ELSEVIER

Biochimica et Biophysica Acta 1415 (1999) 306–322

BIOCHIMICA ET BIOPHYSICA ACTA

BBA

The amino acid/auxin:proton symport permease family[1]

Gregory B. Young, Donald L. Jack, Douglas W. Smith, Milton H. Saier Jr. *

Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA

Received 2 July 1998; received in revised form 2 October 1998; accepted 7 October 1998

Abstract

Amino acids and their derivatives are transported into and out of cells by a variety of permease types which comprise several distinct protein families. We here present a systematic analysis of a group of homologous transport proteins which together comprise the eukaryotic-specific amino acid/auxin permease (AAAP) family (TC #2.18). In characterizing this family, we have (1) identified all sequenced members of the family, (2) aligned their sequences, (3) identified regions of striking conservation, (4) derived a family-specific signature sequence, and (5) proposed a topological model that appears to be applicable to all members of the family. We have also constructed AAAP family phylogenetic trees and dendrograms using six different programs that allow us to trace the evolutionary history of the family, estimate the relatedness of proteins from dissimilar organismal phyla, and evaluate the reliability of the different programs available for phylogenetic studies. The TREE and neighbor-joining programs gave fully consistent results while CLUSTAL W gave similar but non-identical results. Other programs gave less consistent results. The phylogenetic analyses reveal (1) that many plant AAAP family proteins arose recently by multiple gene duplication events that occurred within a single organism, (2) that some plant members of the family with strikingly different specificities diverged early in evolutionary history, and (3) that AAAP family proteins from fungi and animals diverged from the plant proteins long ago, possibly when animals, plants and fungi diverged from each other. The *Neurospora* protein nevertheless exhibits overlapping specificity with those found in plants. Preliminary evidence is presented suggesting that proteins of the AAAP family are distantly related to proteins of the large ubiquitous amino acid/polyamine/choline family (TC #2.3) as well as to those of two small bacterial amino acid transporter families, the ArAAP family (TC #2.42) and the STP family (TC #2.43).

Keywords: Transport; Amino acid; Auxin; Indole acidic acid; Plant; Fungus; Animal; Proton symport

1. Introduction

Amino acids serve as primary sources of organic nitrogen for the growth of many eukaryotic cells [1]. They additionally serve as neurotransmitters and hormones, allowing communication between cells and tissues within multicellular organisms [2,3]. Other amino acids allow adaptation to environmental change, especially to those causing organismal stress [4]. Such functions require the presence of transport systems that catalyze uptake into and release from specialized cell types [5].

Transport proteins catalyzing release of amino acids and amino acid-like hormones from the tissues of synthesis, and those catalyzing uptake into target

* Corresponding author. Fax: +1 (619) 534-7108; E-mail: msaier@ucsd.edu

[1] The accompanying review paper 'Phylogenetic characterization of novel transport protein families revealed by genome analysis' by M.H. Saier Jr. et al. will be published in Biochim. Biophys. Acta, Vol. 1422/1, February 1999 issue.

0005-2736/99/$ – see front matter
PII: S0005-2736(98)00196-5

tissues, probably differ with respect to their polarities, structures and energy coupling mechanisms [6,7]. In fact, even in simple bacteria, distinct transporters are known to catalyze amino acid uptake and release [8]. Permeases belonging to several universal families of primary and secondary carriers have been shown to catalyze amino acid uptake [9–11]. Among the universal secondary carrier families is the amino acid/polyamine/choline (APC) family (TC #2.3 [12,13]). In plants, members of this and other transport protein families allow secretion from sites of synthesis (e.g., the roots and leaves) as well as uptake into tissues that rely on external sources of amino acids for growth and development (e.g., seeds) [1,7]. Auxins such as indole acetic acid are specifically secreted from shoot apical tissues where they are synthesized and taken up into a variety of target tissues within the organism [3]. It is probable that only a small fraction of the essential plant transport proteins have as yet been identified.

Recent studies have resulted in the functional identification of the first animal member of the AAAP family [14]. This protein is responsible for packaging of the inhibitory neurotransmitter γ-aminobutyric acid (GABA) in GABA-specific intraneuronal vesicles of *Caenorhabditis elegans* and *Rattus norvegicus* (see Table 1). A 10 transmembrane spanner (TMS) topological model was proposed for this proton motive force (pmf)-driven, putatively GABA-specific transporter.

Table 1
Sequenced proteins of the amino acid/auxin permease family

Abbreviation[a]	Name or description	Source	Length	Accession
Aux1 Ath	Auxin transport protein	*Arabidopsis thaliana*	485	gbX98772
Aap1 Ath	Amino acid transporter 1	*Arabidopsis thaliana*	485	pirA48187
Aap2 Ath	Amino acid transporter 2	*Arabidopsis thaliana*	493	gbX71787
Aap3 Ath	Amino acid transporter 3	*Arabidopsis thaliana*	475	pirS51168
Aap4 Ath	Amino acid transporter 4	*Arabidopsis thaliana*	466	pirS51169
Aap5 Ath	Amino acid transporter 5	*Arabidopsis thaliana*	480	pirS51170
Aap6 Ath	Amino acid transporter 6	*Arabidopsis thaliana*	481	gbX95736
ProT1 Ath	Proline transporter 1	*Arabidopsis thaliana*	442	gbX95737
ProT2 Ath	Proline transporter 2	*Arabidopsis thaliana*	439	gbX95738
Orf1 Llo	Unidentified open reading frame	*Lilium longiflorum*	518	gbD21814
Aap1 Nsy	Amino acid transporter 1	*Nicotiana sylvestris*	462	gbU31932
Aap1 Rco	Amino acid transporter 1 (fragment)	*Ricinus communis*	284	gbZ68759
Aap1 Stu	Amino acid transporter 1 (fragment)	*Solanum tuberosum*	385	gbY09825
Aap2 Stu	Amino acid transporter 2 (fragment)	*Solanum tuberosum*	376	gbY09826
Aap1 Ncr	Neutral amino acid transporter 1	*Neurospora crassa*	470	pirS47892
Ybi9 Sce	Hypothetical 57.1 kDa protein	*Saccharomyces cerevisiae*	509	spP38176
Yeh4 Sce	Hypothetical 53.3 kDa protein	*Saccharomyces cerevisiae*	480	spP39981
Yeu9 Sce	Hypothetical 48.8 kDa protein	*Saccharomyces cerevisiae*	448	spP40074
Yii8 Sce	Hypothetical 53.7 kDa protein	*Saccharomyces cerevisiae*	490	spP40501
Yjx1 Sce	Hypothetical 65.3 kDa protein	*Saccharomyces cerevisiae*	602	spP47082
Yko6 Sce	Hypothetical 75.5 kDa protein	*Saccharomyces cerevisiae*	692	spP36062
Ynk1 Sce	Hypothetical 80.0 kDa protein	*Saccharomyces cerevisiae*	713	spP50944
Yan9 Spo	Hypothetical 73.1 kDa protein	*Schizosaccharomyces pombe*	656	spQ10074
Orf1 Cel	Cosmid F21D12.3	*Caenorhabditis elegans*	505	gbU23518
Orf2 Cel	Cosmid R02F2.8	*Caenorhabditis elegans*	494	gbU00055
Orf3 Cel	Cosmid C44B7.6	*Caenorhabditis elegans*	434	gbU28928
Ymj2 Cel	Hypothetical 43.2 kDa protein	*Caenorhabditis elegans*	389	spP34479
Unc47 Cel	Vesicular GABA transporter	*Caenorhabditis elegans*	486	spP34579
Unc47 Rno	Vesicular GABA transporter	*Rattus norvegicus*	525	gbAF030253
Orf1 Hsa	Transporter protein	*Homo sapiens*	504	gbU49082
Aap7 Ath	Amino acid transport protein	*Arabidopsis thaliana*	432	gbU39783

[a]The first 28 proteins tabulated were included in the detailed studies reported here. The sequences of the last three proteins were deposited in the database after completing these studies, and they were therefore not included.

In this paper we provide phylogenetic information concerning the AAAP family, one of the most divergent families of transporters found in eukaryotes. Because of the extensive sequence divergence of its members, the AAAP family provides an excellent test bed for evaluation of the various programs and methodologies currently used for generating multiple alignments and either phylogenetic trees (which include branch lengths) or dendrograms (which do not provide branch lengths). We have attempted to standardize the use of these programs to the extent possible by using the typical default gap penalty of 8. The results presented reveal that three of the four programs used for tree construction give remarkably similar results but that one of the tree programs, and both of the dendrogram programs used give substantially different results. Bootstrapping, the process by which identical parts of each sequence are randomized and new trees are generated, applied to two of the programs used, provides some level of confidence for specific branches within a tree, but lacks the ability to evaluate the assumptions upon which these dissimilar programs are based [15]. Based on the results obtained, we suggest that the TREE and neighbor-joining programs provide the most consistent methods for construction of phylogenetic trees.

In this report we present a systematic analysis of the sequences of the members of one functionally well-characterized family which we have designated the amino acid/auxin permease (AAAP) family, a designation based on the currently recognized substrate specificities of AAAP family members [1,3–7,14]. This family has been assigned the transport commission number 2.18 (TC #2.18). We report analyses that allow us to suggest that the common ancestor of all currently recognized extant members of the AAAP family was present in a primordial eukaryote. A primary set of paralogues presumably arose from this primordial protein by early gene duplication events, providing proteins of dissimilar specificities. Orthologues arose due to speciation events that gave rise to the currently recognized eukaryotic kingdoms, animals, plants, fungi and protozoans. Finally, we suggest that as the multicellular state developed and became increasingly complex, proliferation of these systems occurred, particularly in plants, probably due to late gene duplication events that (1) allowed coordination of organismal growth, (2) facilitated communication between the tissues, and (3) paved the way for the development of exquisite tissue-specific regulatory controls. Preliminary evidence is presented suggesting that the AAAP family is a member of the large and ubiquitous APC superfamily.

