UNIVERSITY OF CALIFORNIA, SAN DIEGO

Characterization of the Putative Sulfate Exporter Family

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Biology

by

Vincent Hoang Trong Lam

Committee in charge:

Professor Milton H. Saier, Jr., Chair
Professor Russell F. Doolittle
Professor James Nieh

2010

The Thesis of Vincent Hoang Trong Lam is approved and is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2010

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT OF THE THESIS

Characterization of the Putative Sulfate Exporter Family

by

Vincent Hoang Trong Lam

Master of Science in Biology

University of California, San Diego, 2010

Professor Milton H. Saier, Jr., Chair

The Putative Sulfate Exporter (PSE) family consists primarily of bacterial and some archael family members. It is not yet functionally characterized and is tentatively placed into this family based on its folds. PSE family members are predicted to have 11 transmembrane segments (TMSs). The phylogeny, topology, internal homology, and functionality of the PSE family were analyzed. Evidence suggests that the PSE encoding genes are located on mobile genetic elements, and a 6 + 1 + 6 or 7 x 2 TMS repeat is proposed based on the alignments from SSSearch and GAP. SEED analysis shows that it is likely a sulfate exporter.

Introduction

The putative sulfate exporter (PSE) family (2.A.98) in the Transporter Classification Database (TCDB) is comprised of proteins from both bacterial and archaeal origins. Sulfate uptake has been extensively studied (Brüggemann *et al*., 1966; Pilsyk & Paszewski, 2009), yet its export has yet to garner the same amount of attention (Brüggemann *et al*., 2004).

Sulfate plays many different and important roles in bacteria. Anaerobic species use it as a terminal electron acceptor for energy production; they reduce sulfate to hydrogen sulfide. It can also be reduced to synthesize essential sulfur containing cellular components in various assimilatory pathways. However, the PSE family proteins do not seem to be responsible or related to either of these functions. Although sulfate export has been sparsely researched, it has been hypothesized that PSE proteins are used to maintain osmotic homeostasis within the cell (Rein *et al*., 2005). They may, however, also play protective roles, protecting the organism from inhibitory concentrations of inorganic anionic species.

The presence of a gene possibly coding for a sulfate exporter was revealed while studying the dissimilation of cysteate in *Paracoccus pantorophus* (Rein, 2005). Cysteate is taken into the cell to form 3-sulfolactate. This is then converted into pyruvate, for use in an intermediate metabolic pathway, and sulfite is the byproduct, the two being formed in a 1:1 ratio. The genes, *suyAB*, encode sulfolactate sulfo-lyase, which is responsible for the desulfonation reaction. Sulfite is then converted into sulfate by a sulfite

dehydrogenase. *suyZ*, a gene downstream of *suyAB*, tentatively codes for the PSE. It is hypothesized that this 10 transmembrane segment (TMS) spanning protein is responsible for pumping sulfate out of the cell. It is also theorized that export of sulfate could regulate the ionic balance within the cell. The *suyZ* gene was found to be cotranscribed with *suyB*, and therefore, they are in the same operon. Another analysis of this system showed that the PSE protein is only synthesized when cysteate is present and is metabolized by the cell (Rein *et al*., 2005).

suyZ has striking sequence similarity to *tauZ* (Rein *et al*., 2005). The protein encoded by *tauZ*, TauZ, is also a member of the PSE family (TC# 2.A.98.1.2). The dissimilation pathway of taurine is analogous to that of cysteate (Brüggemann *et al*., 2004) This is not surprising as they both contain sulfonate groups in their structures. Taurine is brought into the cell via a TRAP transporter, in contrast to cysteate which utilizes an ABC transporter. It is acted upon by taurine dehydrogenase (TDH) to yield sulfoacetaldehyde. The enzyme Xsc then converts it to acetyl phosphate and hydrogen sulfite. Sulfite dehydrogenase, however, seems to be absent or in very low concentrations. Because of this, it is not clear if this taurine induced system pumps sulfate, sulfite, or both outside the cell. *suyZ* and *tauZ* share a 73% sequence identity, and both are predicted to encode 10 TMS transporter proteins.

The PSE family has tentatively been suggested to be part of the larger cation:proton antiporter (CPA) superfamily. In this study, we use bioinformatic tools to analyze the topology, phylogenetic and orthologous relationships, as well as the evolutionary pathway taken for the appearance of the PSE family proteins. We also use

the SEED database to provide information about the possible functions of PSE family members. We predict that most PSE family members have 11 TMSs, and not 10 as previously thought. The evidence obtained suggests a 6 + 1 + 6 (-2) or a 7 X 2 (-3) TMS evolutionary pathway, resulting in the final 11 TMS PSE proteins.

Methods

PSI-BLAST searches (Altschul *et al*., 1997) were run using the National Center for Biotechnology Information (NCBI) database to generate a list of homologues. The cysteate-inducible putative sulfate efflux pump, SuyZ, (gi# 51316660) was used to represent the putative sulfate exporter family (TCDB# 2.A.98). A second iteration was run, and a cutoff of e^{-4} was used as the threshold. The BLAST results were converted into TinySeqXML format, which is required for the MakeTable5 script.

The MakeTable5 script (Yen *et al*., 2009) eliminates redundancy in the sequences based on a cutoff specified by the user. A cutoff value of 90% was used to eliminate protein sequences with 90% identity or greater. The script also outputs 3 files: an FAA file containing the remaining protein sequences in FAA format, a 16S file containing the ribosomal RNA sequences, and a tab file containing a table listing the protein abbreviations, descriptions, taxonomic origins, gi numbers, sizes, and organismal domains.

ClustalX (Thompson *et* al., 1997) was used to make a multiple alignment from the set of homologues, analyze for common transmembrane segments (TMS), and identify fragments. Fragments were removed if they aligned poorly with the rest of the homologues or were missing significant TMS regions. An alignment file (.aln) resulted and was used to generate an average hydropathy map via the AveHAS program. ClustalX was also used to generate a neighbor-joining phylogentic tree (.ph) to be viewed with the FigTree program.

The WHAT program (Zhai and Saier, 2001a) creates a hydropathy plot, displaying predicted TMSs in graphical form. HMMTOP (Tusnády and Simon, 1998; Tusnády and Simon, 2001) predicts TMSs in text format and provides the actual residue numbers, their locations, and their orientation within the membrane. The AveHAS program (Zhai and Saier, 2001b) predicts average hydrophobicity, amphipathicity, and similarity based on the multiple alignment generated by ClustalX.

To look for homology and internal repeats, SSSearch (Yen *et al*., 2009) was first used to identify possible regions of similarity. The output of SSSearch revealed a list of protein pairs, revealed their regions of similarity, and provided comparison scores expressed in standard deviations (S.D.). High values from SSSearch were verified with the Global Alignment Program (GAP) ((Devereux et al., 1984) and Global Sequence Alignment Tool (GSAT) programs (Reddy, 2010). Although GAP has been the standard tool used in the Saier lab, GSAT, a variation of GAP, was recently created. It shuffles proteins a variable number of times based on length of the segments being compared, rather than a set number of times as for GAP. It also takes into account the average quality, which is composed of the average score from the shuffles and it's standard deviation. Both, however, use the Needleman Wunsch algorithm to optimally align two protein. Comparison scores of 10 S.D. or higher from GAP/GSAT denote that there is less than a 10^{-24} probability that these sequences are aligning due to chance and is considered sufficient to prove homology between two proteins (Saier, 1994, Saier *et al*., 2009, Yen *et al*., 2009). By our criteria, tested regions also must be at least 60 amino acids in length to establish homology (Saier, 1994).

Additional programs such as GGSearch and HMMER 2.0 were used to confirm the results. GGSearch (http://fasta.bioch.virginia.edu/gasta_www2/fasta_list2.shtml) takes two sets of proteins, aligns them, and provides a score based on similarity between the sets. A score of .01 or smaller provides evidence of homology. HMMER 2.0 (Eddy, 1998; Eddy, 2008) uses the Hidden Markov method to find homology. It takes an aligned set of proteins and generates a profile. The profile is then compared against a target set of sequences in FASTA format. An output file is created, sorting the best scoring pairs of proteins between the profile and the target sequences. Values less than .1 were considered significant. The commands for HMMER are as follows:

hmmbuild <hmm file> <alignment file>

hmmcalibrate <hmm file>

hmmsearch <hmm file> <sequence file>

The SEED database (Overbeek, 2005, Disz, 2010) was used to examine possible functions of the PSE proteins. TauZ (TC# 2.A.98.1.2) and YeiH (TC# 2.A.98.1.3) were put into the BLAST tool incorporated within SEED. A list of similar genomes which contain the gene were listed and examined for possible functionality.

Chapter 1: Identifying Homologs

The cysteate-inducible putative sulfate efflux pump, SuyZ (accession # Q6RH51) protein sequence was run through the BLAST program to generate a set of homologues that serve to represent the PSE family. Position specific iterative (PSI) BLAST was used to increase the sensitivity of the search and find the maximal number of homologues. A second iteration was run with a limit of finding 1000 hits. Sequences that were under the threshold of e^{-5} were placed into the MakeTable5 program to remove redundancies. A 90% cutoff was used to retain sequences that only exhibited less than 90% identity to each other.

The remaining protein sequences were then fed into ClustalX 2.0.12 to generate a multiple alignment. Sequences that were missing an expected TMS or poorly aligned with the rest of the proteins were removed, and the remaining sequences were realigned. This left 340 proteins that consensually aligned based on visual inspection. ClustalX and FigTree were used to draw a neighbor-joining tree. The phylogenetic tree, broken into clusters based on branching patterns, can be seen in Figure 2. The names, abbreviations, descriptions, and gi numbers are presented in Table 1.

Chapter 2: Phylogenetic and Orthologous Analysis

The average size of the PSE homologues is 343 ± 22 amino acids. The averages of each cluster falls within this range with the exception of cluster 10 (367 amino acids). To determine orthologous relationships, the phylogenetic tree was compared to the 16S rRNA tree.

Cluster 1

Phylogenetic trees for the cluster 1 proteins and the corresponding 16S RNAs indicated the occurrence of horizontal gene transfer (HGT). For example, according to the 16S RNA tree, flavobacteria should be more closely related to the bacteriodetes than the fusobacteria. However, in cluster 1, the opposite is true. Further, in the middle of the bacteriodetes cluster is a β-proteobacterium, *Oxalobacter formigenes*, as well as two chlorbi species, these facts are inconsistent with orthology. Similarly, the γ-proteobacteria proteins fall into three major clusters, separated by a variety of bacterial types. Most of the δ-proteobacteria occur in one cluster, but 3 δ-proteobacterial proteins are separated from the others by a Deferribacteres and two γ-proteobacterial proteins. These and other observations clearly suggest extensive HGT during the evolution of cluster 1.

Cluster 2

Cluster 2 proteins exhibit several examples of quite obvious HGT. The vast majority of these proteins are derived from firmicutes, and even among these organisms, the phylogenies of the proteins do not always follow those of the 16S RNAs. However,

most striking is the fact that within this cluster are scattered proteins from other phyla. Examples include Egle1 from an actinobacterium and Meru1 from a euryarchaeon, Funu1 and Sete2 from fusobacteria, and Cume1 from a β-proteobacterium. None of these proteins cluster together, and it seems likely that they were obtained by HGT from various firmicutes to the organisms in the alternative phyla presented in cluster 2 (see Table 1).

Clusters 3 -7, 9-10, 14, 20

Analysis of the phylogenies of the 5 proteins included in cluster 3, and comparison with the 16S rRNA tree, revealed that the two trees do not coincide, suggesting a lack of orthology. Clusters 4, 5, 7, 9, 10, 14, and 20 consist of just one or two proteins, and therefore orthology assignments cannot be made. However, the phylogenetic relationships of the four proteins in cluster 6 are consistent with orthology.

Cluster 8

Cluster 8 is exceptionally diverse, including proteins from euryarcheota (8 proteins), acidobacteria (3), actinobacteria (2), δ-proteobacteria (2), and firmicutes (2). Each of these groups of proteins cluster separately from those of other organismal types. However, within the archaeal subcluster, the phylogenetic positions of the proteins differ from those of the 16S RNAs. In two of these subclusters (actinobacteria and firmicutes), paralogues are present. Because of the small sizes of these subclusters, orthology cannot be assigned.

