Protein Science (1997), 6:2043–2058. Cambridge University Press. Printed in the USA.

REVIEW

Sequence analysis of the AAA protein family

ANDREAS BEYER

Institut für Physiologische Chemie, Medizinische Fakultät, Ruhr-Universität, D-44780 Bochum, Germany

(RECEIVED March 12, 1997; ACCEPTED June 27, 1997)

Abstract

The AAA protein family, a recently recognized group of Walker-type ATPases, has been subjected to an extensive sequence analysis. Multiple sequence alignments revealed the existence of a region of sequence similarity, the so-called AAA cassette. The borders of this cassette were localized and within it, three boxes of a high degree of conservation were identified. Two of these boxes could be assigned to substantial parts of the ATP binding site (namely, to Walker motifs A and B); the third may be a portion of the catalytic center. Phylogenetic trees were calculated to obtain insights into the evolutionary history of the family. Subfamilies with varying degrees of intra-relatedness could be discriminated; these relationships are also supported by analysis of sequences outside the canonical AAA boxes: within the cassette are regions that are strongly conserved within each subfamily, whereas little or even no similarity between different subfamilies can be observed. These regions are well suited to define fingerprints for subfamilies. A secondary structure prediction utilizing all available sequence information was performed and the result was fitted to the general 3D structure of a Walker A/GTPase. The agreement was unexpectedly high and strongly supports the conclusion that the AAA family belongs to the Walker superfamily of A/GTPases.

Keywords: AAA family; ATPase; evolution; secondary structure prediction; Walker family

The AAA protein family ("triple-A family") is a group of ATPases linked by common ancestry. This family is characterized by the presence of one or two copies of a conserved region of 220–250 amino acids, the so-called AAA cassette (Kunau et al., 1993). Some of the AAA proteins consist almost solely of AAA cassette(s), while others contain additional parts in their primary structure—for example, Leu-Zippers and metal-dependent protease domains have been identified in the AAA subfamilies SF 5 and SF 6, respectively (see Table 1). The AAA family does not stand isolated; it belongs to the Walker superfamily of A/GTPases (Walker et al., 1982; Schulz, 1992). Two sequence motifs typical of those proteins, the so-called Walker box A (which is identical to the P-loop) and box B (Saraste et al., 1990; Smith & Rayment, 1996) can be identified in all AAA proteins. And indeed, ATPase activity has been demonstrated and characterized for some members of this family (Peters et al., 1992; Morgan et al., 1994; Whiteheart, 1994; Fröhlich et al., 1995).

AAA proteins are involved in a great variety of processes (see Table 1) and apparently share no common biochemical feature. Very little is known about the molecular function of the AAA cassette yet. As far as can be judged from available data AAA proteins perform a role that may be characterized as being regulatory or biogenetic (Table 1). The AAA cassette has been suggested, amongst other proposals, to harbor protease activity (Dubiel et al., 1992), or to be involved in DNA/RNA unwinding (Makino et al., 1996), but in fact, none of these hypotheses has been substantiated experimentally (Confalonieri & Duguet, 1995). From the viewpoint of sequence analysis, it is only possible to predict ATPase activity on the basis of the conserved Walker motifs. The apparent biochemical diversity prompted Kunau et al. (1993) to suggest the name AAA family for "ATPases associated with *diverse* cellular activities." Some other authors use the acronym CAD family for "conserved ATPase domain" (Choi et al., 1996; Sun et al., 1996).

Reprint requests to: Andreas Beyer, Institut für physiologische Chemie, medizinische Fakultät, Geb. MA2/40, Ruhr-Universität, D-44780 Bochum, Germany. e-mail: andreas.beyer@ruhr-uni-bochum.de.

Abbreviations: A.thal: Arabidopsis thaliana (thale cress); *Cap.an: Capsicum annuum* (bell pepper); *C.eleg: Caenorhabditis elegans* (nematode worm); clpl: chloroplast; *Dro.mel: Drosophila melanogaster* (fruit fly); *E.coli: Escherichia coli* (proteobacteria; gamma subdivision); *Hal.sal: Halobacterium salinarium* (archaea; halobacteriales); *Lac.lac: Lactococcus lactis* (gram-positive eubacteria); *Mar.po: Marchantia polymorpha* (liverwort); *Methb.therm: Methanobacterium thermoautotrophicum* and *Methc.jan: Methanococcus jannaschii* (both archaea; methanococcales); *Myc.lepr: Mycobacterium leprae* and *Myc.tub: Mycobacterium tuberculosis* (both actinomycetes; mycobacteria); *Mypl.gen: Mycoplasma genitalum* (Low G/C gram-positive bacterium); *Od.sin: Odontella sinensis* (diatom); *Par.b* C Vir: *Paramecium bursaria* Chlorella virus; *S.cer: Saccharomyces cerevisiae* (yeast); *Sch.mans: Schistosoma mansoni* (human blood fluke); *S.pombe: Schizosaccharomyces pombe* (yeast); *Sul.ac: Sulfolobus acidocaldarius* (thermo-acidophile archaebacterium); *Syn.sp: Synechocystis* spec. (cyanobacterium); RubPCOAct: Ribulose 1,5 bisphosphate carboxylase/oxygenase activase. SF: Subfamily, EST: expressed sequence tag.

Table 1. *Compilation of all AAA members known so far*[a]

Gene/protein	Organism	Comment
SF 1: Two conserved AAA cassettes (archaea and eucarya)		
YLL034c	*S.cer*;	
Afg2p	*S.cer*;	
T12664	*Zea mays*;	
ESTs	Human, mouse	
SAV	*Sul.ac*;	Cytosolic hexamer,
vat	*Thermoplasma acidophilum*;	Mg^{++}-dependent
cdcH	*Hal.sal*	ATPase, involved in
MJ1156	*Methc.jan*;	homotypic vesicle
AA113717	*Pyrococcus furiosus*	fusion in the
CDC48	*Plasmodium falciparum*,	endomembrane system
CDC48	*A.thal, Glycine max*;	(Confalonieri & Duguet,
Cdc48p	*S.cer*;	1995, and refs. therein)
VCP-1/VCP-2	*C.eleg*;	VCP-1 and VCP-2 are
p97	*Xenopus laevis*;	very similar; obviously
VCP	Mouse, pig, human;	products of recent gene
TER-ATPase	rat;	duplication
ESTs	from many organisms	
T20802	*A.thal*;	ESTs related either to Afg2p or
D15969	Rice;	YLL034c (exact assignment
H62842;H62841	Human;	impossible, yet not enough
C09983	*C.eleg*	sequence information available)
K04G2.3	*C.eleg*	Middle part of cassette 1 aberrant, no orthologue in yeast existent
orf in AD000014 = Z84724	*Myc.tub* (AD000014 contains frameshifts due to sequencing errors!)	Two cassettes; first one poorly conserved even in regions D, F, and H. May have been acquired by horizontal gene transfer from an animal host (see Discussion: phylogenetic history of the AAA family).
SF 2: Biogenesis of peroxisomes—two AAA cassettes; second one conserved (eucarya)		
Pas1p	*S.cer, Candida tropicalis*,	Biogenesis of microbodies,
Pas1p	*C.maltosa, Pichia pastoris*;	first AAA cassette weakly
C11H1.6	*C.eleg*;	conserved, Walker motifs present.
[Now all named "PEX1" (Distel et al., 1996)]		The *C.elegans* protein corresponds only to Pas1p second AAA cassette plus C-terminus
ESTs/fragments	Human, mouse, *A.thal*	
Pas8p	*S.cer*	Biogenesis of microbodies,
Pay4p	*Yarrowia lipolytica*	first AAA cassette weakly
Pas5p	*Pichia pastoris*	conserved with non-
paf-2	Rat	canonical Walker motifs.
PXAAA1	Human	Proteins localized in
X70791	*Zea mays* genom.fragm.	the cytosol.
[now all named "PEX6" (Distel et al., 1996)]		
Sequence similarity between PEX1 and PEX6 can be demonstrated at their C-terminus (see alignments in electronic appendix)		
SFs 3+4: Sf 4: membrane fusion—two AAA cassettes; first one conserved (eucarya)		
Yta7p	*S.cer*	
F11A10.1	*C.eleg*	
Sec18p	*S.cer, Candida albicans*	Heterotypic vesicle fusion
NSF	Tobacco	in the endomembrane
D25240	Rice	system, receptors are
ZK1014.1	*C.eleg*	known (Confalonieri &
NSF-1/NSF-2	*Dro.mel*	Duguet, 1995, and refs
SKD2	Mouse	therein)
NSF	Hamster and human	NSF-1 and NSF-2 are very similar; obviously products of recent gene duplications.
SF 5: Subunits of the 26S Proteasome—one AAA cassette; (archaea and eucarya)		
MJ1176	*Methc.jan*	
Yta1p	*S.cer*	Assignment to certain
LeMA-1	*Lycopersicon esculentum*	eucaryotic proteasomal
YTA1	Rice	subunit not yet available
TBP-1	Different mammals	
TbpM	*Plasmodium falciparum*	Yta3p/Cim5p (and
Yta3p=Cim5p	*S.cer*	orthologues) correspond
C52E4.4	*C.eleg*	to eucaryotic proteasomal
MSS1	*Xenopus laevis*,	subunit 7
MSS1	Different mammals	
Sug1p=Tby-1p	*S.cer*	Sug1p (and orthologues)
let1+	*S.pombe*	correspond to eucaryotic
DdTBP10	*Dictyostelium discoideum*	proteasomal subunit 8
"Tbp homologue"	*Naegleria fowleri*	
18–56 protein	*Manduca sexta*	
SUG1	*Xenopus laevis*,	
trip1	Different mammals	
DdTBP2	*Dictyostelium discoideum*	Yta2p/Ynt1p (and
Yta2p=Ynt1p	*S.cer*	orthologues) correspond to
POTATP1	Potato	eucaryotic proteasomal
POTATP/YTA2	Spinach	subunit 6
DEAD-box-ATPase	*Manduca sexta*	
F23F12.6	*C.eleg*	
TBP7	Different mammals	
Yta5p	*S.cer*	Yta5p (and orthologues)
mts2+	*S.pombe*	correspond to eucaryotic
Subunit 4 of the 26S proteasome	*Dro.mel*, chicken, rice, different mammals	proteasomal subunit 4
Sug2p	*S.cer*	Assignment to certain
CADp44	*Spermophilus tri-decemlineatus*	eucaryotic proteasomal subunit not yet available
p42	Human	
The 20S proteasome (= multicatalytic protease) has various catalytic activities. Together with the 19S cap complex it forms the 26S proteasome, which is responsible for ATP-dependent hydrolysis of ubiquitine-tagged proteins (Rechsteiner et al., 1993). SF 5 members are present in the cap structure. Proteasomal subunits S6, S8, SUG2 as well as TBP-1 contain a Leu-zipper N-terminally adjacent to the AAA cassette		
SF 6: Metal-dependent proteases—one AAA cassette (eubacteria, mitochondria, cyanobacteria, chloroplasts)		
MG39732	*Mypl.gen*	This subgroup contains both
FtsH=mrsC=HflB	*E.coli, Haemophilus influenzae*	eubacterial and eukaryotic members. However, since
HpFtsH	*Helicobacter pylori*	the latter proteins are
tma	*Lac.lac, Bacillus subtilis*	located either in mito-
slr0228, slr1604, slr1390, sll1463	Four paralogues in *Syn.sp*	chondria or in chloroplasts the eukaryotic genes most
ycf25=orf644	*Od.sin clpl*	probably root back to genes
"ATPase"	*Cap.an*	brought from the endosym-
FtsH-protease	*A.thal*	bionts which gave rise to
PFTF	*Cap.an*	the organelles. Determina-
Yta10p=Afg3p =Yph1p	*S.cer*	tion of orthology is very complicated here.
Yta12p=Rca1p	*S.cer*	Yta12p and Yta10p are very similar; obviously products of recent gene duplications.
YME1p=Yta11p =Osd1p	*S.cer*	
A15	*Sch.mans*,	
M03C11.5	*C.eleg*	
ESTs	Human	
ESTs	From many organisms	Unambiguous assignment often impossible
SF 6 proteins contain N-terminal transmembrane-helices and C-terminal Zn-dependent protease domains. Bacterial members have been shown to degrade transcriptional regulator proteins.		