2. Evaluation of phylogenetic tree programs

Progressive alignment distance matrix, parsimony, and maximum likelihood are the three major methods used in phylogenetic tree construction from protein sequences. Progressive alignment is used because it is not practical computationally to use rigorous dynamic programming methods for alignment of more than a few sequences. In progressive alignment, sequences are added sequentially to a growing multiple alignment, starting with alignment of the two most similar sequences. Alignments of sequences are done using distance matrix methods, usually using variations of the Needleman-Wunsch [16] algorithm. Pairwise alignments are initially performed to determine approximate similarity of the sequences, and often an approximate phylogenetic tree called a guide tree is created from these data. This information is used to determine the order of addition of sequences to the multiple alignment. A major problem is how to handle gaps between sequences. Usually gaps, once inserted, are kept throughout the alignment. The Feng-Doolittle [17–19] programs use a standard gap penalty, as does the GCG program PILEUP. Clustal W [20], however, uses position-specific gap penalties and other heuristics to optimize gap introduction and extension.

In distance matrix methods, the multiple sequence alignment is then used to generate a phylogenetic tree. In the Feng-Doolittle [17–19] program TREE, the Fitch-Margoliash [21] method is used to determine branching order of the sequences, and the branch lengths are then calculated using a least-squares approach [22]. The Fitch-Margoliash method allows different mutation rates for each of the tree branches, a distinct advantage over the often-used UPGMA (unweighted pair group method of averages [23]) method which assumes a constant molecular clock throughout the tree. UPGMA nevertheless

often works surprisingly well and is sometimes used for guide tree construction. The GCG program PILEUP uses UPGMA in its tree construction. Clustal W uses rooted neighbor joining [24] for guide tree construction. In neighbor joining, a unique unrooted tree is generated by sequentially joining pairs of neighboring sequences. Neighbors are sequences (or taxa) connected through a single internal node. Two neighbors, once joined, are now a new neighbor to a third sequence. Branch lengths are determined by the Fitch-Margoliash [21] method. This yields the tree with the smallest sum of the branch lengths (minimum evolution), and hence is most likely to exhibit the true branching pattern [25]. The PHYLIP [26] program NEIGHBOR can use either UPGMA or the neighbor-joining method.

Parsimony methods infer the sequence of the ancestral species and deduce a tree by requiring the minimum number of mutational changes throughout the tree. This is generally done residue by residue. Either multiply aligned sequences or unaligned sequences can generally be used. The PAPA (parsimony after progressive alignment) programs [27] use a multiple sequence alignment as input and then consider four residues at a time for a given position in the alignment, selecting the optimal tree among the three possible trees for four taxa. This latter 4-3 approach is commonly used in parsimony programs, but the rules used for what constitutes a mutational change vary between programs. The PHYLIP program PROTPARS, for protein sequence parsimony, insists that any amino acid changes be consistent with the genetic code; a change between two amino acids via a third is counted as at least two mutational changes. Branch lengths are not normally obtained using parsimony methods. However, PAPA3 [27] uses mutational probability parameters for branch length calculations. Parsimony methods are best used for closely related sequences, between which relatively few mutational changes are expected to have occurred. For more distantly related sequences, multiple mutations have likely occurred at a given site, rendering parsimony methods more difficult to use.

Maximum likelihood [28,29] methods use a specific probability model to determine the tree topology which gives the maximum likelihood for obtaining a given set of multiply aligned sequences. The most commonly used probability model for sequence evolution studies is that based on a Poisson process. In the Felsenstein [30] method, both tree topology and branch lengths can be obtained. This method can permit variation of mutation rate between tree branches, as used in the PHYLIP programs DNAML, for DNA sequences, and PROTML, for protein sequences. However, the methods typically are computationally severe, since they examine the likelihood of many tree topologies for the given multiple sequence alignment. Using a fixed molecular clock reduces the computational requirements, as is done in the PHYLIP program DNAMLK. Versions of these programs with reduced computational requirements are available by anonymous ftp (FASTDNAML [31], gopher://megasun.BCH.UMontreal.CA:70/11/CMB/Phylogeny/fastDNAml; NucML and ProtML in the MOLPHY program package [32], ftp://sunmh.ism.ac.jp/pub/molphy/; PUZZLE [33], http://www.zi.biologie.uni-muenchen.de/~strimmer/puzzle.html).

In general, we find that use of the TREE or Clustal W programs to be the most useful in initial tree analyses. Both programs permit variations in the molecular clock between tree branches. Ancillary programs can be used with TREE to crop sequences and find approximate boundaries. Trees can be generated by Clustal W which omit residues in gap regions or which attempt to account for multiple substitutions at a given residue. TREE and Clustal W use different rules for introduction and extension of gaps, making comparison of the trees generated useful. Clustal W includes rules that can use protein tertiary structural information for gap introduction and extension. Clustal W also permits bootstrapping calculations to determine the relative probability for occurrence of internal nodes for the tree. Parsimony methods are difficult if the sequences are distantly related and require additional methods for branch length determination. Maximum likelihood methods suffer from the computational severity of the programs. Results presented here are consistent with these properties of these methods.

3. Proteins of the AAAP family

Table 1 lists the 28 proteins of the AAAP family

that were included in our study, and three more have been sequenced since completion of the reported studies. Ten of these proteins are from *Arabidopsis thaliana*, seven are from *Saccharomyces cerevisiae*, and five are from *Caenorhabditis elegans*. From these observations, it is clear that a single organism may possess numerous AAAP family paralogues. While both *A. thaliana* and *C. elegans* undoubtedly possess additional AAAP family paralogues, the availability of the complete sequence of the *S. cerevisiae* genome suggests that no additional paralogues will be found in this organism. No recognizable AAAP family homologues were identified in bacteria, archaea, or lower eukaryotes.

The protein abbreviations to be used in this study as well as accession numbers that allow easy access to the sequences and primary references are provided in Table 1. Additionally, it can be seen that most of the AAAP family members have similar sizes. Thus, the 10 *A. thaliana* proteins are 432–493 amino acyl residues in length, and the five putative *C. elegans* proteins are 389–505 residues long. The eight sequenced yeast proteins exhibit a much wider spectrum of sizes, varying from 448 to 713 residues. Most of this size variation occurs in the hydrophilic N- and C-terminal regions of these proteins, although variation also occurs towards the C-terminal regions of the hydrophobic domains, particularly in the loops between transmembrane spanners VIII and IX (see below).

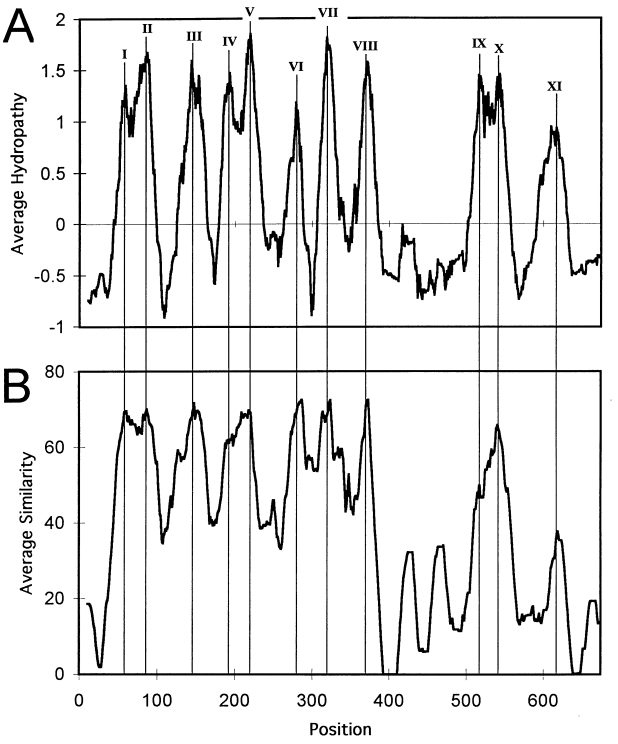

Fig. 1. Average hydropathy (A) and average similarity (B) plots for 28 sequenced members of the AAAP family. N-terminal hydrophilic domains have been artificially removed for the analyses presented. A sliding window of 20 residues was used in both plots. The hydropathy analyses were based on the algorithm of Kyte and Doolittle [41]. Alignment position is provided on the *x*-axis. The program used was a local unpublished C program written by R. Blewitt (Department of Biology, UCSD, La Jolla, CA 92093).

4. Topological analyses

Fig. 1 presents average hydropathy (A) and average similarity (B) plots for the 28 sequenced members of the AAAP family. In both plots, the hydrophilic domains present at the N-terminal regions of some of these proteins were artificially removed before analysis. The average hydropathy plot (Fig. 1A) reveals 11 distinct peaks which correspond to 11 putative transmembrane α-helical spanners (TMSs) (I–XI). When average hydropathy plots were constructed for each of the subfamilies of the AAAP family (see below), similar plots were obtained. We therefore consider it likely that the proteins of the AAAP family uniformly possess 11 rather than the more usual 12 TMSs. An 11 TMS topology has been documented for the Mtr tryptophan permease of *Escherichia coli* [34]. This *E. coli* protein is a member of the aromatic amino acid permease (ArAAP) family (TC #2.43) and could not be shown to be homologous to any protein of the AAAP family using statistical methods to analyze the protein sequences [11,35].

