J Membrane Biol (2011) 240:89–110
DOI 10.1007/s00232-011-9347-9

Evolution of the Oligopeptide Transporter Family

Kenny M. Gomolplitinant · Milton H. Saier Jr.

Received: 14 October 2010 / Accepted: 21 January 2011 / Published online: 24 February 2011

Abstract The oligopeptide transporter (OPT) family of peptide and iron-siderophore transporters includes members from both prokaryotes and eukaryotes but with restricted distribution in the latter domain. Eukaryotic members were found only in fungi and plants with a single slime mold homologue clustering with the fungal proteins. All functionally characterized eukaryotic peptide transporters segregate from the known iron-siderophore transporters on a phylogenetic tree. Prokaryotic members are widespread, deriving from many different phyla. Although they belong only to the iron-siderophore subdivision, genome context analyses suggest that many of them are peptide transporters. OPT family proteins have 16 or occasionally 17 transmembrane-spanning α-helical segments (TMSs). We provide statistical evidence that the 16-TMS topology arose via three sequential duplication events followed by a gene-fusion event for proteins with a seventeenth TMS. The proposed pathway is as follows: 2 TMSs → 4 TMSs → 8 TMSs → 16 TMSs → 17 TMSs. The seventeenth C-terminal TMS, which probably arose just once, is found in just one phylogenetic group of these homologues. Analyses for orthology revealed that a few phylogenetic clusters consist exclusively of orthologues but most have undergone intermixing, suggestive of horizontal transfer. It appears that in this family horizontal gene transfer was frequent among prokaryotes, rare among eukaryotes and largely absent between prokaryotes and eukaryotes as well as between plants and fungi. These observations provide guides for future structural and functional analyses of OPT family members.

Keywords Iron-siderophore · Peptide · Transport · Evolution · Origin · Topology

Introduction

Transport proteins have been classified in the Transporter Classification Database (TCDB, www.tcdb.org; Saier 2000a, b; Saier et al. 2006, 2009). The first class is composed of channels/pores which catalyze facilitated diffusion by an energy-independent process. Electrochemical potential–driven transporters, comprising the second class, are stereospecific carriers catalyzing uniport, antiport and/or symport (Saier 2000c; Busch and Saier 2004). Primary active transporters, class 3, utilize a primary source of energy (chemical, electrical and/or solar) to drive transport of a solute against a concentration gradient (Saier 2000a). Group translocators, class 4, utilize a primary energy source to chemically alter a substrate in a process coupled to transport across a membrane (Mitchell and Moyle 1958).

The oligopeptide transporter (OPT, TC 2.A.67) family consists of electrochemical potential–driven transporters (class 2). All functionally characterized members of this family catalyze uptake of their solutes by a cation:solute symport mechanism (Hauser et al. 2001; Lubkowitz 2006; Yen et al. 2001). Functionally characterized members consist of transporters specific for oligopeptides (three to eight amino acids) and iron-siderophores (Yen et al. 2001). Characterized peptide transporters transport oligopeptides,

Electronic supplementary material The online version of this article (doi:10.1007/s00232-011-9347-9) contains supplementary material, which is available to authorized users.

K. M. Gomolplitinant · M. H. Saier Jr. (✉)
Division of Biological Sciences, University of California
at San Diego, La Jolla, CA 92093-0116, USA
e-mail: msaier@ucsd.edu

glutathione and glutathione conjugates (Kaur et al. 2009; Lubkowitz et al. 1998). Characterized "yellow stripe" (YS) homologues, on the other hand, mediate the uptake of metal-chelating phytosiderophores, including iron-nicotinamine and complexes of iron with secondary amino acid derivatives such as mugineic acid and deoxymugineic acid (Kaur et al. 2009). The biochemical and physiological characteristics of several OPT homologues have been studied (Lubkowitz 2006; Osawa et al. 2006; Stacey et al. 2008; Thakur et al. 2008). Two highly conserved motifs (NPG and KIPPR) have been found in many OPT family proteins (Koh et al. 2002). The OPT family is not to be confused with the proton-dependent oligopeptide transporter (POT or PTR, TC 2.A.17) family (Paulsen and Skurray 1994), the peptide transporters (PepTs) of the ATP-binding cassette (ABC, TC 3.A.1.5) superfamily (Saier 2000a; Busch and Saier 2004) or the peptide/acetyl-CoA transporters (PATs) of the major facilitator superfamily (MFS, TC 2.A.1.25) (Pao et al. 1998).

Oligopeptide transport plays important roles in nitrogen storage and mobilization, quorum sensing, differentiation, sexual induction, mating and pheromone sensing. One of the yeast homologues is the sexual differentiation process (ISP4) protein of *Schizosaccharomyces pombe*. In yeast, OPT family homologues transport oligopeptides, which are commonly tri-, tetra- and/or pentapeptides (Wiles et al. 2006). Recently, it has been found that high-affinity *Saccharomyces cerevisiae* and *Sc. pombe* glutathione transporters, Hgt1p and OPT1, respectively, belong to the OPT family (Dworeck et al. 2009; Kaur et al. 2009).

In *Candida albicans*, eight OPT genes have been identified, encoding putative OPTs. Almost all are represented by polymorphic alleles (Reuss and Morschhauser 2006). OPT1,2,3Δ triple mutants were found to have a severe growth defect, which could be rescued by reintroduction of a single copy of OPT1, OPT2 or OPT3. The various transporters differ in their substrate preferences as shown by the ability of strains expressing specific OPT genes to grow on peptides of defined length and sequence (Reuss and Morschhauser 2006).

In plants, many OPTs appear to be plasma membrane–embedded proteins that import substrates from the apoplasm (the aqueous phase of the cell wall) and the external environment. They may play roles in plant growth and development (Lubkowitz 2006). Unlike many other OPTs, which function in long-distance transport of peptides or metals, YS1, an Fe^{3+}-phytosiderophore uptake system of *Zea mays*, is known to translocate substrates from the rhizosphere (the region of the soil that is directly influenced by root secretions and associated with soil microbes) (Yen et al. 2001; Curie et al. 2001). Expression of the YS1 gene is increased in roots and shoots under iron-deficient conditions (Curie et al. 2001). When YS1 is expressed in mutant yeast lacking its native iron uptake system, it is able to correct the defect, specifically in Fe^{3+} phytosiderophore–containing media.

In *Arabidopsis*, nine OPT paralogues have been identified (Koh et al. 2002), seven of which mediate transport of tetra- and pentapeptides. Cagnac et al. (2004) showed that AtOPT6 can mediate uptake of glutathione derivatives and metal complexes, which led them to suggest that it may also be involved in stress resistance.

Bacterial and archaeal homologues of the OPT family have yet to be characterized biochemically, but as shown here, they are prevalent throughout the prokaryotic world (Kaur et al. 2009). A high-resolution three-dimensional X-ray structure of an OPT family homologue has yet to be solved. We therefore carried out detailed bioinformatic analyses of these transporters, showing that the family is far more widespread than previously recognized and demonstrating the evolutionary relationships of the members of this family to each other. Most surprisingly, we found that these 16-TMS proteins arose from a two-TMS precursor–encoding genetic element which duplicated three times sequentially: 2 TMSs → 4 TMSs → 8 TMSs → 16 TMSs. Although this finding is in principle similar to the origin of animal Na^{+} and Ca^{2+} channel proteins of the voltage-gated ion channel (VIC, TC 1.A.1) family, where a six-TMS precursor twice duplicated to give 24-TMS proteins (Nelson et al. 1999), this is the first demonstration of such an event occurring from a two-TMS element and involving three successive intragenic duplication events.

Methods

PSI-BLAST (Altschul et al. 1997) searches were performed to screen the National Center for Biotechnology Information (NCBI) nonredundant protein database using *C. albicans* Opt1 (gi 74582040), *Sc. pombe* Isp4 (gi 19859374), *Sa. cerevisiae* Opt1 (gi 731969), *Z. mays* YS1 (gi 75168533) and *Myxococcus xanthus* EspB (gi 75421577). The corresponding TinySeq XML format (NCBI) of these proteins was obtained and modified using the script MakeTable5 (Yen et al. 2009) to generate a FASTA file for all of the sequences and a table containing each protein's abbreviation, description, organismal source, size, gi number, organismal kingdom or phylum and organismal domain. MakeTable5 was also used to remove fragments and protein sequences with >90% sequence identity to an included protein.

Multiple alignments of homologous proteins and phylogenetic trees were generated using the CLUSTAL X program (Thompson et al. 1997) followed by the TreeView program (Zhai et al. 2002) with default settings. The WHAT (Zhai and Saier 2001a) and TMHMM (Kall et al. 2007) programs were used to perform topological analyses on single protein sequences. The AveHAS program (Zhai and Saier 2001b) with default settings was used to generate average hydropathy, amphipathicity and similarity plots for multiply aligned sequences. Internal homologous repeat segments in all OPT proteins examined were statistically compared using the IC(Faa2) program (Yen et al. 2009). Segments giving the best comparison scores were further examined using the GAP program with default settings and 500 random shuffles with comparison scores expressed in standard deviations (SDs) (Devereux et al. 1984). A value of 10 SD corresponds to a probability of 10^{-24} that the observed degree of similarity occurred by chance (Dayhoff et al. 1983). To optimize, nonaligned segments were removed, numbers of identities were maximized and numbers of gaps were minimized, maintaining a length of at least 60 residues. The comparison score was then determined again as before. For a stretch of at least 60 amino acyl residues, corresponding to a typical, average-sized protein domain, 10 SD is deemed sufficient to establish homology (Saier 1994; Saier et al. 2009; Yen et al. 2009).

The GGSEARCH (http://fasta.bioch.virginia.edu/gasta_www2/fasta_list2.shtml), HMMER (http://hmmer.janelia.org; Eddy, 2008) and SAM (Yen et al. 2009; Wang et al. 2009) programs were subsequently used to provide confirmatory evidence for homology. The halves, quarters and eighths of these homologues, which showed significant sequence similarity using IC/GAP (Table 2), were subsequently used to generate a profile and a database for each program.

The hmmbuild program was first used to build an HMM profile for each eight- or four-TMS segment. This profile was then calibrated using the hmmcalibrate program to obtain more accurate e-values. The resulting calibrated profile was then used to search a corresponding eight- or four-TMS segment database (FASTA-formatted sequence file) with the hmmsearch program. The resulting output file showed the domain and alignment annotation for each sequence. The HMMER commands used were

```
hmmbuild <hmm file> <alignment file>
hmmcalibrate <hmm file>
hmmsearch <hmm file> <sequence file>
```

The same essential procedures were used for SAM and GGSEARCH. Using the SAM program, the sequence files from the halves and quarters were first trained to build models. The models were subsequently used to search against a database consisting of the corresponding untrained halves and quarters. The SAM commands used were

```
buildmodel <model name> -train <training set> -randseed0
hmmscore <output> -I <model file> -db <target sequence
file? –sw 2 –calibrate 1
```

GGSEARCH of the FASTA package from the University of Virginia (http://fasta.bioch.virginia.edu/fasta_www2?fasta_www.cgi?rm=selectandpgm=gnw) was similarly used to compare the eight-TMS halves and the four-TMS quarters.

Results

Phylogenetic Analysis of OPT Family Members

The 325 proteins included in this study are listed alphabetically in supplementary Table S1 (http://biology.ucsd.edu/~msaier/supmat/OPT/index.html) and according to cluster and position in the phylogenetic tree (Fig. 1) in Table 1. The dendogram corresponding to the tree shown in Fig. 1 can be viewed in supplementary Fig. S2. The tree shown in Fig. 1 reveals five clusters subdivided as follows.

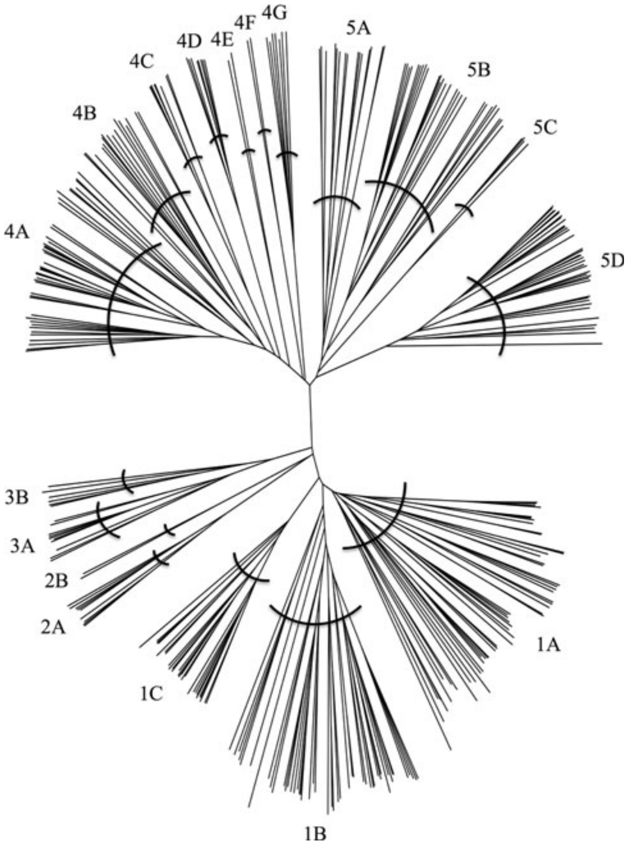

Fig. 1 Phylogenetic tree of 325 OPT superfamily proteins based on the ClustalX multiple alignment shown in Fig. S1 and drawn using the FigTree program. Clusters 1–5 are labeled with their respective subclusters. Subclusters *1A–3B* are putative peptide transporters, while some members of subclusters *4A–5D* are known to be iron-siderophore transporters. Protein abbreviations are presented in Table 1 in the same order as shown in the tree, together with the characteristics of these proteins. The positions of the individual proteins are revealed in the dendrogram shown in Fig. S2