Cluster 11

Cluster 11 consists primarily of ε-proteobacterial proteins which fall into two primary clusters. Within the first of these two clusters, the proteins fall into relationships as expected for orthology with the *Beggiatoa* and *Francisella* proteins being most distant from the other proteins within these subclusters as expected since these two proteins are from γ-proteobacteria. A single *Verrucomicrobia* protein is embedded within the bacteriodetes subcluster, suggestive of HGT. The second larger ε-proteobacterial subcluster contains only proteins from this class of organisms. The three last proteins in cluster 11 are from γ-proteobacteria (2 proteins) and a single □□proteobacterium. It is worth nothing that 5 of the proteins included within this subcluster are *Helicobacter* species, and sandwiched between Heca1 and Hehe1 is a *Campylobacter jejuni* protein (Caje1). Although *Helicobacter* and *Campylobacter* are closely related genuses, it is possible that this represents another example of HGT. Finally, since the □□proteobacterial protein (Mafe1) are sandwiched between two γ-proteobacterial proteins, Pspu1 and Idlo1, this is likely to be still another example of HGT.

Cluster 12

This cluster consists exclusively of β- and γ-proteobacterial proteins. Surprisingly, however, the β-proteobacterial proteins do not represent an outlying subcluster, but instead, are sandwiched in between γ-proteobacterial proteins. By contrast, in the 16S rRNA tree, the β-proteobacteria clearly separate from the γ-proteobacteria. Thus, the

simplest interpretation is that HGT (HGT) has occurred between β- and γ-proteobacteria repeatedly.

Cluster 13

Only 3 of the 5 proteins in cluster 13 are shown in both figures. However, comparison of these two trees renders it highly unlikely that any of these proteins are orthologous.

Cluster 15

Cluster 15 contains a large subcluster of α-proteobacterial and actinobacterial proteins. The α-proteobacterial subcluster proteins all group together in both the 16S rRNA tree and the protein tree. A cyanobacterium, *Synechococcus*, and acidobacterium, *Solibacter usitatus*, follow, clustering together. This is then followed by the last α-proteobacterial protein from *Roseovarius*, which clusters by itself. A large subcluster of actinobacterial proteins are next, followed by twofrom nitrospirae. The diversity of the phyla in this cluster suggest that HGT has occurred, especially since the lone cyanobacterial and acidobacterial proteins are sandwiched between the α-proteobacteria and the actinobacteria. It should be noted that these two phyla are separated on the 16S rRNA tree by the γ- and β- proteobacteria, highly suggestive of HGT in the protein encoding genes.

Cluster 16

Cluster 16 shows an intermixing of proteins from 3 classes of proteobacteria (α, β, and γ) as well as one Chlorobi protein. The α-proteobacterial proteins dominate, but there are 6 β-proteobacterial proteins in 2 clusters, and 6 putative γ-proteobacterial proteins in two sub-clusters. However, the second sub-cluster contains a single unidentified γ-proteobacterial protein. It is possible that this protein is misassigned. Regardless, we find that the α-, β-, and γ-proteobacterial proteins cluster in ways that are inconsistent with orthology. It seems likely that HGT has occurred repeatedly during the evolution of this cluster.

Cluster 17

Cluster 17 consists of 4 actinobacterial proteins which branch separately from but more closely to a group of 7 firmicutes as compared with all other proteins within this cluster. The remaining firmicutes comprise 2 additional subclusters, the first deriving primarily from *Lactobacillus* species, and the second deriving from species of *Bacillus*, *Listeria*, *Geobacillus*, and *Staphylococcus* among other genera. While the *Staphylococcus* proteins cluster tightly together as do the *Geobacillus* proteins, the *Bacillus* and *Listeria* proteins are scattered throughout this subcluster. Three additional proteins that cluster together are each derived from three different bacterial phyla: δ-proteobacteria, Deinococcus-Thermus, and Chloroflexi. It seems highly likely that several proteins in cluster 17 underwent HGT.

Cluster 18

Examination of the protein and 16S RNA trees for cluster 18 reveals that most of these proteins show relationships consistent with orthology. However, 2 interesting exceptions can be found. *Carnobacterium* should be close to the enterococci, but instead, it clusters together with a fusobacterial homologue. Since fusobacteria are gram negative, and in a phylum distant from all firmicutes, we can conclude that the cluster including the fusobacterial and carnobacterial proteins do not conform to orthology. Both proteins may have been obtained from other firmicutes via HGT.

Cluster 19:

All but one of the proteins in cluster 19 are derived from α-proteobacteria. Three of these proteins, Gldi1, Acpa1, and Glox1 cluster together on both the protein and rRNA trees, consistent with orthology. The remaining 13 proteins do not cluster similarly on the two trees, again suggesting the occurence of HGT.

In summary, we have identified a few small clusters which appear to consist of orthologues, but analyses of the majority of the clusters within this tree provide strong evidence for the occurance of HGT, both within a single phylum or order, and between phyla. It is possible that members of the putative sulfate exporter family are often carried on mobile genetic elements that can be easily transferred between organisms.

Chapter 3: Topological Analysis

PSE family members were aligned (ClustalX), and the resulting alignment file was used to produce an average hydropathy plot (AveHAS). Hydrophobic peaks have a positive value and lie above the x axis. The orange peaks at the bottom show the hydropathy profile, generated with a separate program. 11 TMSs were predicted and are labeled numerically as seen in Figure 1.

Notable hydrophilic regions separate the TMSs at three distinct locations. These separations are located between TMS 2-3, 5-6, and 8-9. The last hydrophilic loop between TMSs 8 and 9 is of special importance because it is significantly larger than the other loops, and is characteristic of most PSE homologues. There are a few proteins that contain long hydrophilic stretches and thus elongate the AveHAS plot. These include Kyse1 (*Kytococcus sedentarius* , gi# 256824845) at the C terminus and Chch1 (*Chlorobium chlorochromatii* , gi# 78189222) at the N terminus. Because the rest of their sequences aligned well with the other homologues, they were retained.

The clustering pattern and sizes of the TMSs sometimes suggest how they evolved. TMS 1 and 2 are close together, separated from the rest of the TMSs by a small hydrophilic loop. Other clusters of TMSs include 3-5, 6-8, and 9-11. If the TMSs are split into a 4, 4, 3 TMS arrangement, we can see that TMSs 3, 7, and 10 are smallest in their groups. This pattern could be interpreted to suggest a 4 TMS repeat, a possible that will be tested in the following section.

TMHMM and HMMTOP 2.0 were used to analyze the topology of the protein within the membrane. The small hydrophilic extension that exists before TMS 1 is predicted to face the cytoplasm, and the inter TMS loops follow an alternating inside/outside pattern as expected. The hydrophilic extension after TMS 11 lies external to the cell membrane. In addition, charge analysis was conducted. The number of lysines and arginines was estimated between each TMS, and an average was calculated. As expected, larger numbers of lysines and arginines proved to be in the loops predicted to be located on the inside. This coincides with the positive-inside rule (von Heijne & Gavel, 1988), and the predictions of TMHMM. Thus the different methods used confirmed the results of each other.

Chapter 4: Internal Repeats

The majority of PSE homologues probably contain 11 TMSs, but a few are predicted to have 10 TMSs. To check for internal repeats, the sequences of the homologues were aligned using ClustalX, and clusters were examined using the AveHAS program. 6 TMS regions might have duplicated to form 12 TMSs as the case of many families of 12 TMS carriers. However, SSSearch and GAP/GSAT results suggested a 6 + 1 + 6 (-2) or a 7 + 7 (-3) TMS evolutionary pathway.

SSSearch can be used to detect regions of similarity between two groups of protein sequences. It can also compare specific areas of an alignment if the numerical values of the regions of interest are provided. The region containing the first 6 TMSs (residue positions 250-500 on AveHAS plot) were compared against the last 5 (501-800 on AveHAS). This generated multiple high scoring pairs with TMS 1-4 aligning with TMS 8-11. SSSearch was then run to compare TMS 5-7 (426-530 on AveHAS) against both TMS 1-4 (250-425 on AveHAS) and TMS 9-11(650-800 on AveHAS). No full length alignments were produced when either was compared against TMSs 5-7.

GAP was used to confirm the results of SSSearch. This tool is a standard program used to confirm homology. GAP aligns the sequences and provides a score based on identities, similarities, and gaps. Ideally, sequences have to be at least 60 amino acids in length. Vish1 (TMS 8-11) aligned with Lech1 (TMS 1-4) with a score of 15.2 S.D. (Figure 3). This score far exceeds 10 S.D. and therefore is more than sufficient to prove homology (Saier *et al*., 2009; Yen *et al*., 2009). The new program, GSAT, was also used

to confirm the comparison scores from GAP. When TMSs 8-11 of Vish1 and TMSs 1-4 of Lech1 were compared, a score of 13.8 S.D. was obtained.

Members of the CPA2 family were analyzed to see if a relation with the PSE family could be demonstrated. CPA2 family members are predicted to contain 10-14 TMSs. Homologues were generated using PSI-BLAST, redundancies and close sequences were eliminated using MakeTable5, and an alignment was generated. SSSearch was run between PSE and CPA2 homologues. Stma1 (PSE: TMS 1-6) aligned with Labr2 (CPA2: TMS 8-13) with a comparison score of 14.7 S.D. and 12.7 in GSAT (Figure 4). Errh1 (PSE: TMS 1-8) aligned with Bame1 (CPA2: TMS 1-8) with a comparison score of 12.8 S.D. in GAP and 10.4 in GSAT (Figure 5). Results suggest either a 6 or 7 TMS duplication.

GGsearch and HMMER 2.0 were used to confirm the findings of SSSearch and GAP. GGsearch uses a global-global method of aligning a pair of proteins and gives an e-value to indicate homology. TMS 1-4 vs TMS 8-11 yielded multiple high scoring pairs, with the best having a score of 1.3×10^{-8} (Base1 vs Vish1). However, TMS 5-7 did not reveal appreciable similarity with either TMSs 1-4 or TMSs 8-11. Although there is no set threshold for signifying homology in GGsearch, the high scores from the alignment of TMS 1-4 with TMS 8-11 (relative to TMS 5-7) support the SSSearch and GAP results.

HMMER 2.0 uses the Hidden Markov Model to detect homology. It builds a profile using an alignment file and compares it to a second set of protein sequences (database). E-values smaller than .01 are claimed to be sufficient to suggest homology.

Each set of TMSs (1-4, 5-7, 8-11) were individually aligned and used to build a profile. These profiles were subsequently used to compare the FASTA sequences of the other TMS clusters. For example, the profile of TMSs 1-4 was compared with FASTA sequences of TMSs 5-7 and TMSs 8-11. Significant results were obtained when the profile of TMSs 1-4 and TMSs 8-11 were compared against TMSs 8-11 and TMSs 1-4 with scores of 5.3×10^{-4} and 5.6×10^{-5} respectively. These results were expected as they confirmed the findings of SSSearch and GAP/GSAT.

Chapter 5: Seed Analysis

Use of the SEED database often allows one to predict the function of a gene, based on genome context and the assigned functions of the genes surrounding it. It allows one to view multiple genomes from different organisms that have similar protein sequences. This is especially useful when the gene of interest is part of an operon, increasing chances that its function is related to those surrounding it. Query sequences are colored red and labeled as "1." Successively associated genes are ranked by number according to how often they appear with the query. PSE family members TauZ (TC# 2.A.98.1.2) and YeiH (2.A.98.1.3) were examined in SEED.

TauZ is situated in a predicted 9 gene cluster for the dissimilation of taurine (Bruggerman *et al*., 2004*).* In Figure 8, we see that TauZ is represented as sequence 1 in SEED. Sequences 2 and 11 are part of a TRAP system which is used for taurine uptake. Sequence 9 is the Taurine dehydrogenase (TDH) and 7 is an oxidoreductase. Flanking TauZ are sulfoacetaldehyde acetyltransferase (4) and a phosphate acetyltransferase (6). All the genes that comprise the proposed pathway in Figure 8 could be identified in some of the gene clusters observed with SEED. Topological analyses confirmed that TauZ codes for an 11 TMS protein, a topology similar to those of many secondary carriers, and therefore, it is reasonable to suggest, in agreement with published proposals, that TauZ genes encode sulfate exporters.

SEED analysis of YeiH yielded intriguing results (Figure 9). A transcriptional regulator, YeiE (sequence 2 in the SEED analysis) belongs to the LysR family and is

oriented in the opposite direction and adjacent to YeiH. An Endonuclease IV (sequence 3 in the SEED analysis) is immediately downstream of YeiH. The frequent co-occurrence of these three genes strongly suggests a functional relationship between these three gene products; one almost always accompanies the other two. The co-occurrence of these three adjacent genes leads us to tentatively suggest that in these operons, the YeiH homologues might serve the function of nucleotides or oligonucleotide export. Two homologous putative pyrimidine nucleoside transport proteins (sequence 6), a member of the concentrated nucleotide uptake transporter (CNT) family (TC# 2.A.41), are found downstream from YeiH in positions 3 and 6. These may be associated with pseudouridine catabolism. In fact, an inosine-uridine preferring nucleoside hydrolase could be involved in nucleobase reutilization. A predicted N-ribosyl-nicotinamide CRP-like regulator are sandwiched in between the two CNT family paralogues. PTS components of a fructose-specific catabolic operon (sequences 8, 9, and 10) are sometimes found downstream of YeiH. Further inspection of the upstream genes reveals that in opposite orientation and adjacent to YeiH homologues are putative Lysine permeases (sequence 4), and an outer membrane Colicin I receptor precursor (sequence 5) that may be specific for a catechol iron siderophore.

