices are shown in a red-colored gradient from the N-terminus to the C-terminus. The helices α2-α3 and α4-α5 form the two TPR hairpins. (c) Superimposition of the TPR1 and TPR2 motifs. The vectors represent the helix axes, which were used for the calculation of the packing angle.

the diversity of AAA metalloprotease N-domains is continually evolving.

NMR structure of *Saccharomyces cerevisiae* Yme1 N-domain

The Yme1 protein of *S. cerevisiae* is thus far the best-studied member of this family. Its N-domain comprises a mitochondrial targeting signal of 49 residues. The remaining 180 residues to the single TM helix correspond to a ~19-kDa domain. Preliminary investigation of this segment by NMR and proteolytic stability tests showed a stable and folded core from residue 97 to residue 177. The N-terminal and C-terminal flanking regions were disordered in solution, which is in agreement with secondary structure predictions, performed using IUPred [23]. We therefore used the folded core with a C-terminal His-tag for further structural analysis. Attempts to crystallize the TPR domain were not successful under any of the tested conditions. NMR studies revealed a fold with five helices (Fig. 11a), of which α2-α3 and α4-α5 form TPR hairpins. The hairpins superimpose with an all-backbone RMSD of 0.35 Å and generate, together with the adjacent helix α1, a right-handed superhelix (Fig. 11b). The packing angle between the individual helices of the two repeats is 22° and 18.7°, respectively, which is in agreement with the average packing angle of 24° observed in canonical TPRs (Fig. 11c; see Ref. [24]). While most plants, fungi, and basal metazoans appear to possess a TPR domain similar to the one seen in yeast, in some fungi and plants, the second repeat has an extended loop inserted between the two helices (Fig. 8).

In solution, we found no biochemical evidence for oligomerization of the complete N-domain (without mitochondrial targeting signal) or of the truncated versions of it. Nonetheless, we consider it possible that the highly conserved segment preceding the TM helix (Fig. 8), which is predicted to form an extended β-strand, hexamerizes to form a funnel at the entrance of the TM pore.

Other AAA metalloprotease N-domains

Over the course of our analysis, we became aware of some peripheral clusters of N-domains that show no sequence similarity to other N-domains in the map (Fig. 1). The proteins in these clusters are paralogs of canonical FtsH proteins, as the organisms in which they occur also contain one or more proteins in the central cluster. In one such peripheral cluster, formed by proteins of Bacillales and Clostridiales, a periplasmic N-domain appears to be missing entirely and the AAA domain is preceded solely by a hairpin formed of two TM helices. In another peripheral cluster, formed of proteins from Myxoccocus, no TM helices are detectable, suggesting the existence of soluble AAA metalloprotease forms. The N-terminal region in these proteins is made up of a degenerate AAA ATPase module.

As an exception to the rule that N-domain divergence occurs only in paralogy to canonical forms, we find that the N-domains of Kinetoplastida are all divergent and no canonical forms are observed. These N-domains are found in four clusters, which are scattered in the periphery of the map (Fig. 1) and show no sequence similarity to each other. All Kinetoplastida have paralogs in at least two of the clusters and none in all four.

Conclusion

Proteins of the AAA family have diverse N-domains, corresponding to their broad set of roles

in cellular processes. Within individual branches, however, proteins share similar functional roles and have the same type of N-domain. AAA metalloproteases are unusual in having recruited distinct, non-homologous N-domains several times in the course of their evolution. We have used the principle of maximum parsimony to retrace the evolution of this AAA subfamily, allowing the fewest possible evolutionary changes at each step. Nonetheless, nature often does take more roundabout evolutionary routes, but with the currently available molecular and functional data on AAA metalloproteases, we cannot infer if their diversity may have arisen through such routes.