The average similarity plot shown in Fig. 1B reveals that without exception, each peak of hydrophobicity has a corresponding peak of similarity. Thus, as has been observed for several permease families [35,36], although seldom with the consistency observed for the AAAP family, the transmembrane spanners are better conserved than the hydrophilic loops.

5. Multiple alignment

Fig. 2 shows a portion of the complete multiple alignment upon which the plots shown in Fig. 1 were based. This region encompasses putative TMS VII. As shown in Fig. 2, the sequences are highly divergent, and no single position is fully conserved. In fact, no one residue position proved to be fully conserved in the entire alignment. However, no gaps are found in the 30 residue stretch portrayed, and seven residues are found in the consensus sequence at the end of the region shown. Particularly noteworthy is the terminal G Y–A F G consensus sequence motif. The first G in this six residue sequence is conserved in all but five of the 28 proteins presented; the Y is substituted only by F; the A is found in 15 of the 28 proteins; the F is substituted only by Y except in one protein where a V is found; and the terminal G is found in all but one protein. It is clear from these observations that the TREE program used to generate the multiple alignment has correctly aligned the sequences.

The complete multiple alignment was examined to reveal any potential characteristic features of the 11 putative TMSs. When the sequences of these TMSs were depicted in helical wheel configuration, most of them proved to exhibit weakly amphipathic character with strongly hydrophobic residues predominating on one side and semipolar residues predominating on the other. However, none of these putative TMSs exhibited characteristics that led to the clear suggestion that they serve as channel lining segments.

```
                    1                   11                  21                  30
Aux1 Ath    (263) Q K F K Y I Y L M A T L Y V F T L T I P S A A A V Y W A F G
Aap1 Ath    (278) K A M K R A S L V G V S T T T F F Y I L C G C I G Y A A F G
Aap2 Ath    (286) K T M K K A T K I S I A V T T I F Y M L C G S M G Y A A F G
Aap3 Ath    (269) K T M K K A T L V S V S V T T M F Y M L C G C M G Y A A F G
Aap4 Ath    (259) K T M K I A T R I S I A V T T T F Y M L C G C M G Y A A F G
Aap5 Ath    (274) N T M R K A T F V S V A V T T V F Y M L C G C V G Y A A F G
Aap6 Ath    (276) K A M K R A S L V G V S T T T F F Y M L C G C V G Y A A F G
ProT1 Ath   (254) K N M M K A L Y F Q F T A G V L P M Y A V T F I G Y W A Y G
ProT2 Ath   (251) K N M M K A L Y F Q F T V G V L P M Y A V T F I G Y W A Y G
Orf1 Llo    (326) V P M W R G V K V A Y V L I A F C L F P V A L I G F W S Y G
Aap1 Nsy    (264) G P M W K G V L V A Y I I V A L C Y F P V A I I G Y W I F G
Aap1 Rco    (216) K T M K K A T L I S V A V T T L F Y M L C G C F G Y A A F G
Aap1 Stu    (233) K T M K R A T L I S V A V T T V F Y M L C G C F G Y A A F G
Aap2 Stu    (232) K V M K R A S L A G V S T T T L F Y V L C G T I G Y A A F G
Aap1 Ncr    (270) S D Y K K S I V A L G L I E I F I Y T V T G G V V Y A F V G
Ybi9 Sce    (285) K V I R R I P I F A I V L A Y F L Y I I I G G T G Y M T F G
Yeh4 Sce    (289) A K F T R L T H I S I I I S V I C C A L M G Y S G F A V F K
Yeu9 Sce    (225) E H V M K I P L I A I S L A L I L Y I A I G C A G Y L T F G
Yii8 Sce    (220) E N I T F V I N N S I S L S T A L F L I V G L S G Y L T F G
Yjx1 Sce    (418) D K F K D C L K T T Y K I T S V T D I G T A V I G F L M F G
Yko6 Sce    (511) K H F R P S L S A V M C I V A V I F I S C G L L C Y A A F G
Ynk1 Sce    (512) E K F P L V L A L V I L T A T I L F I S I A T L G Y L A Y G
Yan9 Spo    (485) K N L P K L L T G V M A A I S L L F I S I G L L S Y A A F G
Orf1 Cel    (290) R D F T K S I F A G F L G V V I L Y L P L C I F A F V V Y G
Orf2 Cel    (244) A H F V H S V V L A I I F C T M L Y M C I A V G G Y F V Y G
Orf3 Cel    (249) K G P F G V L S V G V G M V V V I Y S F A G F F G F L T Y G
Ymj2 Cel    (196) L A P F G V L S T T M I I C T A F M T A L G F F G Y T G F G
Ynx6 Cel    (271) A Q F N V M L K W S H I A A A V F K V V F G M L G F L T F G
Consensus:        - - - - - - - - - - - - - - - - - - - Y - - - G - - G Y - A F G

                                                    Putative TMS# VII
```

Fig. 2. A short portion of the complete multiple alignment for 28 sequenced AAAP family proteins analyzed in this study. The region shown encompasses putative TMS VII as indicated by the bar at the bottom of the alignment. The residue number of the first residue depicted in each protein is provided in parentheses after the abbreviated designation of the protein as indicated in Table 1. Alignment position is provided above the alignment, and the consensus sequence (15 of the 28 residues at any one position conserved) is provided below it. The TREE program [18] was used to align the sequences. The complete multiple alignment is available upon request from M.H.S.

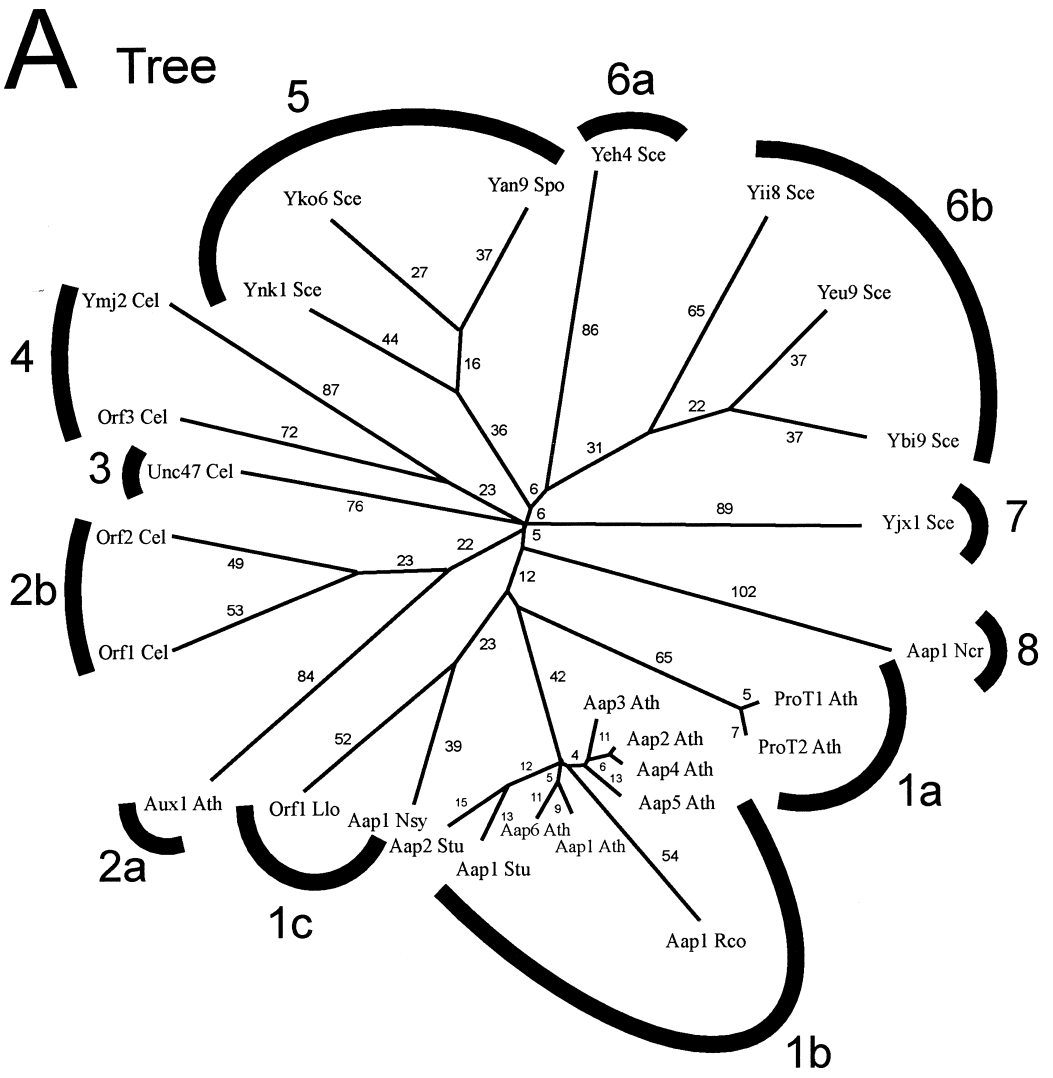

Fig. 3. Phylogenetic trees for the proteins of the AAAP family. Protein abbreviations are provided in Table 1. Phylogenetic distance is approximately proportional to branch length. The trees were generated (A) with the TREE program of Feng and Doolittle [18], (B) with the neighbor-joining program, version 3.5c, of Saitou and Nei [24], (C) the Clustal W program [20], and (D) with the PAPA3 program [27]. In D, Aap1 Rco was omitted due to its being a fragment. In C, bootstrapping has been applied, and the bootstrap values (based on 1000 random runs) provided are underlined. Blowups of two congested regions are provided. These trees are based on nearly complete multiple alignments, a portion of one of which is presented in Fig. 2. Clustering patterns are presented numerically going around the trees in the clockwise direction.