Table 1 OPT protein sequences included in this study

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Subcluster 1A (56 proteins)					
Nfi2	*Neosartorya fischeri* NRRL 181	119471104	Fungi	Eukaryota	757
Acl5	*Aspergillus clavatus* NRRL 1	121709515	Fungi	Eukaryota	761
Aor3	*Aspergillus oryzae*	83768538	Fungi	Eukaryota	751
Ani6	*Aspergillus niger* CBS 513.88	145241488	Fungi	Eukaryota	859
Aor2	*Aspergillus oryzae*	83768389	Fungi	Eukaryota	765
Ani12	*Aspergillus niger* CBS 513.88	145251507	Fungi	Eukaryota	771
Nfi3	*Neosartorya fischeri* NRRL 181	119471211	Fungi	Eukaryota	770
Bfu2	*Botryotinia fuckeliana* B05.10	154313655	Fungi	Eukaryota	779
Aor5	*Aspergillus oryzae*	83768732	Fungi	Eukaryota	770
Nfi6	*Neosartorya fischeri* NRRL 181	119491377	Fungi	Eukaryota	768
Sce2	*Saccharomyces cerevisiae* YJM789	151943695	Fungi	Eukaryota	799
Kla3	*Kluyveromyces lactis* NRRL Y-1140	50307929	Fungi	Eukaryota	793
Pgu6	*Pichia guilliermondii* ATCC 6260	146419361	Fungi	Eukaryota	754
Spo3	*Schizosaccharomyces pombe*	63054465	Fungi	Eukaryota	851
Ani2	*Aspergillus niger* CBS 513.88	67540564	Fungi	Eukaryota	778
Cne4	*Cryptococcus neoformans* var. neoformans B-3501A	134113154	Fungi	Eukaryota	797
Ncr6	*Neurospora crassa* OR74A	164422675	Fungi	Eukaryota	1094
Cgl3	*Chaetomium globosum* CBS 148.51	116193201	Fungi	Eukaryota	1027
Ssc1	*Sclerotinia sclerotiorum* 1980	156039822	Fungi	Eukaryota	1055
Gze5	*Gibberella zeae* PH-1	46125699	Fungi	Eukaryota	1060
Ani1	*Aspergillus niger* CBS 513.88	67516837	Fungi	Eukaryota	792
Aor7	*Aspergillus oryzae*	83770544	Fungi	Eukaryota	778
Mgr2	*Magnaporthe grisea* 70-15	39944474	Fungi	Eukaryota	783
Acl1	*Aspergillus clavatus* NRRL 1	121699197	Fungi	Eukaryota	788
Nfi4	*Neosartorya fischeri* NRRL 181	119477757	Fungi	Eukaryota	772
Pgu2	*Pichia guilliermondii* ATCC 6260	146416527	Fungi	Eukaryota	784
Pst7	*Pichia stipitis* CBS 6054	150864787	Fungi	Eukaryota	782
Dha1	*Debaryomyces hansenii* CBS767	50413511	Fungi	Eukaryota	776
Cal4	*Candida albicans*	68485275	Fungi	Eukaryota	783
Lel2	*Lodderomyces elongisporus* NRRL YB-4239	149235877	Fungi	Eukaryota	804
Kla1	*Kluyveromyces lactis* NRRL Y-1140	50307527	Fungi	Eukaryota	794
Ago5	*Ashbya gossypii* ATCC 10895	45201069	Fungi	Eukaryota	796
Yli1	*Yarrowia lipolytica* CLIB122	50542874	Fungi	Eukaryota	836
Ncr1	*Neurospora crassa* OR74A	9368956	Fungi	Eukaryota	801
Cgl5	*Chaetomium globosum* CBS 148.51	116198757	Fungi	Eukaryota	871
Gze7	*Gibberella zeae* PH-1	46134295	Fungi	Eukaryota	799
Afu2	*Aspergillus fumigatus* Af293	70999364	Fungi	Eukaryota	792
Acl4	*Aspergillus clavatus* NRRL 1	121705906	Fungi	Eukaryota	793
Ate1	*Aspergillus terreus* NIH2624	115397517	Fungi	Eukaryota	788
Aor9	*Aspergillus oryzae*	83775779	Fungi	Eukaryota	768
Ani4	*Aspergillus nidulans* FGSC A4	67901220	Fungi	Eukaryota	794
Ssc4	*Sclerotinia sclerotiorum* 1980	156049297	Fungi	Eukaryota	827
Cim3	*Coccidioides immitis* RS	119194107	Fungi	Eukaryota	812
Pno1	*Phaeosphaeria nodorum* SN15	160705030	Fungi	Eukaryota	845
Mgr5	*Magnaporthe grisea* 70-15	145614314	Fungi	Eukaryota	849
Spo2	*Schizosaccharomyces pombe*	19115899	Fungi	Eukaryota	785

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Cim2	*Coccidioides immitis* RS	119190959	Fungi	Eukaryota	810
Cci1	*Coprinopsis cinerea* okayama7#130	116500528	Fungi	Eukaryota	757
Lbi5	*Laccaria bicolor* S238N-H82	164641826	Fungi	Eukaryota	730
Cci6	*Coprinopsis cinerea* okayama7#130	116510327	Fungi	Eukaryota	772
Uma5	*Ustilago maydis* 521	71020527	Fungi	Eukaryota	807
Cci3	*Coprinopsis cinerea* okayama7#130	116506493	Fungi	Eukaryota	1292
Cci5	*Coprinopsis cinerea* okayama7#130	116509020	Fungi	Eukaryota	771
Lbi4	*Laccaria bicolor* S238N-H82	164640879	Fungi	Eukaryota	757
Cne3	*Cryptococcus neoformans* var. neoformans JEC21	58268358	Fungi	Eukaryota	961
Uma1	*Ustilago maydis* 521	71012856	Fungi	Eukaryota	985
Average protein size ± SD (aas)					825 ± 103
Subcluster 1B (48 proteins)					
Cal1	*Candida albicans*	2367386	Fungi	Eukaryota	945
Lel5	*Lodderomyces elongisporus* NRRL YB-4239	149237448	Fungi	Eukaryota	919
Pst3	*Pichia stipitis* CBS 6054	126139203	Fungi	Eukaryota	917
Dha3	*Debaryomyces hansenii* CBS767	50419775	Fungi	Eukaryota	907
Pgu8	*Pichia guilliermondii* ATCC 6260	146421835	Fungi	Eukaryota	881
Cal5	*Candida albicans* SC5314	87045969	Fungi	Eukaryota	929
Cal6	*Candida albicans*	87045975	Fungi	Eukaryota	904
Cal3	*Candida albicans* SC5314	68476729	Fungi	Eukaryota	921
Lel3	*Lodderomyces elongisporus* NRRL YB-4239	149236581	Fungi	Eukaryota	862
Lel4	*Lodderomyces elongisporus* NRRL YB-4239	149236916	Fungi	Eukaryota	967
Pst4	*Pichia stipitis* CBS 6054	146280790	Fungi	Eukaryota	891
Pst9	*Pichia stipitis* CBS 6054	150866640	Fungi	Eukaryota	913
Pst10	*Pichia stipitis* CBS 6054	150951233	Fungi	Eukaryota	911
Pgu3	*Pichia guilliermondii* ATCC 6260	146416529	Fungi	Eukaryota	922
Pgu7	*Pichia guilliermondii* ATCC 6260	146420005	Fungi	Eukaryota	944
Pgu5	*Pichia guilliermondii* ATCC 6260	146419149	Fungi	Eukaryota	922
Pst8	*Pichia stipitis* CBS 6054	150866635	Fungi	Eukaryota	907
Lel7	*Lodderomyces elongisporus* NRRL YB-4239	149246151	Fungi	Eukaryota	924
Dha2	*Debaryomyces hansenii* CBS767	50417315	Fungi	Eukaryota	850
Pgu9	*Pichia guilliermondii* ATCC 6260	146422868	Fungi	Eukaryota	849
Kla2	*Kluyveromyces lactis* NRRL Y-1140	50307927	Fungi	Eukaryota	869
Sce1	*Saccharomyces cerevisiae*	6325452	Fungi	Eukaryota	877
Vpo1	*Vanderwaltozyma polyspora* DSM 70294	156838884	Fungi	Eukaryota	892
Ago1	*Ashbya gossypii* ATCC 10895	45185387	Fungi	Eukaryota	890
Ago3	*Ashbya gossypii* ATCC 10895	45187474	Fungi	Eukaryota	885
Ago4	*Ashbya gossypii* ATCC 10895	45198503	Fungi	Eukaryota	877
Yli10	*Yarrowia lipolytica* CLIB122	50551841	Fungi	Eukaryota	876
Yli17	*Yarrowia lipolytica* CLIB122	50557248	Fungi	Eukaryota	767
Yli2	*Yarrowia lipolytica* CLIB122	50543154	Fungi	Eukaryota	896
Yli12	*Yarrowia lipolytica* CLIB122	50553458	Fungi	Eukaryota	884
Yli15	*Yarrowia lipolytica* CLIB122	50555966	Fungi	Eukaryota	882
Yli4	*Yarrowia lipolytica* CLIB122	50545932	Fungi	Eukaryota	886
Yli3	*Yarrowia lipolytica* CLIB122	50545745	Fungi	Eukaryota	872
Yli6	*Yarrowia lipolytica* CLIB122	50548489	Fungi	Eukaryota	883
Yli14	*Yarrowia lipolytica* CLIB122	50555666	Fungi	Eukaryota	883

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Yli8	*Yarrowia lipolytica* CLIB122	50549187	Fungi	Eukaryota	882
Yli16	*Yarrowia lipolytica* CLIB122	50556388	Fungi	Eukaryota	948
Yli11	*Yarrowia lipolytica* CLIB122	50553314	Fungi	Eukaryota	879
Yli9	*Yarrowia lipolytica* CLIB122	50549349	Fungi	Eukaryota	903
Yli13	*Yarrowia lipolytica* CLIB122	50555622	Fungi	Eukaryota	874
Yli7	*Yarrowia lipolytica* CLIB122	50549017	Fungi	Eukaryota	1032
Mgr1	*Magnaporthe grisea* 70-15	39941802	Fungi	Eukaryota	926
Cne1	*Cryptococcus neoformans* var. neoformans JEC21	58259793	Fungi	Eukaryota	812
Gze1	*Gibberella zeae* PH-1	46115170	Fungi	Eukaryota	874
Ncr3	*Neurospora crassa* OR74A	85093666	Fungi	Eukaryota	864
Mgr4	*Magnaporthe grisea* 70-15	145602334	Fungi	Eukaryota	870
Gze4	*Gibberella zeae* PH-1	46124369	Fungi	Eukaryota	851
Gze8	*Gibberella zeae* PH-1	46136533	Fungi	Eukaryota	839
Average protein size ± SD (aa)					893 ± 41
Subcluster 1C (27 Proteins)					
Osa3	*Oryza sativa* Indica Group	41053195	Viridiplantae	Eukaryota	755
Osa8	*Oryza sativa* Japonica Group	74267416	Viridiplantae	Eukaryota	751
Vvi12	*Vitis vinifera*	157355114	Viridiplantae	Eukaryota	744
Ath12	*Arabidopsis thaliana*	41352045	Viridiplantae	Eukaryota	729
Mtr1	*Medicago truncatula*	124359202	Viridiplantae	Eukaryota	729
Vvi7	*Vitis vinifera*	157338674	Viridiplantae	Eukaryota	757
Vvi16	*Vitis vinifera*	157359604	Viridiplantae	Eukaryota	739
Vvi9	*Vitis vinifera*	157338676	Viridiplantae	Eukaryota	740
Ath2	*Arabidopsis thaliana*	15218799	Viridiplantae	Eukaryota	734
Osa16	*Oryza sativa* Indica Group	115459700	Viridiplantae	Eukaryota	1278
Vvi5	*Vitis vinifera*	157335739	Viridiplantae	Eukaryota	689
Ath14	*Arabidopsis thaliana*	67460718	Viridiplantae	Eukaryota	766
Osa9	*Oryza sativa* Japonica Group	90265681	Viridiplantae	Eukaryota	763
Osa25	*Oryza sativa* Japonica Group	125540410	Viridiplantae	Eukaryota	766
Osa10	*Oryza sativa* Indica Group	90265683	Viridiplantae	Eukaryota	771
Vvi13	*Vitis vinifera*	157355237	Viridiplantae	Eukaryota	690
Ath16	*Arabidopsis thaliana*	79518939	Viridiplantae	Eukaryota	741
Ath17	*Arabidopsis thaliana*	145359208	Viridiplantae	Eukaryota	736
Ath9	*Arabidopsis thaliana*	18402162	Viridiplantae	Eukaryota	733
Vvi8	*Vitis vinifera*	157338675	Viridiplantae	Eukaryota	731
Ath7	*Arabidopsis thaliana*	15238763	Viridiplantae	Eukaryota	755
Ath15	*Arabidopsis thaliana*	79484897	Viridiplantae	Eukaryota	753
Osa31	*Oryza sativa* Japonica Group	125583075	Viridiplantae	Eukaryota	733
Mac1	*Musa acuminata*	102140021	Viridiplantae	Eukaryota	748
Osa12	*Oryza sativa* Japonica Group	115440825	Viridiplantae	Eukaryota	757
Ath3	*Arabidopsis thaliana*	15234254	Viridiplantae	Eukaryota	737
Ppa2	*Physcomitrella patens* subsp. patens	162689084	Viridiplantae	Eukaryota	733
Average protein size ± SD (aa)					761 ± 105
Subcluster 2A (9 proteins)					
Nfi1	*Neosartorya fischeri* NRRL 181	119467402	Fungi	Eukaryota	788
Ani8	*Aspergillus niger* CBS 513.88	145243688	Fungi	Eukaryota	799
Aor8	*Aspergillus oryzae*	83772997	Fungi	Eukaryota	793