The clearly ancestral form, represented in about 90% of these proteins, possesses two membrane-spanning helices that bracket a small periplasmic domain (Fig. 10). This domain has an unusual α + β fold and hexamerizes to form a ring at the entrance to the central pore of the TM domain through which presumably substrates are translocated. After the origin of eukaryotes and the establishment of mitochondria and chloroplasts as organelles, a paralog of the ancestral form emerged, in which the N-terminal TM helix and periplasmic domain were replaced by a new domain with a TPR fold (Fig. 10). This new form, Yme1, had an inverted membrane topology, with the ATPase and protease domains in the intermembrane space rather than in the matrix of the organelles. The availability of ATP, which is absent from the periplasm in bacteria but present in the intermembrane space of organelles, must have facilitated this inversion. The TPR fold, representing the basal form of Yme1-like N-domains, was elaborated by an additional N-terminal domain (Yme1-NN) before the separation of Cnidaria and Bilateria, as exemplified presently by the N-domain of *Hydra*. Subsequently, the Bilateria substituted the TPR part with a further new domain (Yme1-NC) to give rise to the N-domain of other animals (Fig. 10). Within the Bilateria, Nematoda and Diptera appear to have replaced the Yme1-NN domain by new phylum-specific forms. It would be interesting to complete the picture of N-domain diversity by obtaining structures of these animal domains. Ultimately, it is unclear why so many different N-domains have been harnessed despite the function of the FtsH-like and Yme1-like AAA metalloproteases remaining the same: that of protein quality control.

Materials and Methods

Bioinformatics

To compile the set of AAA metalloprotease N-domains, we gathered their full-length sequences and masked out the AAA and M41 domains. To this end, we obtained the sequences of AAA metalloproteases by searching the non-redundant protein database at NCBI using HMMER3 [12] with the profile hidden Markov model (HMM) of the M41 peptidase domain as query. The profile HMM was calculated from the Pfam seed alignment of the M41 domain (PF01434; see Ref. [25]), comprising 31 sequences, using the *hmmbuild* tool from the HMMER3 package [12].The searches were performed in default settings. We pooled full-length sequences for the resulting matches and filtered out sequences annotated as "partial" and sequences shorter than 400 residues. This was performed to exclude incomplete sequences, as the M41 and AAA domains themselves comprise about 450 residues. The resulting set contained 10,733 sequences.

To detect AAA domains in this dataset, we first built a profile HMM of the AAA domain (helix α0-α12) in the M41 seeds. The AAA domains were aligned using Clustal Omega [26] and the profile HMM was derived using *hmmbuild*. Next, an HMMER3 search was seeded with this profile HMM to detect AAA domains in the pooled sequences. All matches with at least 80% coverage of the AAA profile HMM were grouped to obtain the set of AAA metalloproteases. Sequences that failed the coverage criterion but possessed the Walker A and B motifs were also retained. The metalloproteases of the basal metazoans *T. adhaerens* (NCBI GI number 196013470), *N. vectensis* (156407406), and *M. brevicollis* (167520684) were manually curated and were included in the final set of 10,352 AAA metalloproteases (Supplementary Data S1 and Table S1). For these three proteins, the corresponding genome scaffold regions, with additional 2000 base pairs flanking them, were gathered from the genome portal of the Department of Energy Joint Genome Institute [27] and Augustus gobics [28] was used for gene prediction. For all sequences, the AAA domain and amino acids succeeding it were masked out to obtain the set of N-domains.

The obtained N-domains were clustered in CLANS [13] by their pairwise BLAST *P*-values [29]. Clustering was performed to equilibrium in two-dimensional (2D) space at a *P*-value cutoff of 1.0e-10 with default settings, except for attract value = 20 and attract exponent = 2. We built multiple sequence alignments for each of the 10,352 N-domains using the *buildali.pl* script (with default parameters) from the HHsearch package [30]. Profile HMMs were calculated from the alignments using *hhmake*, also from the HHsearch package. We also built profile HMMs for the N-domains of *E. coli* FtsH and yeast Yme1 and compared them to the profile HMMs of all N-domains using HHsearch. Sequences that matched the N-domain of *E. coli* FtsH or of yeast Yme1 with a probability of ≥50% were assigned to the FtsH-like or to the fungal, plant, and basal metazoan Yme1-like group, respectively. The maps shown in Figs. 2 and 7, as well as in Supplementary Fig. S1 were extracted from the map of all N-domains (Fig. 1) and were also clustered at a *P*-value cutoff of 1.0e-10.