(*continued*)

Table 1. *Continued*

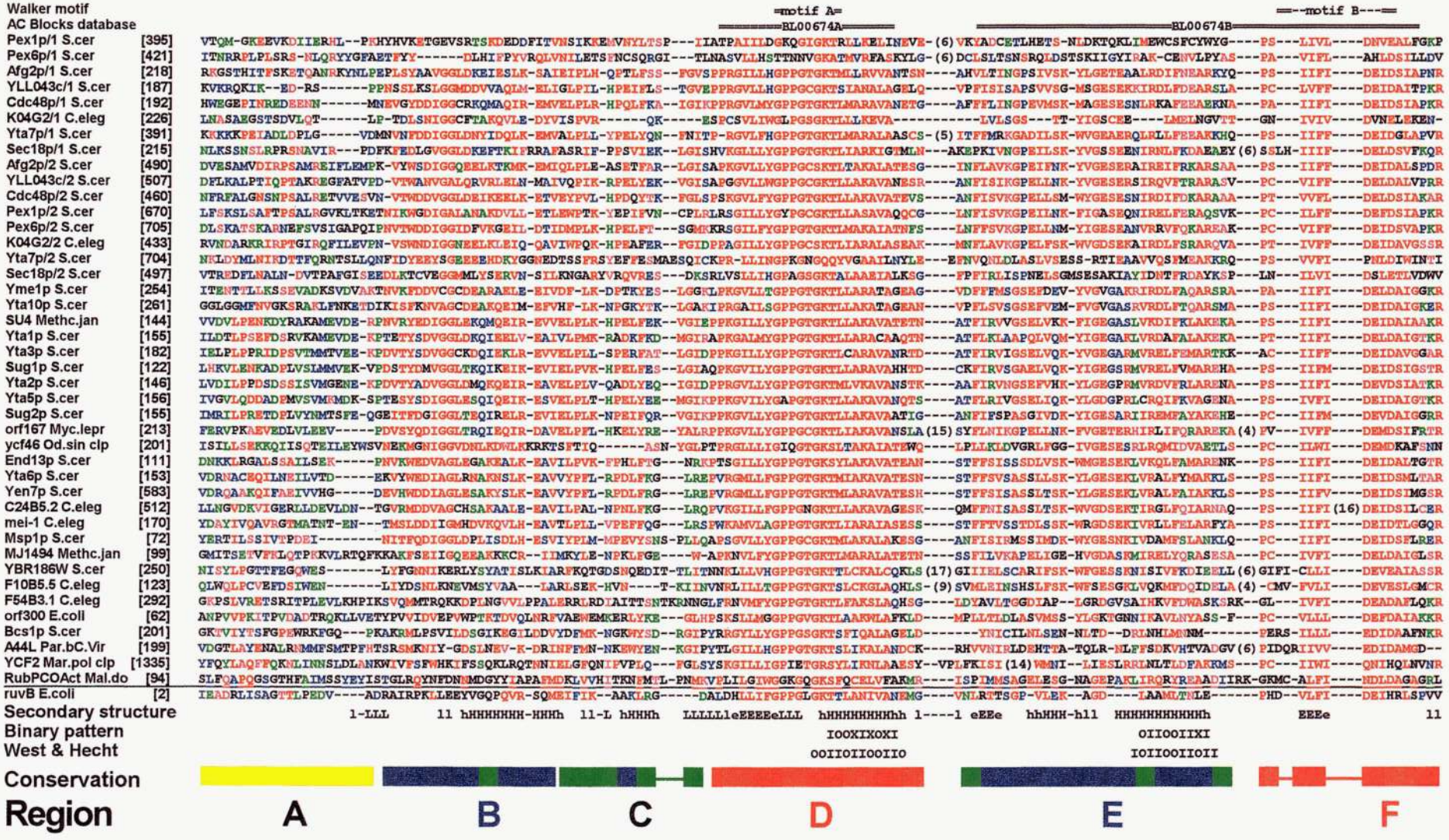
Gene/protein	Organism	Comment
SF 7: Functionally heterogeneous group—one AAA cassette (eucarya)		
End13p=Vsp4p Suppressor Protein SKD1 ESTs	*S.cer* *S.pombe* Mouse; Human and *A.thal*	
Msp1p=Yta4p K04D7.2 DM19DC4Z ESTs	*S.cer* *C.eleg* *Dro.mel* Human	
Yta6p/Yen7p =Sap1p	Both *S.cer*	Both proteins are very similar within their AAA casssettes; obviously products of a more recent duplication event?
C24B5.2 ESTs	*C.eleg* Human	Orthologous to Yta6p/Yen7p?
mei-1	*C.eleg*	Meiotic spindle formation
ESTs	From many organisms	Unambiguous assignment often impossible
SF 8: One AAA cassette (archaea)		
orf1 MJ1494	*Methb.therm* *Methc.jan*	Maybe linked to SF 7. MJ1494 has been labeled as proteasomal subunit 8 (unpublished, see annotations in EMBL database)
SF 9: One AAA cassette (cyanobacteria and chloroplasts)		
slr0374=orf376 slr0480 ycf46 orf491 Z35718 Z47179	Two paralogues in *Syn.sp* *Od.sin* clpI *Porphyra purpurea* clpI *Olisthodiscus luteus* fragm. *Calothrix* D253 fragm.	This subfamily contains members from plant chloroplasts and cyanobacteria. However, since the latter proteins are located in chloroplasts the eukaryotic genes most probably root back to genes brought from the endosymbiont which gave rise to the organelle.
SF 10: One AAA cassette (so far only in Mycobacterium—eubacterium)		
orf C1_167	*Myc.lepr*	Unknown function
SF 11: One AAA cassette (eucarya)		
YBR186W F10B5.5	*S.cer* *C.eleg*	Unknown function
SF 12: One AAA cassette (eucarya)		
F54B3.1/.3 D40223, D40410	*C.eleg* Rice ESTs	Unknown function
SF 13: Assembly of respiratory chain protein—one AAA cassette (eucarya)		
Bcs1p F54C9.6 ESTs	*S.cer* *C.eleg* Different mammals, *A.thal*	Essential for assembly of Rieske Iron Sulfur Protein (subunit of the bc1 complex of the respiratory chain). A role as chaperone has been proposed (Nobrega et al., 1992). There are two paraloguous genes in plants.
SF 14: One AAA cassette—(so far only in *E.coli*–eubacterium)		
orf 300	*E.coli*	Unknown function
SF 15: One AAA cassette (viral)		
A44L	*Parb* C Vir	First known viral AAA sequence
SF 16: Ribulose 1,5 bisphosphate carboxylase/oxygenase activase—one AAA cassette (cyanobacteria & chloroplasts)		
RubPCOAct; (*rca-genes*)	Chloroplasts of many plants and algae, several cyanobacteria	Activates Ribulose 1,5 bisphosphate carboxylase/oxygenase by phosphorylation.
SF 17: YCF2 = orf 2280 proteins—one AAA cassette (chloroplasts)		
YCF2; YCFX, orf2280	Chloroplasts of green plants	Unknown function (Wolfe 1994).
ESTs that cannot be assigned to a distinct SF		
Z34761; Z34739	*A.thal*	
F14313	*A.thal*	
T45469	*A.thal*	

[a]Protein names are in plain text. If the proteins have not been named yet the accession numbers of the respective ESTs (expressed sequence tags) or genomic fragments are listed (underlined) or the names of the open reading frames are given (doubly underlined). Orthologues are listed in one column (exception: SF 6, where orthology is difficult to determine). Some ESTs are mentioned only in summary because of their high number. SFs 11–17 constitute distant members of the AAA family.[2] A detailed list of AAA members, files containing the figures, and detailed alignments can be obtained from the author via e-mail. A continuously updated listing of AAA members as well as a (reduced) phylogenetic tree is also offered by Kai-Uwe Fröhlich and is available in the internet (http://yeamob.pci.chemie.uni-tuebingen.de/AAA/CloseNFar.html). The last search was performed in February 1997.

The family has been growing fast. The first members described were Sec18p from yeast and its mammalian orthologue NSF (Eakle et al., 1988; Wilson et al., 1989, respectively). Nine years later, a data base search using a AAA cassette as query sequence yields more than 700 authentic AAA matches (including expressed sequence tags) belonging to more than 200 individual sequences. Representatives from all kingdoms are known: five (non-orthologous) proteins in eubacteria, three in archaebacteria, and 22 in the yeast *Saccharomyces cerevisiae* alone.[1] Due to the extensive work of Feldmann and co-workers (Schnall et al., 1994), only one member in the meanwhile completely sequenced genome of *S.cer* had been overlooked. Interestingly, while pairs of orthologues in archaea and eukarya can be determined, their relationships to the eubacterial members remain unclear. Not only for this reason it is worth investigating the phylogeny of the AAA family. This theoretical approach might also contribute to the elucidation of the development of biogenesis of cellular structures in the different kingdoms. In which processes are AAA proteins involved in archaea and eukarya? Are there analogous functions in eubacteria, and which proteins do perform them?

In general, the degree of sequence conservation in genes reflects the importance of their biological functions (Wilson et al., 1977; Kimura, 1987; MacIntyre, 1994), this relationship obviously also holds true for parts of proteins. One interesting feature of the AAA family is the widely variable degree of sequence similarity between orthologues within distinct branches. The first AAA cassettes in SAV of *Sulfolobus acidocaldarius* (archaea) and CDC48 of *Glycine max* (plant) share 60% identical residues, while the first AAA cassettes in Pas1p/Pex1p form *Pichia pastoris* and *S.cer* (both yeast) display only 23% sequence identity. In this paper, the

[1]See Table 1: archaebacterial members in SFs 1, 5, and 8; eubacterial members in SFs 6, 9, 10, 14, and 16; eucaryotic members in SFs 1–7, 9, 11–13, and 16).