Discussion

The PSE family of proteins has both archaeal and bacterial members but not eukaryotic members. They have been characterized and some preliminary functional analyses have been performed. It was initially thought that there was a 4 + 4 + 3 TMS evolutionary pathway. However, extensive evidence using our bioinformatics tools suggests an evolutionary pathway of either a 6 + 1 + 6 or a 7 + 7 TMS repeat, which, following the loss of 2 or 3 C-terminal TMSs, resulted in the present day 11 TMS protein.

The clustering of the TMSs suggested a 4 + 4+ 3 TMS evolutionary pathway. Topological analyses showed that TMSs 9-11 were separated from the other TMSs by a large hydrophilic loop. Such large loops frequently are found between repeat units. This could have been generated in a process that accompanied the deletion of a TMS before TMS 9. SSSearch showed multiple high scoring pairs containing the alignment of TMSs 2-4 with TMSs 9-11, and one pair showing the alignment of TMSs 1-4 with TMSs 5-8 with a comparison score over 10 S.D. The lone alignment does not inspire enough confidence to continue the theory of a 4 + 4 + 3 evolutionary pathway.

GAP alignments strongly indicate the possibility of a 6 or 7 TMS repeat unit. There were numerous high scoring pairs of proteins, showing the alignment of TMSs 1-4 with TMSs 8-11 (Figure 3). It is interesting to note the gap in the alignment between TMSs 1 and 2 (Lech1) corresponds to hydrophilic stretch between TMSs 8 and 9 (Vish1). TMSs 5-7 were tested against both clusters and failed to produce a full alignment or a comparison score over 9. The CPA2 family has been tentatively placed in the same

superfamily as PSE, and seems to have the same internal repeat structure with TMSs 1-4 aligning with TMSs 8-11. It was compared against PSE, which showed an alignment of TMSs 1-6 of PSE with TMSs 8-13 of CPA2 (Figure 4). Comparison with CPA2 did not reveal any significant alignments between any members of PSE and TMS 14 of CPA homologues. However, multiple high scoring alignments show TMSs 1-6 aligning with TMSs 8-13, confirming the possibility that there is at least a 6 TMS repeat (Figure 5). In either case, these two proposed pathways suggest but do not prove an inversion occurred during duplication. Although uncommon, it has been shown to occur (Povolotsky, 2010).

GSAT, GGsearch, and HMMER 2.0 were used to complement the comparison scores reported by GAP. GSAT uses a more refined method of calculating the comparison score, by shuffling the alignment based on its length. In addition, the comparison score reported is less variable than that obtained using GAP, resulting from the inclusion of the average quality in its calculation. GGsearch and HMMER 2.0's results both agreed with GAP that there is strong evidence for homology between TMSs 1-4 and TMSs 8-11. However, no evidence links TMSs 5-7 to either of the established repeat element.

Although some phylogenetic clusters of PSE homologues showed evidence of orthology, nearly all clusters in the phylogenetic tree included proteins likely to have been subject to horizontal gene transfer (HGT). The majority of archaeal proteins were situated in cluster 8. Cluster 8 contains proteins from a diverse range of bacterial phyla, highly indicative of HGT. Looking at all the clusters, HGT has probably occurred in most of them in this regard, it is interesting to note that no PSE homologues were identified in

eukaryotes. HGT of PSE member-encoding genetic elements apparently occurred to a greater extent than we have observed for most other families of transporters. In fact, HGT occurred both within single phyla or orders, as well as between phyla. Possibly genes encoding the PSE related proteins are often carried on mobile genetic elements that can be easily transferred between organisms. Other means of gene transfer could also have been utilized.

SEED analysis proved to be useful in predicting the functions of PSE members. TauZ and its homologue, SuyZ, are both purported to export sulfate, as seems reasonable in view of the operon contents associated with taurine dissimilation. There are 5 different genomes that exhibit this pattern. YeiH and several related operons, however, do not seem to exhibit this relationship. This protein is encoded by a gene that is most commonly associated with YeiE, a LysR family transcriptional regulator oriented in the opposite direction and adjacent to YeiH, as well as an Endonuclease IV, encoded within at least 10 different genomes included in SEED. Possible functions for YeiH could be to export nucleotides or oligonucleotides, released by the action of the endonucleases. Numerous other transport genes surround these three genes, such as those encoding a putative pyrimidine nucleoside uptake transporter, a fructose-like phosphotransferase system (PTS), and lysine permeases. These may or may not be relevant to the functions of the transporters under study. Based on our findings, we tentatively suggest a role for YeiH as an oligonucleotide exporter. We agree that TauZ/SuyZ are likely to be sulfate exporters. Further analyses will be necessary to establish the functions of PSE family members.

Appendix

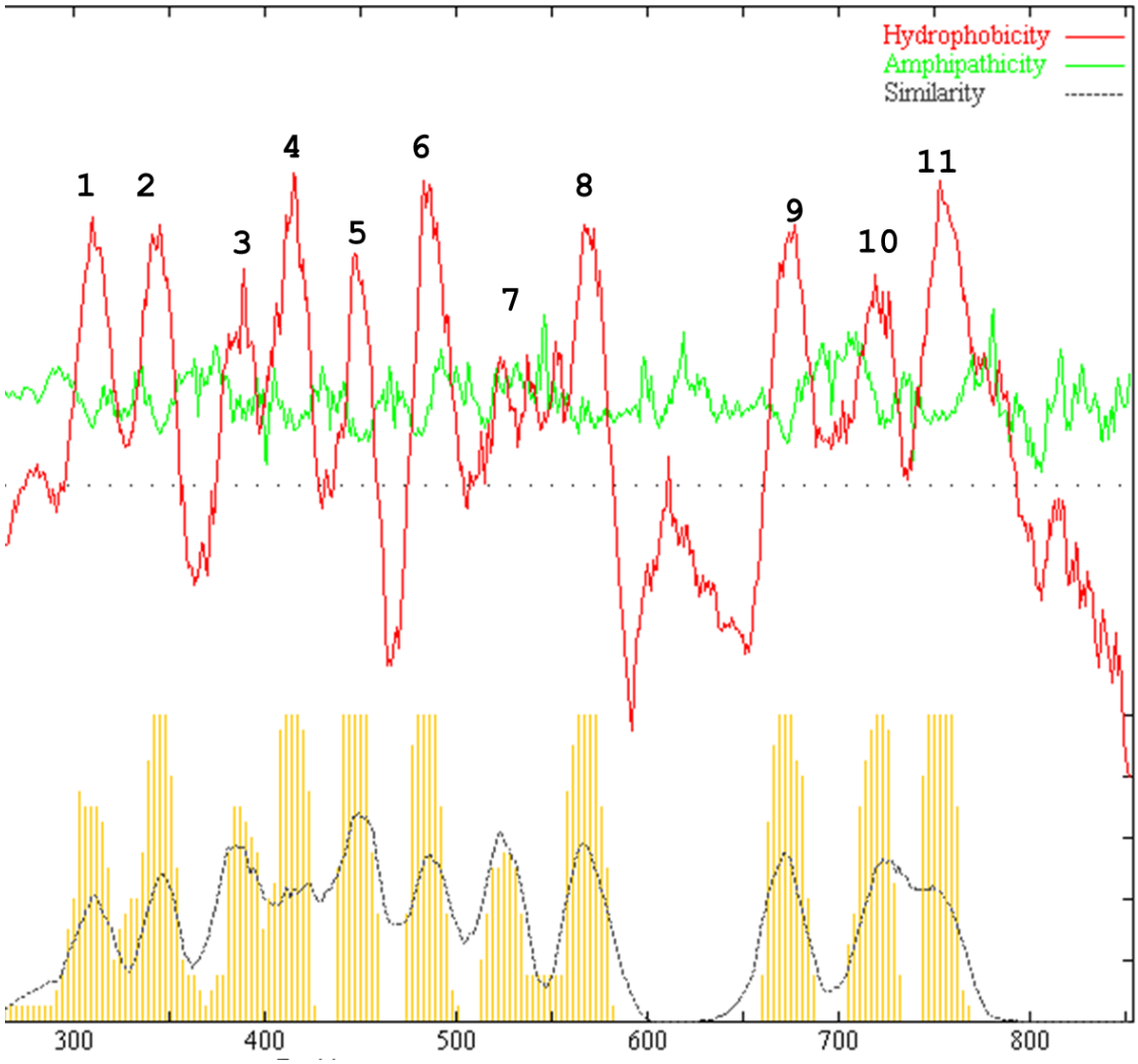

Figure 1. The average hydropathy, amphipathicity, and similarity plot constructed using AveHAS. PSE homologues were first aligned by ClustalX. There are 11 predicted peaks, with a large hydrophilic loop or gap between TMSs 8 and 9.

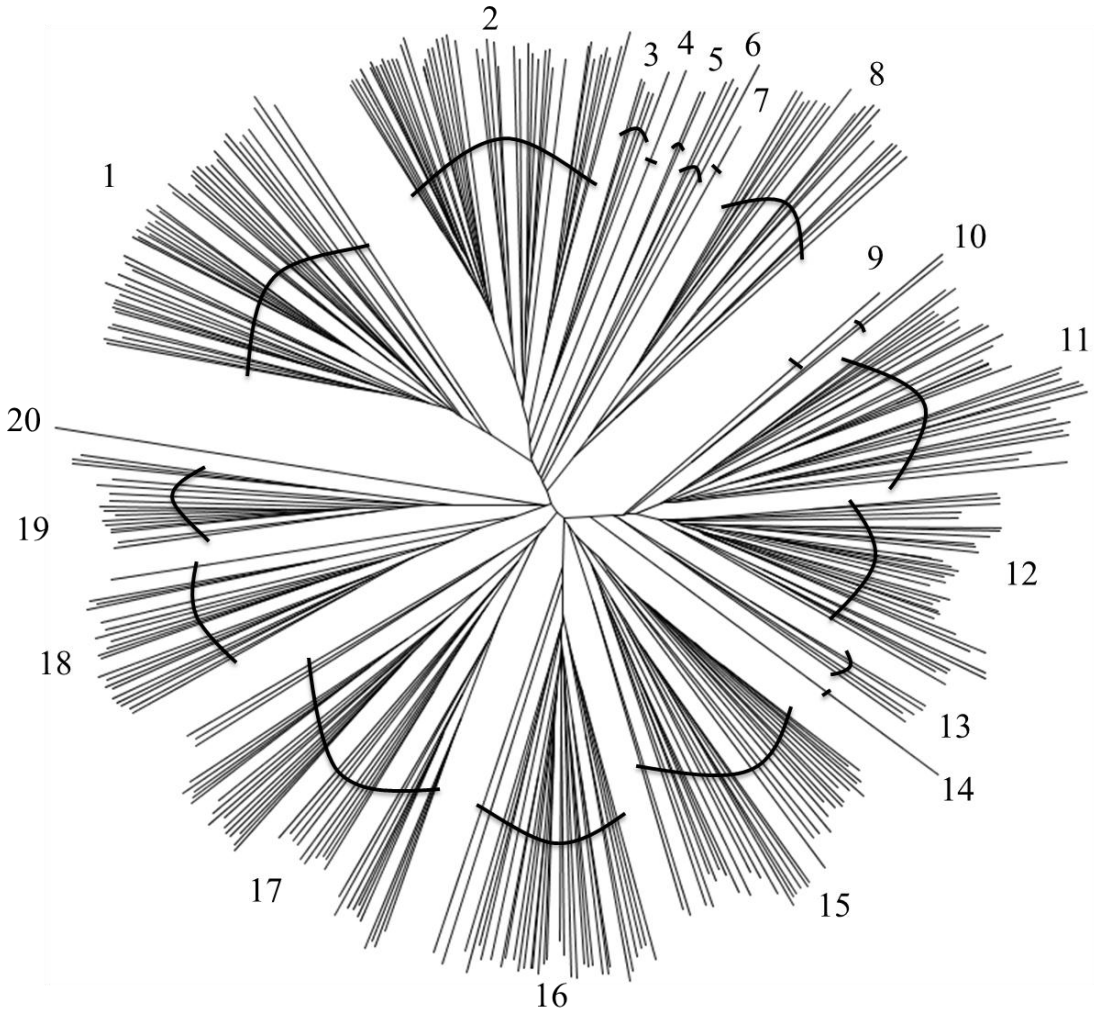

Figure 2. A phylogenetic tree of PSE homologues were created from the alignment in ClustalX. The tree was broken up into 20 clusters based on their branching patterns. Tip labels were removed to improve visibility and a complete breakdown of each cluster can be seen in Table 1.