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Ssc3	*Sclerotinia sclerotiorum* 1980	156046206	Fungi	Eukaryota	812
Cal7	*Candida albicans*	87045979	Fungi	Eukaryota	747
Lel6	*Lodderomyces elongisporus* NRRL YB-4239	149246053	Fungi	Eukaryota	765
Pst5	*Pichia stipitis* CBS 6054	150864397	Fungi	Eukaryota	765
Pst2	*Pichia stipitis* CBS 6054	126139089	Fungi	Eukaryota	771
Pgu4	*Pichia guilliermondii* ATCC 6260	146417045	Fungi	Eukaryota	760
Average protein size ± SD (aa)					778 ± 21
Subcluster 2B (2 proteins)					
Ncr4	*Neurospora crassa* OR74A	85107500	Fungi	Eukaryota	1094
Bfu1	*Botryotinia fuckeliana* B05.10	154292901	Fungi	Eukaryota	767
Average protein size ± SD (aa)					931 ± 231
Subcluster 3A (10 proteins)					
Ssc2	*Sclerotinia sclerotiorum* 1980	156046040	Fungi	Eukaryota	790
Gze2	*Gibberella zeae* PH-1	46115236	Fungi	Eukaryota	789
Uma3	*Ustilago maydis* 521	71016547	Fungi	Eukaryota	797
Acl3	*Aspergillus clavatus* NRRL 1	121701255	Fungi	Eukaryota	775
Nfi5	*Neosartorya fischeri* NRRL 181	119488556	Fungi	Eukaryota	757
Ani3	*Aspergillus nidulans* FGSC A4	67542049	Fungi	Eukaryota	746
Aor4	*Aspergillus oryzae*	83768691	Fungi	Eukaryota	774
Ate2	*Aspergillus terreus* NIH2624	115401822	Fungi	Eukaryota	780
Sco1	*Schizophyllum commune*	6716399	Fungi	Eukaryota	777
Lbi8	*Laccaria bicolor* S238N-H82	164643810	Fungi	Eukaryota	749
Average protein size ± SD (aa)					773 ± 17
Subcluster 3B (6 proteins)					
Gze3	*Gibberella zeae* PH-1	46120458	Fungi	Eukaryota	782
Ncr5	*Neurospora crassa* OR74A	85113749	Fungi	Eukaryota	788
Bfu3	*Botryotinia fuckeliana* B05.10	154321612	Fungi	Eukaryota	829
Uma4	*Ustilago maydis* 521	71019889	Fungi	Eukaryota	860
Ncr7	*Neurospora crassa* OR74A	164423970	Fungi	Eukaryota	793
Cci2	*Coprinopsis cinerea* okayama7#130	116504373	Fungi	Eukaryota	824
Average protein size ± SD (aa)					813 ± 30
Subcluster 4A (41 proteins)					
Csp1	*Caulobacter* sp. K31	113935253	Alphaproteobacteria	Bacteria	662
Ccr1	*Caulobacter crescentus* CB15	16126881	Alphaproteobacteria	Bacteria	666
Swi1	*Sphingomonas wittichii* RW1	148555886	Alphaproteobacteria	Bacteria	658
Neu1	*Nitrosomonas eutropha* C91	114332234	Betaproteobacteria	Bacteria	676
Ssp1	*Sphingomonas* sp. SKA58	94496206	Alphaproteobacteria	Bacteria	655
Nar1	*Novosphingobium aromaticivorans* DSM 12444	87199977	Alphaproteobacteria	Bacteria	650
Mtu1	*Mycobacterium tuberculosis* H37Rv	15609532	Actinobacteria	Bacteria	667
Msm1	*Mycobacterium smegmatis* str. MC2 155	118470017	Actinobacteria	Bacteria	663
Cdi1	*Corynebacterium diphtheriae* NCTC 13129	38232950	Actinobacteria	Bacteria	658
Pac1	*Propionibacterium acnes* KPA171202	50842040	Actinobacteria	Bacteria	662
Aod1	*Actinomyces odontolyticus* ATCC 17982	154508464	Actinobacteria	Bacteria	666
Cup1	*Campylobacter upsaliensis* RM3195	57506152	Epsilonproteobacteria	Bacteria	657
Cco1	*Campylobacter coli* RM2228	57168345	Epsilonproteobacteria	Bacteria	668
Cla1	*Campylobacter lari* RM2100	57241657	Epsilonproteobacteria	Bacteria	661
Bbr1	*Bordetella bronchiseptica* RB50	33602645	Betaproteobacteria	Bacteria	693
Bpe1	*Bordetella petrii* DSM 12804	163856141	Betaproteobacteria	Bacteria	689

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Bav1	*Bordetella avium* 197N	115422286	Betaproteobacteria	Bacteria	677
Rpi1	*Ralstonia pickettii* 12J	121528839	Betaproteobacteria	Bacteria	684
Rso1	*Ralstonia solanacearum* GMI1000	17548014	Betaproteobacteria	Bacteria	683
Reu3	*Ralstonia eutropha* JMP134	113869213	Betaproteobacteria	Bacteria	679
Reu4	*Ralstonia eutropha* H16	116696492	Betaproteobacteria	Bacteria	679
Rme1	*Ralstonia metallidurans* CH34	94312045	Betaproteobacteria	Bacteria	676
Pae1	*Pseudomonas aeruginosa* UCBPP-PA14	116051974	Gammaproteobacteria	Bacteria	678
Hso1	*Haemophilus somnus* 2336	32029457	Gammaproteobacteria	Bacteria	668
Asu1	*Actinobacillus succinogenes* 130Z	152977801	Gammaproteobacteria	Bacteria	670
Hdu1	*Haemophilus ducreyi* 35000HP	33152874	Gammaproteobacteria	Bacteria	669
Apl1	*Actinobacillus pleuropneumoniae* L20	126209177	Gammaproteobacteria	Bacteria	668
Msu1	*Mannheimia succiniciproducens* MBEL55E	52424073	Gammaproteobacteria	Bacteria	668
Hin1	*Haemophilus influenzae* R2866	53733327	Gammaproteobacteria	Bacteria	662
Ngo1	*Neisseria gonorrhoeae* FA 1090	59802215	Betaproteobacteria	Bacteria	672
Gdi1	*Gluconacetobacter diazotrophicus* PAl 5	162148874	Alphaproteobacteria	Bacteria	659
Gox1	*Gluconobacter oxydans* 621H	58038663	Alphaproteobacteria	Bacteria	648
Rgr1	*Rickettsiella grylli*	160871957	Gammaproteobacteria	Bacteria	654
Lpn1	*Legionella pneumophila* str. Corby	148360634	Gammaproteobacteria	Bacteria	666
Rgr2	*Rickettsiella grylli*	160872420	Gammaproteobacteria	Bacteria	669
Xfa1	*Xylella fastidiosa* Ann-1	71899907	Gammaproteobacteria	Bacteria	653
Sma2	*Stenotrophomonas maltophilia* R551-3	126466290	Gammaproteobacteria	Bacteria	654
Nmo1	*Nitrococcus mobilis* Nb-231	88812607	Gammaproteobacteria	Bacteria	655
Pho2	*Pyrococcus horikoshii* OT3	14590884	Euryarchaeota	Archaea	626
Tko1	*Thermococcus kodakarensis* KOD1	57641714	Euryarchaeota	Archaea	624
Sde1	*Saccharophagus degradans* 2-40	90020298	Gammaproteobacteria	Bacteria	672
Average protein size ± SD (aa)					665 ± 14
Subcluster 4B (16 proteins)					
Ade1	*Anaeromyxobacter dehalogenans* 2CP-C	86156672	Deltaproteobacteria	Bacteria	690
Asp5	*Anaeromyxobacter* sp. Fw109-5	163767022	Deltaproteobacteria	Bacteria	706
Hsp1	*Halobacterium* sp. NRC-1	16120189	Euryarchaeota	Archaea	655
Csp2	*Clostridium* sp. L2-50	160894507	Firmicutes	Bacteria	632
Eve1	*Eubacterium ventriosum* ATCC 27560	154484314	Firmicutes	Bacteria	649
Rgn1	*Ruminococcus gnavus* ATCC 29149	154504363	Firmicutes	Bacteria	631
Rto1	*Ruminococcus torques* ATCC 27756	153813838	Firmicutes	Bacteria	633
Rob1	*Ruminococcus obeum* ATCC 29174	153810748	Firmicutes	Bacteria	632
Hor1	*Halothermothrix orenii* H 168	89210028	Firmicutes	Bacteria	636
Cno1	*Clostridium novyi* NT	118445126	Firmicutes	Bacteria	679
Cbo1	*Clostridium botulinum* F str. Langeland	153941447	Firmicutes	Bacteria	651
Tte1	*Thermoanaerobacter tengcongensis* MB4	20806685	Firmicutes	Bacteria	647
Chy1	*Carboxydothermus hydrogenoformans* Z-2901	78045182	Firmicutes	Bacteria	640
Dre1	*Desulfotomaculum reducens* MI-1	134300485	Firmicutes	Bacteria	656
Sus1	*Solibacter usitatus* Ellin6076	116620777	Acidobacteria	Bacteria	674
Aba2	*Acidobacteria bacterium* Ellin345	94971229	Acidobacteria	Bacteria	675
Average protein size ± SD (aa)					655 ± 23
Subcluster 4C (8 proteins)					
Bun1	*Bacteroides uniformis* ATCC 8492	160890502	Bacteroidetes	Bacteria	663
Bfr1	*Bacteroides fragilis* YCH46	53713327	Bacteroidetes	Bacteria	662
Bvu1	*Bacteroides vulgatus* ATCC 8482	150005284	Bacteroidetes	Bacteria	663

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Pdi1	*Parabacteroides distasonis* ATCC 8503	150008072	Bacteroidetes	Bacteria	665
Pme1	*Parabacteroides merdae* ATCC 43184	154492906	Bacteroidetes	Bacteria	666
Pgi1	*Porphyromonas gingivalis* W83	34540265	Bacteroidetes	Bacteria	659
Aba1	*Acidobacteria bacterium* Ellin345	94969462	Acidobacteria	Bacteria	664
Sus2	*Solibacter usitatus* Ellin6076	116622365	Acidobacteria	Bacteria	667
Average protein size ± SD (aa)					664 ± 3
Subcluster 4D (8 proteins)					
Lca1	*Lactobacillus casei* ATCC 334	116495639	Firmicutes	Bacteria	641
Ppe1	*Pediococcus pentosaceus* ATCC 25745	116491982	Firmicutes	Bacteria	639
Lsa1	*Lactobacillus sakei* subsp. sakei 23K	81427933	Firmicutes	Bacteria	645
Ckl1	*Clostridium kluyveri* DSM 555	153954672	Firmicutes	Bacteria	639
Cbe1	*Clostridium beijerinckii* NCIMB 8052	150016123	Firmicutes	Bacteria	640
Cba1	*Clostridium bartlettii* DSM 16795	164687644	Firmicutes	Bacteria	648
Cdi2	*Clostridium difficile* 630	126699006	Firmicutes	Bacteria	642
Cpe1	*Clostridium perfringens* str. 13	18310260	Firmicutes	Bacteria	638
Average protein size ± SD (aa)					642 ± 3
Subcluster 4E (2 proteins)					
Cae1	*Collinsella aerofaciens* ATCC 25986	139438467	Actinobacteria	Bacteria	558
Cce1	*Clostridium cellulolyticum* H10	118726871	Firmicutes	Bacteria	537
Average protein size ± SD (aa)					548 ± 15
Subcluster 4F (2 proteins)					
Orf1	uncultured methanogenic archaeon RC-I	147920129	Euryarchaeota	Archaea	553
Orf2	uncultured methanogenic archaeon RC-I	147920131	Euryarchaeota	Archaea	552
Average protein size ± SD (aa)					553 ± 1
Subcluster 4G (7 proteins)					
Bsp1	*Bacillus* sp. B14905	126653239	Firmicutes	Bacteria	524
Vei1	*Verminephrobacter eiseniae* EF01-2	121610237	Betaproteobacteria	Bacteria	524
Spr1	*Serratia proteamaculans* 568	157369266	Gammaproteobacteria	Bacteria	524
Bcl1	*Bacillus clausii* KSM-K16	56962356	Firmicutes	Bacteria	526
Pho1	*Pyrococcus horikoshii* OT3	14590271	Euryarchaeota	Archaea	527
Mth1	*Moorella thermoacetica* ATCC 39073	83589078	Firmicutes	Bacteria	519
Rob2	*Ruminococcus obeum* ATCC 29174	153812663	Firmicutes	Bacteria	558
Average protein size ± SD (aa)					529 ± 13
Subcluster 5A (15 proteins)					
Asp1	*Anaeromyxobacter* sp. K	153003141	Deltaproteobacteria	Bacteria	540
Ade2	*Anaeromyxobacter dehalogenans* 2CP-C	86158243	Deltaproteobacteria	Bacteria	540
Mxa4	*Myxococcus xanthus* DK 1622	108763515	Deltaproteobacteria	Bacteria	592
Sau2	*Stigmatella aurantiaca* DW4/3-1	115377255	Deltaproteobacteria	Bacteria	637
Mxa5	*Myxococcus xanthus* DK 1622	108763588	Deltaproteobacteria	Bacteria	631
Asp3	*Anaeromyxobacter* sp. Fw109-5	153005805	Deltaproteobacteria	Bacteria	605
Mxa2	*Myxococcus xanthus* DK 1622	108762092	Deltaproteobacteria	Bacteria	606
Sau4	*Stigmatella aurantiaca* DW4/3-1	115378283	Deltaproteobacteria	Bacteria	625
Psy1	*Pseudomonas syringae* pv. syringae B728a	66044430	Gammaproteobacteria	Bacteria	581
Ppu1	*Pseudomonas putida* W619	119857963	Gammaproteobacteria	Bacteria	585
Pst1	*Pseudomonas stutzeri* A1501	126134803	Gammaproteobacteria	Bacteria	570
Spe1	*Shewanella pealeana* ATCC 700345	157963678	Gammaproteobacteria	Bacteria	577
Sse1	*Shewanella sediminis* HAW-EB3	157373494	Gammaproteobacteria	Bacteria	576
Asp2	*Anaeromyxobacter* sp. K	153003206	Deltaproteobacteria	Bacteria	583