Walker motif
AC Blocks database
=motif A=
BL00674A
BL00674B
=--motif B---=
Pex1p/1 S.cer [395]
Pex6p/1 S.cer [421]
Afg2p/1 S.cer [218]
YLL043c/1 S.cer [187]
Cdc48p/1 S.cer [192]
K04G2/1 C.eleg [226]
Yta7p/1 S.cer [391]
Sec18p/1 S.cer [215]
Afg2p/2 S.cer [490]
YLL043c/2 S.cer [507]
Cdc48p/2 S.cer [460]
Pex1p/2 S.cer [670]
Pex6p/2 S.cer [705]
K04G2/2 C.eleg [433]
Yta7p/2 S.cer [704]
Sec18p/2 S.cer [497]
Yme1p S.cer [254]
Yta10p S.cer [261]
SU4 Methc.jan [144]
Yta1p S.cer [155]
Yta3p S.cer [182]
Sug1p S.cer [122]
Yta2p S.cer [146]
Yta5p S.cer [156]
Sug2p S.cer [155]
orf167 Myc.lepr [213]
ycf46 Od.sin clp [201]
End13p S.cer [111]
Yta6p S.cer [153]
Yen7p S.cer [583]
C24B5.2 C.eleg [512]
mei-1 C.eleg [170]
Msp1p S.cer [72]
MJ1494 Methc.jan [99]
YBR186W S.cer [250]
F10B5.5 C.eleg [123]
F54B3.1 C.eleg [292]
orf300 E.coli [62]
Bcs1p S.cer [201]
A44L Par.bC.Vir [199]
YCF2 Mar.pol clp [1335]
RubPCOAct Mal.do [94]
ruvB E.coli [2]
Secondary structure
Binary pattern
West & Hecht
Conservation
Region
A
B
C
D
E
F

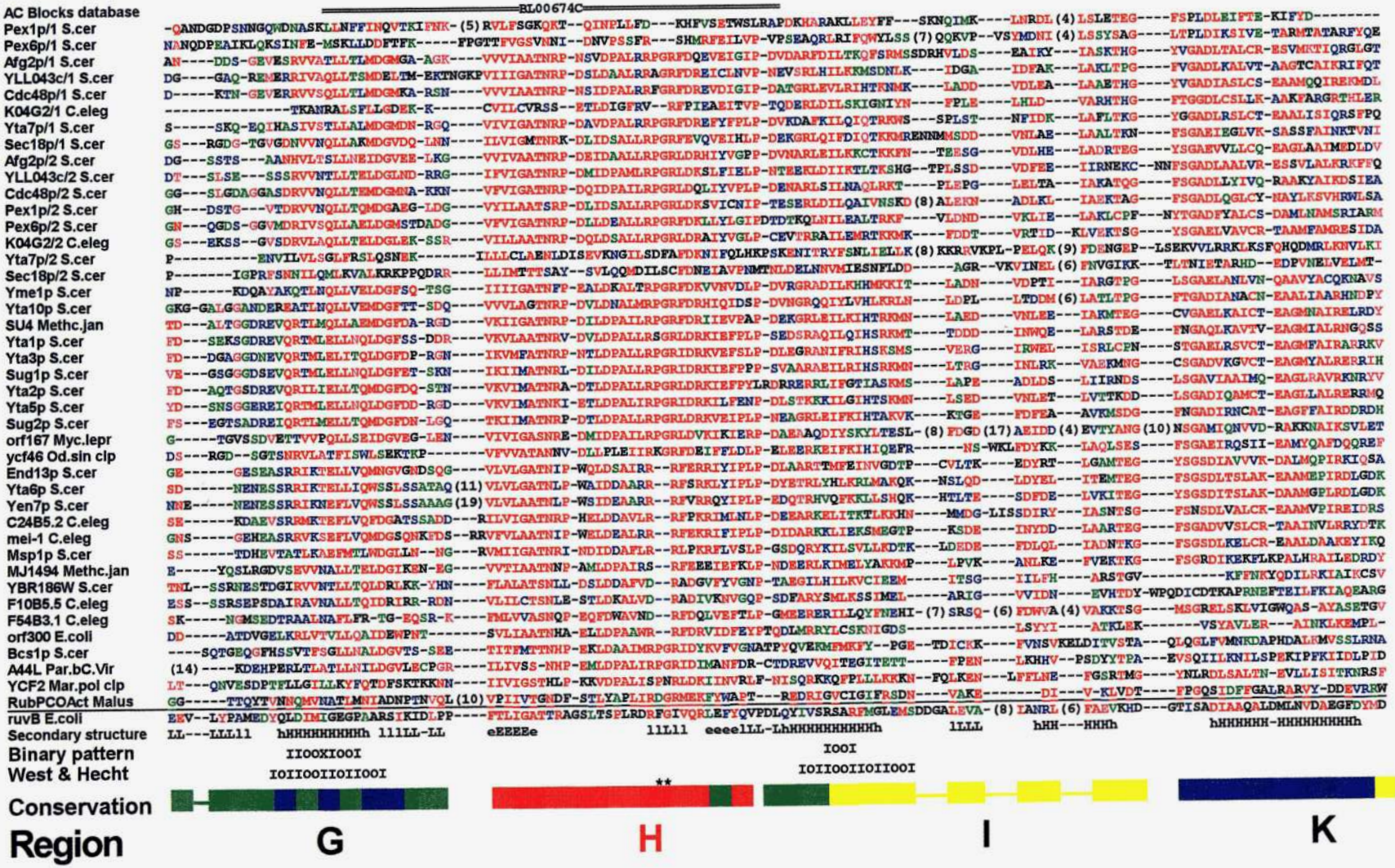


Fig. 1. Multiple sequence alignment of the AAA cassettes from representatives from all subgroups in comparison to a non-AAA Walker ATPase. The following groups of amino acids were regarded as beeing conservatively substituted: R=K=H; D=E=Q=N; G=A; C=S=T=A; V=I=L=M=C=F=Y=W (compare to Kunau et al., 1993). Colors denote the degree of conservation—Red: single or at least dominant group of amino acids in the given position. Blue; green; violet: different groups in one position. Black: no or poor conservation. Red bars: most conserved regions of the AAA cassette, little differences between subgroups, also similarities to other Walker proteins. Blue bars: AAA-typical regions: similar parts present in almost all AAA members but not in other Walker proteins. Green bars: subgroup-specific conservation is much stronger than similarity over the entire family (comparison to orthologues not shown, a detailed alignment can be obtained from the author via e-mail). Yellow bars: parts poorly conserved, even within subfamilies. Different regions were defined within the AAA cassette based on their different conservation pattern. For gene names see Table 1. RuvB does not belong to the AAA family but is related to it. "Name"/1 and "Name"/2: in cases where two AAA cassettes are present, the first and second, respectively. Figures in brackets: length of insertions in the given position in amino acids. Figures in angle brackets: number of the first residue given in the alignment in the respective protein. AC Blocks database: accession-number in the blocks database [http://www.blocks.fhcrc.org/; Henikoff & Henikoff (1991)]. Secondary structure: three states are discriminated: helix, extended, and loop. The method gives probability values between 0 and 9. Capital and small letters: Mean values for the given position range between 8–9 and 6.5–8, respectively. Binary pattern: hydrophilicity patterns of the predicted helices where they match with the binary patterns of amphiphilic α-helices according to West and Hecht (1995). I: hydrophilic, O: hydrophobic, X: intermediate or not conserved. The complete consensus pattern (West & Hecht) is IOIIOOIIOIIOOI. A sequence logo has been calculated from this alignment using alpro and makelogo (online: URL http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi; Schneider & Stephens, 1990). It is available in the electronic appendix (AAALOGO.GIF; AAALOGO.PS).

different molecular clock rates will be discussed in relationship to the proteins' functions.

Also presented here is an extensive sequence analysis. The aims of this analysis are fourfold: (1) to identify the outer boundaries of the AAA cassette and to determine its substructural components because so far no consensus exists in the literature even about the length of the AAA cassette; (2) to determine how a protein can be identified as a AAA family member; (3) a secondary structure prediction will be employed to fit the AAA cassette to the general Walker-fold; and (4) a model of the phylogeny of the family will be developed.

Results

AAA members and subfamilies

Table 1 shows a compilation of all AAA members known to date. Sequences are sorted into subfamilies according to their similarity (compare also to the dendrogram in Fig. 6). Within each subfamily the level of sequence identity between members is much higher than to all the other AAA representatives. Moreover, in most cases sequence similarity in regions outside the AAA cassettes—in many cases over the entire length of the respective proteins—can be demonstrated within the subfamilies but not between them (data not shown).

There are at least two subfamilies that contain both eukaryotic and archaebacterial members (SFs 1 and 5). On the other hand, SFs 1, 6, and 9 contain eubacterial as well as eukaryotic representatives. This clustering, however, must be regarded with caution (see comment in Table 1 and discussion).

Architecture of the AAA cassette

The AAA cassettes of all known family members were incorporated into a multiple alignment (Fig. 1). A number of different programs and algorithms were used and, in each case, the result was checked by pairwise dot matrix analysis. ClustalW was found to do best. Figure 1 contains only one member of each group; additional orthologues are omitted. All AAA representatives from *S.cer* are included. Conservation is demonstrated by colors and similarities have been defined according to Kunau et al. (1993).

The alignment was used to deduce a scheme for the general architecture of the AAA cassette (Fig. 2). Ten regions were distinguished, three of which are extremely well conserved (regions D, F, and H—red bars) and contain very few insertions or deletions. Regions D and F contain the Walker boxes A (=P-loop) and B, respectively. It must be emphasized that the AAA-typical consensus outlined in Figure 2 in those regions is much longer and much more defined than the Walker minimal consensus (Walker et al., 1982). While Walker motifs A and B can be identified in all and most Walker proteins, respectively (Fig. 3), weaker but nevertheless significant similarity between the AAA family and a small subset of Walker proteins can be detected in region H (see Figs. 1, 3, and 4). Besides these extremely well-conserved parts, the AAA cassette contains sequence motifs unique to this family (regions B, E, parts of G and K: blue bars). Moreover, in some regions, subgroup-specific conservation is very pronounced (regions C, G, and small parts of other regions; green bars). Finally, some poorly conserved boxes can be identified (regions A and I, yellow bars).

How can a AAA member be identified? The case of the poorly conserved AAA cassettes[2]

The alignment in Figure 1 includes some sequences that share only 10% sequence identity with the majority of other AAA proteins in the region of the cassette. Although this is even below Dayhoff's proposed "twilight zone" where homology cannot be postulated but suspected (Dayhoff et al., 1983; Vogt et al., 1995), there are good reasons to assume that these regions are AAA cassettes, although poorly conserved ones. In Figures 3 and 4 arguments are presented that support this conclusion. Figure 3 shows a block alignment performed with MACAW. Walker motifs A and B are detected in each sequence. With the very degenerate motif B, significance over the entire Walker superfamily is lowest. High similarities link the AAA cassettes including even the poorly conserved ones. Lower similarity is detected between the AAA family and some other ATPases—namely, ruvB/ClpB, Hsp104p first cassette, and lon in region H. The only similarities that can be demonstrated in all Walker proteins are restricted to the Walker motifs. The significance values of linked blocks are outlined in Table 2.

Figure 4 shows pairwise dot matrix comparisons between well and poorly conserved AAA cassettes[2] as well as distant members.[2] In a common one-dimensional sequence alignment it is often difficult to judge whether the level of similarity is significant. In a two-dimensional dot matrix analysis, however, the ratio of the number of dots in the diagonal to the number of dots in the remainder of the field provides an internal control concerning the significance of the similarity. A comparison of two highly related AAA members, mei1 vs. SAV/first cassette, and a comparison of a AAA member to a non-AAA Walker protein, SAV/first cassette vs. RNA helicase, are included as positive and negative controls, respectively. Significant similarities are evident in the comparisons to all distant members and most poorly conserved cassettes at first glance. The first cassette of YTA7 is the least similar sequence included in this analysis, at this stringency only region H is detected. Moreover, the similarities between the AAA cassette and some other non-AAA ATPases are evident.

Further evidence for the inclusion of poorly conserved cassettes as well as distant members to the AAA family comes from homology searching of a protein database. When used as query sequence, even these representatives detect predominantly other AAA family members (Table 3). Searches can also be performed with blocks extracted from the poorly conserved cassettes. However, only when taking together the first AAA cassettes of the pex1 and pex6 proteins, enough sequence information is available to allow detection of significant blocks out of the sequences. These blocks were employed to search a protein database (http://www.sdsc.edu/MEME/meme/website/meme.html; Bailey & Elkan, 1994). Indeed, the 31 best matches were AAA proteins, their significance values ranging from 1.2e-7 to 3.1e-3. The best non-AAA match,

[2]The distinction between 'distant members' and 'poorly conserved cassettes' is made as follows: the AAA cassettes of distant members are less similar to the remainder of the family than well-conserved cassettes are. Nevertheless, AAA-typical regions can easily be detected and the Walker motifs are present. Poorly conserved cassettes are derived from well-conserved ones but here, even the AAA-typical regions and the Walker motifs are no longer conserved. As far as can be judged from sequence data, distant members harbor ATPase activity while poorly conserved cassettes do not.