The lack of distinctive character of these TMSs may explain the flexibility of many of AAAP proteins to accommodate substrates of strikingly different structures.

6. Signature sequence for the AAAP family

The segment of the multiple alignment shown in Fig. 2 represents one of the most highly conserved portions of the AAAP proteins (Fig. 1). From this region, a potential signature sequence was derived. This signature sequence is: [PFYMLIV]-[LIVKTRMPWNF]-[RKHYLIVPFGD]-[LIVSACGYM]-[LIVYTSP]-X-[LIVWAFGMTN]-[AGSTLIVQ]-[LIVMFYAGT]-[LIVSACTGK]-[LIVMATGFY]-X-[LIVSTAFY]-[LIVFATM]-[LIVFCPT]-X-[LIVMAFYST]-[LIVFGAPSTC]-[LIVFMASCT]-[ACGT]-[LIVMFY-

ACGST]-[LIVMFAGTS]-[GCSAV]-[FY]-[LIVFWAMT]-[LIVASTGMF]-[FYV]-[GK] (X = any residue; alternative residues at any one position are presented in brackets).

This sequence, unusual with respect to its degeneracy, was screened against the SwissProt database and was found to retrieve only members of the AAAP family. It is thus a bona fide signature sequence and can be used to identify new members of the family as they become sequenced.

7. Phylogenetic trees and dendrograms for the AAAP family

A phylogenetic tree, derived using the TREE program of Feng and Doolittle [18] is shown in Fig. 3A. Several points are noteworthy. First, the six general amino acid permeases (AAP1–6) of *A. thaliana* as well as the two permeases of *Solanum tuberosum* and that of *Ricinus communis* cluster tightly together (cluster 1b). The long branch length of the *R. com-*

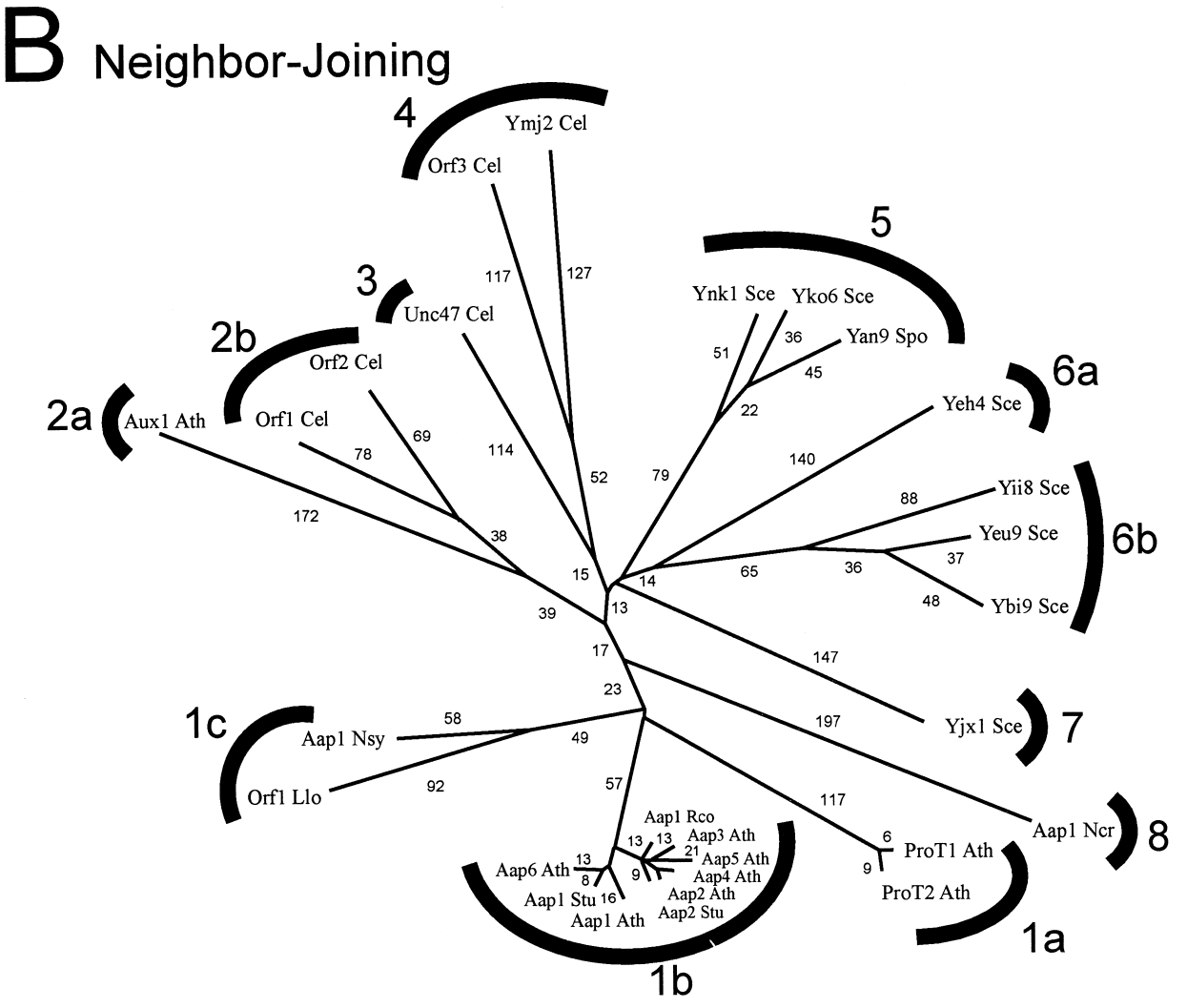

Fig. 3 (*Continued*)

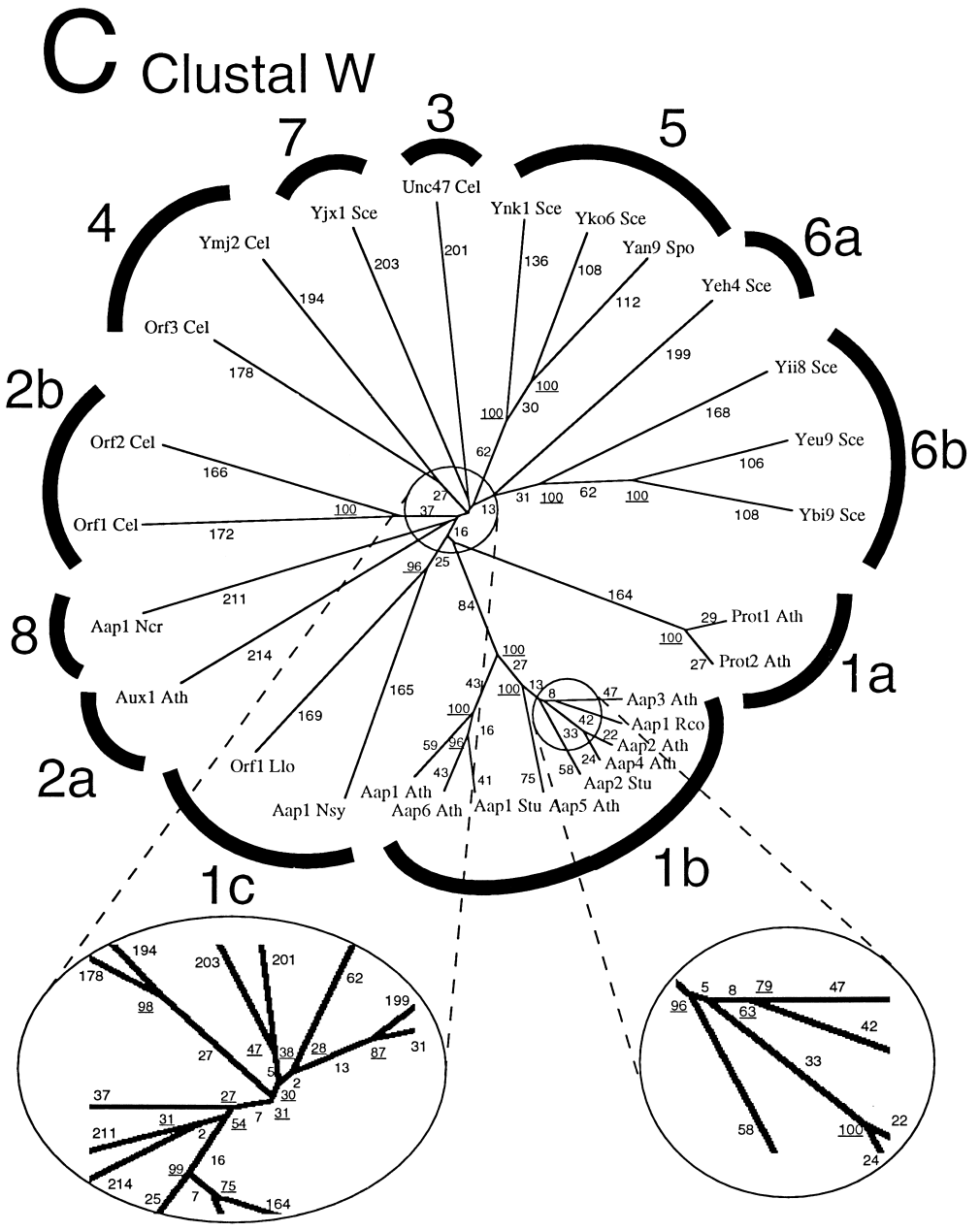

Fig. 3 (*Continued*)

munis protein is artifactual, resulting from the fact that the full sequence is not available (see Table 1). The tight clustering and short branch lengths reflect the high degree of sequence similarity among these proteins and indicate that all of these paralogues and orthologues arose recently in evolutionary time.