The structural alignment shown in Fig. 3c was generated interactively in Swiss-PDB viewer [31] and the corresponding structure-based sequence alignment is shown in Fig. 3d. The multiple sequence alignments shown in Figs. 8 and 9 were generated manually, guided by pairwise alignments obtained from HHsearch. Secondary structure prediction shown in these two figures was calculated using the Quick2D tool from the MPI Bioinformatics Toolkit [32]. The evolutionary conservation analysis shown in Fig. 5 was carried out using ConSurf in default settings, with the "Clean UniProt" as the reference

database [21]. The homology model of the Afg3L2-N hexamer was generated by superimposing the structure of Afg3L2-N monomer (PDB code 2LNA) onto our hexameric assembly of *E. coli* FtsH (PDB code 4V0B) in Swiss-PDB viewer. TM helices were predicted using Phobius [33] and TMHMM [34].

Cloning and expression

The *E. coli* FtsH-N gene, amino acid residues 25–96, was amplified and isolated by PCR from genomic DNA (NCBI GI number 388476123; forward primer: 5′-CATGCCATGG-CAAGCGAGTCTAATGGCCGTAAGGTGGATTAC-3′, reverse primer: 5′-CCGCTCGAGCGGTTCTTCAGGCGGTT CACCGACAACCTT-3′) and was cloned into pET28b for expression of a protein with a C-terminal hexa-histidine tag. For expression, *E. coli* C41(DE3) cells were transformed with the vector.

The gene encoding the Yme1 N-domain of *S. cerevisiae* (amino acids 49–226; NCBI GI number 418575) was purchased in the pUC57 vector (GenScript). Primers were designed for ligation into the pET30a vector for the production of a C-terminal hexa-histidine-tagged fusion protein, comprising residues 97–176. For expression, the plasmid was transformed in the *E. coli* C41(DE3) strain.

For expression of FtsH-N and Yme1-N, respective *E. coli* cultures were grown at 37 °C in LB medium, supplemented with kanamycin (100 μg/ml), induced with 1 mM isopropyl-β-D-thiogalactoside when OD_{600} reached 0.4–0.6, and harvested after 4 h of induction.

For ^{15}N sample labeling and $^{15}N/^{13}C$ sample labeling, *E. coli* cells were grown in M9 minimal medium with $^{15}NH_4Cl$ and ^{13}C uniformly labeled glucose (Eurisotop) as the sole nitrogen and carbon source, respectively.

Purification of FtsH-N

After resuspension in 50 mM Tris–HCl (pH 7.0), 1 mM PMSF (*p*henyl*m*ethyl*s*ulfonyl *f*luoride), and protease inhibitor mix (Serva), cells were lysed by a French press. The soluble fraction of the lysate was loaded onto a QHP anion-exchange column [GE Healthcare; 25 mM Tris–HCl (pH 7.4), 20 mM to 1 M NaCl gradient, and 2% glycerol]. Sample-containing fractions were mainly found in the flow through and were further applied to nickel-nitrilotriacetic acid affinity chromatography [20 mM Tris–HCl (pH 7.4), 200 mM NaCl, and 0–0.5 M imidazole gradient]. The purified protein was concentrated by a flow filtration system (Amicon Ultra, Millipore) and dialyzed against 15 mM Mops (pH 7.2) and 75 mM NaCl for crystallization and against phosphate-buffered saline for NMR studies.

Purification of Yme1-N

Cell pellets were resuspended in 20 mM Tris–HCl (pH 7.9), 30 mM NaCl, 4 mM $MgCl_2$, 1 mM PMSF, and protease inhibitor mix (Serva) and were lysed by a French press. Soluble fractions of the sample were subjected to nickel-nitrilotriacetic acid affinity chromatography [20 mM Tris–HCl (pH 7.9), 300 mM NaCl, and 0–0.5 M imidazole gradient]. Further purification of the sample was achieved by Superdex 75 gel filtration [GE Healthcare; 0.1 M $NaHCO_3$ (pH 8.6)]. The Amicon Ultra flow filtration system (Millipore) was used to concentrate the protein for structural analyses.