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Asp4	*Anaeromyxobacter* sp. Fw109-5	163766993	Deltaproteobacteria	Bacteria	583
Average protein size ± SD (aa)					589 ± 29
Subcluster 5B (27 Proteins)					
Cal2	*Candida albicans* SC5314	68475797	Fungi	Eukaryota	718
Pst6	*Pichia stipitis* CBS 6054	150864483	Fungi	Eukaryota	722
Pgu1	*Pichia guilliermondii* ATCC 6260	146416523	Fungi	Eukaryota	658
Dha4	*Debaryomyces hansenii* CBS767	50423315	Fungi	Eukaryota	723
Sce4	*Saccharomyces cerevisiae* YJM789	162453039	Fungi	Eukaryota	725
Kla4	*Kluyveromyces lactis* NRRL Y-1140	50311091	Fungi	Eukaryota	732
Vpo2	*Vanderwaltozyma polyspora* DSM 70294	156848856	Fungi	Eukaryota	733
Cgl2	*Candida glabrata* CBS 138	116182960	Fungi	Eukaryota	724
Ago2	*Ashbya gossypii* ATCC 10895	45185483	Fungi	Eukaryota	704
Acl2	*Aspergillus clavatus* NRRL 1	121699721	Fungi	Eukaryota	800
Afu3	*Aspergillus fumigatus* Af293	71002356	Fungi	Eukaryota	843
Ani11	*Aspergillus nidulans* FGSC A4	145249626	Fungi	Eukaryota	754
Aor6	*Aspergillus oryzae*	83770379	Fungi	Eukaryota	851
Cim1	*Coccidioides immitis* RS	119186699	Fungi	Eukaryota	797
Ncr2	*Neurospora crassa*	85075374	Fungi	Eukaryota	738
Cne2	*Cryptococcus neoformans* var. neoformans JEC21	58265596	Fungi	Eukaryota	740
Lbi3	*Laccaria bicolor* S238N-H82	164637207	Fungi	Eukaryota	646
Uma2	*Ustilago maydis* 521	71016340	Fungi	Eukaryota	740
Ddi1	*Dictyostelium discoideum* AX4	66802892	Slime Mold	Eukaryota	777
Aor1	*Aspergillus oryzae*	83766128	Fungi	Eukaryota	725
Aca1	*Ajellomyces capsulatus* NAm1	154279250	Fungi	Eukaryota	759
Mgr3	*Magnaporthe grisea* 70-15	39955178	Fungi	Eukaryota	740
Cgl1	*Chaetomium globosum* CBS 148.51	50287709	Fungi	Eukaryota	753
Gze9	*Gibberella zeae* PH-1	46138015	Fungi	Eukaryota	743
Cci4	*Coprinopsis cinerea* okayama7#130	116509017	Fungi	Eukaryota	726
Lbi7	*Laccaria bicolor* S238N-H82	164643762	Fungi	Eukaryota	706
Uma6	*Ustilago maydis* 521	71023771	Fungi	Eukaryota	751
Average protein size ± SD (aa)					742 ± 45
Subcluster 5C (4 proteins)					
Reu1	*Ralstonia eutropha* H16	73539143	Betaproteobacteria	Bacteria	592
Reu2	*Ralstonia eutropha* JMP134	73542650	Betaproteobacteria	Bacteria	593
Rme2	*Ralstonia metallidurans* CH34	94314714	Betaproteobacteria	Bacteria	634
Sau3	*Stigmatella aurantiaca* DW4/3-1	115377807	Deltaproteobacteria	Bacteria	606
Average protein size ± SD (aa)					606 ± 20
Subcluster 5D (37 proteins)					
Vvi1	*Vitis vinifera*	147765903	Viridiplantae	Eukaryota	665
Vvi4	*Vitis vinifera*	147843808	Viridiplantae	Eukaryota	665
Ath13	*Arabidopsis thaliana*	42568235	Viridiplantae	Eukaryota	688
Tca3	*Thlaspi caerulescens*	82468795	Viridiplantae	Eukaryota	693
Vvi6	*Vitis vinifera*	157335740	Viridiplantae	Eukaryota	713
Vvi10	*Vitis vinifera*	157354855	Viridiplantae	Eukaryota	713
Ath8	*Arabidopsis thaliana*	15241078	Viridiplantae	Eukaryota	724
Tca2	*Thlaspi caerulescens*	82468793	Viridiplantae	Eukaryota	716
Osa20	*Oryza sativa* Japonica Group	115466102	Viridiplantae	Eukaryota	708

Table 1 continued

Abbreviation	Organism	GenBank index	Kingdom	Domain	Protein size (aa)
Osa11	*Oryza sativa* Indica Group	115435562	Viridiplantae	Eukaryota	771
Osa30	*Oryza sativa* Indica Group	125562004	Viridiplantae	Eukaryota	717
Osa26	*Oryza sativa* Japonica Group	125549198	Viridiplantae	Eukaryota	724
Osa13	*Oryza sativa* Japonica Group	115455379	Viridiplantae	Eukaryota	882
Osa19	*Oryza sativa* Japonica Group	115462865	Viridiplantae	Eukaryota	694
Vvi2	*Vitis vinifera*	147778971	Viridiplantae	Eukaryota	677
Nta1	*Nicotiana tabacum*	126567465	Viridiplantae	Eukaryota	675
Tca1	*Thlaspi caerulescens*	82468791	Viridiplantae	Eukaryota	672
Ath6	*Arabidopsis thaliana*	15238761	Viridiplantae	Eukaryota	675
Ath1	*Arabidopsis thaliana*	15218331	Viridiplantae	Eukaryota	664
Osa14	*Oryza sativa* Indica Group	115459506	Viridiplantae	Eukaryota	716
Ath4	*Arabidopsis thaliana*	15236800	Viridiplantae	Eukaryota	673
Vvi15	*Vitis vinifera*	157356740	Viridiplantae	Eukaryota	661
Osa2	*Oryza sativa* Japonica Group	38347209	Viridiplantae	Eukaryota	674
Osa15	*Oryza sativa* Japonica Group	115459698	Viridiplantae	Eukaryota	726
Zma1	*Zea mays*	162460137	Viridiplantae	Eukaryota	682
Osa7	*Oryza sativa* Japonica Group	57834124	Viridiplantae	Eukaryota	672
Hvu1	*Hordeum vulgare* subsp. vulgare	84453180	Viridiplantae	Eukaryota	678
Ath5	*Arabidopsis thaliana*	15236912	Viridiplantae	Eukaryota	670
Ath11	*Arabidopsis thaliana*	25083021	Viridiplantae	Eukaryota	677
Osa1	*Oryza sativa* Indica Group	28144882	Viridiplantae	Eukaryota	678
Ppa3	*Physcomitrella patens* subsp. patens	162697041	Viridiplantae	Eukaryota	661
Osa28	*Oryza sativa* Japonica Group	125553884	Viridiplantae	Eukaryota	724
Osa5	*Oryza sativa* Japonica Group	49387869	Viridiplantae	Eukaryota	708
Osa22	*Oryza sativa* Indica Group	116309354	Viridiplantae	Eukaryota	717
Osa23	*Oryza sativa* Japonica Group	116310949	Viridiplantae	Eukaryota	683
Osa21	*Oryza sativa* Japonica Group	115466104	Viridiplantae	Eukaryota	679
Osa4	*Oryza sativa* Japonica Group	42409160	Viridiplantae	Eukaryota	686
Average protein size ± SD (aa)					697 ± 40

Proteins are listed based on position in the phylogenetic tree (Fig. 1, clockwise direction) according to cluster and subcluster. The average sizes of the members of each subcluster are presented below the list of these proteins

Cluster 1 includes three subclusters, 1A–1C; clusters 2 and 3 have two subclusters each, A and B; cluster 4 includes seven subclusters, labeled 4A–4G; and cluster 5 contains four subclusters, 5A–5D (Fig. 1).

The data presented in Table 1 reveal the organismal types and size distributions of these proteins according to subcluster. Thus, for example, subclusters 1A (56 proteins) and 1B (48 proteins) are derived exclusively from fungi, but subcluster 1C (27 proteins) is derived exclusively from plants. Subcluster 1C is more distantly related to 1A and 1B than these latter two subclusters are to each other (Fig. 1). The average sizes of the proteins in subclusters 1A–1C are 825 ± 103 amino acids (aas), 893 ± 41 aas and 761 ± 105 aas, respectively. These size differences are statistically significant and suggest fundamental differences between these three groups of proteins. Plant proteins on average are 11% smaller than fungal proteins. This corresponds to the same average size differences observed between plant and fungal homologues of several other ubiquitous families of transporters, as reported by Chung et al. (2001).

The variations in size within each of these subclusters are also of considerable interest. For example, in subcluster 1A, the four proteins Ncr6, Cgl3, Ssc1 and Gze5 cluster tightly together and are roughly 250 aas larger than most of the other homologues. BLAST searches revealed that the extra amino acids in these proteins are at the N termini, do not comprise a domain recognized by the Conserved Domain Database (CDD) and, although probably homologous, are very diverse in sequence. Another protein of even greater size is Cci3, with 1,292 aas. This protein also exhibits a long N-terminal extension that proved to similarly represent a CDD nonrecognizable domain. It showed similarity to only a few other fungal proteins. Finally, two moderately large fungal proteins, Cne3 and Uma1, have 961–985 aas. The extensions again proved to be at the N termini, and these sequences showed little similarity to other protein sequences in the NCBI database. When these large homologues were removed from the list of subcluster 1A proteins, the average size proved to be 790 ± 30 aas. Thus, we conclude that the basic size of these proteins is about 790 aas, and all of the larger homologues have extra N-terminal hydrophilic extensions.

The variation in size within subcluster 1B is minimal. Several proteins have sizes within the range 900–967 aas, but one protein, Yli7, contains 1,032 aas. This protein was also examined and proved to have an N-terminal extension that was not homologous to anything in the NCBI database. When this protein was removed from subcluster 1B proteins, the average size was 890 ± 36 aas.

Subcluster 1C includes proteins with sizes that vary between 689 and 771 aas with one exception, Osa16. This plant protein shows a long C-terminal hydrophilic extension of about 530 aas. CDD recognized this domain as a member of the pepsin (protease) superfamily. It makes physiological sense that a protease would be fused to a peptide transporter, and thus, it appears likely that this fusion is not artifactual. Two programs, TMHMM (Krogh et al. 2001) and HMMTOP (Tusnady and Simon 2001), were used to determine the orientation of this protein in the membrane. Both programs indicated that the protease domain is located on the cytoplasmic side of the membrane. In fact, these programs showed agreement that most 16-TMS members of the OPT family have both their N and C termini on the inside. Excluding Osa16, the average size for all remaining proteins in this subcluster is 742 ± 20 aas.

Clusters 2 (11 proteins) and 3 (16 proteins) are close together on the phylogenetic tree, and both derive exclusively from fungi. Both clusters can be subdivided into two subclusters; the subclusters in cluster 2 are deep-branching, while those in cluster 3 are not. Cluster 3 proteins have an average size of 788 ± 30 aas, and all proteins occur within the range 746–860 aas. Cluster 2 is of even greater size uniformity except for one protein (Ncr4), which is about twice as large (1,619 aas) as the others. The OPT family homology region begins at about residue 920 with the expected ~16 TMSs, while the first 900 residues exhibit characteristics of a water-soluble protein. A BLAST search against the NCBI database of this region retrieved fungal peptidases from the S41 family. It was therefore clear that Ncr4 is the second OPT family protein identified which has a fused protease domain. However, in contrast to Osa16, which had a C-terminal pepsin fusion, Ncr4 has an N-terminal peptidase S41 homologue fusion. Again, the two programs, TMHMM and HMMTOP, were used to estimate the orientation of this protein in the membrane. Surprisingly, and contrary to results of most other members of the OPT family, these two programs predicted that the N terminus of Ncr4 is on the outside. We therefore examined the distribution of lysine and arginine residues within the transmembrane domain of this protein as well as all members present in the multiple alignment shown in Supplementary Figure S1, which can be viewed on our Web site. In both cases, the results clearly suggested that the N termini are on the cytoplasmic side of the membrane. The mistake made by the two programs may have resulted from incorrect assignments of four cytoplasmic regions that the programs considered transmembrane. Once again, fusion of a peptidase with a peptide transporter makes excellent physiological sense. As expected, based on topological and charge distribution analyses, the cytoplasmic peptidase would hydrolyze the peptides brought in by the transporter in a sequential or coupled process (Saier et al. 2005; Merdanovic et al. 2005; Black and DiRusso 2007).

