UNIVERSITY OF CALIFORNIA, SAN DIEGO

P-type ATPase Analysis in Cyanobacteria and Six Miscellaneous Bacterial Phyla with Fully Sequenced Genomes

A thesis submitted in partial satisfaction of the requirements for the degree

Master of Science

in

Biology

by

Vardan Babayan

Committee in charge:

Professor Milton H. Saier, Jr., Chair
Professor Immo E. Scheffler
Associate Professor Joseph Pogliano

2010

The Thesis of Vardan Babayan is approved and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

Chair

University of California, San Diego
2010

DEDICATION

I dedicate this thesis to my friends and family and to Dr. Saier and the members of Saier Lab. A giant thank you goes out to Rita and Ruben Abagyan for being kind enough to take me under their wing and help me get through college in one piece. A huge thank you also goes to Dorjee and Jeenie for their everlasting support and guidance.

Most importantly, I'd like to thank Milton Saier for his superior guidance throughout this research. Dr. Saier has been a beacon of light to many students and I would like to thank him for standing by his brood and always making the best of it. Thank you to all of Saier Lab for helping and being there.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I'd like to acknowledge Dr. Milton H. Saier, Jr., for his support as the chair of my committee. Throughout my years in the lab he has never failed to guide and support the research performed by his students. He puts long hours into the laboratory and accomplishes more than anyone else I have met. His hungry enthusiasm for learning pushes everyone in the lab to better themselves and increase their awareness.

I would also like to acknowledge Dorjee G. Tamang, and Ming-Ren Yen. Dorjee provided everlasting guidance when it came to the technical aspect of the research. His great grasp on computer knowledge provided valuable in the use of many bioinformatical programs and websites. His input and effort is unmatched. Dr Ming contributed a great deal by providing the Make_table program which is used by the whole laboratory. I would especially like to mention the crucial help he provided by gradually changing the program to keep aiding the research in this thesis. I would also like to acknowledge Immo E. Scheffler and Joseph Pogliano for their involvement as committee members and for their understanding and communication in the process.

This thesis, including the Abstract, Introduction, Methods, Results, Discussion, and Appendix, in part, is being prepared for publication. The thesis author will be a co-author and co-investigator for this paper. Co-authors include: Milton H. Saier, Henry Chan, Charmy Gandhi, Kunal Hak, Danielle Harake, Kris Kumar, Perry Lee, Tze T. Li, Hao Yi Liu, Tony Chung Tung Lo, Cynthia J. Meyer, Steven Stanford and Krista S. Zamora.

ABSTRACT OF THE THESIS

P-type ATPase Analysis in Cyanobacteria and Six Miscellaneous Bacterial Phyla with Fully Sequenced Genomes

by

Vardan Babayan

Master of Science in Biology

University of California, San Diego, 2010

Professor Milton H. Saier, Jr., Chair

Nine cyanobacteria and 14 genera of bacteria from six phyla, with fully sequenced genomes, were analyzed for genes encoding P-type ATPase catalytic alpha-subunits. These miscellaneous phylas are *Aquificae, Chlamydia, Chlorobi,*

Chloroflexi, Deinococcus-Thermus and *Thermotogae*. In cyanobacteria 59 such proteins were identified, and in the miscellaneous phyla, a total of 31 ATPases were identified. The P-type ATPase representation in genomes of closely related species was greatly varied, suggesting a gain or loss of genes. Along with the examined phyla, an integrative approach was taken to examine the P-type ATPase representation across all the major kingdoms of life. The Clustal X and TreeView programs were used to construct multiple alignments and phylogenetic trees, respectively.

The families of proteins identified were either functionally characterized or previously uncharacterized. The alignments of the families were then used to examine motif conservation. Novel toplogies were discovered for families unique to both the prokaryotes and the eukaryotes, while the known toplogoical types were reinforced by the integrative approach to the reasearch. The constructed 16S rRNA trees were used along with the family trees to gain evidence for orthology, and further shed light on the evolutionary descendency of the proteins. The same approaches taken at the individual phyla level were then applied to the integrated data. Overall, these studies reveal the nature of cyanobacterial P-type ATPases, and the ATPases from the six miscellaneous phyla. The integrated approach was also used to reinforce known ideas about the proteins as they are represented across the major kindgoms of life, and also to reveal novel information about the P-type ATPase proteins in general.

INTRODUCTION

The phylum Cyanobacteria includes photosynthetic bacterial organisms that are very crucial to our environment. By removing carbon dioxide and producing oxygen these bacteria have immensely contributed to the equilibrium of the earth's atmosphere and the evolution of the present biosphere (Tandeau et al, 1993). Interestingly, cyanobacteria are the only phototrophic prokaryotes to perform photosynthesis in this plant-like manner, and prove to be an early evolutionary ancestor to the higher plant and algal chloroplasts (Garcia-Pichel, 1997). In this study nine cyanobacterial organisms were examined.