NMR structure determination

For FtsH-N, spectra were recorded at 298 K on Bruker spectrometers at 600, 750, or 900 MHz. Backbone sequential assignments were made using standard triple-resonance experiments. An HNHA experiment was used to derive $^3J_{HNHA}$ coupling constants and an HNHB experiment was acquired to assist in rotamer and stereospecific assignments. For Yme1-N, spectra were recorded at 298 K on Bruker spectrometers at 600 or 800 MHz. Backbone sequential assignment was performed using a strategy based on three-dimensional (3D) HN(CA)NNH [35] and HNCA spectra. HNHA and HNHB spectra were acquired as for FtsH, but these were combined with a 3D HA[HB,HN](CACO)NH spectrum [36], both to resolve any ambiguities in sequential assignment and to provide more definitive rotamer and stereospecific assignment. For both proteins, assignment of aliphatic side chains was completed using standard ^{13}C-based total correlated spectroscopy spectra and assignment of aromatic side chains could be largely completed using contacts in a 2D nuclear Overhauser enhancement spectroscopy (NOESY) spectra.

Structure calculations were based on distance data derived from 3D ^{15}N heteronuclear single quantum coherence NOESY and 3D NNH-NOESY spectra acquired on ^{15}N-labeled samples, as well as 3D ^{13}C heteronuclear single quantum coherence NOESY and 3D CCH-NOESY and 3D CNH-NOESY spectra [37] on a $^{15}N,^{13}C$-labeled sample. For Yme1-N, aromatic contacts were observed in a ^{12}C-filtered 2D NOESY spectrum acquired on ^{15}N-labeled sample. Structural restraints were compiled using a protocol aimed at high local definition whereby expectation NOESY spectra are used to test local conformational hypotheses (in-house software). Chemical shift similarity searches using the TALOS+ server [38] were used to generate hypotheses for backbone conformations, while side-chain rotamers not defined during the process of stereospecific assignment (e.g., χ1/χ2 for leucine and isoleucine) were searched exhaustively. Conformations identified in this manner were applied via dihedral restraints, using the TALOS-derived tolerances for backbone and ±30° for side chains. Further nuclear Overhauser enhancement contacts were assigned iteratively using back-calculation of expectation NOESY spectra from preliminary structures.

Structures were calculated with Xplor (NIH version 2.9.4) using a three-stage simulated annealing protocol based on standard scripts. A first stage calculated raw simulated annealing structures based on all experimental data. Subsequent stages were used to apply a conformational database potential and to relax potentials specifying covalent geometry (e.g., planarity of the peptide bond). The force field used was modified to allow hydrogen bond restraints via pseudo-covalent bonds. These were applied for amide protons in secondary structure where water exchange rates were low and where hydrogen bond acceptors were consistently identified in preliminary calculations. Sets of 100 structures were calculated and a subset was chosen on the basis of lowest restraint violations (19 structures for FtsH-N and 22 for Yme1-N). An average structure was calculated and regularized to

give a structure representative of the ensemble. Tables of solution structure statistics for the two structures are presented in Tables S3 and S4.

X-ray crystallography

Crystallization trials of FtsH-N were performed with a protein concentration of 10 mg/ml in 15 mM Mops (pH 7.2) and 75 mM NaCl. Hanging drops were prepared with each 1 ml of protein and reservoir solution and were equilibrated against 500 µl reservoir solution at 297 K. Best-diffracting crystals grew with a reservoir solution containing 0.1 M Hepes (pH 7.5) and 2 M ammonium sulfate. Data were collected at 100 K and at a wavelength of 0.976 Å at beamline X10SA of the Swiss Light Source on a MAR225 detector (Mar Research). The best dataset was processed and scaled to a resolution of 2.55 Å in space group *P*6 using XDS [39]. Molecular replacement was carried out with MOLREP [40] and the monomeric NMR structure as a search model. Three copies were located in the asymmetric unit, which belong to two hexameric rings that are built by crystallographic symmetry. After initial rigid-body refinement using REFMAC5 [41], the model was completed by cyclic manual modeling with Coot [42] and refinement with PHENIX [43]. Data collection and refinement statistics are summarized in Table S5. The structure was deposited in the PDB under accession code 4V0B.