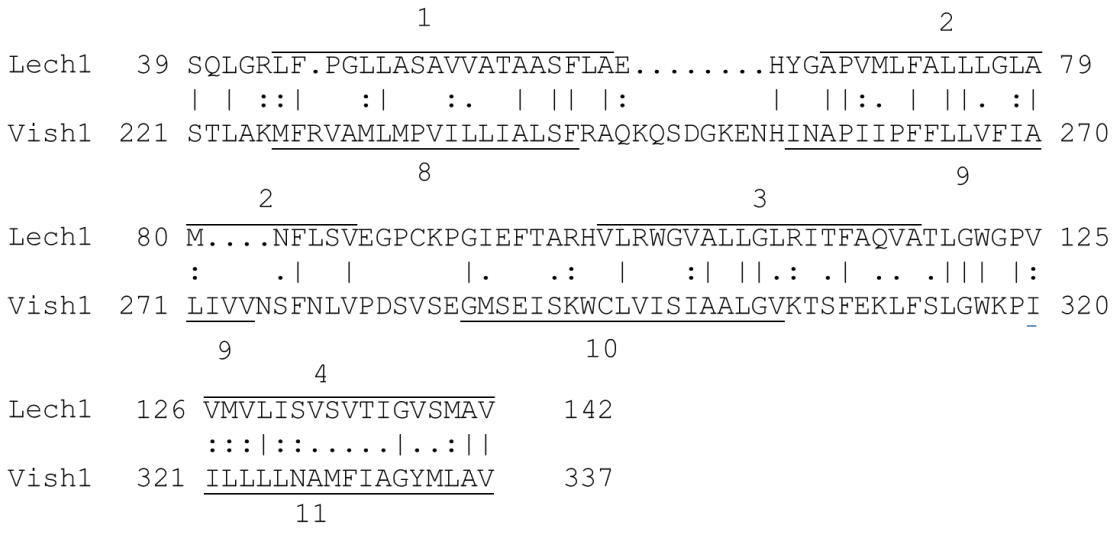

Figure 3. Alignment of PSE homologues. TMSs 1-4 of Lech1 (*Leptothrix cholodnii*, gi# 171058852) with TMSs 8-11 of Vish1 (*Vibrio shilonii*, gi# 149187943). This is one of many pairs that showed this alignment pattern and had a high comparison score. The gap between TMSs 1 and 2 of Lech1 is caused by the hydrophilic region between TMSs 8 and 9 of Vish1. Numbers on the left and right sides correspond to the residue numbers in the actual protein. TMS positions were predicted with HMMTOP. This alignment gave a comparison score of 15.2 S.D. with GAP and 13.8 S.D. with GSAT.

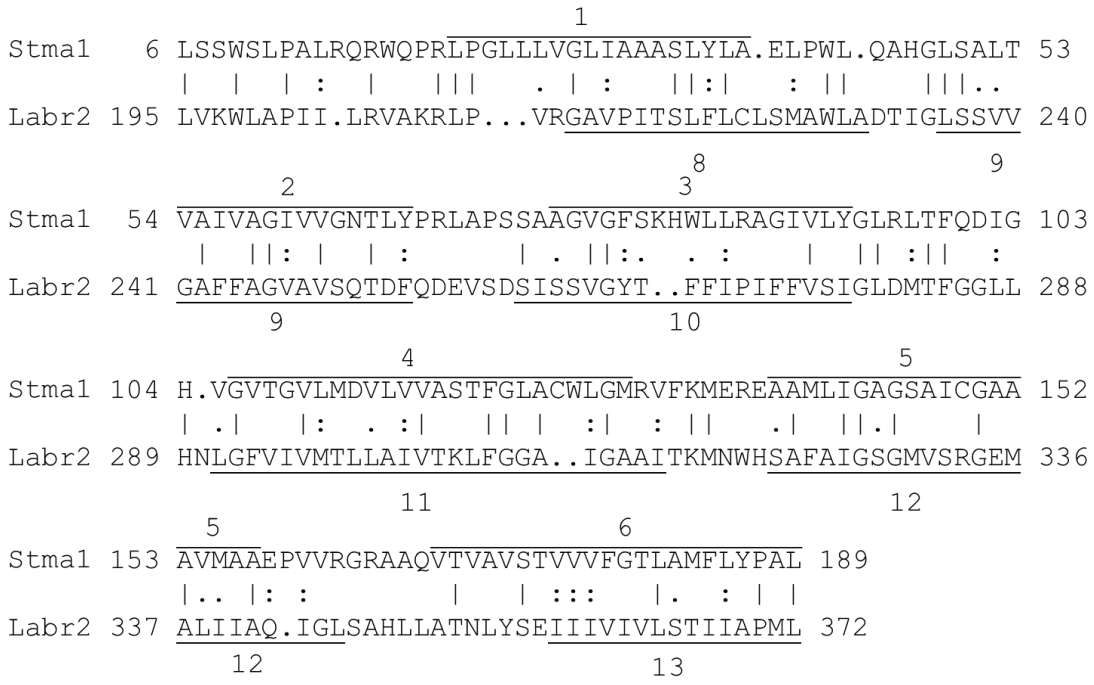

Figure 4. The alignment of TMSs 1-6 of Stma1 (PSE homologue, *Stenotrophomonas maltophilia*, gi# 194365408) with TMSs 8-13 of Labr2 (CPA2 homologue, *Lactobacillus brevis*, gi# 227509830) This alignment received a comparison score of 14.7 S.D. from GAP and 12.7 from GSAT.

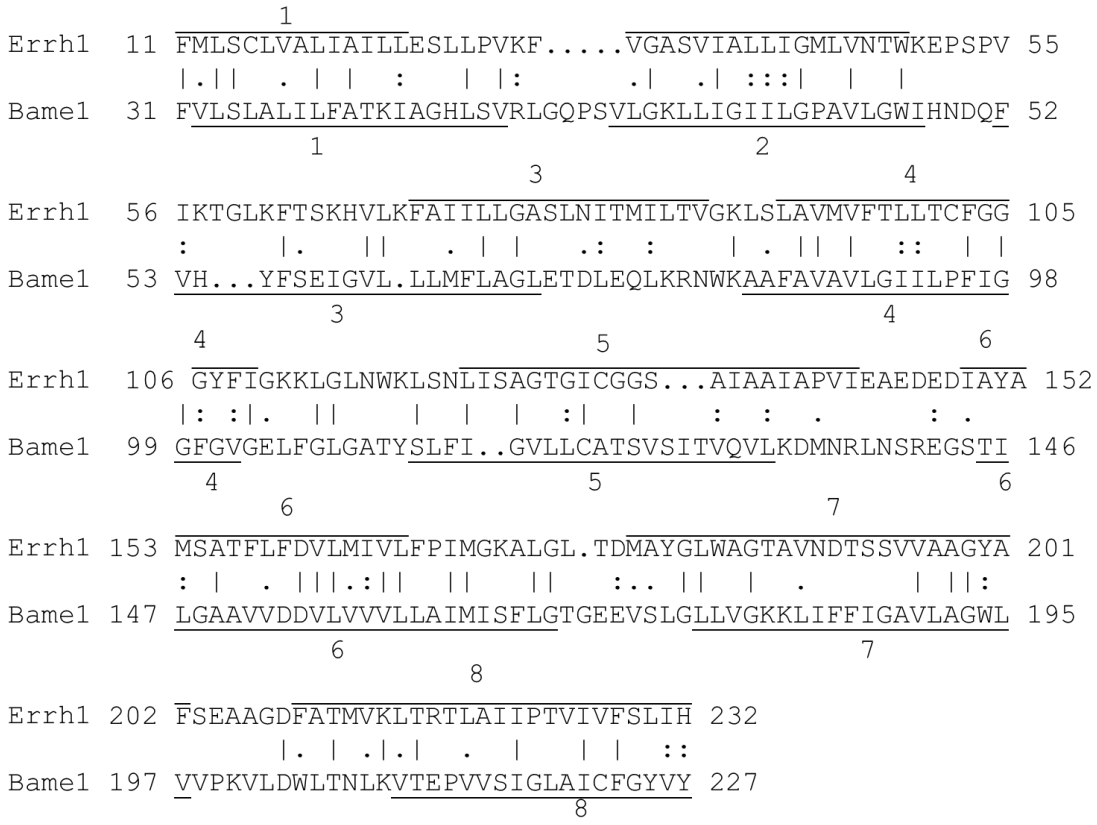

Figure 5. The alignment of TMSs 1-8 of Errh1 (PSE homologue, *Erysipelothrix rhusiopathiae*, gi# 259504620) with TMSs 1-8 of Bame1 (CPA2 homologue, *Bacillus megaterium*, gi# 294509009). This alignment gave a comparison score of 12.8 in GAP and 10.4 in GSAT.

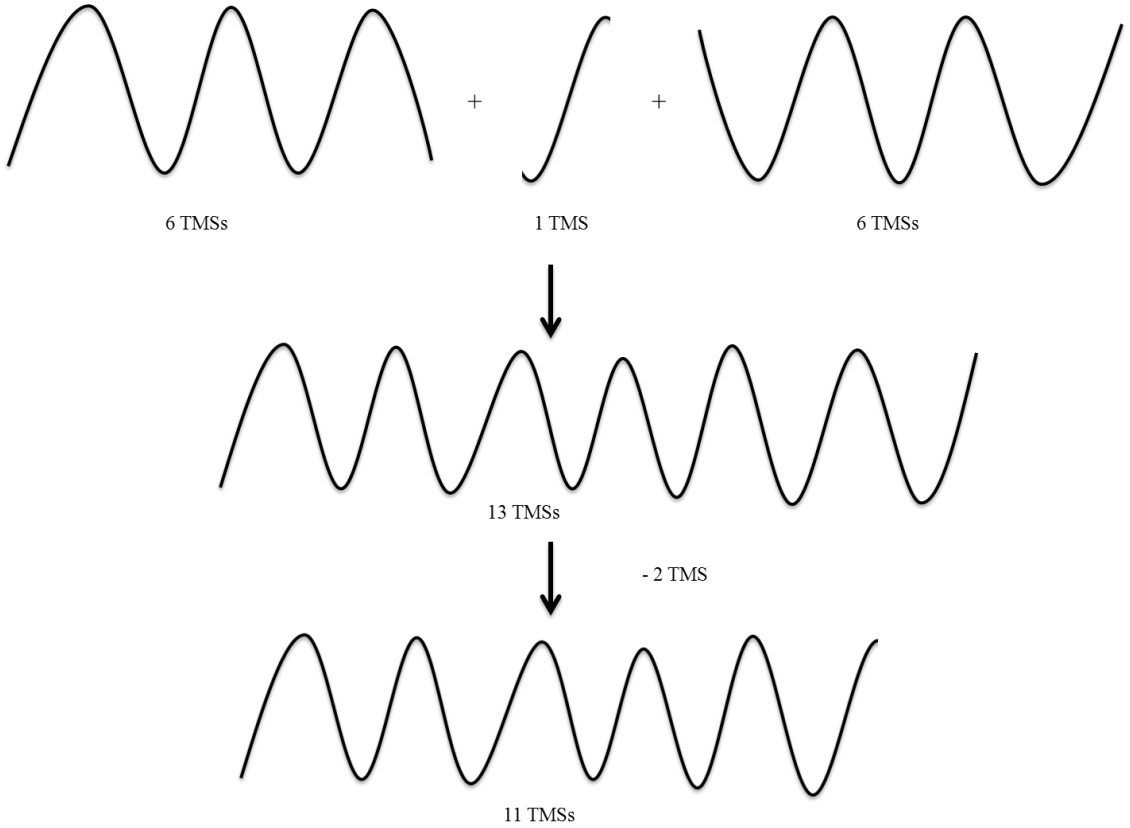

Figure 6. First proposed evolutionary pathway of PSE homologues. 6 + 6 + 6 TMSs combined to form 13 TMSs. A loss of 2 TMSs at the C terminus could have given rise to the present day 11 TMS PSE proteins.