Acaryochloris marina was isolated in 1996 from a cell suspension that came from the shallow coast of Republic of Palau in the west Pacific Ocean (Miyashita et al, 2003). In this experiment the actual genome examined belongs to the strain MBIC11017. *Anabaena variabilis* is a prokaryotic alga that is very similar to the chloroplast and possesses a thylakoid membrane (Sato et al, 1980). The strain of *Anabaena* examined is ATCC 29413. A very close relative of the *Anabaena* is the genus *Nostoc*. *Nostoc sp.,* just as the *Anabaena*, are filamentous cyanobacteria, and are common on both terrestrial and aquatic habitats (Dodds et al, 1995). In fact, the terrestrial *Nostoc* is known to stay desiccated for months or even years and upon exposure to liquid water, can retain metabolic activity within a few hours (Dodds et al, 1995). *Gloeobacter violaceus* PCC 7421 is a rod-shaped unicellular bacterium that was isolated from a rock in Switzerland (Nakamura et al, 2003). The *Gloeobacter* cells lack thylakoid membranes, unlike other cyanobacteria, and from molecular

phylogenetic analysis, are known to have diverged in the earliest stages of the radiation of cyanobacteria and chloroplasts (Nakamura et al, 2003).

Prochlorococcus marinus is a marine cyanobacterium. Interestingly, they are known to dominate phytoplankton communities in most temperate, tropical and open-ocean ecosystems and thus are the dominant photosynthetic organism in the ocean (Dufresne et al, 2003). It can be seen what a grand role these and other cyanobacteria play in the wellbeing of our biosphere. The actual strain of *Prochlorococcus* used in this study is MIT 9303. *Synechococcus sp.* are unicellular bacteria that, like other cyanobacteria, require inorganic nutrients and the presence of light for growth (Collier et al, 1992). It is interesting to note that like other cyanobacteria *Synechococcus* go through interesting morphological and physiological changes when they are deprived of an essential element (Collier et al, 1992). The actual strain of *Synechococcus* used in this study is JA-2-3B. *Synechocystis* strain PCC 6803 is another member of cyanobacteria included in this study. This strain of *Synechocystis* is known to be able to grow photoheterotrophically, and thus is a great model for the study of oxygenic photosynthesis (Fulda et al, 2000).

Thermosynechococcus elongatus BP-1 is a thermophilic unicellular cyanobacterium. The thermophilic nature of the bacterium makes it very interesting for research and it is known to inhabit hot springs, having an optimum growth temperature of approximately 55 degrees Celsius (Nakamura et al, 2002). It is also interesting to note that in previous studies, *T. elongatus* displayed branching that is very close to the origin of cyanobacteria, which also points to

the interest of studying this bacterium. *Trichodesmium erythraeum* is the last cyanobacterial organism included in this study. The genus *Trichodesmium* was originally described in 1830 by Ehrenberg (Baalen et al, 1969). More specifically *T. erythraeum* is a planktonic alga that is frequently observed floating in great masses at the surface of the ocean in temperate areas (Baalen et al, 1969). Due to the early discovery of this bacterium, a lot of literature has been published describing the inner working and mechanisms of its physiology. The actual strain of *Trichodesmium eythraeum* studied is strain IMS101.

Cyanobacteria all share certain morphological and physiological similarities along with certain essential differences that make them more responsive to the stresses of their respective niches. Dating back to the Pre-Cambrian era, cyanobacteria have a long history of the adaptation to the Earthly biosphere. In fact, cyanobacteria are found in almost all of the ecosystems examined thus far (Tandeau et al, 1993). Indeed, this adaptive radiation across numerous habitats and changes in morphology and physiology are displayed in the above listing of the organisms included in this study. This divergence and historic adaptation are both beneficial and harmful to a general study, and can prove to be very valuable pieces of information when approached from a broad perspective.

Along with the cyanobacteria, six miscellaneous bacterial phyla were examined. *Aquifex aeolicus* is known to be one of the most thermophilic bacteria known. In fact, it is thought that this heat tolerance is a legacy from the earliest of bacterial forms (Deckert et al, 1998). The only organism studied from the

phylum Aquifex was *Aquifex aeolicus* VF5. The phylum Chlorobi includes three organisms, *Chlorobium chlorochromatii* CaD3, *Chlorobium tepidum* TLS and *Pelodictyon luteolum* DSM273. Chlorobi are green sulfur bacteria and are obligate photolithoautotrophs and obligate anaerobes and are capable of growth at light intensities that will not support the growth of any other photosynthetic organisms on Earth (Overmann et al, 2000; Overmann et al 2002; Tuschak et al 1999). The only organism from the phylum Thermotogae included in the study is *Thermotoga maritima* MSB8. *T. maritima,* a rod shaped bacterium, was originally isolated from geothermal deated marine sediment at Vulcano, Italy. This bacterium is known to metabolize many simple and complex sugars, making it interesting for applications in renewable energy (Nelson et al, 1999).