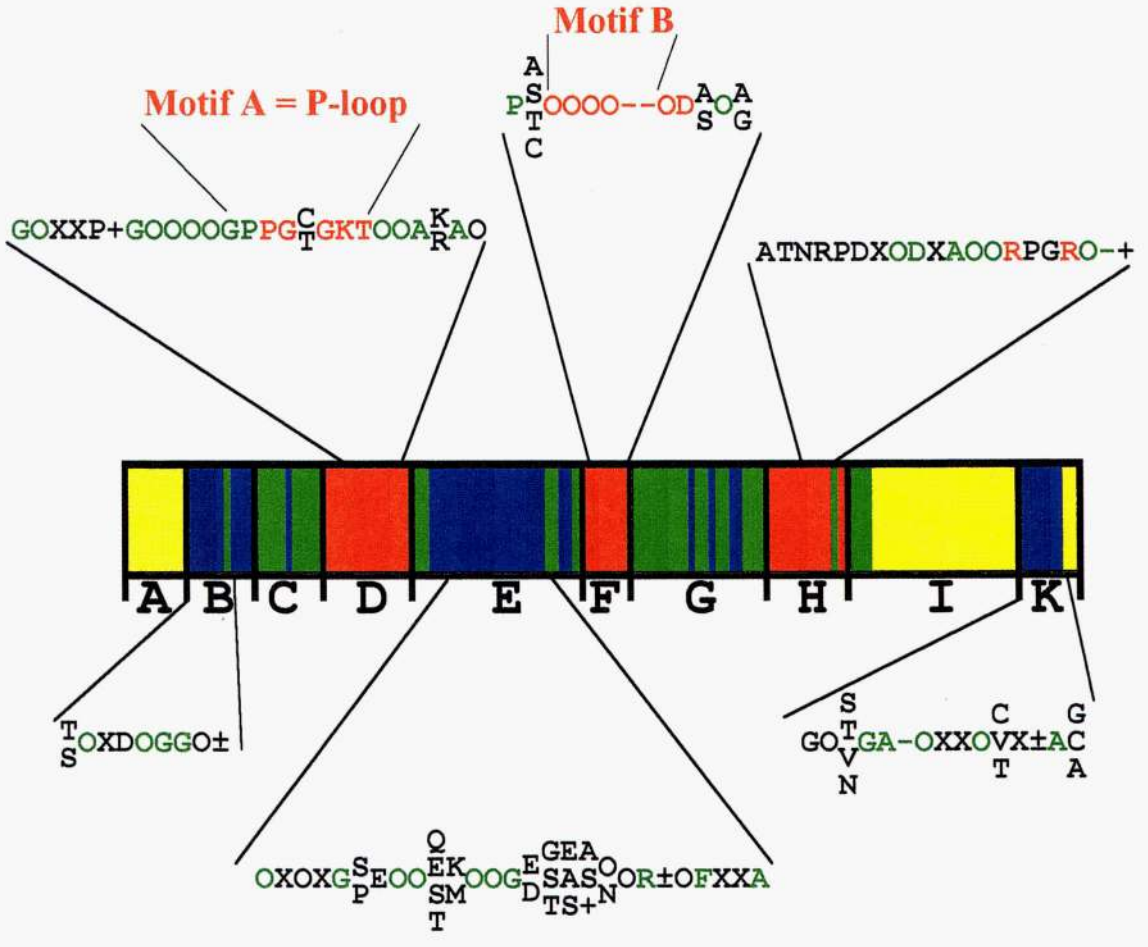


Fig. 2. Scheme: architecture of the AAA cassette. Different regions within the AAA cassette are denoted by coloured boxes—for definition see legend to Figure 1. Sequence patterns typical of the AAA family are listed in detail. In addition to the standard one-letter code for amino acids the following symbols are used: − = D/E, + = R/K/H, ± = charged, O = large and hydrobhobic. Colors denote conservation: red, green, black means almost absolutely conserved, well-conserved, and conserved, respectively. The Walker consensus motifs A (GX_4GKS/T) and B (4xhydrophobic D/E) are contained in regions D and F, respectively.

SWISSPROT P38323; YB77_YEAST, was detected at 4.1e-3 (data not shown).

Secondary structure prediction of the AAA cassette

The multiple alignment presented in Figure 1 was used for prediction of secondary structure. Accuracy values of available algorithms range about 60%, i.e., they are too low for a reliable prediction (Rooman & Wodak, 1988). For this reason, mean values were taken to eliminate stochastic errors caused by the compositions of individual representatives. For this purpose, the alignment was divided into windows of a length of five amino acids and for each interval, mean values of the results of the prediction were calculated (Fig. 5). The result of the network prediction is incorporated into Figure 1.

Phylogeny of the AAA family

For phylogenetic comparison, only homologous regions of proteins can be used. Thus, the analysis was restricted to the cassette region. Nevertheless, because the AAA family contains both very conserved and very dissimilar members, calculation of a reliable dendrogram is no simple task. Several methods for dendrogram construction were used and their results analyzed to determine whether the predicted branching order remained stable regardless of the number of family members included in the analysis and the presence or absence of single sequences. Of the programs examined, employing matrix as well as maximum parsimony methods, the neighbor-joining method (Saitou & Nei, 1987) proved to be the most robust. Significance was calculated by bootstrap resampling (Fig. 6); because this method has the inherent problem that the samples are not independent from each other, an even-odd analysis was performed. In such an analysis, two independent calculations are performed with the even and odd positions of the alignment, respectively. Both dendrograms had the same overall branching order, except for some minor differences at places where the bootstrap values are low (data not shown). The clustering of the sequences in the upper right part of the dendrogram (from first cassettes of SF 1 to SF 8) is artificial. Closer examination of pairwise similarities reveals that those sequences are joined only because of their common dis-

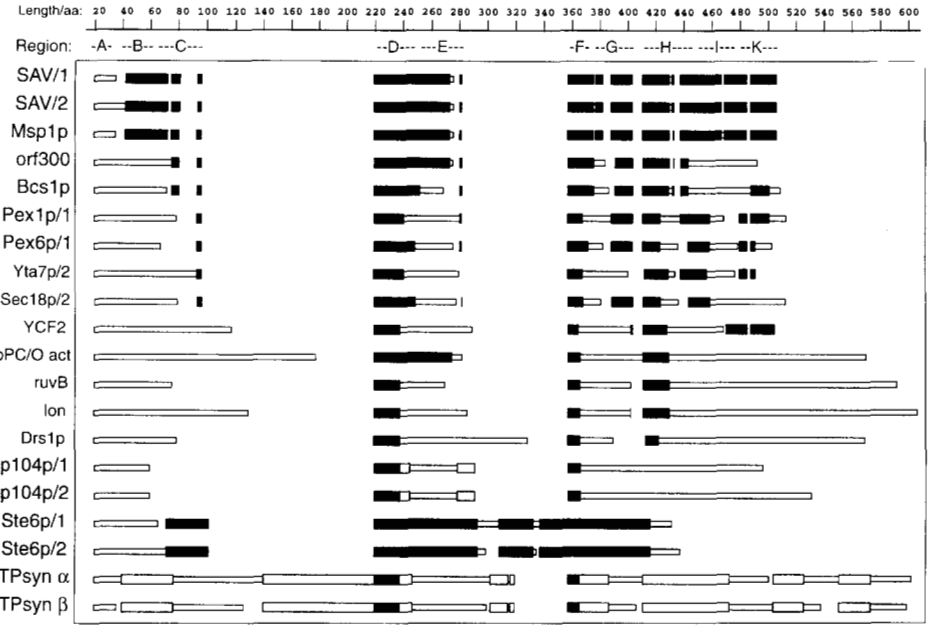


Fig. 3. Block alignment of AAA proteins in comparison to some other Walker ATPases. All proteins are from *S.cer* unless otherwise noted. Strongly conserved representatives: SAV/1&2 *Sul.ac*, Msp1p; distant members[2]: Bcs1p, orf 300 *E.coli*, YCF2 *Mar.pol* clpI, RubPCOAct *Mal.do*; poorly conserved cassettes[2]: Pex1p/1, Pex6p/1, Yta7p/2, Sec18/2; non-AAA Walker ATPases: ruvB (holliday junction DNA helicase) and lon (ATP-dependent protease) from *E.coli* as well as the other yeast proteins: Drs1p (ATP-dependent RNA helicase), Hsp104p (heat-shock protein 104; belongs to the CLPA/CLPB protease family), Ste6p (ABC-transporter), ATPsyn (mitochondrial ATP synthase; α and β subunit). Only the Walker cassette regions were used in the alignment. "Name"/1 and "Name"/2: In cases where two AAA cassettes are present, the first and second, respectively. Black boxes: similar parts in AAA members, some of which can also be identified in other Walker proteins. Boxes shaded in grey: similarities that can be detected only in the respective proteins but not in others (and, hence, not in AAA members).

similarity to the remainder of the family. To prevent Figure 6 from becoming even more complicated, orthologous sequences are represented by triangles rather than individual branches. Different branch lengths are indicative for differences in molecular clock rates and, hence, different degrees of conservation: the length of each triangle shaded in dark gray from its base to the inner tip corresponds to the time that has passed since the divergence of plants, yeasts, and animals. Note that SFs 1, 6, and 9 contain both eubacterial and eukaryotic members; SFs 1 and 5 contain archaebacterial and eukaryotic ones.

Discussion

Definition of the AAA family on the basis of protein sequence—the problem of distant and poorly conserved members[2]

A multiple alignment of AAA members (Fig. 1) reveals the architecture of the AAA cassette (Fig. 2). It must be admitted that the choice of groups of conservatively substituted residues is arbitrary (Fig. 1). However, even when groups are defined more or less stringently (e.g., E=D/Q=N or E=D=Q=N=S=T, respectively) the distribution of regions with a qualitatively different conservation pattern remains the same and is clear enough to justify the scheme presented in Figure 2. Nevertheless, one must keep in mind that the distinctions between regions conserved within a protein family are, of necessity, generalizations. AAA region G, for example, contains subgroup-specific sites as well as sites conserved in a AAA-typical manner in close proximity. Moreover, even in the most conserved regions there are some variant positions.

Cassette borders should be defined in such a way that the whole conserved, AAA-typical region is covered. Hence, it is reasonable to define the AAA cassette as the stretch including regions B to K because those parts can be identified with certainty in the vast majority of AAA representatives. The C-terminal end of the cassette can reliably be determined as a sharply defined border at the end of region K, beyond which similarity can be detected between different subfamilies only in one case (in the C-terminal regions of SFs 1 and 2). The N-terminal end of the cassette, however, is more nebulous, and the inclusion of region A within the cassette is a matter of definition. The "true" AAA cassette, that is, that part of the polypeptide that constitutes one defined three-dimensional domain (see below) may begin within the region A.

Taking this into account it makes sense to define an AAA protein as one that contains (a) all three of the most conserved AAA-typical regions (namely regions D, F, and H—note that the presence of the pure Walker consensus in regions D and F is necessary but not sufficient); (b) as well as at least one of the AAA-typical regions B, E, G, or K.