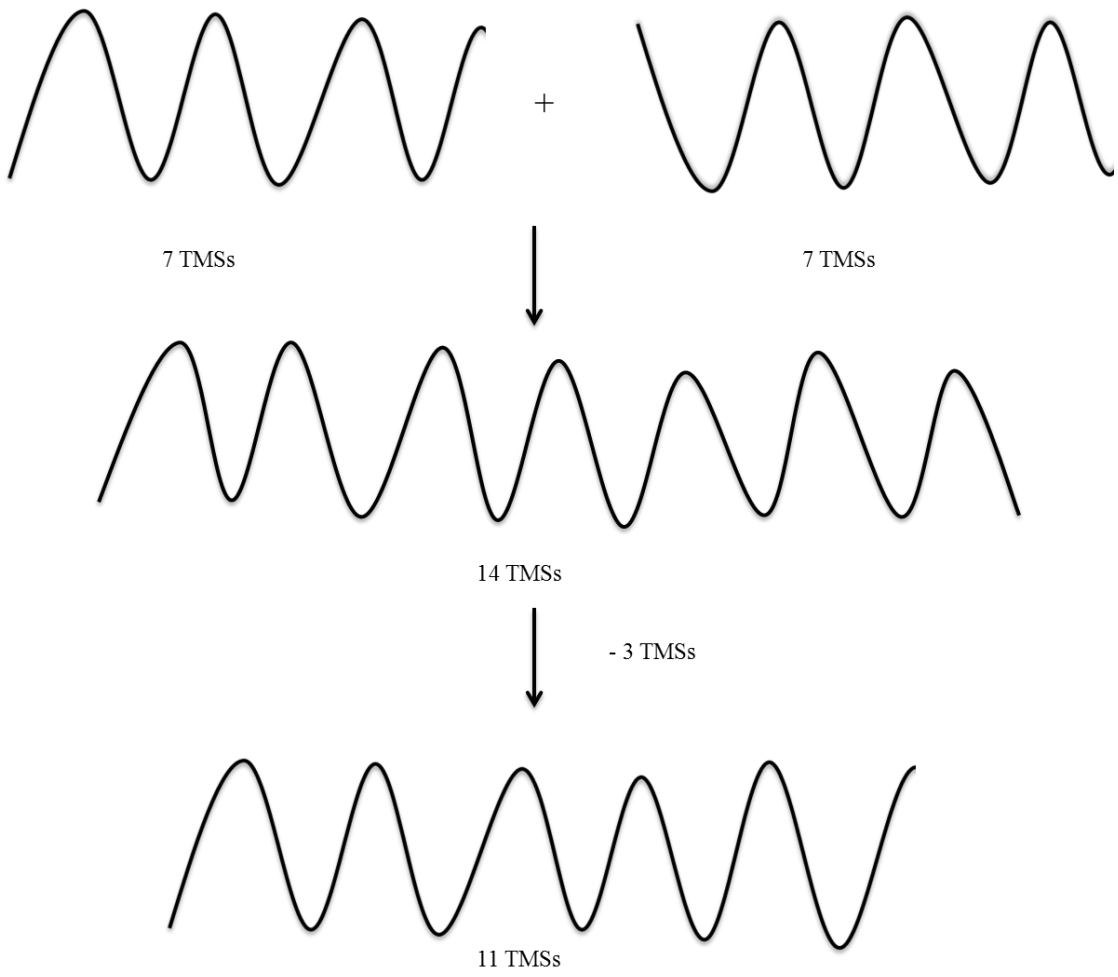

Figure 7. Second proposed pathway for PSE homologues. A 7 TMS duplication resulted in 14 TMSs. A loss of 3 TMSs at the C terminus could have resulted in the 11 TMS proteins.

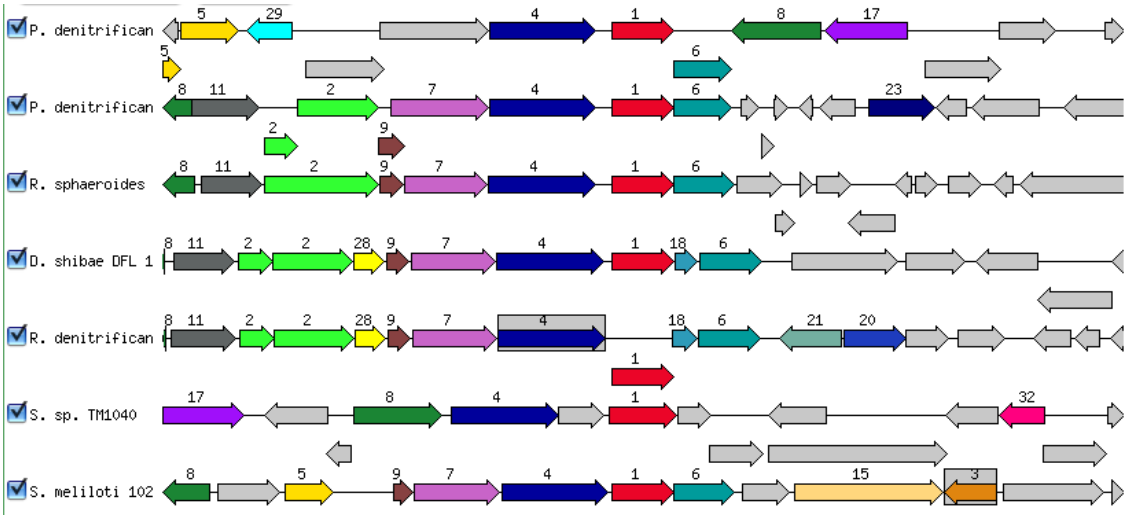

Figure 8. SEED results retrieved from the search of TauZ. Sequence 1 in red shows the query, TauZ. Notable sequences include TRAP components (2 and 11), oxidoreductase (7), sulfoacetaldehyde acetyltransferase (4), and phosphate acetyltransferase (6). These components are involved in taurine uptake.

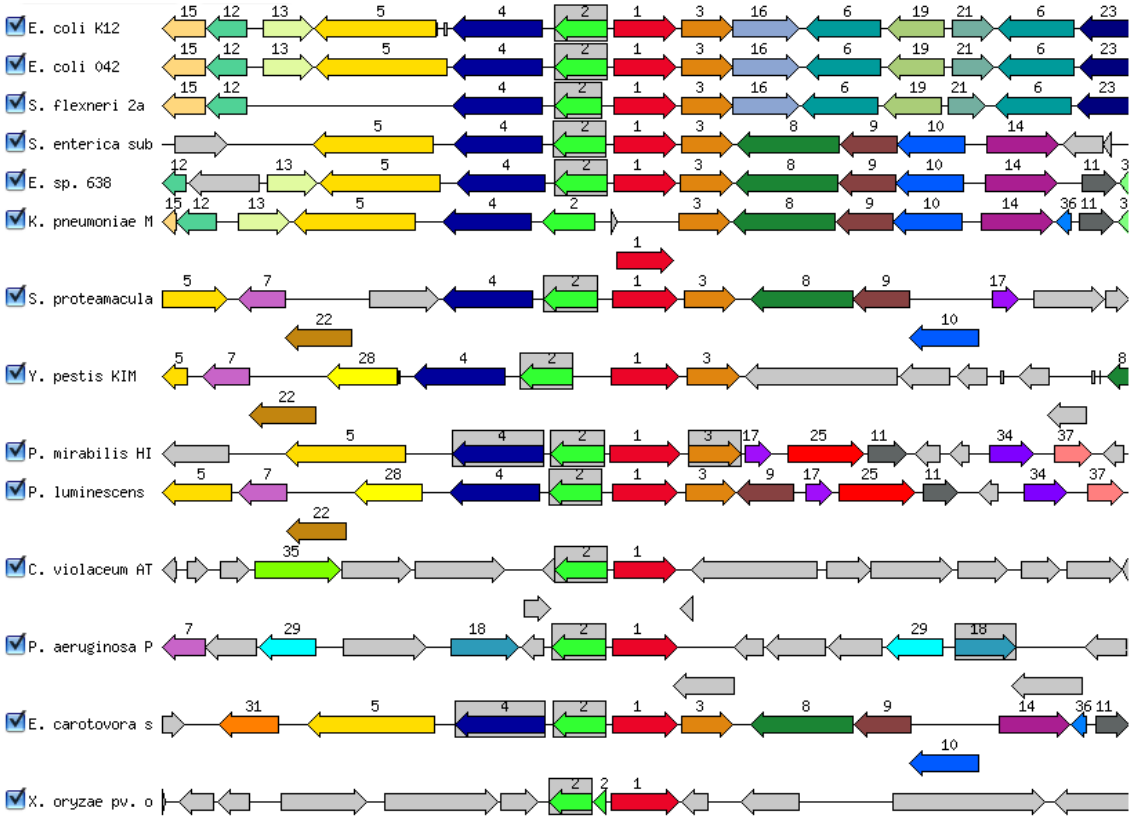

Figure 9. SEED results retrieved from the search of YeiH. Sequenced 1 in red represents the query, YeiH. Notable sequences are as follows: YeiE of the LysR family(2), Endonuclease IV (3), lysine permeases (4), and Colicin I receptor precursor (5).

Table 1. PSE family homologues generated through a PSI BLAST search. Abbreviations were created through the MakeTable5 script. Proteins were separated into clusters based on branching patterns in the phylogenetic tree. The average and standard deviation of the proteins in each cluster was calculated.

Cluster 1

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Bath1	Bacteroides thetaiotaomicron VPI-5482	29350017	Bacteroidetes	Bacteria	330
Pogi1	Porphyromonas gingivalis W83	34541597	Bacteroidetes	Bacteria	330
Alpu1	Alistipes putredinis DSM 17216	167752263	Bacteroidetes	Bacteria	317
Baor1	Bacteroidetes oral taxon 274 str. F0058	298373695	Bacteroidetes	Bacteria	308
Oxfo1	Oxalobacter formigenes HOxBLS	237745582	Betaproteobacteria	Bacteria	315
Pehe1	Pedobacter heparinus DSM 2366	255530183	Bacteroidetes	Bacteria	311
Pesp2	Pedobacter sp. BAL39	149280122	Bacteroidetes	Bacteria	328
Spli1	Spirosoma linguale DSM 74	284005865	Bacteroidetes	Bacteria	334
Chpi1	Chitinophaga pinensis DSM 2588	256423984	Bacteroidetes	Bacteria	333
Fljo1	Flavobacterium johnsoniae UW101	146300633	Bacteroidetes	Bacteria	333
Chch1	Chlorobium chlorochromatii CaD3	78189222	Chlorobi	Bacteria	573
Chth1	Chloroherpeton thalassium ATCC 35110	193215532	Chlorobi	Bacteria	332
Zupr1	Zunongwangia profunda SM-A87	295134482	Bacteroidetes	Bacteria	325
Acsp4	Acidobacterium sp. MP5ACTX8	299138684	Acidobacteria	Bacteria	343
Acca1	Acidobacterium capsulatum ATCC 51196	225873793	Acidobacteria	Bacteria	313
Kove2	Candidatus Koribacter versatilis Ellin345	94971365	Acidobacteria	Bacteria	319
Psin1	Psychromonas ingrahamii 37	119945393	Gammaproteobacteria	Bacteria	315
Mosp1	Moritella sp. PE36	149907988	Gammaproteobacteria	Bacteria	315
Shsp1	Shewanella sp. MR-4	113972180	Gammaproteobacteria	Bacteria	316
Shpi1	Shewanella piezotolerans WP3	212637638	Gammaproteobacteria	Bacteria	310
Phsp1	Photobacterium sp. SKA34	89075598	Gammaproteobacteria	Bacteria	316
Shde1	Shewanella denitrificans OS217	91791585	Gammaproteobacteria	Bacteria	315
Shlo1	Shewanella loihica PV-4	127514596	Gammaproteobacteria	Bacteria	318
Pssp1	Psychromonas sp. CNPT3	90409054	Gammaproteobacteria	Bacteria	314
Visp1	Vibrio sp. Ex25	262394797	Gammaproteobacteria	Bacteria	305
Vifi1	Vibrio fischeri ES114	59713388	Gammaproteobacteria	Bacteria	325

Table 1. (Continued)