Cluster 4 (84 proteins) and cluster 5 (83 proteins) are the two largest clusters of OPT family members (about half of the total proteins included), as shown in the top half of the tree in Fig. 1. While cluster 4 can be conveniently divided into seven subclusters, we have divided cluster 5 into 4 subclusters. All cluster 4 proteins are derived from prokaryotes, very few of which are derived from archaea (two in subcluster 4A, one in subcluster 4B, two in subcluster 4F and one in subcluster 4G). Only subcluster 4F lacks bacterial homologues. Within each of these subclusters there is little size variation; thus, the average sizes of subclusters 4A–4D vary between 642 and 665 aas. By contrast, the proteins in subclusters 4E–4G are much smaller (average subcluster size of 529–553 aas). Not even a single protein within these seven subclusters is substantially outside of its subcluster size range. The difference in size between these two groups of subclusters, about 110 residues, proved to be due to a C-terminal extension present in every one of the former proteins but lacking in the latter as well as the loss of several short sequences within the loop regions between transmembrane domains of the latter. This 110-aa extension proved to be unrelated to anything else in the NCBI nr-protein databank.

Cluster 5 is much more divergent with respect to organismal type and size, but each of the four subclusters exhibits a surprising degree of uniformity. Thus, subcluster 5A (15 proteins) derives exclusively from δ- and γ-proteobacteria, and these proteins exhibit an average size of 589 ± 29 aas; no protein is appreciably outside of this range. Subcluster 5B (27 proteins) derives from fungi with one exception, a protein from the slime mold *Dictyostelium discoideum*. The average size is 742 ± 45 aas, and two *Aspergillus* proteins are substantially larger than the others (Afu3, 843 aas; Aor6, 851 aas). Examination of the multiple alignment revealed that these latter two proteins have neither N- nor C-terminal extensions. Instead, both have internal insertions near their N termini immediately preceding TMS 1. These inserts are found only in these two proteins. The other insert is near the C termini of these proteins, immediately preceding the last TMS. Homologous sequences are found in a few other proteins, mostly from species of *Aspergillus*. Neither of these 40-residue inserts shows appreciable sequence similarity with other proteins in the NCBI Protein Database.

Subcluster 5C (four proteins) derives from three β-proteobacteria and one δ-proteobacterium. The average size is 606 ± 20 aas, similar to that of subcluster 5A, also derived from proteobacteria. These proteins are much shorter than the eukaryotic proteins of subclusters 5B and 5D. Subcluster 5D (37 proteins) is derived exclusively from plants and has an average size of 697 ± 40 aas. Only one protein is substantially larger than the others, Osa13 (882 aas). It has an approximately 150-residue C-terminal hydrophilic extension found in no other member of this subcluster. This region of the protein showed a low degree of sequence similarity with chloride transporters of the ClC family (TC 2.A.49). However, the functional significance of this observation is questionable.

One member of each subcluster was used as the query sequence to search TCDB using TC-BLAST. All subclusters in clusters 1–3 (lower half of the tree) proved to bring up peptide transporters, while all of the subclusters from clusters 4 and 5 brought up the iron-complex transporters. The phylogenetic segregation between these two functional types is considerable, suggesting that, in general, function correlates with phylogeny. However, genome context analyses reported below suggest otherwise.

Orthologous Relationships Within Subclusters of the OPT Family Tree

The phylogenetic tree for the 16S/18S rRNAs is shown in Fig. 2. The bacteria appear at the top of this tree, the archaea in the small cluster on the right-hand side and the eukaryotes at the bottom. Every genus included in our study of OPT family members is represented in this tree with the exceptions of *Acidobacteria*, *Ashbya*, *Cryptococcus* and *Thlaspi*. The tree shows that all of the γ- and β-proteobacteria cluster most closely together followed by the α-, δ- and ε-proteobacteria on the upper left-hand side. Surprisingly, in this tree, the ε-proteobacteria cluster loosely with the bacteroidetes, distantly from the other proteobacteria. The cluster on the upper right-hand side of the tree includes a single member of the acidobacteria, a single cluster of actinobacterial rRNAs and two distinct clusters of firmicutes. The eukaryotic branch of the tree shows the slime mold *Dictyostelium* closer to the center of the tree, with the fungal and plant RNAs clustering more closely to each other but much more distantly from the slime mold at the bottom of the tree.

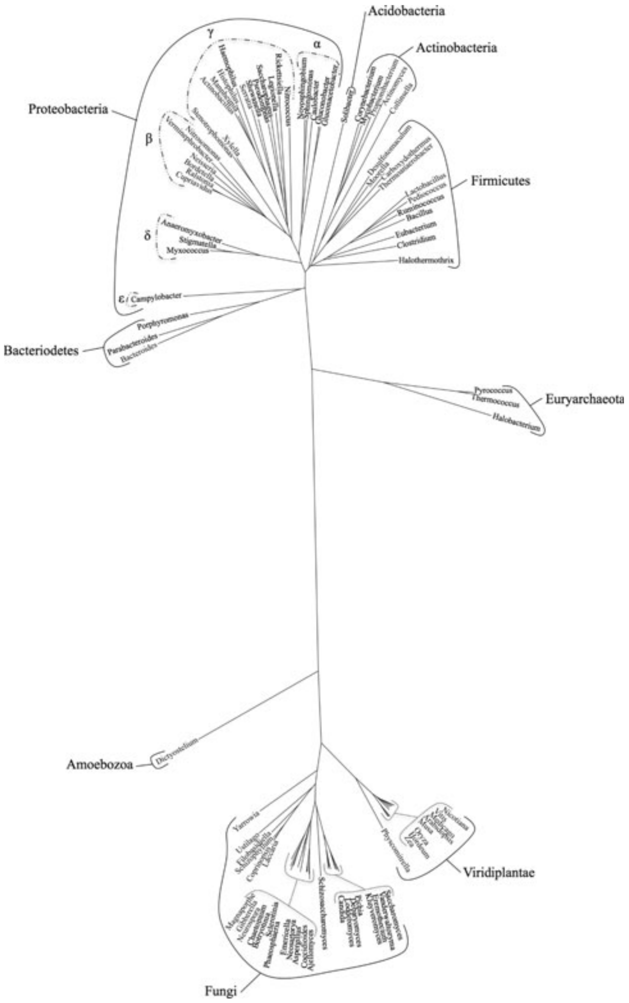

Fig. 2 Phylogenetic tree of 16S/18S rRNAs from all genera represented in this study with the exceptions of *Acidobacteria*, *Ashbya*, *Cryptococcus* and *Thlaspi*. All bacterial rRNAs appear at the top of the tree, the eukaryotic rRNAs are at the bottom of the tree and the three archaeal genera are positioned on the central branch on the right-hand side of the tree. The phylum/kingdom is indicated for each of the clusters, while the genus is shown at the end of each branch

Orthologues are defined as homologues (derived from a single common ancestor) that arose purely as a result of speciation. That is, they arose via vertical, rather than horizontal, transmission, from parent cell to daughter cell throughout their evolutionary histories. This is reflected by approximately the same phylogenetic relationships observed for the proteins under consideration and the 16S rRNAs. The 16S rRNAs are assumed to have arisen solely by vertical transmission. Any set of proteins that exhibit the same relationships to each other as to the 16S rRNAs that were derived from the same species are considered to exhibit orthologous relationships.

Comparing the protein tree (Figs. 1 and S2) with the RNA tree (Fig. 2), we see that in some, but not other, cases orthologous relationships are difficult to establish. This is true for the large cluster 1. For example, subcluster 1C can be subdivided into five sub-subclusters, all but one of which contain paralogues from a single organism. In the largest sub-subcluster, for example, we find five paralogues from *Vitis vinifera*, two from *Oryza sativa* of the Indica group and two from *Arabidopsis thaliana*. The only sub-subcluster that lacks paralogues is the uppermost sub-subcluster with four proteins from four different organisms. Based on the comparison between Figs. 1 and 2, only in this sub-subcluster are the results consistent with orthology.

In the adjacent sub-subcluster, where we find three proteins, one from rice (*Oryza)* and two from thale cress (*Arabidopsis*), it appears that the two thale cress proteins arose by gene duplication after these two organisms diverged from each other. The same situation is observed for the next sub-subcluster, where three *Arabidopsis* proteins cluster tightly together, with a single *V. vinifera* protein being the outlier. We interpret these results to mean that after *Arabidopsis* diverged from *Vitis*, two gene-duplication events in the former organism gave rise to the three paralogues Ath9, Ath16 and Ath17. Similar observations were made for subclusters 1A and 1B.

Cluster 2 shows relationships which suggest orthology. Thus, in both trees, we find the proteins and rRNAs from *Neosartorya*, *Aspergillus* and *Sclerotinia* clustering together; *Candida*, *Lodderomyces* and *Pichia* clustering together; and *Neurospora* and *Botryotinia* clustering together. Even within each of these three groups, the phylogenetic order in both trees is the same. We conclude that cluster 2 probably represents a collection of pure orthologues, with no evidence for paralogues or horizontal gene transfer. This observation suggests that these proteins all serve a single unified function in all of these organisms.

In contrast to cluster 2, cluster 3 contains a number of nonadjacent paralogues and shows clear nonorthologous relationships. The obvious paralogues include two proteins each from *Gibberella zeae* and *Ustilago maydis* in two different subclusters that are shared by this pair of paralogues from these two organisms. Additionally, based on the comparison between Figs. 1 and 2 (see also the dendogram in Fig. S2), Uma4 from *U. maydis* does not show orthologous relationships with the other members of this subcluster. Furthermore, the two *Neurospora crassa* proteins, Ncr5 and Ncr7, are two paralogues within the same sub-subcluster. On the other hand, the three *Aspergillus* proteins and the one from *Neosartorya fisceri* form a sub-subcluster on the protein tree as well as the RNA tree, and the same is true for the two *Schizophyllum* and *Laccaria* proteins and RNAs which form a distinct sub-subcluster in both trees. The relationships of all of these proteins are similar to the corresponding relationships in the rRNA tree and are therefore consistent with orthology.

The prokaryotic proteins were similarly analyzed. Starting with subcluster 4A, we find seven distinct sub-subclusters. Progressing in the clockwise direction, sub-subcluster 1 includes proteins from α- and β-proteobacteria as well as actinobacteria. As a single β-proteobacterial protein is flanked by α-proteobacterial proteins, it is possible that this one β-proteobacterial protein (Neu1) was obtained by horizontal transfer. However, the α-proteobacterial proteins do not show orthologous relationships. The actinobacterial proteins show relationships consistent with orthology.

Sub-subcluster 2 is derived exclusively from *Campylobacter* species. Sub-subcluster 3 contains β-proteobacterial proteins with a single outlier (Pae1) from a γ-proteobacterium. The members of this small sub-subcluster could be orthologous. However, in sub-subclusters 4, 6 and 7, orthology is not possible. For example, in sub-subcluster 4 *Haemophilus* and *Actinobacillus* proteins are interspersed, while in sub-subcluster 7 γ-proteobacterial and archaeal proteins are interspersed. It would appear that the precursor of the two archaeal proteins were obtained from γ-proteobacteria via horizontal transfer, but this remains speculative.

Analyses of subclusters 4B–4G allowed us to come to similar conclusions. Thus, for example, subcluster 4B contains proteins from highly divergent organisms including δ-proteobacteria, acidobacteria, firmicutes and archaea; subcluster 4C includes proteins from two different bacterial phyla, the bacteroidetes and the acidobacteria; subcluster 4E includes just two proteins from two different bacterial phyla; subcluster 4G contains proteins from firmicutes, β- and γ-proteobacteria, and an archaeon. It seems likely that in all of these subclusters horizontal gene transfer was rampant during the evolution of these proteins.

The four cluster 5 subclusters (A–D) were similarly analyzed. Subcluster 5A, derived from δ- and γ-proteobacteria, includes paralogues with little indication of orthology. Subcluster 5B derives from fungi with the exception of one slime mold protein. It also exhibits relationships suggestive of horizontal gene transfer (especially the slime mold protein Ddi1, which probably derived from a fungus) as well as distant paralogues from three different genera. Even the small subcluster 5C shows signs of the existence of horizontal gene transfer since the δ-proteobacterial protein (Sau3) is unexpectedly closely related to the β-proteobacterial proteins. Finally, subcluster 5D shows many paralogous proteins (e.g., at least 12 probable *O. sativa* [Japonica group] paralogues and at least seven *A. thaliana* paralogues). In this case, it is difficult to know if horizontal gene transfer has occurred as all of these proteins could have arisen by vertical transmission from multiple precursor paralogues in the primordial plant.

Topological Analyses of OPT Family Proteins

Figure 3 shows the average hydropathy (top) and average similarity (bottom) plots for all 325 members of the OPT family included in this study. This plot reveals 16 peaks of hydropathy that in general correspond to peaks of similarity. The first four TMSs (labeled 1–4) cluster loosely together. TMSs 4 and 5 are separated by a substantial hydrophilic loop, but again, the next four TMSs (5–8) cluster together. Between TMSs 8 and 9 is an even larger hydrophilic loop, but the remaining eight TMSs cluster tightly together. It is interesting to note that peak 3 and peak 11 appear to divide into two small peaks, possibly due to a misalignment. In fact, there appears to be a gap within the region designated as peak 3 and a smaller gap within the region designated as peak 11. Based on the appearance of this plot, it seemed possible that TMSs 1–8 are repeated in TMSs 9–16. Further, the clustering pattern suggested that these proteins might have arisen from a four-TMS precursor peptide that duplicated twice to give the present-day 16-TMS proteins. In this regard, it should be noted that in all four apparent quadrants the first two TMSs (1–2, 5–6, 9–10 and 13–14) are always close together, while the subsequent two TMSs in each quadrant are separated by greater distances. Following TMS 16 is a poorly conserved region that exhibits moderate hydrophobicity.