On the same main branch with the large plant AAAP cluster (cluster 1b) are found the two stress related proline-specific permeases (ProT1 and 2) of *A. thaliana* (cluster 1a). Branching from a point even closer to the trunk of the tree, are found the proteins of *Nicotiana sylvestris* and *Lilium longiflorum* (cluster 1c). These proteins must have diverged from the other general amino acid permeases relatively early during the evolution of the AAAP family.

The putative auxin transporter, Aux1 of *A. thali-*

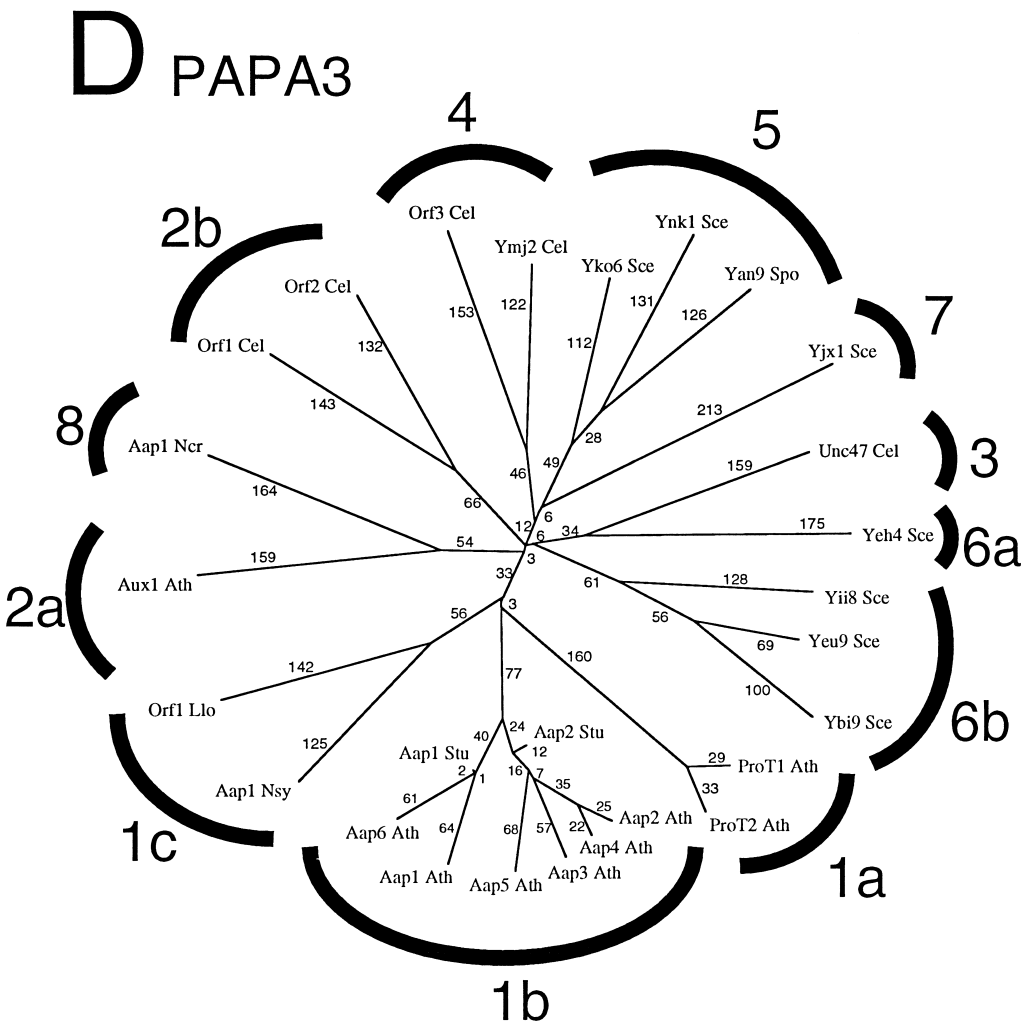

Fig. 3 (*Continued*)

ana (cluster 2a), surprisingly clusters loosely with two uncharacterized Orfs from *C. elegans* (cluster 2b). The branch bearing these three proteins stems from a point close to the center of the unrooted tree.

All remaining proteins are from the worm, *C. elegans* (clusters 3–4), the two yeast species, *S. cerevisiae* and *Schizosaccharomyces pombe* (clusters 5–7) and the fungus, *Neurospora crassa* (cluster 8). None of these proteins has been functionally characterized. Because of the deep branching pattern, the specificities of these proteins cannot be surmised. It is obvious, however, that (1) Orf1 and Orf2 of *C. elegans*, (2) Ynk1, Yko6 and Yan9 of yeast, and (3) Yeu9 and Ybi9 of yeast each comprises a tight cluster of proteins which may serve the same function or closely related functions. The general clustering of proteins derived from organisms of the same kingdoms (plants, animals or fungi) is particularly worthy of note.

The phylogenetic tree shown in Fig. 3B was derived using a distinct program, the neighbor-joining program [24]. It is apparent that the two trees shown in Fig. 3A and B are very similar in configuration. In fact, the primary branching orders (except in protein cluster 1b) are the same. These two trees differ primarily with respect to their branch lengths. Variation in branch length is seldom more than 2-fold.

The phylogenetic trees shown in Fig. 3C and D were derived using the Clustal W and PAPA3 programs. The latter program uses maximum parsimony after progressive alignment of protein sequences [27]. Clustering patterns shown in Fig. 3C and D are sim-

ilar but not identical to those observed for Fig. 3A and B. Specifically, clusters 1a, 1b, 1c, and 2a as well as clusters 2b, 4, 5, 6a and 6b appear similar, but cluster 8 is loosely associated with cluster 2a. Additionally, cluster 3 is loosely associated with cluster 7 in Fig. 3C, and with cluster 6a in Fig. 3D. Finally, cluster 7 is more closely related to clusters 4 and 5 in Fig. 3D than in Fig. 3A, 3B or 3C. Thus, the Clustal W and PAPA3 programs display greater divergence in clustering patterns than observed for the TREE and neighbor-joining programs. It should be noted that most of these differences reflect short branch lengths found deep within the trees.

The tree generated for the AAAP protein sequences using Clustal W shown in Fig. 3C provides both branch lengths (plain numbers) and bootstrap values (underlined numbers). The distant region of short branch lengths and the cluster containing five of the 1b external nodes are shown as expanded figures. The bootstrap values, based on 1000 random runs, provide a measure of uniqueness for each internal tree node. Thus, a bootstrap value of 100 means this internal node appeared in all 1000 trees generated by the bootstrap procedure. Conversely, a bootstrap value of 20 means this node appeared in only 20% of the 1000 trees so generated. Note that all of the bootstrap values of less than 50 appear in the distant region of short branch lengths. This means that a variety of tree topologies agree equally well with the Clustal W generated multiple sequence alignment upon which the phylogenetic tree is based. This argues that the alignment data are insufficient to predict a unique phylogenetic tree for these sequences. This is equivalent to arguing that the short branch lengths involved are essentially zero, and that the branches involved diverged essentially at the same time. When viewed in this manner, all of the trees shown in Fig. 3A–D are nearly equivalent.

Fig. 4A,B presents dendrograms for the same 28 proteins obtained using the PILEUP (A) [23] and PROTPARS (B) [37,38] programs. In the dendrogram shown in Fig. 4A, clustering patterns are similar to but somewhat different from those revealed in Fig. 3A–D. Moreover, the two dendrograms shown in Fig. 4A,B differ significantly from each other. Although bootstrap values for each node are provided in Fig. 4B, the differences in dendrogram configuration primarily reflect the different assumptions that were made in designing the two programs. Bootstrapping [39] does not provide a means of evaluating the reliability of the assumptions upon which a particular program is based (see Section 9). It should be noted, however, that all of the major differences in tree and dendrogram topology correspond to those nodes with low bootstrap values.