All of the subfamilies listed in Table 1 meet these criteria. The situation is more complicated in the proteins of SFs 2, 3, and 4, where one well-conserved and, according to the hypothesis discussed now, one poorly conserved cassette is present. The block alignment and dot matrices (Figs. 3 and 4) present evidence that the respective regions can indeed be called AAA cassettes. However, block alignments can be misleading (Henikoff, 1991). Espe-

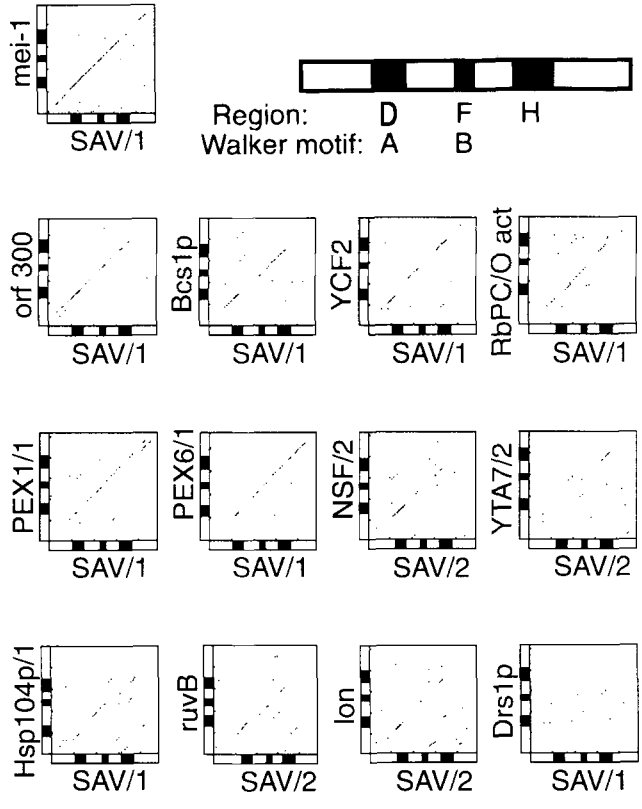


Fig. 4. Dot matrix analysis of Walker cassette regions. Because the two AAA cassettes of SAV are the most conserved of the entire family (Confalonieri et al., 1994), they have been used for comparison to improve sensitivity. First line: comparison of two typical AAA members. Second line: comparison of distant AAA members[2] to SAV. Third line: comparison of poorly conserved AAA cassettes[2] to SAV. Fourth line: comparison of non-AAA Walker cassettes to SAV. Parameters used: window = 25; matches = 5; score = 88/10 residues (log-odds matrix, Dayhoff), k-tuple = 1. Note that the program places one dot at the beginning of the interval that passes over the threshold chosen. "Name"/1 and "Name"/2: first and second cassette of the given protein, respectively. For protein names refer to Table 1.

cially in the case of the program MACAW it is not sufficient to calculate probabilities for joined blocks of putatively related sequences. It must be shown that the significance value increases or remains constant when a new sequence is joined. This has been done for the poorly conserved cassettes, and indeed, significance values for the block increase by orders of magnitude when one of the poorly conserved members is joined (data not shown). This kind of analysis was not originally performed for AFG1 and in this way, this protein was falsely identified as an AAA member (Lee & Wickner, 1992). Additional reasons to include these poorly conserved cassettes in the AAA family are as follows:

1. All of these cassettes are found in proteins that contain an additional, well-conserved cassette. In a dendrogram (see Fig. 6), the latter one clusters in the case of SF 2 with the second AAA cassettes of SF 1 and in the case of SF 3 with their first conserved cassettes. Hence, it is reasonable to assume that SFs 2 and 3 have been split off SF 1 *after* the internal duplication event. For SF 2 this hypothesis is further supported by the fact that it branches off *within* the group of second conserved cassettes (see Fig. 6) and that similarities between SFs 1 and 2 can be demonstrated beyond the AAA cassette in the vicinity of the C-terminus (data not shown). Indeed, a second set of Walker motifs, or at least similar sequences, can be detected in these proteins at the expected locations.
2. When the poorly conserved cassettes are used in homology searches of protein data bases with BLASTP, AAA members are detected, in all cases, much more often and, in most cases, with scores higher than other ATPases (Table 3). This result in principal remains the same when other algorithms are used.
3. As previously discussed, dot matrix analysis reveals similarities in AAA-typical regions (see Fig. 4).
4. Secondary structure prediction yields a similar profile for conserved and poorly conserved members (see below).

Relationship between the AAA family and other Walker proteins

There is no clear-cut division between the AAA family and other Walker families. Comparison of typical AAA representatives such as the proteasomal subunits to other members of the Walker superfamily yields a list of continuously less similar proteins. Hence, it is a problem to define non-arbitrary rules for membership in the AAA family.

High conservation within protein families is often found in substrate binding sites and catalytic centers (Wilson et al., 1977). Regions D and F of the AAA cassette contain the Walker motifs and are widespread among the families of the Walker superfamily. The third most conserved part of the AAA cassette is region H, a fact that hints that it might form a portion of the catalytic center in a way that could obviously also hold true for some other Walker families (a supposition that is supported by secondary structure prediction, see below). This is why the presence of regions D, F, and H is regarded as necessary but not sufficient to identify a AAA member (see above). In addition, at least one of the AAA-typical regions outlined in Figures 1 and 2 must be detectable. According to this definition, SFs 11–17 (see Table 1) do belong to the AAA family, although as distant members. The similarity between YCF2 proteins and the AAA family has already been described (Wolfe, 1994). The RNA helicases, HSP104-proteins, the ClpB/ruvB proteins, and the lon proteases, however, share no similarity with the AAA family except in regions D, F, and less pronounced in region H. Thus, according to the definition above they may at best be regarded as sister groups to the AAA family but not as members. Indeed, the similarity in region H strongly suggests a common phylogenetic stem and possibly a common *molecular reaction mechanism*. However, it must be emphasized that this does not imply that the proteins perform a similar *molecular function*, e.g., in DNA/RNA unwinding. This hypothesis has previously been suggested (Mian, 1993; Makino et al., 1996), but Mians alignment was based on very degenerate similarities that cannot be substantiated in a closer examination. Moreover, the assumption of a common molecular function is obviously false because some AAA members are known to be subunits of the 26S proteasome, some are membrane-associated proteases, others are involved in vesicle fusion processes. Thus, the AAA cassette probably performs a more general function (see below).

Table 2. *Significance values of linked blocks*[a]

Query sequence	Organism	Best pairwise probability				
	Region:	D	F	G	H	K
PEX1 (1st cass.)	Human	8.4 e-5	7.1 e-7		8.9 e-2	5.5 e-4
PEX6 (1st cass.)	Rat	1.2 e-7	4.4 e-2		1.8 e-3	2.6 e-8
NSF(2nd cass.)	Hamster	6.1 e-8	4.9 e-2		1.4 e-3	
YTA7 (2nd cass.)	*C.eleg*		4.5 e-2		2.0 e-3	
orf 300	*E.coli*	3.6 e-8	1.2 e-5	5.2 e-4	5.9 e-5	
Bcs1p	*S.cer*	7.7 e-9	3.6 e-2	8.6 e-2	2.7 e-8	
YCF2 = orf2280 protein	*Mar.pol* clpl	9.6 e-7	2.0 e-3		8.0 e-9	2.7 e-3
RubPCOAct	Apple tree	3.8 e-1	2.0 e-2	6.9 e-2	1.0 e-2	6.2 e-1

[a]Poorly conserved AAA cassettes (first four sequences) and cassettes of distant members (last four sequences) were compared pairwise to all conserved AAA members using MACAW. The best pairwise significance values are given as error probabilities.

Have all AAA subfamilies been identified? Problems of assignment

Within the AAA family 17 subfamilies can be distinguished, five of which contain two or more paralogues (see Table 1, SFs 1, 2, 5, 6, and 7): it follows that in these cases more recent gene duplication events have given rise to genes that display high similarity and, quite likely, possess similar functions. Very recent duplications probably lead to two copies of VCP in *Caenorhabditis elegans* and two copies of NSF in *Drosophila melanogaster*. Indeed, classification on the basis of overall similarities is equivalent to a grouping by function, at least as far as information about the respective gene functions is available (Table 1, right column).

The identification of orthologous proteins is not always straight forward. SUG1 (SF 5), for example, was originally described as the yeast gene corresponding to human TBP-1 (also SF 5, but not

Table 3. *Homology search of a protein database with different Walker cassettes as query sequences*[a]

Query sequence	Organism	Number of AAA members found within the given interval			
	Interval:	1–10	11–20	21–30	31–40
orf 300	*E.coli*	10	10	10	10
Bcs1p	*S.cer*	10	10	10	8
YCF2	*Mar.pol* clpl	10	10	10	10
RubPCOAct	Apple tree	10	9	8	6
Pex1p (1st cass.)	*S.cer*	4	2	1	4
PEX1 (1st cass.)	Human	10	10	7	4
PEX6p (1st cass.)	*S.cer*	9	3	5	3
PEX6 (1st cass.)	Rat	10	10	10	10
Sec18p (2nd cass.)	*S.cer*	4	6	7	10
NSF(2nd cass.)	Hamster	10	10	7	8
Yta7p(2nd cass.)	*S.cer*	9	4	3	3
YTA7 (2nd cass.)	*C.eleg*	3	2	1	2
lon	*E.coli*	0	2	2	1
ruvB	*E.coli*	0	0	8	7
Hsp104p	*S.cer*	0	1	1	1

[a]First four sequences: Distant members of the AAA family; next eight sequences: poorly conserved AAA cassettes; last three sequences: non-AAA Walker ATPases. With each sequence, a homology search of the non-redundant OWL Database (Unix-Server at EMBL, Heidelberg, Germany) was performed. Subsequently, in each case the list of matches, sorted according to the score value, was divided into intervals of 10 subjects. Figures denote the number of AAA members present in the respective interval. Orthologues of the respective query sequence have been left out, very close relatives have only been counted once.

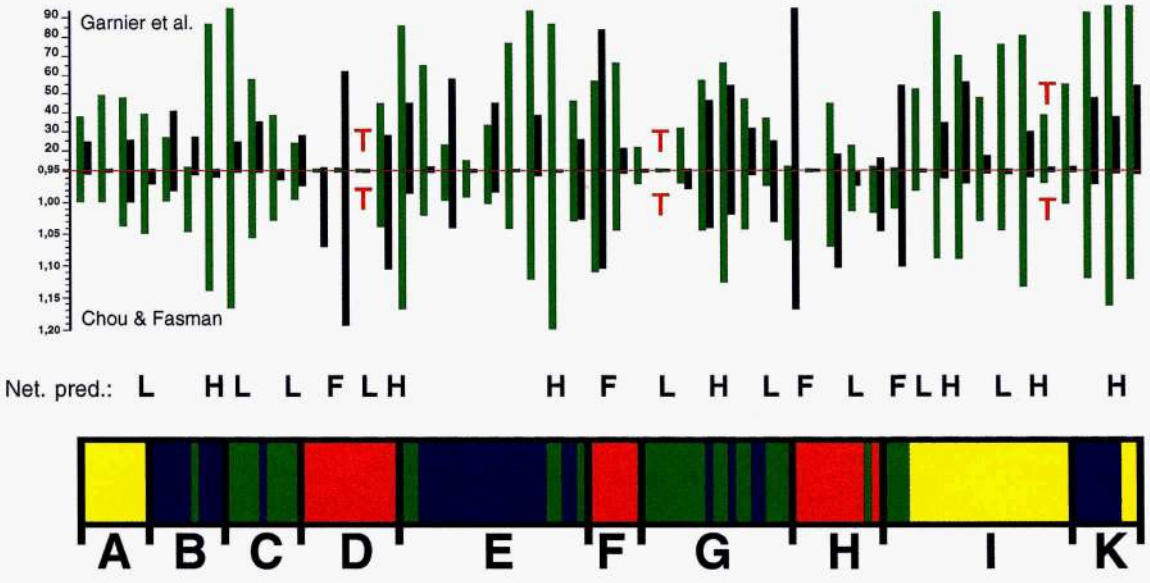


Fig. 5. Secondary structure prediction of the AAA cassette. The AAA cassette is outlined in colors as described in Figures 1 and 2. From each subfamily listed in Table 1 one representative was taken and a prediction was performed. Subsequently, mean values were calculated for a window of five residues: green and black bars: α-helix and β-fold potential, respectively. Results using the methods of Chou and Fasman (1974) and Garnier et al. (1978) are shown in the lower and upper panel, respectively. For the latter method, which works with a dynamical threshold value, the ratio of representatives where the threshold has been passed over is denoted in %. Red Ts: predicted β-turns. Net pred.: result of the prediction performed by a neuronal network (outlined in detail in Fig. 1)—H: α-Helix; F: β-Fold; L: Loop. Predicted helices were considered only if they were long enough to build at least two turns.

orthologous to SUG1) (Goyer et al., 1992). Later, as more sequences of proteasomal paralogues were published, the true mammalian orthologue of SUG1 was identified, and verified by functional complementation, as trip1 (Sun et al., 1996). The problem of recognizing true orthologues is known and has been described by Pamilo and Nei (1988). There are several lines of evidence that affirm genes from different organisms as very likely being orthologues: (1) In a phylogenetic dendrogram, the sequences are placed into the same order as phylogenetically expected for the respective species; (2) orthologues are, in many cases, able to heterologically complement for each other; (3) orthologues should stand on corresponding places in the biochemical network of the organisms, that is, the proteins are involved in the same processes and interact with orthologue partners; (4) compared to paralogues, orthologues in many cases display unique features in their molecular organization, for example, additional domains or a certain conservation pattern.