When the individual subclusters shown in Fig. 1 were analyzed for average hydropathy and average similarity as shown in Fig. 3 for all members of the family, we found that almost all subclusters exhibit the typical 16-TMS topology. However, the proteins within subclusters 4A–4D appeared to have a seventeenth transmembrane segment that was not part of the C-terminal four-TMS repeat. Also, in these four subclusters TMS 13 showed only moderate hydrophobicity as revealed by the AveHAS program. The origin of putative TMS 17 in these proteins is unknown, but it could have arisen as a result of a gene-fusion event. The long N- and C-terminal hydrophilic extensions have been discussed above, and two of them proved to be homologues of functionally recognizable proteases.

Establishment of Internal Repeats in OPT Family Proteins

As noted above, most members of the OPT family contain 16 putative TMSs, although a few appear to have 17 TMSs, the extra one being at the C terminus of each of the cluster 4A–4D proteins. In order to confirm TMS assignment and establish the evolutionary origins of these proteins, we conducted analyses of potential internal repeats. Although initially analyzed assuming different numbers of TMSs per repeat unit, we were able to show with relative ease that these proteins include an eight-TMS duplication. Thus, when using the IC/GAP programs to compare the first halves of these proteins with the second halves, comparison scores of up to 12.6 SD were obtained (see Table 2, Fig. 4). This value is substantially greater than that required to establish homology (Saier 1994; Yen et al. 2009; Wang et al. 2009; Matias et al. 2010).

We next examined the possibility that the eight-TMS halves themselves arose by an earlier intragenic duplication event from a four-TMS precursor. The results from these analyses are also presented in Table 2, and the alignment upon which the best comparison score was based is shown in Fig. 5. In Table 2, we summarize the results obtained using the IC and GAP programs with 500 random shuffles and default settings. All four quarters of these proteins were compared with each other. Only the top two scores are reported, and these were averaged. For all comparisons, values in excess of 10 SD were obtained, clearly indicating homology. However, the best scores were obtained when A vs. C and B vs. D were compared (12.2 and 13.2 SD, respectively). The fact that higher values were obtained for these two comparisons than for any of the others provides evidence that these two duplication events, giving rise to

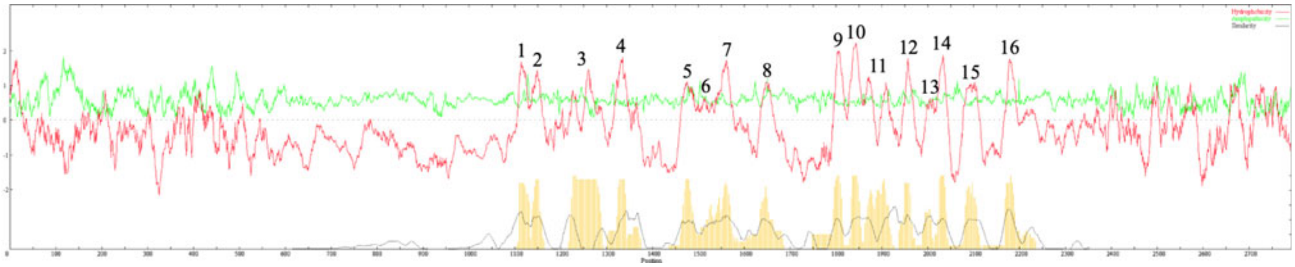

Fig. 3 Average hydropathy, amphipathicity and similarity plots for the 325 OPT family proteins included in this study. The majority of OPT proteins contain 16 TMSs, which correspond to the 16 conserved peaks labeled *1–16*. The central portion of this plot includes all 16 peaks of hydrophobicity which comprise the transporter domain. Functional assignments for the N- and C-terminal hydrophilic domains are discussed in the text. *Upper graphs* show average hydropathy (*dark line*) and average amphipathicity (*light line*), while the *bottom graph* shows average similarity (*continuous line*) as well as average hydropathy using a different program (*vertical lines*)

Table 2 Comparison of different segments within OPT proteins using the GAP and IC programs

Comparison	Segment	Protein-1	Amino acids	TMS	Protein-2	Amino acids	TMS	IC/GAP score (SD)	Average score (SD)
1. 1–8 vs. 9–16	AB vs. CD	Spr1	16–241	1–8	Lsa1	358–589	9–16	12.6	12.0
	AB vs. CD	Zma1	51–216	1–4	Chy1	358–505	9–12	11.3	
2. 1–4 vs. 5–8	A vs. B	Hso1	41–139	1–3	Sde1	174–270	5–7	11.9	11.3
	A vs. B	Ngo1	45–143	1–3	Sde1	174–270	5–7	10.7	
3. 1–4 vs. 9–12	A vs. C	Zma1	51–159	1–3	Chy1	358–455	9–11	12.5	12.2
	A vs. C	Mth1	14–123	1–4	Mgr3	467–577	9–12	11.9	
4. 1–4 vs. 13–16	A vs. D	Gze4	139–266	1–2	Sus1	532–662	13–14	12.1	11.9
	A vs. D	Mxa5	54–147	1–3	Ckl1	512–604	13–15	11.8	
5. 5–8 vs. 9–12	B vs. C	Sco1	327–427	7–8	Mtu1	366–461	11–12	12.2	11.6
	B vs. C	Sco1	320–435	6–8	Ath5	414–531	10–12	10.9	
6. 5–8 vs. 13–16	B vs. D	Osa28	315–421	6–8	Asu1	550–649	14–16	14.1	13.2
	B vs. D	Osa4	202–331	6–8	Msu1	494–621	14–16	12.3	
7. 9–12 vs. 13–16	C vs. D	Vvi4	370–470	9–11	Ath9	602–706	13–15	10.3	10.2
	C vs. D	Pgi1	385–469	10–11	Ani11	606–689	14–15	10.1	
8. 1–2 vs. 3–4	A	Cim2	104–162	1–2	Acl1	176–236	3–4	9.1	8.9
	A	Cim2	118–162	1–2	Pgu9	248–292	3–4	8.7	
9. 5–6 vs. 7–8	B	Nfi3	251–291	5	Yli4	411–450	7	11.5	11
	B	Ani11	210–260	5	Tko1	244–294	7	10.5	
10. 9–10 vs. 11–12	C	Sus2	351–394	9–10	Cco1	388–431	11–12	8.6	8.6
	C	Asu1	313–369	9–10	Pdi1	421–475	11–12	8.5	

Entry 1 presents comparisons for the first eight-TMS half versus the second eight-TMS half. *Entries 2–7* present comparisons for the four four-TMS quarters compared to each other. *Entries 8–10* present comparisons for four representative adjacent 2 TMS hairpin structures

the 16-TMS proteins, were separated by a substantial period of evolutionary time. Thus, we suggest that the primordial four-TMS-encoding genetic element duplicated once to give the eight-TMS precursor and then, later, the second duplication occurred, giving rise to the 16-TMS proteins. Alternatively, segments A and C may share a structure/function that is substantially different from the structure/function shared by segments B and D (see "Discussion" section).

As the final step, we examined the possibility that within each of the four-TMS quadrants of these proteins we could detect two two-TMS repeat sequences. Much to our surprise and delight, this possibility could be demonstrated. As shown in Table 2 and Fig. 6, comparing the first two TMSs with the second two TMSs of the first of these four four-TMS repeats gave a maximal value of 8.9 SD, which was insufficient to establish homology. However, when comparing the two two-TMS segments of the second of these four repeats, we were able to get comparison scores in excess of 10 SD, thus establishing homology. In this case, the alignment giving this value included all of TMS 5 compared to TMS 7. When the same was done with the third of these four repeats, a maximal value of 8.6 SD was obtained. The same procedure with the fourth of these four repeats did not give values above 7 SD. Applying the superfamily principle, the values obtained clearly indicate that these proteins arose from an initial two-TMS precursor. We therefore conclude that members of the OPT superfamily arose in three steps: duplication of two TMSs to give four, duplication of four-TMSs to give eight and duplication of eight-TMSs to give 16. The addition of a seventeenth TMS to a small fraction of these proteins presumably occurred as a result of a late gene-fusion event in just one phylogenetic cluster of these proteins.

Use and Evaluation of Programs to Detect Similarity and Establish Homology

To confirm the results obtained using the IC/GAP programs, three other programs capable of identifying sequence similarity between repeat segments were used. These programs were GGSEARCH, HMMER and SAM (Table 3). All three programs substantiated the conclusions obtained with IC/GAP. For example, when the two halves were compared with GGSEARCH, a value of 1.7e^{-8} was obtained. The best value resulting from the use of the HMMER program was 4e^{-4}. When SAM was used, the best value was 4e^{-3}. All of these values confirm our conclusion of homology.

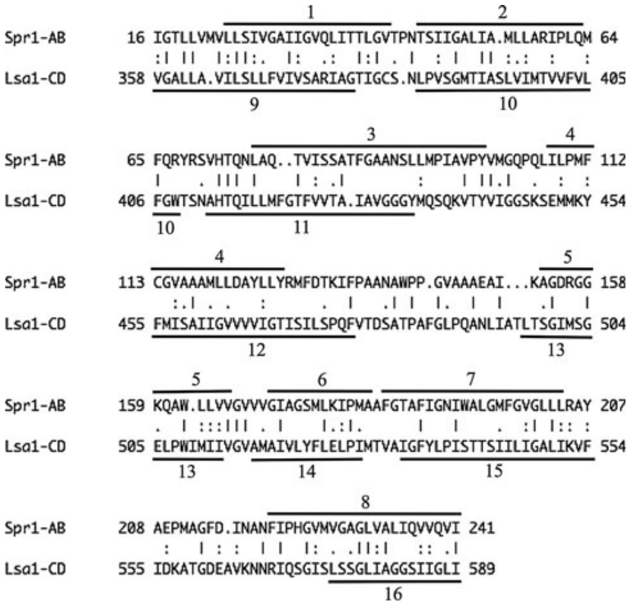
1
2
Spr1-AB 16 IGTLLVMVLLSIVGAIIGVQLITTLGVTPNTSIIGALIA.MLLARIPLQM 64
Lsa1-CD 358 VGALLA.VILSLLFVIVSARIAGTIGCS.NLPVSGMTIASLVIMTVVFVL 405
9
10
3
4
Spr1-AB 65 FQRYRSVHTQNLAQ..TVISSATFGAANSLLMPIAVPYVMGQPQLILPMF 112
Lsa1-CD 406 FGWTSNAHTQILLMFGTFVVTA.IAVGGGYMQSQKVTYVIGGSKSEMMKY 454
10
11
4
5
Spr1-AB 113 CGVAAAMLLDAYLLYRMFDTKIFPAANAWPP.GVAAAEAI...KAGDRGG 158
Lsa1-CD 455 FMISAIIGVVVVIGTISILSPQFVTDSATPAFGLPQANLIATLTSGIMSG 504
12
13
5
6
7
Spr1-AB 159 KQAW.LLVVGVVVGIAGSMLKIPMAAFGTAFIGNIWALGMFGVGLLLRAY 207
Lsa1-CD 505 ELPWIMIIVGVAMAIVLYFLELPIMTVAIGFYLPISTTSIILIGALIKVF 554
13
14
15
8
Spr1-AB 208 AEPMAGFD.INANFIPHGVMVGAGLVALIQVVQVI 241
Lsa1-CD 555 IDKATGDEAVKNNRIQSGISLSSGLIAGGSIIGLI 589
16

Fig. 4 Alignment of OPT TMSs 1–8 of Spr1 (*Serratia proteamaculans*, gi 157369266) with OPT TMSs 9–16 of Lsa1 (*Lactobacillus sakei*, gi 81427933). The IC program was used to identify the two internal segments exhibiting the greatest statistical similarity. The GAP program was used to generate the alignment with default settings and 500 random shuffles. Numbers at the beginning and end of each line indicate the residue numbers in the proteins. The *vertical line* represents an identity, the *colon* represents a close similiarity and the *period* represents a more distant similarity. This convention of presentation is also used in Figs. 5 and 6. In all three figures, positions of the TMSs were predicted using the TMHMM program. This alignment gave a comparison score of 12.6 SD

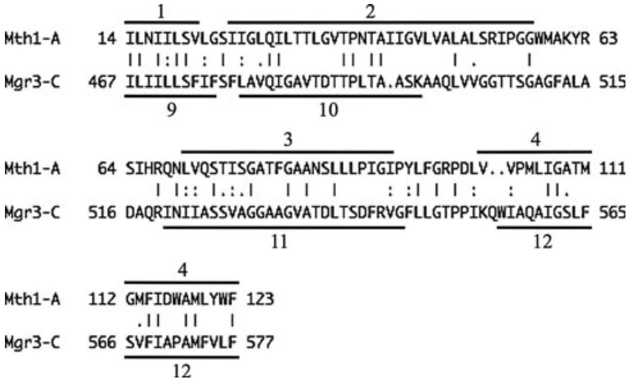
1
2
Mth1-A 14 ILNIILSVLGSIIGLQILTTLGVTPNTAIIGVLVALALSRIPGGWMAKYR 63
Mgr3-C 467 ILIILLSFIFSFLAVQIGAVTDTTPLTA.ASKAAQLVVGGTTSGAGFALA 515
9
10
3
4
Mth1-A 64 SIHRQNLVQSTISGATFGAANSLLLPIGIPYLFGRPDLV..VPMLIGATM 111
Mgr3-C 516 DAQRINIIASSVAGGAAGVATDLTSDFRVGFLLGTPPIKQWIAQAIGSLF 565
11
12
4
Mth1-A 112 GMFIDWAMLYWF 123
Mgr3-C 566 SVFIAPAMFVLF 577
12

Fig. 5 Alignment of OPT TMSs 1–4 of Mth1 (*Moorella thermoacetica*, gi 83589078) with OPT TMSs 9–12 of Mgr3 (*Magnaporthe grisea*, gi 39955178). This alignment gave a comparison score of 11.9 SD

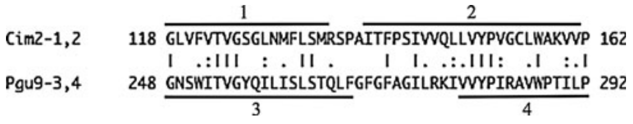
1
2
Cim2-1,2 118 GLVFVTVGSGLNMFLSMRSPAITFPSIVVQLLVYPVGCLWAKVVP 162
Pgu9-3,4 248 GNSWITVGYQILISLSTQLFGFGFAGILRKIVVYPIRAVWPTILP 292
3
4

Fig. 6 Alignment of OPT TMSs 1 and 2 of Cim2 (*Coccidioides immitis*, gi 119190959) with OPT TMSs 3 and 4 of Pgu9 (*Pichia guilliermondii*, gi 146422868). This alignment gave a comparison score of 8.7 SD

When the four quarters of the OPT family proteins were compared, again the best values were usually obtained when segments A were compared with segments C and when segments B were compared with segments D. Thus, when using GGSEARCH, the values for these two comparisons were 8.6e^{-6} and 3.9e^{-8}. When using HMMER, the best values were 0.03 and 0.006. With SAM, the best values were 0.002 and 0.001, respectively (Table 3). As revealed by the data in Table 3, only in two instances were values obtained in the other comparisons comparable to these. These results confirm that (1) the four four-TMS quarters of OPT family proteins are all homologous and therefore derive from a common origin, (2) the first and third four-TMS segments are more similar to each other than they are to the second and fourth TMS segments and (3) the second and fourth TMS segments are more similar to each other than they are to the first and third segments.