Atomic structure fit

A model of *E. coli* FtsH was built using our experimental structure of the N-domain hexamer and a homology model of the AAA ATPase and protease domains based on the crystal structures of *T. maritima* built with Modeller [44]. These structures and models were fitted manually into the cryo-EM density map of FtsH. To model the TM domain, we first sharpened the cryo-EM map using a non-negative deconvolution algorithm [45]. The sharpened density map clearly shows 12 regularly arranged, rod-like densities forming six-membered inner and outer rings into which the TM helices were placed. The structure of entire FtsH was obtained by connecting the modeled parts based on evolutionary and biochemical considerations. The N-terminal TM helices form the inner ring of the TM domain followed by the periplasmic N-domain ring. The outer TM ring is formed by the TM helices C-terminal to the N-domain. These helices then extend into the AAA ATPase domain. The structure of a single FtsH monomer was refined using ISD [46] by fitting it flexibly into the cryo-EM map assuming *C*6 symmetry. In addition to the density fitting score, a purely repulsive non-bonded force field was imposed to solve van der Waals clashes. Note that the cryo-EM map contains six additional areas of knob-shaped density, which appear docked to the N-domain ring and remain unaccounted for by our model.

Accession codes

Coordinates and structure factors for the FtsH-N structures and Yme1-N have been deposited in the PDB (the PDB accession code of the FtsH-N NMR structure is 2MUY, that of the hexameric X-ray structure is 4V0B, and that of Yme1-N is 2MV3).

Acknowledgements

We are grateful to Ines Wanke and Iris Asen for crystallographic sample preparation and data collection, as well as to the staff of beamline X10SA/Swiss Light Source for their continuous support. The work was supported by institutional funds from the Max Planck Society.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2014.12.024.

Received 1 October 2014;
Received in revised form 1 December 2014;
Accepted 29 December 2014
Available online 8 January 2015

Keywords:
AAA proteins;
FtsH;
Yme1;
m-AAA;
i-AAA

Present address: M. Habeck, Institute for Mathematical Stochastics, Georg August University Göttingen, 37073 Göttingen, Germany.

Abbreviations used:
TPR, tetratricopeptide repeat; TM, transmembrane; 2D, two dimensional; 3D, three-dimensional; PDB, Protein Data Bank; EM, electron microscopy; NOESY, nuclear Overhauser enhancement spectroscopy; NCBI, National Center for Biotechnology Information; HMM, hidden Markov model.

References

[1] Erdmann R, Wiebel FF, Flessau A, Rytka J, Beyer A, Fröhlich KU, et al. PAS1, a yeast gene required for peroxisome biogenesis, encodes a member of a novel family of putative ATPases. Cell 1991;64:499–510.
[2] Ammelburg M, Frickey T, Lupas AN. Classification of AAA+ proteins. J Struct Biol 2006;156:2–11.
[3] Lupas AN, Martin J. AAA proteins. Curr Opin Struct Biol 2002;12:746–53.
[4] Frickey T, Lupas AN. Phylogenetic analysis of AAA proteins. J Struct Biol 2004;146:2–10.
[5] Tomoyasu T, Gamer J, Bukau B, Kanemori M, Mori H, Rutman AJ, et al. *Escherichia coli* FtsH is a membrane-bound, ATP-dependent protease which degrades the heat-shock transcription factor sigma 32. EMBO J 1995;14:2551–60.