Most likely, true orthologues are joined in the following subfamilies—the genes' functions and the representatives in yeast are given: SF 1 (homotypic vesicle fusion, Cdc48p: 1,3,4); SF 2 (biogenesis of peroxisomes, Pex1p and Pex6p: 1,3,4); SF 4 (heterotypic vesicle fusion, Sec18p: 1,2,3,4); SF 5 (26S proteasomal subunits, Yta1p, Yta2p, Yta3p, Yta5p, Sug1p, and Sug1p: 1,2,3,4). Figures in parentheses refer to the arguments listed above.

With SFs 6 and 7 (metal-dependent proteases and the "functionally heterogeneous group," respectively) the situation is very complicated. Within SF 6, two groups can be distinguished. The first one (including the bacterial members) contains closely related sequences (see also Fig. 6). Their relationship to the second group (Yme1p, A15, M03C11.5—no bacterial member) is, however, unclear. This group probably arose early in eukaryote evolution by gene duplication of the FtsH gene brought by an endosymbiont. While one copy retained its original function—giving rise to Yta10p/Yta12p and orthologues—the other gained a new function and became the ancestor of Yme1p and orthologues. This hypothesis is supported by the fact that the latter group has longer branches in a dendrogram suggesting that the rate of sequence divergence may have changed under different selective pressure. Elucidation of relationships between SF 6 members in plants is even more complicated. There are four FtsH-like genes in the cyanobacterium *Synechocystis* spec., and sequence comparison of ESTs reveals the existence of at least five paralogues in plants. The exact relationship between cyanobacterial and plant FtsH-like proteins is difficult to evaluate. Sequence comparisons of whole proteins as well as parts show that most likely cyanobacterial slr0228 is orthologous to ycf25 from *Porphyra purpurea* chloroplast and slr1604 corresponds to the ATPase from *Capsicum annuum* (data not shown). This is a strong hint that the duplication events of FtsH in cyanobacteria predate the endosymbiosis event that gave rise to chloroplasts. PFTF (also *Cap.an*) and ycf25 (*Od.sin* clpl), however, cannot be clearly assigned as an orthologue to one of the cyanobacterial members. This may be due to gene conversion events by which parts have been exchanged between these very similar genes.

SF 7 is unique in that its members are involved in such different processes that no functional assignment is typical of the group. Moreover, different eukaryotes clearly contain different numbers of SF 7 members, for example there is no orthologue of the C.eleg protein mei-1 in yeast. The only subgroups that display full-length similarity and, hence, quite likely are orthologues are End13p= Vsp4p (*S.cer*)—Suppressor protein (*Schizosaccharomyces pombe*)—SKD1 (mouse) as well as, less pronounced, Msp1p=Yta4p (*S.cer*)—K04D7.2 (*C.eleg*)—DM19DC4Z (*Dro.mel*).

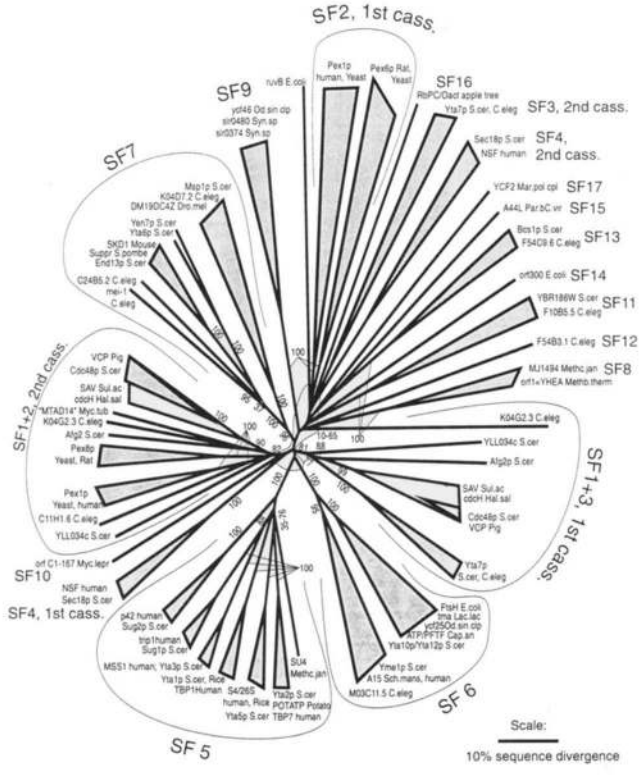


Fig. 6. Neighbor-joining dendrogram of the AAA family calculated from the alignment shown in Figure 1. For genes and numbering of subfamilies ("SF") refer to Table 1. Figures denote percent bootstrap probability (1000 trials). The line in the middle of the dendrogram encloses a region where branching orders are completely insignificant. Orthologues are represented by shaded triangles. Dark grey: plants/fungi/animals; light grey: eukaryotes/archaea. The bar given as scale corresponds to 10% sequence diversity, not corrected for multiple substitutions.

Due to the extremely poor quality of sequence data, a number of expressed sequence tags could not be identified. Moreover, in many cases sequence information was too restricted to allow an unambiguous assignment. Hence, the question of the total number of AAA subfamilies can hardly be answered on the basis of ESTs and genomic fragments alone. Additional information comes from the complete genome sequencing projects. This data gives an impression which AAA members are essential for the given cellular and biochemical organization. Moreover, one can be sure that no AAA member has been overlooked in the respective organism.

Because no orthologue to *C.eleg* mei-1 can be found within the completely sequenced genome of *S.cer*, the orf 300-protein from *Escherichia coli* cannot be identified in *Haemophilus influenzae*, *Mypl.gen*, and *Syn.sp.*, and SF 9 proteins most likely are restricted to cyanobacteria and chloroplasts, there are obviously AAA subgroups that are present only in smaller taxonomic units (see Table 1). It follows that novel subgroups might still be discovered. Thus it is possible that the non-assigned ESTs from plants (Table 1, last group) might represent new subgroups.

Suggestion of a 3D structure of the AAA cassette

Because the tertiary structure of a number of Walker proteins has been elucidated, the general fold of this superfamily is known (Schulz, 1992, and references therein) and the predicted secondary structure profile of the AAA cassette could be fitted to this template (Fig. 7).

Several lines of evidence support the model outlined in Figure 7.

1. There is a high agreement in the results of all methods used (compare Figs. 1 and 5).
2. The Walker motifs are invariably connected to secondary structure elements. Motif A is located in a loop, the so-called p-loop, which always is preceded by a β-strand and followed by an α-helix. Motif B consists of a more or less hydrophobic stretch, which forms a β-strand, followed by at least one negatively charged residue. Indeed, all methods employed correctly predicted these structural units (the AAA-typical versions of the Walker motifs are outlined in Fig. 2, their relationships to secondary structure elements within the general tertiary structure of a Walker protein are shown in Fig. 7).
3. The predicted structure is compatible with the distribution of Gly and Pro (data not shown). For structural reasons, these amino acids rarely occur in β-strands and α-helices because Pro

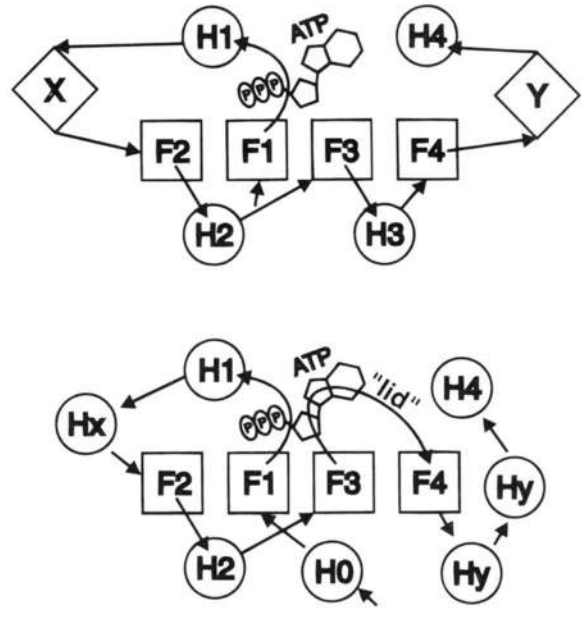


Fig. 7. The AAA cassette fit to the Walker fold. Upper panel: the general fold of a Walker A/GTPAse according to Schulz (1992). Lower panel: the predicted secondary structure profile of the AAA cassette fit to this general fold. Squares: β-folds; circles: α-helices; X,Y: sites where insertions frequently are observed in various Walker families. The corresponding elements in the AAA cassette consequently are called "Hx" and "Hy," the putative helix preceding motif A is denoted as "H0." The following assignments between regions (bold capitals, see Fig. 1) and secondary structure elements in the AAA cassette (lower panel) can be made: B: H0; D: F1 + P-loop (Walker motif A) + H1; E: Hx; F: F2 + motif B (at the C-terminal end of F2); G: H2; H: F3 + "lid"(= loop connecting F3 to F4) + F4; I: first and second Hy; K: H4. The arrows pointing to F1 and H0, respectively, mark the N-terminal ends of the polypeptide chains. In both cases, the cassette ends up with H4. Note that in this model helix 3 is replaced by a loop in the AAA cassette that covers the ATP binding cleft like a lid. Although modeling of the core is very well supported (see Discussion), no suggestion can be made about the orientation of the outer elements of the cassette (H0, Hy, and Hy).

hardly is compatible with the spatial demands a residue must fulfill in secondary structures and Gly is very small and flexible so that it much more often occurs in loops and turns. Only 4% of the prolines and 8% of the glycines occur in the predicted helices and β-folds; most of these are located at the boundaries of those secondary structure elements.

4. Insertions and deletions are found much more often in solvent-exposed turns than in core regions of proteins (Bajaj & Blundell, 1984). The secondary structure elements predicted by this model are nearly free of sequence alignment gaps. Insertions and deletions in the AAA cassette occur almost exclusively outside the predicted secondary structure elements. In those cases where gaps in the alignment appear within a predicted α-helix or β-fold the sequences concerned are not well conserved at the respective location; thus, these gaps may well be artificial (see Fig. 1).