Functional Predictions Based on Genome Context

Each subcluster was examined using the SEED database in order to allow prediction of potential substrates. These analyses were conducted only for prokaryotic clusters found in clusters 4 and 5. These subclusters will be analyzed sequentially.

Ccr1 in subcluster 4A (see Table 1) is present within a gene cluster that includes an acetyl transferase of the GNAT family (position 2), many members of which are aminoacyl and aminoglycoside transferases. Also present is a PhoH-like ATPase with a Rossman fold similar to ArsA of *Escherichia coli*. In the same operon with the oligopeptide transporter gene, we also find a glutathione *S*-transferase-like protein, which undoubtedly uses glutathione as a substrate for nucleophilic addition reactions involving electrophiles. Another nearby gene encodes a protein with a peptidoglycan-binding domain, presumably to anchor the protein or a protein complex to the cell wall. These observations suggest that this particular OPT family permease may be a peptide transporter specific for glutathione. Also within subcluster 4A, Nmo1 was examined and proved in several genomes to be directly upstream of and transcribed in the same direction as genes encoding dipeptidyl aminopeptidases (position 2). Again, it appears likely that Nmo1 is a peptide uptake porter. A third protein examined was Reu3, which brought up operons in various organisms, several of which encoded peptidases of different designations. Again, the most likely function for this protein appears to be a peptide transporter. We therefore conclude that many or perhaps all of the OPT family members in subcluster 4A are peptide transporters.

Table 3 Comparison of different segments within OPT proteins using the GGSEARCH, HMMER and SAM programs (The format of presentation is the same as for Table 2)

Comparison	Superfamily	Family, TC	Profile		Database		GGSEARCH	HMMER	SAM
			Protein-1	Acc	Protein-2	Acc	(e-value)	(e-value)	(e-value)
1	OPT AB vs. CD	2.A.67.3	Spr1	YP_001477255.1	Lsa1	YP_394932.1	$1.7e^{-8}$	$4.0e^{-4}$	0.1
	OPT CD vs. AB	2.A.67.4	Lsa1	YP_394932.1	Spr1	YP_001477255.1	$7.7e^{-7}$	0.004	0.004
2	OPT A vs. B	2.A.67.4	Ngo1	YP_208927.1	Sde1	YP_526125.1	$5.8e^{-6}$	0.06	0.5
	OPT B vs. A	2.A.67.4	Sde1	YP_526125.1	Ngo1	YP_208927.1	$3.2e^{-5}$	0.2	0.09
3	OPT A vs. C	2.A.67.2	Zma1	NP_001104952.1	Chy1	YP_361078.1	$8.6e^{-6}$	0.03	0.002
	OPT C vs. A	2.A.67.4	Chy1	YP_361078.1	Zma1	NP_001104952.1	$9.2e^{-6}$	0.03	0.02
4	OPT A vs. D	2.A.67.1	Gze4	XP_389463.1	Sus1	YP_822933.1	$8.0e^{-4}$	0.09	2
	OPT D vs. A	2.A.67.4	Sus1	YP_822933.1	Gze4	XP_389463.1	$1.4e^{-4}$	0.03	0.2
5	OPT B vs. C	2.A.67.1	Sco1	AAF26618.1	Mtu1	NP_216911.1	$3.6e^{-2}$	0.07	0.01
	OPT C vs. B	2.A.67.4	Mtu1	NP_216911.1	Sco1	AAF26618.1	$1.9e^{-3}$	0.08	0.003
6	OPT B vs. D	2.A.67.2	Osa28	CAE02279.2	Asu1	YP_001343430.1	$3.9e^{-8}$	0.006	0.02
	OPT D vs. B	2.A.67.4	Asu1	YP_001343430.1	Osa28	CAE02279.2	$3.7e^{-4}$	0.007	0.001
7	OPT C vs. D	2.A.67.4	Pgi1	NP_904744.1	Ani11	XP_658304.1	$2.4e^{-4}$	0.2	2
	OPT D vs. C	2.A.67.2	Ani11	XP_658304.1	Pgi1	NP_904744.1	$2.0e^{-4}$	0.05	0.5

Cno1 within subcluster 4B proved to be related to operons which encode ornithine carbamoyl transferases, alanine symporters, potential *N*-acetyl muramoyl-L-alanine amidases and enzymes involved in glutamate metabolism. Because of the association of amino acid metablic enzymes, we again predict that these proteins take up peptides. When Aba2 was examined, several operons appeared to encode dipeptidyl aminopeptidases downstream of the OPT family transporter. Thus, we conclude that subcluster 4B proteins also transport peptides.

When Bfr1 of subcluster 4C was examined using SEED, a frequently cotranscribed gene encodes an endo-1, 4-*β*-xylanase, which may be anchored to the outer membrane. The transcriptional regulator for this operon appears to be a member of the AraC family. Similar results were obtained when Pdi1 was examined. We interpret these results to suggest that subcluster 4C OPT family proteins may be concerned with uptake of xylan-oligosaccharides.

Lca1 of subcluster 4D proved to be present in a gene cluster which also includes genes encoding catabolic threonine dehydrotase, a dipeptidase and an alanine dehydrogenase. In fact, these proteins appear to be in a single operon in the closely related species *Pediococcus pentosaceus*. Similar results were obtained when Cbe1 was used as the query sequence. We therefore conclude that these proteins are peptide transporters.

Only two proteins comprise subcluster 4E. In the gene cluster with Cce1, we identified genes encoding a pantothenate kinase as well as phospholipases. The other member of this subcluster is from an organism that is not included in the SEED database. These results may suggest that the substrate of this and related transporters could be a phospholipid, but the data are insufficient to make such a prediction with confidence.

Subcluster 4F could not be examined as representation was not present in SEED. However, subcluster 4G included Vei1 in a gene cluster that appeared to be involved in aromatic amino acid metabolism. While we might therefore predict that these transporters are also peptide uptake systems, we again do not believe the evidence is sufficient to make this prediction with confidence.

Cluster 5 proteins include four subclusters. Subclusters 5A and 5C include proteins derived from prokaryotes, while subclusters 5B and 5D include proteins only from eukaryotes. We therefore examined the former two clusters. Examining Ade2 of subcluster 5A, we observed a probable regulatory serine/threonine kinase (position 2) as well as components of a pyruvate/α-ketoglutarate dehydrogenase complex. We also identified an octanoate-[acyl-carrier-protein]-protein-n-octanoyl transferase, a deoxyribonuclease, a protein that recognizes phosphothreonine residues in proteins as well as an aspartokinase involved in threonine and homoserine biosynthesis. Another protein in this subcluster, Ppu1, brought up in position 2 a glutathione *S*-transferase as well as a putative transcriptional regulator of the LysR type. Finally, Asp4 brought up a glycosyl transferase as well as an NADPH-dependent reductase. We are therefore hesitant to make predictions for the members of this subcluster.

Subcluster 5C includes Reu2, which proved to be encoded by a gene that colocalizes with a zinc-binding protein encoding gene (position 2) and a mutT mutator protein (7,8-dihydro-8-oxoguanine-triphosphatase), with all three probably in a single operon. This operon may be regulated by an AsnC-type transcriptional regulator. Nearby genes also encode a putative ATP/GTP-binding protein, a dephospho-CoA kinase and components of either pyruvate or α-ketoglutarate dehydrogenase complexes. We tentatively suggest that these transporters might be nucleoside or oligonucleotide transporters.

Discussion

In this article, we have described the OPT family of peptide and iron-siderophore uptake transporters and have defined the evolutionary pathway by which these proteins arose. This pathway is illustrated in Fig. 7. A genetic element encoding a two-TMS precursor duplicated to give four TMSs, this duplicated again to give eight TMSs and this also duplicated to give the final 16-TMS topology. In few instances has it been possible to trace back the evolutionary history as far as we have done for the OPT family (Saier 2003). Furthermore, in no other instance has this particular pathway been demonstrated for any other family of transport proteins (Saier 2003 and unpublished observations).

We could demonstrate greater similarities between TMSs 1–4 and TMSs 9–12, as well as between TMSs 5–8 and TMSs 13–16, than for other quadrants compared, suggesting that there was a reasonable period of evolutionary time between these two last duplication events. However, the fact that similar maximal values were obtained for the eight-TMS halves, the four-TMS quarters and the two-TMS eighths suggests that all three of these duplication events happened in a relatively short period of evolutionary time. These two apparent inconsistencies could be resolved if the first and third quadrants serve a common structure/function that differs from that of quadrants 2 and 4. In an analogous situation where a six-TMS voltage-gated ion channel has four six-TMS repeats, this last possibility seemed unlikely (Nelson et al. 1999).

A similar situation has been suggested for members of the mitochondrial carrier family which underwent triplication of a primordial two-TMS-encoding genetic element (Kuan and Saier 1993a, b). This family of proteins appears to have undergone rapid intragenic and extragenic duplication events, giving rise not only to the six-TMS porters but also to the main functional types or subfamilies within a relatively short period of time (Kuan and Saier 1993a). Interestingly, in the mitochondrial carriers, the third thirds of these proteins diverged in sequence more than the first two thirds (Kuan and Saier 1993a). The explanation for this observation is not yet clear, but possibly, the last two TMSs are of less functional importance than the first four.

Many transporters have been shown to arise from a two-TMS precursor, but in no case has it been possible to demonstrate three sequential duplication events. Other families in which a two-TMS element duplicated to give four TMSs include the voltage-gated ion channel (VIC, TC 1.A.1) family, the c-subunits of F-type ATPases (F-ATPase, TC 3.A.2) which both duplicated and triplicated and the YiaAB family (TC 9.B.44) (Saier 2003d). Several examples of 4 TMS transmembrane proteins that arose from duplication of a simple 2 TMS hairpin structure have been documented (Sawhney M, Tamang DG and Saier MH Jr., unpublished observations).

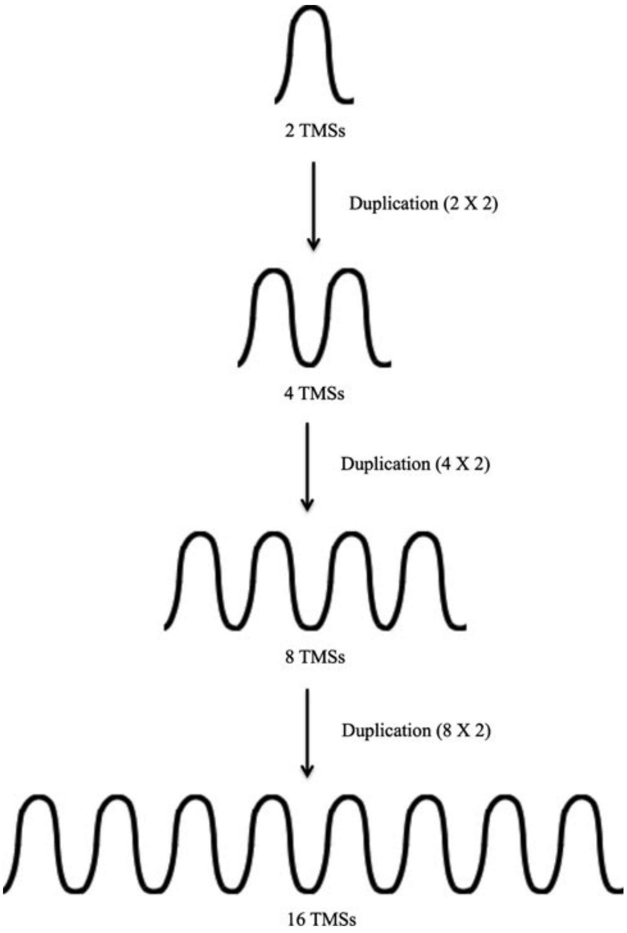

Fig. 7 Proposed pathway for the evolutionary appearance of present-day OPT family proteins. Evidence is presented that the ultimate precursor of the 16- (and sometimes 17-) TMS proteins was a two-TMS hairpin structure (*top*). This then duplicated three times: first to give the four-TMS intermediate, second to give the eight-TMS intermediate and last to give the present-day 16-TMS proteins. Evidence was presented that either the duplication of four TMSs to give eight TMSs occurred substantially before the duplication of eight-TMSs that gave rise to the 16-TMS permeases or segments 1 and 3 share functional/structural features not shared by segments 2 and 4 (see "Discussion" section). In the 17-TMS proteins, the extra TMS is at the C termini of these homologues

A surprising observation was that all members of the OPT family have either 16 or 17 TMSs. The vast majority have 16 TMSs, while a smaller fraction (subclusters 4A–4D in the phylogenetic tree shown in Fig. 1) have 17 putative TMSs. In fact, no 17-TMS protein was found outside of subclusters 4A–4D, and only 17-TMS proteins were found in these four subclusters. The extra TMS at the C termini of these proteins most probably arose only once during the evolution of this family. The only additional variations apparently resulted from the fusion of these integral membrane proteins with soluble domains, two of which could be recognized on the basis of homology searches. In these two cases the fused domains proved to correspond to two different families of peptidases. Since the transporters were predicted to function in peptide uptake and the peptidase domains were predicted to be localized to the cytoplasmic side of the membrane, the fusion of these two catalytic proteins made excellent physiological sense. The peptidase domain probably hydrolyzes peptides upon entry into the cell, possibly in a tightly or loosely coupled process. If tightly coupled, this could be a novel example of group translocation where chemical modification of the substrate is coupled to its transport (Herbert et al. 2003; Hirsch et al. 1998; Merdanovic et al. 2005; Saier et al. 2005).