In comparing the dendrograms shown in Fig. 4A,B with the trees shown in Fig. 3, the following

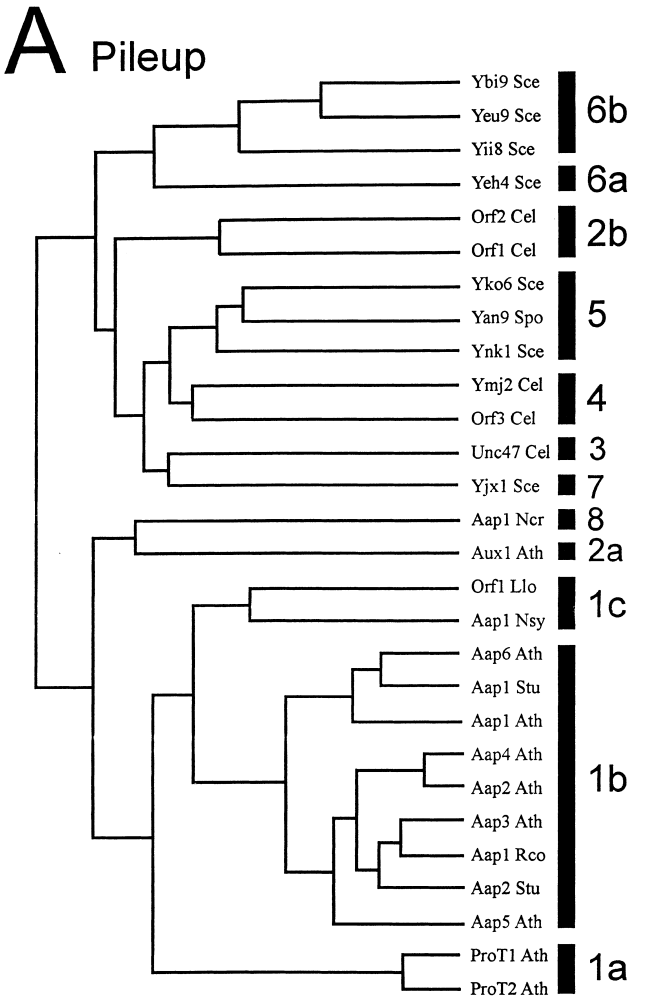

Fig. 4. Dendrograms for the proteins of the AAAP family. The dendrograms were generated using (A) the PILEUP program [23] in the GCG package [42] and (B) the PROTPARS parsimony method [37,38]. In the latter dendrogram, bootstrap values (expressed in percentage) for each node are provided. The multiple alignment was bootstrapped 500 times using SEQBOOT. It was then analyzed by PROTPARS. A consensus tree was then constructed through use of the program CONSENSE. All three programs are available in the PHYLIP package [43]. Clusters are indicated numerically as shown in Fig. 3.

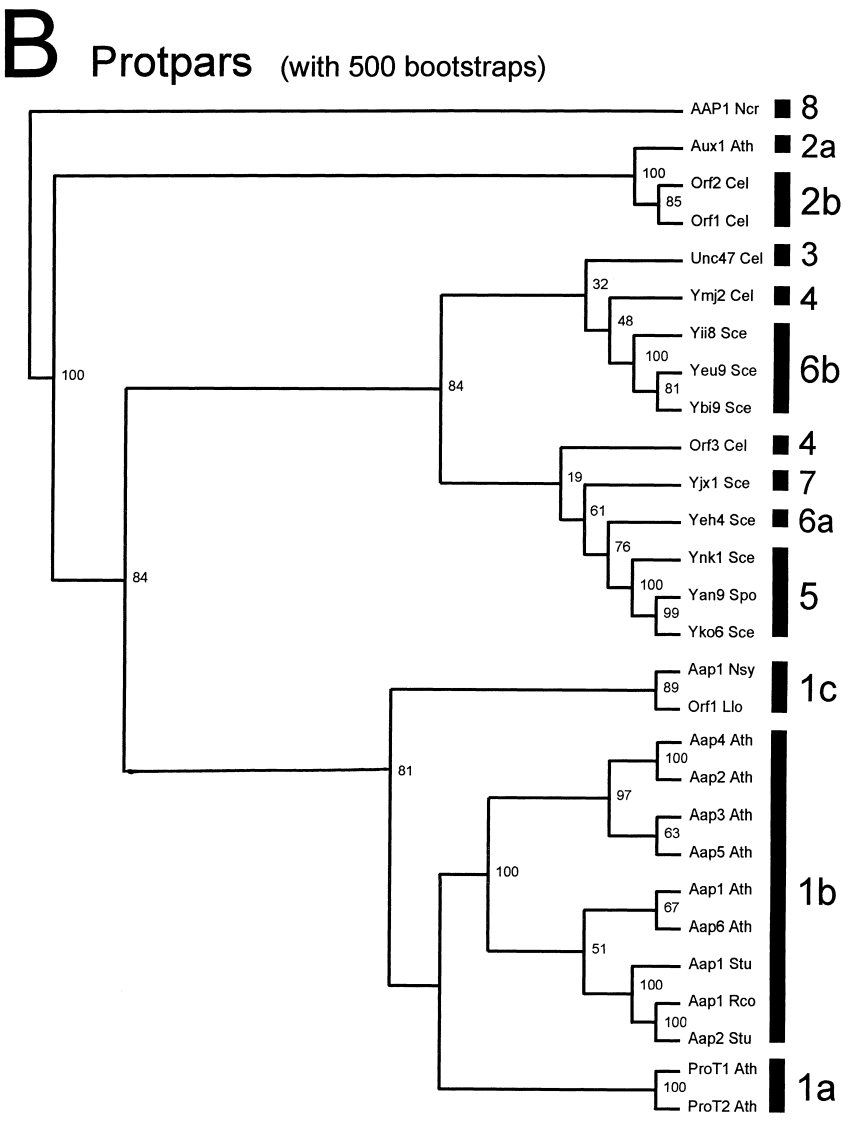

Fig. 4 (*Continued*)

similarities and differences are worthy of note. First, clusters 1a, 1b and 1c contain the same proteins in the two dendrograms and in the four trees shown in Fig. 3. While clusters 1a and 1c are the same in all figures, 1b differs in the clustering patterns observed. Second, clusters 2a and 2b are next to each other in the dendrogram shown in Fig. 4B (as observed in the trees shown in Fig. 3A,B), but they are widely separated from each other in the dendrogram shown in Fig. 4A. Third, cluster 3 groups loosely with cluster 7 in Fig. 4A but more tightly with clusters 4 and 6b in Fig. 4B. In Fig. 3, none of these proteins cluster tightly together. Fourth, the proteins of cluster 4 are grouped very loosely with cluster 5 proteins in Fig. 4A but with clusters 3, 5, 6 and 7 in Fig. 4B. Fifth, cluster 5 proteins are grouped together in the two dendrograms shown in Fig. 4 as in the trees shown in Fig. 3. Sixth, clusters 6a and 6b are together in Fig. 4A in agreement with the trees shown in Fig. 3. However, the proteins of clusters 6a and 6b are distant from each other in the dendrogram shown in Fig. 4B. Seventh, Yjx1 Sce, which alone comprises cluster 7, is portrayed as being distantly related to Unc47 Cel (cluster 3) in Fig. 4A but to some of the proteins in clusters 4, 5 and 6a in Fig. 4B. Finally, Aap1 Ncr (cluster 8) is distantly related

Table 2
Representative members of four suggested families of the amino acid-polyamine-choline (APC) superfamily

Family abbreviation	TC#	Protein abbreviation	Name or description of protein	Organism	Size (no. residues)	Database and accession number
AAAP	2.18	Orf Cel	Cosmid C44B7.6	*Caenorhabditis elegans*	434	gbU28928
AAAP	2.18	Ynk1 Sce	Hypothetical 80.0 kDa protein	*Saccharomyces cerevisiae*	713	spP50944
APC	2.3	CadB Eco	Cadaverine:lysine antiporter	*Escherichia coli*	444	spP23891
APC	2.3	PotE Eco	Putrescine:ornithine antiporter	*Escherichia coli*	439	spP24170
STP	2.43	TdcC Eco	Threonine permease	*Escherichia coli*	443	spP11867
STP	2.43	SdaC Eco	Serine permease	*Escherichia coli*	429	spP36559
ArAAP	2.42	TyrP Hin	Tyrosine-specific transport protein	*Haemophilus influenzae*	400	spP44727
ArAAP	2.42	TyrP Eco	Tyrosine permease	*Escherichia coli*	403	spP18199

to all other members of the family in all of the figures except Fig. 3D.

The results presented in Figs. 3 and 4 lead us to suggest that the trees shown in Fig. 3 provide a more reliable index of the phylogenetic relationships of the proteins in the AAAP family than do the dendrograms shown in Fig. 4 because they provide branch lengths. The dendrograms tend to group more proteins together than is justified by their degrees of sequence similarity. The major discrepancies between the two dendrograms shown in Fig. 4 and between these two dendrograms and the phylogenetic trees shown in Fig. 3 can be attributed to artificial clustering of proteins which in fact should not cluster together at all. It should be noted that bootstrap values presented in both Fig. 3C and Fig. 4B are generally low (see comments above) when observable differences are noted between the tree clustering patterns.

8. Possible common ancestry of the AAAP family to other amino acid transporter families

Several currently recognized families of transporters exhibit exclusive specificities for amino acids and their derivatives ([10]; see our web site (http://www-biology.ucsd.edu/~msaier/transport/titlepage.html). The members of each of these families exhibit sufficient sequence similarity to allow establishment of homology for all members of the family, but insufficient sequence similarity to establish homology with any member of the other families (see [11] for criteria for establishing homology). We have examined these amino acid-specific families for sequence, topological and functional similarities and have found that some of them exhibit similarities that convince us of a probable common evolutionary origin. Representative members of four such families are presented in Table 2. These families are the AAAP family (TC #2.18) analyzed here, the amino acid/polyamine/choline (APC) family (TC #2.3), the serine/threonine permease (STP) family (TC #2.43) and the aromatic amino acid permease (ArAAP) family (TC #2.42). While the AAAP family is eukaryotic-specific, the APC family is ubiquitous, and the STP and ArAAP families are so far restricted to bacteria. In the present study, two members of each family were selected for comparison. As summarized in Table 2, all of these proteins are 400–443 amino acyl residues in length except for one selected yeast member of the AAAP family which possesses a 300 residue N-terminal hydrophilic extension. This extension was removed in the analyses reported.