5. In most of the putative loop regions, high flexibility or high β-turn potential is predicted (Figs. 1 and 5).

6. The succession of predicted secondary structure elements is very well compatible with the standard Walker protein fold, that is, there are no elements missing or in the wrong order. There is only one discrepancy between the predicted AAA profile and the Walker general fold. Region H should build the unit [β-sheet → α-helix → β-sheet] (compare upper to lower panel in Fig. 7). Indeed, the two β-sheets are predicted correctly, but there can be no helix between them (helix H3; see also Figs. 1 and 5). Moreover, it cannot be understood why a portion of the protein that is located at its surface rather than in the vicinity of the bound ATP should be that extremely conserved. This contradiction can be easily resolved by assuming that helix H3 is omitted in the AAA cassette and that the respective part of region H forms a structure resembling a lid that can cover the bound ATP. This assumption is not far-fetched; a similar situation is found in the adenylate kinases (Abele & Schulz, 1995). In case this hypothesis is correct, one might speculate that the two extremely well-conserved Arg-residues constitute a part of the catalytic center that forms when the flexible loop they are lying in closes up the ATP binding cleft.

7. Hydrophobicity also is very well compatible with the fit (compare the binary patterns in Fig. 1 to the model in Fig. 7). The putative loop regions are rather hydrophilic (data not shown). β-Folds 1–3 are very hydrophobic and, hence, most likely buried in the protein's core, whereas β-fold 4 is rather hydrophilic and thus well suited to build a more exposed edge of the entire β-sheet. According to the general Walker fold at least helices H1, H2, and H3 should be located on the surface of the domain with one face exposed to the solvent and, therefore, they should be amphiphilic. West and Hecht (1995) have developed a binary pattern for amphiphilic α-helices and β-folds, and indeed, this pattern matches the distribution of residues found in the helices H1, Hx, H2, and the first of two Hy in the AAA cassette. Helices H0, second Hy, and H4 are rather hydrophilic; thus, they might be more exposed. If the prediction in general is correct, the AAA domain may have a very compact side (left half in Fig. 7, lower panel), composed of the hydrophobic outer β-fold F2 surrounded by amphiphilic helices and a more open and flexible side composed of the more hydrophilic elements outer fold F4 and helices second Hy and H4.

8. Independent secondary structure predictions were performed on poorly conserved cassettes and distant members of the AAA family. There were only minor differences to the AAA prediction profile outlined in Figure 1: the location of Helix H1 may be shifted in YCF2 and RubPCOAct, Helix Hy may be omitted in Bcs1p, and surprisingly, the canonical Walker Helix H3 is predicted for F10B5.5 and F54B3.1 (*C.eleg*); however, because there are only few close relatives of these two genes known, the reliability of the predicted structure is unclear. The helical extension of the cassette (Hy-Hy-H4) is in all cases as predicted as expected.

Though the 3D model is well supported, it contributes little to our knowledge concerning the function of the cassette. As already pointed out, the intriguing diversity of processes in which the AAA proteins are involved makes it difficult to argue for a certain molecular function. Moreover, many AAA representatives show a pleiotropic phenotype. For example, it was first proposed that the proteasomal regulatory subunits function as transcriptional activators (Swaffield et al., 1992), and only later, it was shown that the apparent effect on transcriptional control can be explained by proteolytic degradation performed by the proteasome (Dubiel et al., 1992; for discussion see Confalonieri & Duguet, 1995). Taking all available data, the only common feature may be the mediation or control of protein–protein interactions, as suggested by Confalonieri and Duguet (1995).

Phylogenetic history of the AAA family

Based on the dendrogram in Figure 6, a hypothetical phylogenetic tree was constructed (Fig. 8). In addition to the points already discussed, it contains the following implications:

1. Two out of three archaebacterial AAA proteins can be grouped together with eukaryotic representatives (Table 1, SFs 1 and 5). The third archaebacterial AAA group (orf 1 and MJ1494 in SF 8) may be related to SF 7, because both groups share a highly specific two-amino acid deletion in region H (marked by two asterisks above the bar representing region H in Fig. 1); however, this is speculative because the overall similarity within the cassette is not higher between both groups than to the remainder of the whole family. Nevertheless, it can be argued that since SF 7 proteins are much less conserved than those of SFs 1 and 5 all traces of a common root linking SF 7 to SF 8 have been lost. Indeed, the SF 7 proteins share extremely low similarity outside their cassettes. The assignment for MJ1494 by TIGR (annotation of MJ1494 in the EMBL database) as being a proteasomal subunit (Table 1) may well be wrong; this notion cannot be substantiated by sequence similarity to SF 5.

2. The intragenic duplication of AAA domains happened only once, very early in archaebacterial evolution in the common ancestor of extant archaea and eukarya. This assumption seems reasonable because there is growing evidence for archaea and eukarya being closely related (Cavalier-Smith, 1987; Iwabe et al., 1989; Barinaga, 1994).

3. Duplications of that double cassette happened early in eukaryote evolution and gave rise to the genes of SFs 1–4. However, it cannot be ruled out that SF 4 is derived from an independent duplication event because it does not show a particularly closer

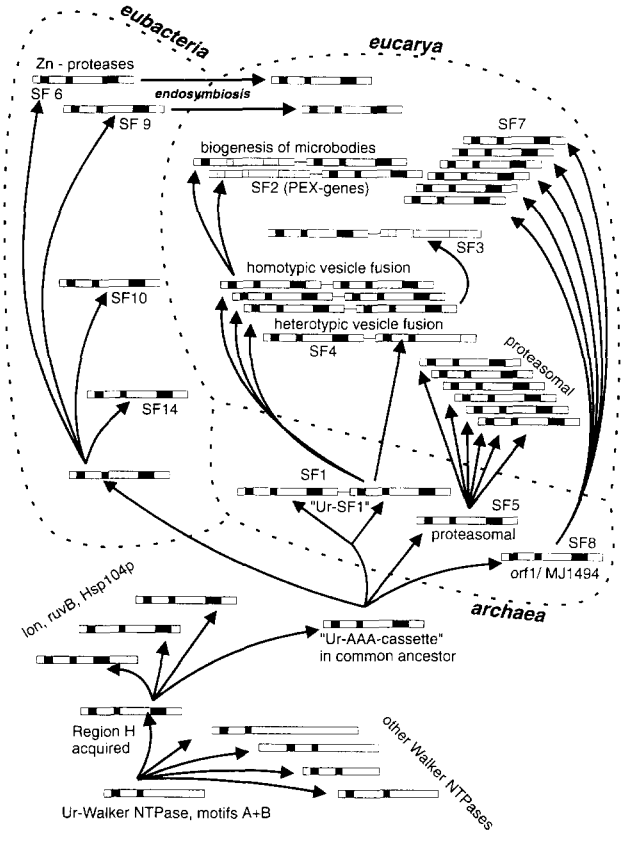


Fig. 8. Hypothetical evolutionary tree of the AAA family. Only cassette regions are shown; additional parts of the protein molecules are left out. Walker motifs A and B as well as the putative loop in region H are represented by filled boxes (left, middle, and right, respectively). Non-conserved parts are shaded in gray. "Ur-AAA cassette" denotes the primordial AAA protein, containing all features typical of the family. "Ur-SF1" means the gene stemming from the internal duplication leading to SFs 1–4. "Ur-SF1" may have been similar to present-day SAV. Because there is not enough information to trace the history of SFs 11–13 and 15–17, they are omitted.

relationship to SFs 1–3 (see Fig. 6). Nevertheless, the fact that both SF 1 and 4 are involved in membrane fusion events supports the idea of common ancestry of SFs 1–4. In SFs 2–4, one of the two AAA domains became dispensable and, hence, less conserved. This conclusion is supported by the clustering of the conserved cassettes of the respective proteins (see Fig. 6) and by weak but significant similarities outside the cassette regions (data not shown). The continued presence of these poorly conserved cassettes may indicate that they perform a structural or regulatory role rather than an enzymatical one; a similar situation can be observed in F1/F0-ATPases (Nelson, 1992).

4. Three subfamilies contain both eubacterial and eukaryotic AAA representatives. This clustering is probably due to horizontal transfer events rather than to vertical inheritance. As far as it is known today, the SF 6 proteins are restricted to eubacteria, mitochondria, and chloroplasts; the SF 9 proteins are further restricted to cyanobacteria and chloroplasts. Thus, these genes were most likely acquired from the endosymbionts early in eukaryote evolution. A genomic fragment from *Mycobacterium tuberculosis* contains an orf belonging to SF 1 (Table 1). This is so far the only eubacterial example for a protein with two AAA cassettes; *Mypl.gen*, *Haemophilus influenzae*, and *Syn.sp* do not possess an orthologue of this orf. Thus, it is tempting to speculate that it was acquired by horizontal gene transfer from a mammalian host. Such an event does not seem unlikely because *Myc.tub* lives in close association to the putative gene donor. This hypothesis is strengthened by a further argument: the first AAA cassette of the *Myc.tub* orf is weakly conserved, even within the Walker motifs A, B, and region H altered and, hence, is likely non-functional. If SF 1 was archaic, dating back to the common ancestor of present day live, this would imply that the first cassette diverged, for example, due to the release of selective pressure, but was nevertheless maintained as a recognizable AAA cassette for billions of years. Moreover, independent loss of the orthologues must have occurred in the linages leading to *Haemophilus influenzae*, *Mypl.gen*, and *Syn.sp*. On the other hand, a recent horizontal transfer event fits the data quite well: selective pressure on the first cassette may have been released and this part of the protein may now be in the process of disappearing. The publication of additional bacterial genomes will clarify this matter.

The unequal clock rate in different subfamilies is an interesting feature of the AAA family. This is observed in the dendrogram (Fig. 6) as variable depth of the triangles representing orthologues and, moreover, leads to an artificial clustering of poorly conserved and distant members. Not only do clock rates *between* subfamilies vary drastically, but also *within* SF 1: the duplicated AAA cassettes of the archaebacterial members are more conserved than their eukaryotic orthologues (Confalonieri et al., 1994). Unequal clock rates in different branches of one and the same protein family have been reported (Goodman, 1981), although this interpretation is controversial (Wilson et al., 1977). The phylogenetic scenario suggested for SFs 1–4 might provide an explanation for this finding. Only one AAA member with duplicated cassettes is present in *Methanococcus jannaschii*, while there are seven in *S.cer*, belonging to SFs 1–4. The assumption that a SAV-like gene is indeed present in all archaebacteria is supported by the known existence of such a gene in five representatives (Table 1). Perhaps this single SF 1 protein of archaea is more important because it performs tasks that in eukarya are distributed among several proteins. The function of the archaebacterial representatives, however, is mysterious. In eukarya at least some of the members of SFs 1–4 are involved in membrane/vesicle fusion events (see Table 1). Yet there is no hint of such events in archaea. Hence, the compelling question is: "For what indispensable fuction does an archaebacterium need a SF 1 protein?"

As far as data is available, sequence conservation of AAA proteins correlates well with functional indispensability. Membrane fusion events in eukaryotic cell cycle (SF 1) and vesicular transport (SF 4) as well as proteasomal functions (SF 5) are essential for the cell, whereas the cell may, under certain circumstances, survive impairment of certain peroxisomal (SF 3) or even mitochondrial functions (SFs 6 and 13).

Conclusions

Sequence analysis provides insight into the organization of the AAA cassette and allows the establishment of rules that distinguish AAA-related sequences: cassette regions B, G, and K are very indicative for a AAA protein. However, it must be emphasized that

there is no clear-cut border line between the AAA family and the remainder of the Walker superfamily. The division of the AAA family into subgroups was discussed and a good correlation was found between functional grouping, dendrogram, and certain conserved parts of the cassette: cassette regions C and G were found to be conserved in a predominantly subgroup-specific manner that enables these regions to be used as fingerprints for the categorization of new AAA members.

Insights into the evolutionary history of the AAA family have been obtained, although the future analysis of additional complete genomes will be helpful. A three-dimensional model of the AAA cassette was deduced from secondary structure prediction and homology comparisons. In this manner, a portion of the AAA cassette—the middle part of region H—was identified as a candidate for the catalytic center or at least a substantial part thereof. This prediction may be used as a template for the construction of mutants to elucidate the molecular function of the cassette, a function that still remains obscure.