[6] Akiyama Y, Kihara A, Tokuda H, Ito K. FtsH (HflB) is an ATP-dependent protease selectively acting on SecY and some other membrane proteins. J Biol Chem 1996;271: 31196–201.
[7] Kihara A, Akiyama Y, Ito K. Host regulation of lysogenic decision in bacteriophage lambda: transmembrane modulation of FtsH (HflB), the cII degrading protease, by HflKC (HflA). Proc Natl Acad Sci USA 1997;94:5544–9.
[8] Akiyama Y, Kihara A, Mori H, Ogura T, Ito K. Roles of the periplasmic domain of *Escherichia coli* FtsH (HflB) in protein interactions and activity modulation. J Biol Chem 1998;273: 22326–33.
[9] Leonhard K, Herrmann JM, Stuart RA, Mannhaupt G, Neupert W, Langer T. AAA proteases with catalytic sites on opposite membrane surfaces comprise a proteolytic system for the ATP-dependent degradation of inner membrane proteins in mitochondria. EMBO J 1996;15: 4218–29.
[10] Mann NH, Novac N, Mullineaux CW, Newman J, Bailey S, Robinson C. Involvement of an FtsH homologue in the assembly of functional photosystem I in the cyanobacterium *Synechocystis* sp. PCC 6803. FEBS Lett 2000;479: 72–7.
[11] Wagner R, Aigner H, Funk C. FtsH proteases located in the plant chloroplast. Physiol Plant 2012;145:203–14.
[12] Finn RD, Clements J, Eddy SR. HMMER Web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39:W29–37.
[13] Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 2004;20:3702–4.
[14] Janska H, Kwasniak M, Szczepanowska J. Protein quality control in organelles—AAA/FtsH story. Biochim Biophys Acta 1833;2013:381–7.
[15] Ferro M, Brugière S, Salvi D, Seigneurin-Berny D, Court M, Moyet L, et al. AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins. Mol Cell Proteomics 2010;9: 1063–84.
[16] Janska H. ATP-dependent proteases in plant mitochondria. Physiol Plant 2005;123:399–405.
[17] Urantowka A, Knorpp C, Olczak T, Kolodziejczak M, Janska H. Plant mitochondria contain at least two i-AAA-like complexes. Plant Mol Biol 2005;59:239–52.
[18] Yu F, Park S, Rodermel SR. The *Arabidopsis* FtsH metalloprotease gene family: interchangeability of subunits in chloroplast oligomeric complexes. Plant J 2004;37:864–76.
[19] Rodrigues Ricardo AO, Silva-Filho MC, Cline K. FtsH2 and FtsH5: two homologous subunits use different integration mechanisms leading to the same thylakoid multimeric complex. Plant J 2011;65:600–9.
[20] Ramelot TA, Yang Y, Sahu ID, Lee H, Xiao R, Lorigan GA, et al. NMR structure and MD simulations of the AAA protease intermembrane space domain indicates peripheral membrane localization within the hexaoligomer. FEBS Lett 2013;587: 3522–8.
[21] Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, et al. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 2005;33:W299–302.
[22] Lee S, Augustin S, Tatsuta T, Gerdes F, Langer T, Tsai FT. Electron cryomicroscopy structure of a membrane-anchored mitochondrial AAA protease. J Biol Chem 2011;286: 4404–11.
[23] Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–4.
[24] D'Andrea LD, Regan L. TPR proteins: the versatile helix. Trends Biochem Sci 2003;28:655–62.
[25] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res 2012;40:D290–301.
[26] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.
[27] Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res 2014;42:D26–31.
[28] Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res 2006;34:W435–9.
[29] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–402.
[30] Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244–8.
[31] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PDB viewer. Electrophoresis 1997;18:2714–23.
[32] Biegert A, Mayer C, Remmert M, Söding J, Lupas A. The MPI Toolkit for protein sequence analysis. Nucleic Acids Res 2006;34:W335–9.
[33] Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004; 338:1027–36.
[34] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–80.
[35] Weisemann R, Rüterjans H, Bermel W. 3D triple-resonance NMR techniques for the sequential assignment of NH and ^{15}N resonances in ^{15}N- and ^{13}C-labelled proteins. J Biomol NMR 1993;3:113–20.
[36] Löhr F. Simultaneous measurement of $^{3}J_{HN,H\alpha}$ and $^{3}J_{H\alpha,H\beta}$ coupling constants in ^{13}C,^{15}N-labeled proteins. J Am Chem Soc 1999;121:11821–6.
[37] Diercks T, Coles M, Kessler H. An efficient strategy for assignment of cross-peaks in 3D heteronuclear NOESY experiments. J Biomol NMR 1999;15:177–80.
[38] Shen Y, Delaglio F, Cornilescu G, Bax A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 2009;44: 213–23.
[39] Kabsch W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. Appl Crystallogr Online 1993;26:795–800.
[40] Vagin A, Teplyakov A. An approach to multi-copy search in molecular replacement. Acta Crystallogr 2000;56:1622–4.
[41] Murshudov GN, Vagin AA, Lebedev A, Wilson KS, Dodson EJ. Efficient anisotropic refinement of macromolecular structures using FFT. Acta Crystallogr 1999;55:247–55.
[42] Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60:2126–32.

[43] Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX. Acta Crystallogr D Biol Crystallogr 2010;66:213–21.
[44] Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
[45] Hirsch M, Schölkopf B, Habeck M. A blind deconvolution approach for improving the resolution of cryo-EM density maps. J Comput Biol 2011;18:335–46.
[46] Rieping W, Habeck M, Nilges M. Inferential structure determination. Science 2005;309:303–6.