Fig. 5 shows the average hydropathy (A) and average similarity (B) plots for the eight representative members of the proposed APC superfamily, based on the complete multiple alignment for these proteins as generated with the TREE program. The striking similarity of the hydropathy plot with that shown for the AAAP family in Fig. 1A is worthy of note. Thus, in both Fig. 1A and Fig. 5A, (a) peaks I and II are barely separated from each other, (b) the distances between peaks II and III and the depths of these troughs are greater than for other parts of these plots; (c) the spacings between all of the first eight hydrophobic peaks are nearly the same; (d) large relatively hydrophilic regions separate peaks VIII and IX, and (e) peaks IX are closer to peaks X

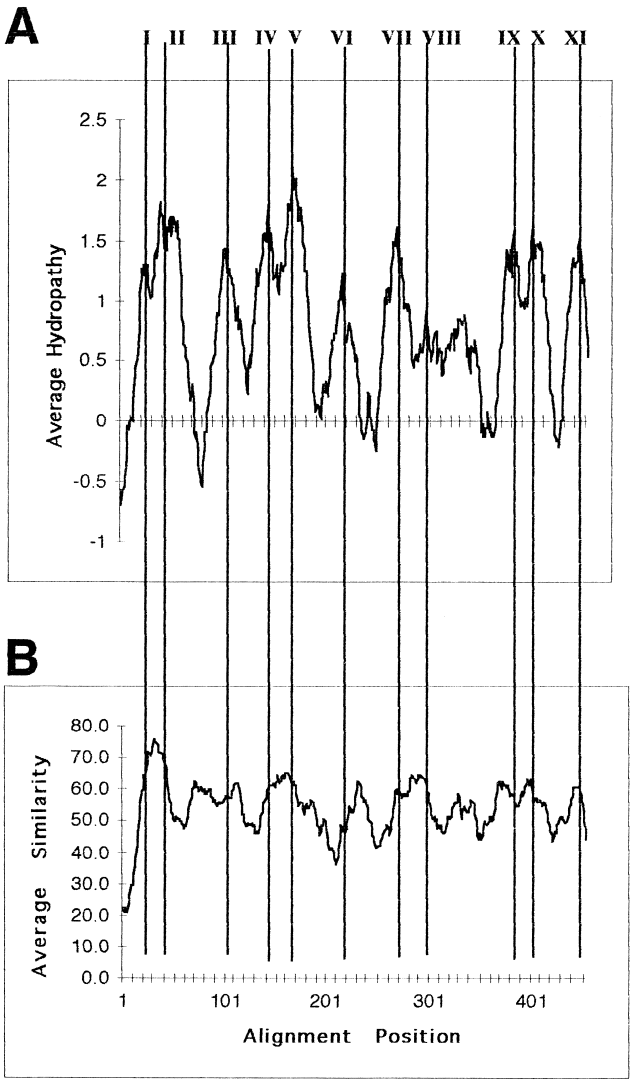

Fig. 5. Average hydropathy (A) and average similarity (B) plots for the eight representative members of the APC superfamily (see Table 2). The N-terminal hydrophilic domain of the yeast AAAP family member Ynk1 was artificially removed for the analyses presented. A sliding window of 21 residues was used in both plots. Programs used and format of presentation are as for Fig. 1.

than peaks X are to peaks XI. Thus, the topologies of these proteins are very similar. It should be noted that while in Fig. 1 there is no evidence for a hydrophobic peak between peaks VIII and IX, there may be a weakly hydrophobic peak at this position in Fig. 5. The average similarity plot in Fig. 1B is substantially different from that in Fig. 5B.

As shown in Fig. 5B, the first putative TMS is by far the best conserved portion of these proteins. Fig. 6 presents the portion of the complete multiple alignment corresponding to this peak of similarity. The sequences aligned without the introduction of a single gap, and four residues (three Gs and one P) are fully conserved. Moreover, in 12 of the 27 positions shown, the nature of the residue at any one position is constant (i.e., either hydrophobic or semipolar). This striking degree of sequence similarity is not likely to have arisen by chance.

Fig. 7 shows a phylogenetic tree for the eight proteins presented in Table 2. As expected, each pair of proteins representing one of the four sequence divergent families cluster together. The tree suggests that the two prokaryotic families, the STP and ArAAP families, are more closely related to each other than they are to the AAAP or APC families. The AAAP and APC families are more distant from each other than from either of the bacterial-specific families.

9. Conclusions

The AAAP family exhibits minimal sequence conservation in spite of the fact that all of its members are derived from eukaryotes. Hydropathy analyses suggested that these proteins uniformly possess an unusual 11 TMS topology instead of the more common 12 TMS topology. Interestingly, this feature is shared by a small family of bacterial aromatic amino acid permeases (ArAAPs; TC #2.42) that include the high and low affinity tryptophan permeases of *E. coli*, Mtr and TnaB respectively, as well as the *E. coli* tyrosine-specific permease, TyrP [34]. The strikingly similar topologies of the prokaryotic ArAAP family members and the eukaryotic AAAP family members as well as the overlapping specificities of these groups of permeases, particularly of the Mtr protein of *N. crassa* and the Mtr protein of *E. coli*, led to the possibility that these proteins might share a common evolutionary origin. This possibility, however, could not be established on the basis of statistical analyses of sequence similarities (see [11]).

In this paper we present evidence that the AAAP and ArAAP families are related to each other, and also to two other families, the prokaryotic STP family (which also appears to exhibit 11 TMSs) and the large and ubiquitous APC family (which appears to exhibit 10–14 TMSs [12]). The evidence comes from

Family	Protein	Residue	Sequence alignment
			* * * *
ArAAP	TyrP Hin	(17)	**I G A G M L A M P** L A A A G V **G** F S V T L I L **L I** G L
ArAAP	TyrP Eco	(08)	**I G A G M L A M P** L T S A G I **G** F G F T L V L **L L** G L
STP	TdcC Eco	(33)	**I G A G V L F F P** I R A G F G **G** L I P I L L M **L V** L A
STP	SdaC Eco	(34)	**I G A G V L F L P** I N A G V G **G** M I P L I I M **A I** L A
AAAP	Ynk1 Sce	(312)	**I G T G V L F L P** N A F H N G **G** L F F S V S M **L A** F F
AAAP	Orf Cel	(35)	**L G A G C F S V P** L A F K Q S **G** Y V S G L V I **I V** V L
APC	CadB Eco	(21)	**M G S G I A L L P** A N L A S I **G** G I A I W G W **I I** S I
APC	PotE Eco	(23)	**M G S G I I M L P** T K L A E V **G** T I S I I S W **L V** T A
Concensus:			I G A G V L F L P L A A A - G G - I - I L - M L V - L

Fig. 6. Partial multiple alignment of representative members of the proposed APC superfamily. Two members of each of the four families (AAAP, APC, STP and ArAAP) were arbitrarily selected for inclusion. The family and protein abbreviations are as specified in Saier [10,11] and our web site (see Table 2). The residue numbers present in parentheses represent the first residue shown in each of the eight proteins. Fully conserved residues are indicated by asterisks above the aligned sequences, while the dominant residue (at least three positions conserved) (Consensus) is indicated below the aligned sequences. Positions in which the amino acid type is conserved are presented in bold print. The TREE program of Feng and Doolittle [18] was used to generate the alignment.

several distinct lines of reasoning. First, all of these permeases are specific for amino acids and their derivatives. Second, all appear to function by essentially the same mechanism (i.e., proton symport). Third, they exhibit very similar topologies based on hydropathy analyses. Fourth, all of these proteins exhibit detectable sequence similarity, particularly in a region near their N-termini as shown in the multiple alignment presented in Fig. 6. It is important to note that the highest degree of sequence similarity is observed in the N-terminal regions of many sequence divergent families of secondary carriers [35,36]. Finally, using the PSI BLAST program [40], motif searches provided additional evidence that these four families might be distantly related to each other, and therefore might comprise a sequence diverse superfamily. The evidence that these (and possibly other amino acid specific transporter families) are related is therefore very substantial.

A startling observation concerns the variation in the substrate specificities of the AAAP family members. Thus, some of these proteins exhibit exceptionally broad specificity, being capable of transporting all 20 natural amino acids found in proteins (e.g., AAP1–AAP5 of *A. thaliana* [1,6,7]) while other members apparently exhibit absolute specificity for a single amino acid (e.g., the ProT1 and ProT2 proline transporters of *A. thaliana* [4]). These extreme examples of broad versus narrow specificity within a single family of transport proteins provide an interesting model system for understanding the phenomenon of substrate recognition and transport among homologous sets of proteins.