Materials and methods

Database searching

Homology searches were performed online at the UNIX server of the European molecular biology laboratory (EMBL) in Heidelberg, Germany, using the programs FASTA (Pearson, 1990) and BLASTP for proteins and BLASTN (Altschul et al., 1990) for nucleic acid sequences.

Alignments

ClustalW (DOS) (Higgins et al., 1992, Thompson et al., 1994) was employed for multiple protein alignments. Default parameters were used except that window size = 25, top diagonals = 20, and gap penalty = 9, were used to improve sensitivity. MACAW (Windows) (Schuler et al., 1991) was used to create block alignments to calculate statistical significance of given similarities. Expressed sequence tags were assembled using LASERGENE (MacIntosh) (DNASTAR Inc., Madison, Wisconsin). Frameshifts in single sequences were detected manually by comparison of all reading frames to the complete multiple alignment.

Dot matrices

Dot matrix analysis was carried out with GENEPRO (DOS) (Riverside Scientific Enterprise, Seattle, Washington).

Calculation of dendrograms

Dendrograms were generated using several computer algorithms. The neighbor-joining method (Saitou & Nei, 1987), performed by ClustalW (DOS) (Higgins et al., 1992) proved most robust. Statistical significance values were evaluated with the bootstrapping method (Felsenstein, 1985; Wu, 1986) included in the program. Additionally, an even-odd analysis (Wu, 1986) was used to create independent samples for dendrogram calculation and bootstrapping.

Secondary structure prediction

Three different methods were employed to predict the secondary structure of the AAA cassette: GENEPRO (DOS) using the algorithm of Chou and Fasman (1974), PREDICT (DOS) using the method developed by Garnier et al. (1978), and an on-line prediction performed at the EMBL in Heidelberg (Rost et al., 1994), using a profile network prediction procedure (Rost & Sander, 1993a, 1993b). To eliminate stochastic errors caused by individual sequence variation, members of each AAA subfamily were subjected to a prediction and mean values were subsequently calculated. The following sequences were included in the prediction: Afg2p and Cdc48p, both first and second cassette; Yta7p and Sec18p, first cassette; Pex1p second cassette; Yta1p, Sug2p, Yta10p, Yme1p, End13p, Msp1p, YBR186W; Bcs1p (all *S.cer*); orf1 (*Methanobacterium thermoautotrophicum*); C1_167 (*Mycobacterium leprae*); F54B3.1 (*C.eleg*); orf300 (*E.coli*).

Electronic appendix

The following files are available in the electronic appendix: AAAALIGN.ASC: detailed alignments (AscII-code); AAALIST.ASC: a list of AAA members and accession numbers (AscII-code); AAALOGO.PS and AAALOGO.GIF: sequence logo of the AAA cassette (GIF and postscript file, respectively), TABLE1.DOC: color version of Table 1 (WinWord 6.0 file).

Acknowledgments

I am thankful for criticism, support, and helpful advise from Dr. K.U. Fröhlich, Dr. W. Hanstein, Dr. L.M.G. Heilmeyer, Jr., Dr. C. Jaquet, Dr. W.-H. Kunau, and Dr. A. T. Maichele.

References

Abele U, Schulz GE. 1995. High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap5A, showing the pathway of phosphoryl transfer. *Protein Sci 4*:1262–1271.

Altschul SF, Gish W Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, California: AAAI Press.

Bajaj M, Blundell T. 1984. Evolution and the tertiary structure of proteins. *Annu Rev Biophys Bioeng 13*:453–492.

Barinaga M. 1994. Archaea and eukaryotes grow closer. *Science 264*:1251.

Cavalier-Smith T. 1987. The origin of eukaryote and archaebacterial cells. *Ann NY Acad Sci USA 503*:17–54.

Choi HS, Seol W, Moore DD. 1996. A component of the 26S proteasome binds on orphan member of the nuclear hormone receptor superfamily. *J Steroid Biochem Mol Biol 56*:23–30.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:22–245.

Confalonieri F, Duguet M. 1995. A 200-amino acid ATPase module in search of a basic function. *Bioessays 17*:639–650.

Confalonieri F, Marsault J, Duguet M. 1994. SAV, an archaebacterial gene with extensive homolgy to a family of highly conserved eukaryotic ATPases. *J Mol Biol 235*:396–401.

Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. *Methods Enzymol 91*:524–545.

Distel B, Erdmann R, Gould SJ, Blobel G, Crane DI, Cregg JM, Dodt G, Fujiki Y, Goodman JM, Just WW, Kiel JAKW, Kunau WH, Lazarow PB, Mannaerts GP, Moser HW, Osumi T, Rachubinski RA, Roscher A, Subramani S, Tabak HF, Tsukamoto T, Valle D, Vanderklei I, Vanveldhoven PP, Veenhuis M. 1996. Unified nomenclature for peroxisome biogenesis factors. *J Cell Biol 135*:1–3.

Dubiel W, Ferrell K, Pratt G, Rechsteiner MC. 1992. Subunit 4 of the 26S protease is a member of a novel eukaryotic ATPase family. *J Biol Chem 267*:22699–22702.

Eakle KA, Bernstein M, Emr SD. 1988. Characterization of a component of the yeast secretion machinery: Identification of the SEC18 gene product. *Mol Cell Biol 8*:4098–4109.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution 39*:783–791.

Fröhlich KU, Fries HW, Peters JM, Mecke D. 1995. The ATPase activity of purified Cdc48p from *Saccharomyces cerevisiae* shows complex dependence on ATP-, ADP-, and NADH-concentrations and is completely inhibited by NEM. *Biochim Biophys Acta 1253*:25–32.

Garnier J, Osgusthorpe DJ, Robson B. 1978. Identification of predictive sequence motifs limited by protein structure data base size. *J Mol Biol 120*:97–120.

Goodman M. 1981. Globin evolution was apparently very rapid in early vertebrates: A reasonable case against the rate-constancy hypothesis. *J Mol Evol 17*:114–120.

Goyer C, Lee HS, Malo D, Sonenberg N. 1992. Isolation of a yeast gene encoding a protein homologous to the human Tat-binding protein TBP-1. *DNA Cell Biol 11*:579–585.

Henikoff S. 1991. Playing with blocks: Some pitfalls of forcing multiple alignments. *New Biol 3*:1148–1154.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res 19*:6565–6572.

Higgins DJ, Bleasby AJ, Fuchs R. 1992. CLUSTAL V: Improved software for multiple sequence alignment. *Comput Appl Biosci 8*:189–191.

Iwabe N, Kuma KI, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA 86*:9355–9359.

Kimura M. 1987. Molecular evolutionary clock and the neutral theory. *J Mol Evol 26*:24–33.

Kunau W-H, Beyer A, Franken T, Götte K, Marzioch M, Saidowsky J, Skaletz-Rorowski A, Wiebel FF. 1993. Two complementary approaches to study peroxisome biogenesis in *Saccharomyces cerevisiae*: Forward and reversed genetics. *Biochimie 75*:209–224.

Lee YJ, Wickner RB. 1992. AFG1, a new member of the SEC18-NSF, PAS1, CDC48-VCP, TBP-1 family of ATPases. *Yeast 8*:787–790.

MacIntyre RJ. 1994. Molecular evolution: Codes, clocks, genes and genomes. *Bioessays 16*:699–703.

Makino Y, Yogosawa S, Kanemaki M, Yoshida T, Yamano K, Kishimoto T, Moncollin V, Egly JM, Muramatsu M,, Tamura T. 1996. Structures of the rat proteasomal ATPases—Determination of highly conserved structural motifs and rules for their spacing. *Biochem Biophys Res Commun 220*:1049–1054.

Mian I. 1993. Sequence similarities between cell regulation factors, heat shock proteins and RNA helicases. *Trends Biochem Sci 18*:125–127.

Morgan A, Dimaline R, Burgoyne RD. 1994. The ATPase activity of N-ethylmaleimide-sensitive fusion protein (NSF) is regulated by soluble NSF attachment proteins. *J Biol Chem 269*:29347–29350.

Nelson N. 1992. Evolution of organellar proton-ATPases. *Biochim Biophys Acta 1100*:109–124.

Nobrega FG, Nobrega MP, Tzagoloff A. 1992. BCS1, a novel gene required for the expression of functional Rieske iron-sulfur protein in *Saccharomyces cerevisiae*. *EMBO J 11*:3821–3829.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol 5*:568–583.

Pearson WR. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol 183*:63–98.

Peters JM, Harris JR, Lustig A, Müller S, Engel A, Volker S, Franke WW. 1992. Ubiquitous soluble Mg^{2+}-ATPase complex—A structural study. *J Mol Biol 223*:557–571.

Rechsteiner M, Hoffmann L, Dubiel W. 1993. The multicatalytic and 26S proteases. *J Biol Chem 268*:6065–6068.

Rooman MJ, Wodak SJ. 1988. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Nature 335*:45–49.

Rost B, Sander C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA 90*:7558–7562.

Rost B, Sander C. 1993b. Prediction of protein structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Rost B, Sander C, Schneider R. 1994. An automatic mail server for protein secondary structure prediction. *Comput Appl Biosci 10*:53–60.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol 4*:406–425.

Saraste M, Sibbald PR, Wittinghofer A. 1990. The P-loop a common motif in ATP- and GTP-binding proteins. *Trends Biochem Sci 15*:430–434.

Schnall R, Mannhaupt G, Stucka R, Tauer R, Ehnle S, Schwarzlose C, Vetter I, Feldmann H. 1994. Identification of a set of yeast genes coding for a novel family of putative ATPases with high similarity to constituents of the 26S protease complex. *Yeast 10*:1141–1155.

Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res 18*:6097–6100.

Schuler GD, Altschul SF, Lipman DJ. 1991. A workbench for multiple alignment construction and analysis. *Proteins 9*:180–190.

Schulz GE. 1992. Binding of nucleotides by proteins. *Curr Opin Struct Biol 2*:61–67.

Smith CA, Rayment I. 1996. Active site comparisons highlight structural similarities between myosin and other p-loop proteins. *Biophys J 70*:1590–1602

Sun DH, Sathyanarayana UG, Johnston SA, Schwartz LM. 1996. A member of the phylogenetically conserved CAD family of transcriptional regulators is dramatically up-regulated during the programmed cell death of skeletal muscle in the tobacco hawkmoth *Manduca Sexta*. *Dev Biol 173*:499–509.

Swaffield JC, Bromberg JF, Johnston SA. 1992. Alterations in a yeast protein resembling HIV Tat-binding protein relieve requirement for an acidic activation domain in GAL4. *Nature 357*:698–700.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res 22*:4673–4680.

Vogt G, Etzold T, Argos P. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J Mol Biol 249*:816–831.

Walker JE, Saraste M, Runswick MJ, Gay NJ. 1982. Distantly related sequences in the alpha- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J 1*:945–951.

West MW, Hecht MH. 1995. Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci 4*:2032–2039.

Whiteheart SW. 1994. N-ethylmaleimide-sensitive fusion protein: A trimeric ATPase whose hydrolysis of ATP is required for membrane fusion. *J Cell Biol 126*:945–954.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem 46*:573–639.

Wilson DW, Wilcox CA, Flynn GC, Chen E, Kuang WJ, Henzel WJ, Block MR, Ullrich A, Rothman JE. 1989. A fusion protein required for vesicle-mediated transport in both mammalian cells and yeast. *Nature 339*:355–359.

Wolfe KH. 1994. Similarity between putative ATP-binding sites in land plant plastid ORF2280 proteins and the FtsH/CDC48 family of ATPases. *Curr Genet 25*:379–383.

Wu CFJ. 1986. Jackknife, bootstrap and other resampling plans in regression analysis. *Ann Stat 14*:1261–1295.