Amco1	Aminobacterium colombiense DSM 12261	294101527	Synergistetes	Bacteria	311
Deac3	Denitrovibrio acetiphilus DSM 12809	291286952	Deferribacteres	Bacteria	316
Depe1	Dethiosulfovibrio peptidovorans DSM 11002	288573915	Synergistetes	Bacteria	313
Pepr1	Pelobacter propionicus DSM 2379	118580595	Deltaproteobacteria	Bacteria	310
Gesu1	Geobacter sulfurreducens PCA	39997911	Deltaproteobacteria	Bacteria	322
Peca1	Pelobacter carbinolicus DSM 2380	77919197	Deltaproteobacteria	Bacteria	310
Gelo1	Geobacter lovleyi SZ	189424018	Deltaproteobacteria	Bacteria	310
Deac4	Desulfuromonas acetoxidans DSM 684	95928330	Deltaproteobacteria	Bacteria	307
Dede1	Deferribacter desulfuricans SSM1	291279816	Deferribacteres	Bacteria	313
Psha1	Pseudoalteromonas haloplanktis TAC125	77360511	Gammaproteobacteria	Bacteria	308
Pstu1	Pseudoalteromonas tunicata D2	88859227	Gammaproteobacteria	Bacteria	311
Ande1	Anaeromyxobacter dehalogenans 2CP-C	86156779	Deltaproteobacteria	Bacteria	328
Bdba1	Bdellovibrio bacteriovorus HD100	42522953	Deltaproteobacteria	Bacteria	325
Stau1	Stigmatella aurantiaca DW4/3-1	115378988	Deltaproteobacteria	Bacteria	315
Opte1	Opitutus terrae PB90-1	182413263	Verrucomicrobia	Bacteria	331
Viva1	Victivallis vadensis ATCC BAA-548	281355009	Lentisphaerae	Bacteria	323
Lego1	Leptotrichia goodfellowii F0264	262037255	Fusobacteria	Bacteria	310
Kako1	Kangiella koreensis DSM 16069	256823212	Gammaproteobacteria	Bacteria	338
Orf1	marine gamma proteobacterium HTCC2148	254482236	Gammaproteobacteria	Bacteria	315
				Average Protein Size	325
				Standard Deviation	39

Cluster 2

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Ruto1	Ruminococcus torques ATCC 27756	153813821	Firmicutes	Bacteria	335
Anca1	Anaerostipes caccae DSM 14662	167745780	Firmicutes	Bacteria	389
Suva1	Subdoligranulum variabile DSM 15176	261368967	Firmicutes	Bacteria	343
Egle1	Eggerthella lenta DSM 2243	257792256	Actinobacteria	Bacteria	380
Clas1	Clostridium asparagiforme DSM 15981	225386363	Firmicutes	Bacteria	336
Euve1	Eubacterium ventriosum ATCC 27560	154484939	Firmicutes	Bacteria	344

Table 1. (Continued)

Clsc1	Clostridium scindens ATCC 35704	167758659	Firmicutes	Bacteria	343
Lade1	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	116514580	Firmicutes	Bacteria	341
Acfe1	Acidaminococcus fermentans DSM 20731	284048012	Firmicutes	Bacteria	339
Pean1	Peptostreptococcus anaerobius 653-L	289423369	Firmicutes	Bacteria	342
Fima1	Finegoldia magna ATCC 29328	169824985	Firmicutes	Bacteria	341
Acsp2	Acidaminococcus sp. D21	227824734	Firmicutes	Bacteria	349
Meru1	Methanobrevibacter ruminantium M1	288561386	Euryarchaeota	Archaea	346
Stsu1	Streptococcus suis 89/1591	223934172	Firmicutes	Bacteria	340
Anla1	Anaerococcus lactolyticus ATCC 51172	227485395	Firmicutes	Bacteria	357
Stpy1	Streptococcus pyogenes MGAS8232	19746028	Firmicutes	Bacteria	339
Stpy2	Streptococcus pyogenes M1 GAS	15675048	Firmicutes	Bacteria	339
Steq1	Streptococcus equi subsp. zooepidemicus	225868366	Firmicutes	Bacteria	339
Stag1	Streptococcus agalactiae 2603V/R	22537283	Firmicutes	Bacteria	335
Stpn1	Streptococcus pneumoniae TIGR4	15899980	Firmicutes	Bacteria	336
Stth1	Streptococcus thermophilus LMG 18311	55820910	Firmicutes	Bacteria	335
Funu1	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	19703868	Fusobacteria	Bacteria	346
Vesp1	Veillonella sp. 3_1_44	294794844	Firmicutes	Bacteria	346
Eusi1	Eubacterium siraeum DSM 15702	167751223	Firmicutes	Bacteria	361
Mege1	Megasphaera genomosp. type_1 str. 28L	290968583	Firmicutes	Bacteria	328
Sesp1	Selenomonas sputigena ATCC 35185	260888101	Firmicutes	Bacteria	327
Eusa1	Eubacterium saphenum ATCC 49989	255994551	Firmicutes	Bacteria	344
Clkl1	Clostridium kluyveri DSM 555	153955000	Firmicutes	Bacteria	331
Etha1	Ethanoligenens harbinense YUAN-3	289639841	Firmicutes	Bacteria	330
Stsa1	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	73662050	Firmicutes	Bacteria	326
Maca1	Macrococcus caseolyticus JCSC5402	222152052	Firmicutes	Bacteria	320
Stpa1	Streptococcus parasanguinis ATCC 15912	296876165	Firmicutes	Bacteria	332
Stsa2	Streptococcus sanguinis SK36	125717717	Firmicutes	Bacteria	326
Grel1	Granulicatella elegans ATCC 700633	260584340	Firmicutes	Bacteria	329
Clce1	Clostridium cellulolyticum H10	220928081	Firmicutes	Bacteria	335
Clbo1	Clostridium botulinum A str. ATCC 3502	148378366	Firmicutes	Bacteria	336

Table 1. (Continued)

Casp2	Carnobacterium sp. AT7	163789482	Firmicutes	Bacteria	327
Cume1	Cupriavidus metallidurans CH34	94312800	Betaproteobacteria	Bacteria	328
Clte1	Clostridium tetani E88	28211482	Firmicutes	Bacteria	345
Sete2	Sebaldella termitidis ATCC 33386	269118696	Fusobacteria	Bacteria	341
Tusp1	Turicibacter sp. PC909	293375738	Firmicutes	Bacteria	334
Errh1	Erysipelothrix rhusiopathiae ATCC 19414	259504620	Firmicutes	Bacteria	337
Anpr1	Anaerococcus prevotii DSM 20548	257066632	Firmicutes	Bacteria	337
Fapr1	Faecalibacterium prausnitzii A2-165	257438125	Firmicutes	Bacteria	346
Aevi1	Aerococcus viridans ATCC 11563	295397120	Firmicutes	Bacteria	344
Acla1	Acholeplasma laidlawii PG-8A	162447967	Tenericutes	Bacteria	334
				Average Protein Size	340
				Standard Deviation	13

Cluster 3

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Stsp1	Streptomyces sp. AA4	256668217	Actinobacteria	Bacteria	350
Krfl1	Kribbella flavida DSM 17836	284028838	Actinobacteria	Bacteria	354
Frsp1	Frankia sp. EuI1c	280964447	Actinobacteria	Bacteria	363
Beca1	Beutenbergia cavernae DSM 12333	229819022	Actinobacteria	Bacteria	372
Nofa1	Nocardia farcinica IFM 10152	54026360	Actinobacteria	Bacteria	335
				Average Protein Size	355
				Standard Deviation	14

Cluster 4

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Camo1	Catonella morbi ATCC 51271	229824457	229824457	Firmicutes	339

Cluster 5

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**

Table 1. (Continued)

Basp1	Bacillus sp. SG-1	149182048	Firmicutes	Bacteria	337
Basp2	Bacillus sp. NRRL B-14911	89097939	Firmicutes	Bacteria	411
				Average Protein Size	374
				Standard Deviation	52

Cluster 6

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Pesp1	Candidatus Pelagibacter sp. HTCC7211	254455444	Alphaproteobacteria	Bacteria	360
Peub1	Candidatus Pelagibacter ubique HTCC1002	91762037	Alphaproteobacteria	Bacteria	357
Base1	Bacillus selenitireducens MLS10	297583165	Firmicutes	Bacteria	379
Clph1	Clostridium phytofermentans ISDg	160879921	Firmicutes	Bacteria	344
				Average Protein Size	360
				Standard Deviation	14

Cluster 7

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Spth1	Sphaerobacter thermophilus DSM 20745	269838330	Chloroflexi	Bacteria	360

Cluster 8

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Hawa1	Haloquadratum walsbyi DSM 16790	110669489	Euryarchaeota	Archaea	335
Hasa1	Halobacterium salinarum R1	169237800	Euryarchaeota	Archaea	338
Nama1	Natrialba magadii ATCC 43099	289582206	Euryarchaeota	Archaea	348
Hatu1	Haloterrigena turkmenica DSM 5511	284166442	Euryarchaeota	Archaea	327
Haut1	Halorhabdus utahensis DSM 12940	257052879	Euryarchaeota	Archaea	334
Hala1	Halorubrum lacusprofundi ATCC 49239	222478859	Euryarchaeota	Archaea	331
Havo1	Haloferax volcanii DS2	292656618	Euryarchaeota	Archaea	339
Orf3	uncultured haloarchaeon	148507971	Euryarchaeota	Archaea	346
Acsp1	Acidobacterium sp. MP5ACTX8	299137343	Acidobacteria	Bacteria	365
Kove1	Candidatus Koribacter versatilis Ellin345	94969403	Acidobacteria	Bacteria	390

Table 1. (Continued)

Acsp6	Acidobacterium sp. MP5ACTX8	299138210	Acidobacteria	Bacteria	364
Myav1	Mycobacterium avium subsp. paratuberculosis	8927423	Actinobacteria	Bacteria	422
Stsp2	Streptomyces sp. AA4	256670250	Actinobacteria	Bacteria	359
Thte1	Thermobaculum terrenum ATCC BAA-798	269925163	none	Bacteria	352
Devu1	Desulfovibrio vulgaris str. Hildenborough	46578540	Deltaproteobacteria	Bacteria	360
Depi1	Desulfovibrio piger ATCC 29098	212704746	Deltaproteobacteria	Bacteria	355
Dere1	Desulfotomaculum reducens MI-1	134299714	Firmicutes	Bacteria	357
Dere2	Desulfotomaculum reducens MI-1	134299719	Firmicutes	Bacteria	374
				Average Protein Size	355
				Standard Deviation	23

Cluster 9

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Dear1	Dechloromonas aromatica RCB	71908113	Betaproteobacteria	Bacteria	336

Cluster 10

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Raso1	Ralstonia solanacearum GMI1000	17545830	Betaproteobacteria	Bacteria	366
Cuta1	Cupriavidus taiwanensis	194290140	Betaproteobacteria	Bacteria	368
				Average Protein Size	367
				Standard Deviation	1

Cluster 11

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Arni1	Arcobacter nitrofigilis DSM 7299	296272734	Epsilonproteobacteria	Bacteria	342
Arbu1	Arcobacter butzleri RM4018	157737300	Epsilonproteobacteria	Bacteria	340
Sude1	Sulfurospirillum deleyianum DSM 6946	268678926	Epsilonproteobacteria	Bacteria	340
Came1	Caminibacter mediatlanticus TB-2	149194709	Epsilonproteobacteria	Bacteria	341
Sude2	Sulfurimonas denitrificans DSM 1251	78778222	Epsilonproteobacteria	Bacteria	348

Table 1. (Continued)

Nisp1	Nitratiruptor sp. SB155-2	152989917	Epsilonproteobacteria	Bacteria	351
Besp1	Beggiatoa sp. PS	153872787	Gammaproteobacteria	Bacteria	356
Frtu1	Francisella tularensis subsp. holarctica	89256458	Gammaproteobacteria	Bacteria	350
Pema1	Persephonella marina EX-H1	225850223	Aquificae	Bacteria	339
Suaz1	Sulfurihydrogenibium azorense Az-Fu1	225847856	Aquificae	Bacteria	336
Prsp1	Prevotella sp. oral taxon 299 str. F0039	288801300	Bacteroidetes	Bacteria	348
Poen1	Porphyromonas endodontalis ATCC 35406	229495309	Bacteroidetes	Bacteria	352
Akmu1	Akkermansia muciniphila ATCC BAA-835	187735208	Verrucomicrobia	Bacteria	340
Prco1	Prevotella copri DSM 18205	281421027	Bacteroidetes	Bacteria	338
Prme1	Prevotella melaninogenica ATCC 25845	252122021	Bacteroidetes	Bacteria	348
Bath2	Bacteroides thetaiotaomicron VPI-5482	29347329	Bacteroidetes	Bacteria	336
Baco1	Bacteroides coprocola DSM 17136	189460393	Bacteroidetes	Bacteria	358
Hewi1	Helicobacter winghamensis ATCC BAA-430	237752356	Epsilonproteobacteria	Bacteria	346
Hebi1	Helicobacter bilis ATCC 43879	237751625	Epsilonproteobacteria	Bacteria	339
Hemu1	Helicobacter mustelae 12198	291277613	Epsilonproteobacteria	Bacteria	342
Heca1	Helicobacter canadensis MIT 98-5491	224419042	Epsilonproteobacteria	Bacteria	370
Caje1	Campylobacter jejuni RM1221	57237824	Epsilonproteobacteria	Bacteria	365
Hehe1	Helicobacter hepaticus ATCC 51449	32266660	Epsilonproteobacteria	Bacteria	345
Wosu1	Wolinella succinogenes DSM 1740	34558479	Epsilonproteobacteria	Bacteria	333
Cacu1	Campylobacter curvus 525.92	154174967	Epsilonproteobacteria	Bacteria	331
Caco1	Campylobacter concisus 13826	157165450	Epsilonproteobacteria	Bacteria	327
Cala1	Campylobacter lari RM2100	222823305	Epsilonproteobacteria	Bacteria	347
Cafe1	Campylobacter fetus subsp. fetus 82-40	118474923	Epsilonproteobacteria	Bacteria	342
Pspu1	Pseudomonas putida GB-1	167033248	Gammaproteobacteria	Bacteria	353
Mafe1	Mariprofundus ferrooxydans PV-1	114775399	Zetaproteobacteria	Bacteria	342
Idlo1	Idiomarina loihiensis L2TR	56459735	Gammaproteobacteria	Bacteria	331
				Average Protein Size	344
				Standard Deviation	10