The comparative phylogenetic analyses conducted

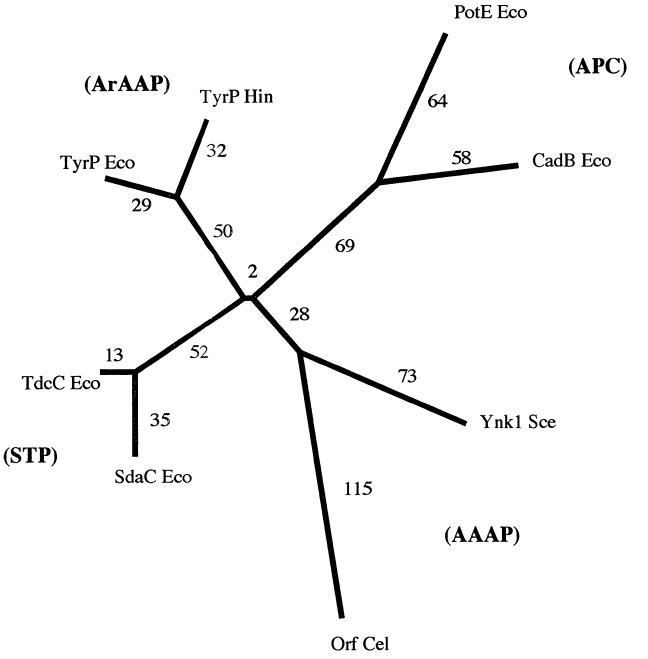

Fig. 7. Phylogenetic tree for the putative APC superfamily. The TREE program was used to generate the tree which was based on the complete multiple alignment used to also generate the average hydropathy and similarity plots shown in Fig. 5. Abbreviations of the families and proteins are as presented in Table 2.

with six different programs have led us to suggest that the phylogenetic trees generated with the TREE and neighbor-joining programs give the most consistent results and are more reliable than the dendrograms generated with the PILEUP and PROTPARS programs. The PAPA and Clustal W programs give results of an intermediate degree of consistency. Moreover, bootstrapping, which *does* provide a measure of confidence at specific nodes, *does not* increase the confidence levels obtained using distinct programs such as Clustal W and PROTPARS or allow evaluation of one program relative to another because this technique does not evaluate the reliability of the algorithms used in the programs. That is, bootstrapping does not evaluate the assumptions made in designing the various programs. Algorithmic differences give rise to the major differences in tree or dendrogram configuration [44]. The dendrograms shown in Fig. 4 tend to artificially cluster proteins that, in fact, are too distant in their sequence similarities to warrant clustering. This fact contributes significantly to the differences observed between the two dendrograms shown in Fig. 4, and between any one of these dendrograms, and the four trees shown in Fig. 3.

Acknowledgements

We thank Mary Beth Hiller for assistance in the preparation of the manuscript. Work in the authors' laboratory was supported by USPHS Grants 2RO1 GM55434 from the National Institute of General Medical Sciences, 2R01 AI 14176 from The National Institute of Allergy and Infectious Diseases, and by the M.H. Saier, Sr. Memorial Research Fund.

References

[1] W.-N. Fischer, M. Kwart, S. Hummel, W.B. Frommer, Substrate specificity and expression profile of amino acid transporters (AAPs) in *Arabidopsis*, J. Biol. Chem. 270 (1995) 16315–16320.

[2] J.A. Clark, S.G. Amara, Amino acid neurotransmitter transporters: Structure, function, and molecular diversity, BioEssays 15 (1993) 323–332.

[3] M.J. Bennett, A. Marchant, H.G. Green, S.T. May, S.P. Ward, P.A. Millner, A.R. Walker, B. Schulz, K.A. Feldmann, *Arabidopsis AUX1* gene: A permease-like regulator of root gravitropism, Science 273 (1996) 948–950.

[4] D. Rentsch, B. Hirner, E. Schmelzer, W.B. Frommer, Salt stress-induced proline transporters and salt stress-repressed broad specificity amino acid permeases identified by suppression of a yeast amino acid permease-targeting mutant, Plant Cell 8 (1996) 1437–1446.

[5] L.E. Williams, J.A. Bick, A. Neelam, K.N. Weston, J.L. Hall, Biochemical and molecular characterization of sucrose and amino acid carriers in *Ricinus communis*, J. Exp. Bot. 47, (Special Issue) (1996) 1211–1216.

[6] M. Kwart, B. Hirner, S. Hummel, W.B. Frommer, Differential expression of two related amino acid transporters with differing substrate specificity in *Arabidopsis thaliana*, Plant J. 4 (1993) 993–1002.

[7] W.B. Frommer, M. Kwart, B. Hirner, W.N. Fischer, S. Hummel, O. Ninnemann, Transporters for nitrogenous compounds in plants, Plant Mol. Biol. 26 (1994) 1651–1670.

[8] R. Krämer, Systems and mechanisms of amino acid uptake and excretion in prokaryotes, Arch. Microbiol. 162 (1994) 1–13.

[9] I.T. Paulsen, M.K. Sliwinski, M.H. Saier Jr., Microbial genome analyses: Global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities, J. Mol. Biol. 277 (1998) 573–592.

[10] M.H. Saier, Jr., Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya, in: R.K. Poole (Ed.), Advances in Microbial Physiology, Academic Press, San Diego, CA, 1998, pp 81–136.

[11] M.H. Saier, Jr., Classification of transmembrane transport systems in living organisms, in: L. Van Winkle (Ed.), Biomembrane Transport, Academic Press San Diego, CA, 1998, in press.

[12] J. Reizer, K. Finley, D. Kakuda, C.L. MacLeod, A. Reizer, M.H. Saier Jr., Mammalian integral membrane receptors are homologous to facilitators and antiporters of yeast, fungi, and eubacteria, Protein Sci. 2 (1993) 20–30.

[13] J. Reizer, A. Reizer, M.H. Saier Jr., A functional superfamily of sodium/solute symporters, Biochim. Biophys. Acta 1197 (1994) 133–166.

[14] S.L. McIntire, R.J. Reimer, K. Schuske, R.H. Edwards, E.M. Jorgensen, Identification and characterization of the vesicular GABA transporter, Nature 389 (1997) 870–876.

[15] M.K. Kuhner, J. Felsenstein, A simulation comparison of phylogenetic algorithms under equal and unequal evolutionary rates, Mol. Biol. Evol. 11 (1994) 459–468.

[16] S.B. Needleman, C. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.

[17] D.-F. Feng, R.F. Doolittle, Progressive sequence alignment as a prerequisite to correct phylogenetic trees, J. Mol. Evol. 25 (1987) 351–360.

[18] D.-F. Feng, R.F. Doolittle, Progressive alignment and phylogenetic tree construction of protein sequences, Methods Enzymol. 183 (1990) 375–387.

[19] D.-F. Feng, R.F. Doolittle, Progressive alignment of amino acid sequences and construction of phylogenetic trees from them, Methods Enzymol. 266 (1996) 368–382.
[20] D.G. Higgins, J.D. Thompson, T.J. Gibson, Using CLUSTAL for multiple sequence alignments, Methods Enzymol. 266 (1996) 383–402.
[21] W.M. Fitch, E. Margoliash, Construction of phylogenetic trees, Science 155 (1967) 279–284.
[22] L.C. Klotz, R.L. Blanken, A practical method for calculating evolutionary trees from sequence data, J. Theor. Biol. 91 (1981) 261–272.
[23] P.H.A. Sneath, R.R. Sokal, Numerical Taxonomy, The Principles and Practice of Numerical Classification, W.H. Freeman and Company, San Francisco, CA, 1973.
[24] N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[25] A. Rzhetsky, M. Nei, Theoretical foundation of the minimum-evolution method of phylogenetic inference, Mol. Biol. Evol. 10 (1993) 1073–1095.
[26] J. Felsenstein, PHYLIP – Phylogeny Inference Package (Version 3.2), Cladistics 5 (1989) 164–166.
[27] R.F. Doolittle, D.-F. Feng, Nearest neighbor procedure for relating progressively aligned amino acid sequences, Methods Enzymol. 183 (1990) 659–669.
[28] L.L. Cavalli-Sforza, A.W.F. Edwards, Phylogenetic analysis: Models and estimation procedures, Am. J. Hum. Genet. 19 (1967) 233.
[29] W.M. Fitch, On the problem of discovering the most parsimonious tree, Am. Natur. 111 (1977) 223–257.
[30] J. Felsenstein, Evolutionary trees from DNA sequences: A maximum likelihood approach, J. Mol. Evol. 17 (1981) 368–376.
[31] G.J. Olsen, H. Matsuda, R. Hagstron, R. Overbeek, FastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood, Comput. Appl. Biosci. 10 (1994) 41–48.
[32] J. Adachi, M. Hasegawa, MOLPHY: Programs for Molecular Phylogenetics, Version 2.2, Institute of Statistical Mathematics, Tokyo, 1994.
[33] K. Strimmer, Maximum Likelihood Methods in Molecular Phylogenetics, PhD Thesis, 1997.
[34] J.P. Sarsero, A.J. Pittard, Membrane topology analysis of *Escherichia coli* K-12 Mtr permease by alkaline phosphatase and β-galactosidase fusions, J. Bacteriol. 177 (1995) 297–306.
[35] M.H. Saier Jr., Phylogenetic approaches to the identification and characterization of protein families and superfamilies, Microb. Comp. Genomics 1 (1996) 129–150.
[36] M.H. Saier Jr., Computer-aided analyses of transport protein sequences: Gleaning evidence concerning function, structure, biogenesis, and evolution, Microbiol. Rev. 58 (1994) 71–93.
[37] R.V. Eck, M.O. Dayhoff, Atlas of Protein Sequence and Structure 1966, National Biomedical Research Foundation, Silver Spring, MD, 1966.
[38] W.M. Fitch, Toward defining the course of evolution: Minimum change for a specified tree topology, Syst. Zool. 20 (1971) 406–416.
[39] J. Felsenstein, Confidence limits on phylogenies: An approach using the bootstrap, Evolution 39 (1985) 783–791.
[40] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[41] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132.
[42] Wisconsin Package Version 8.1, Genetics Computer Group (GCG), Madison, WI.
[43] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle, WA, 1993.
[44] M. Nie, Phylogenetic analysis in molecular evolutionary genetics, Annu. Rev. Genet. 30 (1996) 371–403.