Uniformity of topology is found in some families, while others show tremendous variation. For example, all recognized proteins in the mitochondrial carrier family (TC 2.A.29) have six TMSs, and no exception has yet been reported (Kuan and Saier 1993a and unpublished results). Another example is the largest superfamily of secondary carriers, the major facilitator superfamily (TC 2.A.1). All recognized members of this superfamily have either 12 or 14 TMSs, where the extra two TMSs in the 14-TMS proteins are present in the center between the two six-TMS repeat units, and they occur only in three of the 70 currently recognized MFS families. This situation is to be contrasted with families that show tremendous topological variations. These include the integral membrane cytochrome *c* biogenesis proteins of the heme handling protein family (TC 9.B.14) (Lee et al. 2007) and the SdpI family of receptor/signal-transduction proteins (TC 9.A.32) (Povolotsky et al. 2010). In both of these cases, the families include proteins having a wide variety of topological types with numbers of TMSs ranging anywhere from three to 12. Further, they can have segments present in inverted order in some of the proteins relative to other members of the same family. In the SdpI family, this is understood because the different three-TMS repeat segments within these proteins probably serve distinct subfunctions (Povolotsky et al. 2010).

OPT family members were found in both eukaryotes and prokaryotes. The vast majority of the eukaryotic proteins were derived from fungi (subclusters 1A, 1B and 5B as well as clusters 2 and 3) and plants (subclusters 1C and 5D). The only exception is a single slime mold homologue found in subcluster 5B, a cluster otherwise derived entirely from fungi. We hypothesize that this one homologue from *D. discoideum* was acquired by horizontal transfer from a fungus, a suggestion that is not surprising since slime molds eat other microorganisms (Eichinger et al. 2005). However, we obtained no evidence for horizontal transfer within and between fungi and plants. In view of the fact that homologues of these proteins are found in many bacterial and archaeal phyla, it is surprising that these proteins are not found within the animal kingdom or any of the unicellular eukaryotes except for slime molds.

Prokaryotic homologues of the OPT family are found in subclusters 4A–4G as well as 5A and 5C. In contrast to the situation with eukaryotes, apparent horizontal transfer within and between prokaryotic phyla has been rampant. For example, in subcluster 4A, proteins are derived from four of the five common classes of proteobacteria, the only exception being the δ-proteobacteria. However, this subcluster also contains proteins from actinobacteria and even euryarchaeota. Similarly, subcluster 4B includes proteins from δ-proteobacteria, firmicutes, acidobacteria and euryarchaeota. Subcluster 4C has protein representation only from bacteroidetes and acidobacteria. Subcluster 4D is one of the few “pure” prokaryotic subclusters where all of the proteins derive from firmicutes. Subcluster 4G, a small subcluster of seven proteins, is surprisingly diverse, having members from firmicutes, β- and γ-proteobacteria and euryarchaeota. Finally, subcluster 5A has representation only from γ- and δ-proteobacteria, while subcluster 5C has representation only from β- and δ-proteobacteria. These observations can be interpreted to suggest that horizontal transfer between phyla has occurred in all but two of the prokaryotic subclusters identified in this study.

The large OPT family consists of peptide and iron-siderophore uptake porters, and based on functionally characterized eukaryotic members of this family, iron-siderophore transporters (clusters 4 and 5) segregate from peptide transporters (clusters 1–3). Our operon and genomic context analyses, however, suggest that prokaryotic members of the OPT family are often peptide transporters. This was true for subclusters 4A, 4B and 4D and possibly for 4G and 5A. However, the small subcluster 4C appears more likely to be specific for oligosaccharides, specifically for β-xylan-oligosaccharides. Furthermore, weak evidence suggests that subcluster 5C proteins might be nucleoside or oligonucleotide transporters. At least one eukaryotic OPT can transport both peptides and iron-siderophores. Further, some of the phytosiderophores and mugineic acids resemble peptides in structure. Thus, although the OPT family includes members capable of taking up both types of substrates, there is a need to provide functional analyses of prokaryotic OPTs of the various subclusters in order to establish the range of substrates transported by members of this family.

We have no clear explanation as to why OPT family members appear to be lacking in the animal kingdom as well as many eukaryotic protists. It is possible that these proteins entered the eukaryotic domain from prokaryotes late by horizontal transfer rather than early by vertical descent and that they were either obtained only by fungi and plants (our preferred explanation) or lost from the animal kingdom as well as many eukaryotic protists. If further genome sequencing reveals the presence of these homologues in other types of eukaryotes, this will raise the question of whether these arose by horizontal gene transfer from fungi, plants or slime molds. This may be an important question since in this study we found very little evidence for horizontal transfer between eukaryotic phyla. Future functional analyses and further sequencing efforts are likely to provide eventual answers to these questions. We hope that the analyses reported here will serve as useful guides for molecular biological and bioinformatic analyses of this important family of transporters.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Black PN, DiRusso CC (2007) Vectorial acylation: linking fatty acid transport and activation to metabolic trafficking. Novartis Found Symp 286:127–138 (discussion 138–141, 162–163, 196–203)

Busch W, Saier MH Jr (2004) The IUBMB-endorsed transporter classification system. Mol Biotechnol 27:253–262

Cagnac O, Bourbouloux A, Chakrabarty D, Zhang MY, Delrot S (2004) AtOPT6 transports glutathione derivatives and is induced by primisulfuron. Plant Physiol 135:1378–1387

Chung YJ, Krueger C, Metzgar D, Saier MH Jr (2001) Size comparisons among integral membrane transport protein homologues in bacteria, Archaea, and Eucarya. J Bacteriol 183: 1012–1021

Curie C, Panaviene Z, Loulergue C, Dellaporta SL, Briat JF, Walker EL (2001) Maize yellow stripe1 encodes a membrane protein directly involved in Fe(III) uptake. Nature 409:346–349

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

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–395

Dworeck T, Wolf K, Zimmermann M (2009) SpOPT1, a member of the oligopeptide family (OPT) of the fission yeast *Schizosaccharomyces pombe*, is involved in the transport of glutathione through the outer membrane of the cell. Yeast 26:67–73

Eddy SR (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol 4:e1000069

Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Madan Babu M, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail MA, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox EC, Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel AA, Barrell B, Kuspa A (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43–57

Hauser M, Narita V, Donhardt AM, Naider F, Becker JM (2001) Multiplicity and regulation of genes encoding peptide transporters in *Saccharomyces cerevisiae*. Mol Membr Biol 18:105–112

Herbert M, Sauer E, Smethurst G, Kraiss A, Hilpert AK, Reidl J (2003) Nicotinamide ribosyl uptake mutants in *Haemophilus influenzae*. Infect Immun 71:5398–5401

Hirsch D, Stahl A, Lodish HF (1998) A family of fatty acid transporters conserved from mycobacterium to man. Proc Natl Acad Sci USA 95:8625–8629

Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res 35:W429–W432

Kaur J, Srikanth CV, Bachhawat AK (2009) Differential roles played by the native cysteine residues of the yeast glutathione transporter, Hgt1p. FEMS Yeast Res 9:849–866

Koh S, Wiles AM, Sharp JS, Naider FR, Becker JM, Stacey G (2002) An oligopeptide transporter gene family in *Arabidopsis*. Plant Physiol 128:21–29

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Kuan J, Saier MH Jr (1993a) The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. Crit Rev Biochem Mol Biol 28:209–233

Kuan J, Saier MH Jr (1993b) Expansion of the mitochondrial carrier family. Res Microbiol 144:671–672

Lee JH, Harvat EM, Stevens JM, Ferguson SJ, Saier MH Jr (2007) Evolutionary origins of members of a superfamily of integral membrane cytochrome *c* biogenesis proteins. Biochim Biophys Acta 1768:2164–2181

Lubkowitz M (2006) The OPT family functions in long-distance peptide and metal transport in plants. Genet Eng (NY) 27:35–55

Lubkowitz MA, Barnes D, Breslav M, Burchfield A, Naider F, Becker JM (1998) *Schizosaccharomyces pombe* isp4 encodes a transporter representing a novel family of oligopeptide transporters. Mol Microbiol 28:729–741

Matias MG, Gomolplitinant KM, Tamang DG, Saier MH Jr (2010) Animal Ca^{2+} release-activated Ca^{2+} (CRAC) channels are homologous to and derived from the ubiquitous cation diffusion facilitators. BCM Res Notes 3(1):158

Merdanovic M, Sauer E, Reidl J (2005) Coupling of NAD^{+} biosynthesis and nicotinamide ribosyl transport: characterization of NadR ribonucleotide kinase mutants of *Haemophilus influenzae*. J Bacteriol 187:4410–4420

Mitchell P, Moyle J (1958) Group-translocation: a consequence of enzyme-catalysed group-transfer. Nature 182:372–373

Nelson RD, Kuan G, Saier MH Jr, Montal M (1999) Modular assembly of voltage-gated channel proteins: a sequence analysis and phylogenetic study. J Mol Microbiol Biotechnol 1:281–287

Osawa H, Stacey G, Gassmann W (2006) ScOPT1 and AtOPT4 function as proton-coupled oligopeptide transporters with broad but distinct substrate specificities. Biochem J 393:267–275

Pao SS, Paulsen IT, Saier MH Jr (1998) Major facilitator superfamily. Microbiol Mol Biol Rev 62:1–34

Paulsen IT, Skurray RA (1994) The POT family of transport proteins. Trends Biochem Sci 19:404

Povolotsky TL, Orlova E, Tamang DG, Saier MH Jr (2010) Defense against cannibalsim: the SpdI family of bacterial immunity/signal transduction proteins. J Membr Biol 235:145–162

Reuss O, Morschhauser J (2006) A family of oligopeptide transporters is required for growth of *Candida albicans* on proteins. Mol Microbiol 60:795–812

Saier MH Jr (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol Rev 58:71–93

Saier MH Jr (2000a) A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 64:354–411

Saier MH Jr (2000b) Families of proteins forming transmembrane channels. J Membr Biol 175:165–180

Saier MH Jr (2000c) Vectorial metabolism and the evolution of transport systems. J Bacteriol 182:5029–5035

Saier MH Jr (2003) Tracing pathways of transport protein evolution. Mol Microbiol 48:1145–1156

Saier MH Jr, Hvorup RN, Barabote RD (2005) Evolution of the bacterial phosphotransferase system: from carriers and enzymes to group translocators. Biochem Soc Trans 33:220–224

Saier MH Jr, Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and informaiton. Nuceic Acids Res 34:181–186

Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. Nucleic Acids Res 37:D274–D278

Stacey MG, Patel A, McClain WE, Mathieu M, Remley M, Rogers EE, Gassmann W, Blevins DG, Stacey G (2008) The *Arabidopsis* AtOPT3 protein functions in metal homeostasis and movement of iron to developing seeds. Plant Physiol 146:589–601

Thakur A, Kaur J, Bachhawat AK (2008) Pgt1, a glutathione transporter from the fission yeast *Schizosaccharomyces pombe*. FEMS Yeast Res 8:916–929

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850

Wang B, Dukarevich M, Sun EI, Yen MR, Saier MH Jr (2009) Membrane porters of ATP-binding cassette transport systems are polyphyletic. J Membr Biol 231:1–10

Wiles AM, Cai H, Naider F, Becker JM (2006) Nutrient regulation of oligopeptide transport in *Saccharomyces cerevisiae*. Microbiology 152:3133–3145

Yen MR, Tseng YH, Saier MH Jr (2001) Maize Yellow Stripe1, an iron-phytosiderophore uptake transporter, is a member of the oligopeptide transporter (OPT) family. Microbiology 147:2881–2883

Yen MR, Choi J, Saier MH Jr (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. J Mol Microbiol Biotechnol 17:163–176

Zhai Y, Saier MH Jr (2001a) A Web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. J Mol Microbiol Biotechnol 3:501–502

Zhai Y, Saier MH Jr (2001b) A Web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. J Mol Microbiol Biotechnol 3:285–286

Zhai Y, Tchieu J, Saier MH Jr (2002) A Web-based Tree View (TV) program for the visualization of phylogenetic trees. J Mol Microbiol Biotechnol 4:69–70