Table 1. (Continued)

Cluster 12

Abbreviation	Organism	gi #	Group	Kingdom	Protein Size
Psat2	Pseudoalteromonas atlantica T6c	109896446	Gammaproteobacteria	Bacteria	368
Alba2	Alteromonadales bacterium TW-7	119471064	Gammaproteobacteria	Bacteria	358
Masp1	Marinobacter sp. ELB17	126666024	Gammaproteobacteria	Bacteria	360
Toau1	Tolumonas auensis DSM 9187	237807100	Gammaproteobacteria	Bacteria	353
Laho1	Laribacter hongkongensis HLHK9	226941856	Betaproteobacteria	Bacteria	352
Dida1	Dickeya dadantii Ech703	242239750	Gammaproteobacteria	Bacteria	359
Dize1	Dickeya zeae Ech1591	251790019	Gammaproteobacteria	Bacteria	361
Peat1	Pectobacterium atrosepticum SCRI1043	50121649	Gammaproteobacteria	Bacteria	357
Pasp2	Pantoea sp. At-9b	258638162	Gammaproteobacteria	Bacteria	353
Paan1	Pantoea ananatis LMG 20103	291618107	Gammaproteobacteria	Bacteria	350
Erta1	Erwinia tasmaniensis Et1/99	188533420	Gammaproteobacteria	Bacteria	351
Yein1	Yersinia intermedia ATCC 29909	238794877	Gammaproteobacteria	Bacteria	336
Yepe1	Yersinia pestis KIM 10	22126755	Gammaproteobacteria	Bacteria	368
Prst1	Providencia stuartii ATCC 25827	183599515	Gammaproteobacteria	Bacteria	359
Prpe1	Proteus penneri ATCC 35198	226330586	Gammaproteobacteria	Bacteria	386
Phlu1	Photorhabdus luminescens subsp. laumondii TTO1	37526746	Gammaproteobacteria	Bacteria	360
Edta1	Edwardsiella tarda EIB202	269138535	Gammaproteobacteria	Bacteria	354
Saen1	Salmonella enterica subsp. arizonae serovar 62:z4,z23:--	161502641	Gammaproteobacteria	Bacteria	391
Chvi1	Chromobacterium violaceum ATCC 12472	34496218	Betaproteobacteria	Bacteria	341
Xaor1	Xanthomonas oryzae pv. oryzae KACC10331	58581823	Gammaproteobacteria	Bacteria	379
Stma1	Stenotrophomonas maltophilia R551-3	194365408	Gammaproteobacteria	Bacteria	359
Psae1	Pseudomonas aeruginosa UCBPP-PA14	116053530	Gammaproteobacteria	Bacteria	355
Acsp3	Acidovorax sp. JS42	121595140	Betaproteobacteria	Bacteria	360
Deac2	Delftia acidovorans SPH-1	160901144	Betaproteobacteria	Bacteria	361
Luni2	Lutiella nitroferrum 2002	224823562	Betaproteobacteria	Bacteria	347
Hese1	Herbaspirillum seropedicae SmR1	300313603	Betaproteobacteria	Bacteria	349
Posp3	Polaromonas sp. JS666	91791006	Betaproteobacteria	Bacteria	365

Table 1. (Continued)

Mesp2	Methylotenera sp. 301	297539025	Betaproteobacteria	Bacteria	346
Psar1	Psychrobacter arcticus 273-4	71066400	Gammaproteobacteria	Bacteria	360
Pssp2	Psychrobacter sp. PRwf-1	148652837	Gammaproteobacteria	Bacteria	411
Hain1	Haemophilus influenzae Rd KW20	16273532	Gammaproteobacteria	Bacteria	338
Agac1	Aggregatibacter actinomycetemcomitans D11S-1	261867728	Gammaproteobacteria	Bacteria	349
Haso1	Haemophilus somnus 129PT	113460490	Gammaproteobacteria	Bacteria	325
Caho1	Cardiobacterium hominis ATCC 15826	258545535	Gammaproteobacteria	Bacteria	326
				Average Protein Size	357
				Standard Deviation	17

Cluster 13

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Chsa1	Chromohalobacter salexigens DSM 3043	92113885	Gammaproteobacteria	Bacteria	343
Maaq1	Marinobacter aquaeolei VT8	120537072	Gammaproteobacteria	Bacteria	336
Cops1	Colwellia psychrerythraea 34H	71277912	Gammaproteobacteria	Bacteria	329
Neca1	Neptuniibacter caesariensis	89092492	Gammaproteobacteria	Bacteria	330
Hach1	Hahella chejuensis KCTC 2396	83647402	Gammaproteobacteria	Bacteria	327
				Average Protein Size	333
				Standard Deviation	7

Cluster 14

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Lisp1	Limnobacter sp. MED105	149925525	Betaproteobacteria	Bacteria	345

Cluster 15

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Xaau1	Xanthobacter autotrophicus Py2	154244222	Alphaproteobacteria	Bacteria	360
Brsp2	Bradyrhizobium sp. ORS278	146337430	Alphaproteobacteria	Bacteria	324

Table 1. (Continued)

Hyde1	Hyphomicrobium denitrificans ATCC 51888	300022361	Alphaproteobacteria	Bacteria	356
Afsp1	Afipia sp. 1NLS2	299132612	Alphaproteobacteria	Bacteria	345
Rhpa1	Rhodopseudomonas palustris CGA009	39934839	Alphaproteobacteria	Bacteria	354
Rhva1	Rhodomicrobium vannielii ATCC 17100	283824404	Alphaproteobacteria	Bacteria	359
Susp1	Sulfitobacter sp. EE-36	83943575	Alphaproteobacteria	Bacteria	342
Azca1	Azorhizobium caulinodans ORS 571	158422295	Alphaproteobacteria	Bacteria	370
Xaau2	Xanthobacter autotrophicus Py2	154246032	Alphaproteobacteria	Bacteria	355
Rhpa2	Rhodopseudomonas palustris BisA53	115523802	Alphaproteobacteria	Bacteria	362
Pade1	Paracoccus denitrificans PD1222	119387075	Alphaproteobacteria	Bacteria	339
Melo1	Mesorhizobium loti MAFF303099	13472054	Alphaproteobacteria	Bacteria	325
Rhru1	Rhodospirillum rubrum ATCC 11170	83591944	Alphaproteobacteria	Bacteria	351
Sysp1	Synechococcus sp. PCC 7335	254421461	Cyanobacteria	Bacteria	341
Sous1	Candidatus Solibacter usitatus Ellin6076	116619968	Acidobacteria	Bacteria	360
Rosp2	Roseovarius sp. TM1035	149204735	Alphaproteobacteria	Bacteria	342
Clmi1	Clavibacter michiganensis subsp. sepedonicus	170782669	Actinobacteria	Bacteria	334
Orf2	marine actinobacterium PHSC20C1	88856864	Actinobacteria	Bacteria	363
Stsp3	Streptomyces sp. C	256768632	Actinobacteria	Bacteria	321
Arau1	Arthrobacter aurescens TC1	119964080	Actinobacteria	Bacteria	349
Arch1	Arthrobacter chlorophenolicus A6	220912636	Actinobacteria	Bacteria	347
Resa1	Renibacterium salmoninarum ATCC 33209	163840827	Actinobacteria	Bacteria	343
Kira1	Kineococcus radiotolerans SRS30216	152967537	Actinobacteria	Bacteria	335
Acmi1	Actinosynnema mirum DSM 43827	256380034	Actinobacteria	Bacteria	288
Saer1	Saccharopolyspora erythraea NRRL 2338	291008791	Actinobacteria	Bacteria	314
Prfr1	Propionibacterium freudenreichii subsp. shermanii CIRM-BIA1	297626815	Actinobacteria	Bacteria	360
Cogl1	Corynebacterium glutamicum ATCC 13032	62388907	Actinobacteria	Bacteria	345
Acod1	Actinomyces odontolyticus ATCC 17982	154508466	Actinobacteria	Bacteria	353
Brfa1	Brachybacterium faecium DSM 4810	257068965	Actinobacteria	Bacteria	361
Rher1	Rhodococcus erythropolis PR4	226303855	Actinobacteria	Bacteria	348
Rhjo1	Rhodococcus jostii RHA1	111025633	Actinobacteria	Bacteria	357
Kyse1	Kytococcus sedentarius DSM 20547	256824845	Actinobacteria	Bacteria	441

Table 1. (Continued)

Brli1	Brevibacterium linens BL2	260904301	Actinobacteria	Bacteria	367
Lefe1	Leptospirillum ferrodiazotrophum	251772776	Nitrospirae	Bacteria	358
Lesp1	Leptospirillum sp. Group II '5-way CG'	206603468	Nitrospirae	Bacteria	343
				Average Protein Size	349
				Standard Deviation	23

Cluster 16

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Casp1	Caulobacter sp. K31	167645244	Alphaproteobacteria	Bacteria	369
Posp2	Polaromonas sp. JS666	91787914	Betaproteobacteria	Bacteria	338
Deac1	Delftia acidovorans SPH-1	160897433	Betaproteobacteria	Bacteria	348
Posp1	Polaromonas sp. JS666	91787056	Betaproteobacteria	Bacteria	340
Luni1	Lutiella nitroferrum 2002	224827331	Betaproteobacteria	Bacteria	345
Vish1	Vibrio shilonii AK1	149187943	Gammaproteobacteria	Bacteria	344
Alba1	Alteromonadales bacterium TW-7	119469732	Gammaproteobacteria	Bacteria	334
Alma1	Alteromonas macleodii ATCC 27126	239994669	Gammaproteobacteria	Bacteria	334
Psat1	Pseudoalteromonas atlantica T6c	109899182	Gammaproteobacteria	Bacteria	331
Acsp5	Acinetobacter sp. DR1	299769466	Gammaproteobacteria	Bacteria	336
Lech1	Leptothrix cholodnii SP-6	171058852	Betaproteobacteria	Bacteria	364
Raeu1	Ralstonia eutropha JMP134	73538367	Betaproteobacteria	Bacteria	342
Chte1	Chlorobium tepidum TLS	21673674	Chlorobi	Bacteria	358
Meno2	Methylobacterium nodulans ORS 2060	220919678	Alphaproteobacteria	Bacteria	346
Papa1	Paracoccus pantotrophus	73917985	Alphaproteobacteria	Bacteria	338
Rhsp2	Rhodobacter sphaeroides 2.4.1	77465158	Alphaproteobacteria	Bacteria	332
Pave1	Paracoccus versutus	90818629	Alphaproteobacteria	Bacteria	338
Rusp1	Ruegeria sp. TM1040	99079994	Alphaproteobacteria	Bacteria	345
Rode1	Roseobacter denitrificans OCh 114	110678179	Alphaproteobacteria	Bacteria	346
Rosp1	Roseobacter sp. AzwK-3b	149912549	Alphaproteobacteria	Bacteria	344
Rosp3	Roseovarius sp. HTCC2601	114762777	Alphaproteobacteria	Bacteria	343

Table 1. (Continued)

Cisp1	Citreicella sp. SE45	260430142	Alphaproteobacteria	Bacteria	346
Thsp1	Thalassiobium sp. R2A62	255261877	Alphaproteobacteria	Bacteria	343
Laal1	Labrenzia alexandrii DFL-11	254500665	Alphaproteobacteria	Bacteria	331
Rhsp1	Rhodobacter sphaeroides KD131	221369709	Alphaproteobacteria	Bacteria	338
Sast1	Sagittula stellata E-37	126730984	Alphaproteobacteria	Bacteria	342
Orf4	marine gamma proteobacterium HTCC2080	119504477	Gammaproteobacteria	Bacteria	334
Fupe1	Fulvimarina pelagi HTCC2506	114705652	Alphaproteobacteria	Bacteria	336
Hoph1	Hoeflea phototrophica DFL-43	163757941	Alphaproteobacteria	Bacteria	335
Ronu1	Roseovarius nubinhibens ISM	83952400	Alphaproteobacteria	Bacteria	328
Pssp3	Pseudovibrio sp. JE062	254472845	Alphaproteobacteria	Bacteria	354
Spja1	Sphingobium japonicum UT26S	294146855	Alphaproteobacteria	Bacteria	340
Meno1	Methylobacterium nodulans ORS 2060	220922665	Alphaproteobacteria	Bacteria	351
				Average Protein Size	342
				Standard Deviation	9

Cluster 17

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Bilo1	Bifidobacterium longum DJO10A	46190801	Actinobacteria	Bacteria	338
Biga1	Bifidobacterium gallicum DSM 20093	261337255	Actinobacteria	Bacteria	372
Gava1	Gardnerella vaginalis ATCC 14019	297585970	Actinobacteria	Bacteria	353
Bian1	Bifidobacterium animalis subsp. lactis AD011	219683482	Actinobacteria	Bacteria	369
Lafe1	Lactobacillus fermentum IFO 3956	184154862	Firmicutes	Bacteria	342
Labr1	Lactobacillus brevis subsp. gravesensis ATCC 27305	227510697	Firmicutes	Bacteria	348
Labr2	Lactobacillus brevis ATCC 367	116333063	Firmicutes	Bacteria	343
Pepe1	Pediococcus pentosaceus ATCC 25745	116492112	Firmicutes	Bacteria	341
Lajo1	Lactobacillus johnsonii NCC 533	42519043	Firmicutes	Bacteria	343
Geha1	Gemella haemolysans ATCC 10379	241889197	Firmicutes	Bacteria	328
Mimu1	Mitsuokella multacida DSM 20544	255658506	Firmicutes	Bacteria	336
Pala1	Paenibacillus larvae subsp. larvae BRL-230010	167463572	Firmicutes	Bacteria	361
Baan1	Bacillus anthracis str. Ames	30265201	Firmicutes	Bacteria	340
Lysp1	Lysinibacillus sphaericus C3-41	169830090	Firmicutes	Bacteria	339

Table 1. (Continued)

Limo1	Listeria monocytogenes EGD-e	16804186	Firmicutes	Bacteria	335
Baam1	Bacillus amyloliquefaciens FZB42	154687883	Firmicutes	Bacteria	325
Ocih1	Oceanobacillus iheyensis HTE831	23100861	Firmicutes	Bacteria	340
Thpo1	Thermincola potens JR	296132062	Firmicutes	Bacteria	346
Cahy1	Carboxydothermus hydrogenoformans Z-2901	78043356	Firmicutes	Bacteria	343
Moth1	Moorella thermoacetica ATCC 39073	83590712	Firmicutes	Bacteria	341
Hemo1	Heliobacterium modesticaldum Ice1	167629030	Firmicutes	Bacteria	347
Bame1	Bacillus megaterium QM B1551	294498493	Firmicutes	Bacteria	348
Brbr1	Brevibacillus brevis NBRC 100599	226315217	Firmicutes	Bacteria	356
Ligr1	Listeria grayi DSM 20601	299821263	Firmicutes	Bacteria	332
Pasp1	Paenibacillus sp. oral taxon 786 str. D14	253574339	Firmicutes	Bacteria	363
Gesp1	Geobacillus sp. Y412MC10	261409828	Firmicutes	Bacteria	347
Geka1	Geobacillus kaustophilus HTA426	56419425	Firmicutes	Bacteria	345
Deha1	Desulfitobacterium hafniense Y51	89896937	Firmicutes	Bacteria	345
Thca1	Thermosinus carboxydivorans Nor1	121535229	Firmicutes	Bacteria	356
Batu1	Bacillus tusciae DSM 2912	295695644	Firmicutes	Bacteria	353
Pasp3	Paenibacillus sp. JDR-2	251799896	Firmicutes	Bacteria	349
Stau4	Staphylococcus aureus	270300222	Firmicutes	Bacteria	331
Stca1	Staphylococcus carnosus subsp. carnosus TM300	224476415	Firmicutes	Bacteria	330
Stau3	Staphylococcus aureus subsp. aureus COL	57652600	Firmicutes	Bacteria	331
Stau2	Staphylococcus aureus	270055359	Firmicutes	Bacteria	322
Stwa1	Staphylococcus warneri L37603	239636543	Firmicutes	Bacteria	330
Myxa1	Myxococcus xanthus DK 1622	108758020	Deltaproteobacteria	Bacteria	347
Mesi1	Meiothermus silvanus DSM 9946	297567155	Deinococcus-Thermus	Bacteria	335
Ktra1	Ktedonobacter racemifer DSM 44963	298249957	Chloroflexi	Bacteria	361
				Average Protein Size	344
				Standard Deviation	12

Table 1. (Continued)

Cluster 18

Abbreviation	Organism	gi #	Group	Kingdom	Protein Size
Fusp1	Fusobacterium sp. 3_1_5R	257452238	Fusobacteria	Bacteria	339
Casp3	Carnobacterium sp. AT7	163791133	Firmicutes	Bacteria	344
Clpe1	Clostridium perfringens str. 13	18309111	Firmicutes	Bacteria	339
Clbu1	Clostridium butyricum 5521	182417819	Firmicutes	Bacteria	335
Clme1	Clostridium methylpentosum DSM 5476	225016941	Firmicutes	Bacteria	340
Clno1	Clostridium novyi NT	118443784	Firmicutes	Bacteria	334
Clbe1	Clostridium beijerinckii NCIMB 8052	150017970	Firmicutes	Bacteria	343
Tusp2	Turicibacter sp. PC909	293376354	Firmicutes	Bacteria	343
Cldi1	Clostridium difficile 630	126697763	Firmicutes	Bacteria	343
Enfa1	Enterococcus faecium DO	69250003	Firmicutes	Bacteria	347
Enfa2	Enterococcus faecalis V583	29375506	Firmicutes	Bacteria	342
Enca1	Enterococcus casseliflavus EC30	257866432	Firmicutes	Bacteria	340
Lapl1	Lactobacillus plantarum WCFS1	28379312	Firmicutes	Bacteria	336
Lala1	Lactococcus lactis subsp. cremoris SK11	116511231	Firmicutes	Bacteria	346
Lasa1	Lactobacillus sakei subsp. sakei 23K	81428923	Firmicutes	Bacteria	337
Sete1	Sebaldella termitidis ATCC 33386	269122099	Fusobacteria	Bacteria	331
Oeoe1	Oenococcus oeni PSU-1	116490807	Firmicutes	Bacteria	331
Leme1	Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	116619128	Firmicutes	Bacteria	330
Wepa1	Weissella paramesenteroides ATCC 33313	241895342	Firmicutes	Bacteria	331
Haor1	Halothermothrix orenii H 168	220931390	Firmicutes	Bacteria	326
				Average Protein Size	338
				Standard Deviation	6

Cluster 19

Abbreviation	Organism	gi #	Group	Kingdom	Protein Size
Rhle1	Rhizobium leguminosarum bv. viciae 3841	116249406	Alphaproteobacteria	Bacteria	353
Agtu1	Agrobacterium tumefaciens str. C58	51316737	Alphaproteobacteria	Bacteria	347

Table 1. (Continued)

Meex1	Methylobacterium extorquens AM1	240137889	Alphaproteobacteria	Bacteria	350
Brsu1	Brucella suis 1330	23500943	Alphaproteobacteria	Bacteria	336
Roce1	Roseomonas cervicalis ATCC 49957	296532888	Alphaproteobacteria	Bacteria	357
Azsp1	Azospirillum sp. B510	288961159	Alphaproteobacteria	Bacteria	365
Mesp1	Methylobacterium sp. 4-46	170744488	Alphaproteobacteria	Bacteria	348
Mech1	Methylobacterium chloromethanicum CM4	218532118	Alphaproteobacteria	Bacteria	356
Brja1	Bradyrhizobium japonicum USDA 110	27378300	Alphaproteobacteria	Bacteria	352
Spwi1	Sphingomonas wittichii RW1	148553873	Alphaproteobacteria	Bacteria	353
Xaal1	Xanthomonas albilineans	285019310	Gammaproteobacteria	Bacteria	335
Bahe1	Bartonella henselae str. Houston-1	49475041	Alphaproteobacteria	Bacteria	348
Chsp1	Chelativorans sp. BNC1	110635301	Alphaproteobacteria	Bacteria	339
Gldi1	Gluconacetobacter diazotrophicus PAl 5	162146484	Alphaproteobacteria	Bacteria	362
Acpa1	Acetobacter pasteurianus IFO 3283-01	258542401	Alphaproteobacteria	Bacteria	365
Glox1	Gluconobacter oxydans 621H	58040588	Alphaproteobacteria	Bacteria	372
				Average Protein Size	352
				Standard Deviation	11

Cluster 20

Abbreviation	**Organism**	**gi #**	**Group**	**Kingdom**	**Protein Size**
Arfu1	Archaeoglobus fulgidus DSM 4304	11499213	Euryarchaeota	Archaea	328
					328

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:**3389-402.

Brüggemann C, Denger K, Cook AM, Ruff J., 2004**,** Enzymes and genes of taurine and isethionate dissimilation in Paracoccus denitrificans. *Microbiology* **150**, 805-816

Devereux, J. Haeberli, P., Smithies, O. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* **12:**387-95.

Disz T, Akhter S, Cuevas D, Olson R, Overbeek R, Vonstein V, Stevens R, Edwards RA. 2010**.** Accessing the SEED genome databases via Web services API: tools for programmers. *BMC Bioinformatics*. Jun 14;11:319.

Eddy, S.R., 1998. Multiple alignment and multiple sequence based searches. http://selab.janelia.org/publications/Eddy98b/Eddy98b-preprint.pdf.

Eddy, S.R., 2008 A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PloS Comput Biol* **4:**e1000069.

Krogh, A., Larsson, B., von Heijne G., Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:**567-80.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. 33:5691-702.

Pardee AB., 1966. Purification and properties of a sulfate-binding protein from Salmonella typhimurium. *J Biol Chem*. **241**: 2886-92

Pearson, W., 2006. *GGsearch*. [Computer program]. http://fasta.bioch.virginia.edu/gasta_www2/fasta_list2.shtml

Piłsyk S, Paszewski A., 2009. Sulfate permeasesphylogenetic diversity of sulfate transport. *Acta Biochimica Polonica* **56**: 375-84

Reddy, V. 2010. *Global Sequence Alignment Tool*. [Computer program]

Rein U, Gueta R, Denger K, Ruff J, Hollemeyer K, Cook AM., 2005. Dissimilation of cysteate via 3-sulfolactate sulfo-lyase and a sulfate exporter in Paracoccus pantotrophus NKNCYSA. *Microbiology* **151**, 737-747.

Saier, M.H., Jr. 1994. Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol Rev* **58:**71-93.

Saier, M.H., Jr., Yen, M.R., Noto, K., Tamang, D.G., Elkan, C. 2009. The Transporter Classification Databse: Recent Advances. *Nucleic Acids Res.* **37:**D274-8.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. 1997, The Clustal_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **37:**D274-8.

Tusnády, G.E., Simon, I. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* **283:**489-506.

Tusnády, G.E., Simon, I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17:**849-850.

von Heijne, G., 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J*. **5:**3021-7

von Heijne, G. 1992. Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule. *J Mol Biol.* **225:**487-94.

Yen, M.R., Choi, J., Saier, M.H., Jr. 2009. Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. *J Mol Microbiol Biotechnol* **17:**163-76.

Zhai, Y., Saier, M.H., 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:**501-2.

Zhai, Y., Saier, M.H., 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:**285-6.

Zhai Y, Saier M.H., Jr. 2002. A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins. *J Mol Microbiol Biotechnol* **4:**375-7.

Zhai, Y., Tchieu, J. Saier, M.H., Jr. 2002. A web-based Tree View (TV) program for the visualization of phylogenetic trees. *J Mol Microbiol Biotechnolol* **4:**69-70.