The phylum Chloroflexi includes three organisms, *Dehalococcoides sp.* CBDB1, *Dehalococcoides ethenogenes* 195 and *Thermomicrobium roseum* DSM 5159. Of these, the bacteria included in the genus *Dehalococcoides* are most interesting due to their unique ability to detoxify compounds in the environment, that are otherwise present for decades (Kube et al, 2005). There are other bacteria in the environment that are able to perform this task but it is known that *Dehalococcoides* is particularly adapted for this detoxification (Kube et al, 2005). The phylum Chlamydia includes four organisms. These are: *Chlamydia muridarum Nigg, Chlamydophila abortus* S26/3, *Chlamydophila caviae* GPIC and*Chlamydophila pneumoniae* TW-183. These bacteria are interesting due to their pathogenic nature, as in the case of *C. pneumoniae,* which is known to be a major cause of pneumonia among humans (Frikha-Gargouri et al, 2008). Along

with their ability to cause pneumonia links between infection of *C. pneumoniae* heart attacks and atherosclerosis have been found (Blasi et al, 1996). The last miscellaneous phylum studied is Deinococcus-Thermus. The organisms from this phylum included in this study are *Deinococcus radiodurans* R1 and *Thermus thermophilus* HB27. This phylum includes many species of bacteria that are known to be resistant to extreme radiation (Griffiths et al, 2007). They are interesting to investigate due to the fact that not a single unique biochemical or physiological characteristic of this group is known (Griffiths et al, 2007).

Within the above-mentioned phylas, P-type ATPase proteins were analyzed. These P-type ATPases are essential ion transporting pumps that goare present in nearly all living cells. Nine functionally characterized P-type ATPase families have been identified, and most have homologues in both eukaryotes and prokaryotes: only one (phospholipid flippases: TC# 3.A.3.8) is found only in eukaryotes, and only one (Kdp-type K^+-ATPase complexes; TC# 3.A.3.7) is found only in prokaryotes (Haupt et al, 2005; Poulsen et al, 2008). The others catalyze uptake and/or efflux of various cations in all domains of life. Evidence for P-type ATPases catalyzing transport of other ions in animals such as chloride has been presented (Gerencser et al, 2003), but the molecular identities of these enzymes have not been achieved.

The activities of all P-type ATPases depend on a cycle of autophosphorylation and dephosphorylation, and these pumps have two main conformations, E1 and E2 as illustrated in Fig. 1 (Gadsby et al, 2007, Kühlbrant et al, 2004). The ion-binding sites are located deep within the intramembrane

region of the pumps, and these sites become accessible to ions from the cytoplasm in the E1 conformation (Gadsby et al, 2007) and the extracellular medium in the E2 conformation. Cytoplasmic ion binding promotes phosphorylation of the pump, and the resulting E1P state occludes the bound ions. ADP is then released, relaxing the pump to the E2P conformation. The ion binding sites are exposed to the external surface of the cell, and an exchange of ions usually occurs. Dephosphorylation to the E2 conformation causes the newly bound ions to become occluded. The pump then relaxes back to the E1 state, and the entire cycle repeats (Fig. 1; Gadsby et al, 2007; Kühlbrant et al, 2004). Studies on the opening and closing of P-type ATPases suggest that they act like ion channels with two gates opening and closing alternately. The occurrence of the occluded state ensures that one gate closes before the other one opens (Artigas et al, 2003).

Most functionally characterized P-type ATPases share nine well-conserved motifs as shown in Fig. 2 (Moller et al, 1996). There are three primary hydrophilic structural entities in P-type ATPases: the B-domain for actuation, the large cytosolic C-domain for nucleotide binding, protein phosphorylation and enzyme-phosphate hydrolysis, and the junctional J-domain for phosphorylation control (Moller et al, 1996; Okkeri et al, 2002). The relative positions of the C and J domains change during catalysis allowing the bound ATP and the critical aspartyl phosphorylation site to interact during phosphoryl transfer (Clarke et al, 1990; Okkeri et al, 2002). Considerable information about their probable functions is available. Their order of appearance, in progression from the N-terminus to the

C-terminus of each sequence, has been defined as follows: (i) PGD, (ii) PAD, (iii) TGES, (iv) PEGL, (v) DKTGTLT, (vi) KGAPE, (vii) DPPR, (viii) MVTGD, and (ix) VAVTGDGVNDSPALKKADIGVAM (Moller et al, 1996).

In this paper, phylogenetic trees for the proteins and the 16S rRNAs are constructed to observe family representation and orthology of the P-type ATPases. Multiple alignments and hydropathy profiles are created and analyzed to locate conserved motifs and determine topology of the functionally characterized and uncharacterized families. Additionally, an integrative approach to the analysis is taken. P-type ATPases from Eukaryotes, Archaea and Bacteria have been integrated and studied together to analyze similarities, differences and unique characteristics. A uniform methodical approach is taken to analyze the P-type ATPases in the specific bacterial phylas, and to integrate the data on these crucial proteins.

Parts of the Introductory section are being prepared for publication. The thesis author will be a co-investigator and co-author of this paper. Co-authors include: Milton H. Saier, Henry Chan, Charmy Gandhi, Kunal Hak, Danielle Harake, Kris Kumar, Perry Lee, Tze T. Li, Hao Yi Liu, Tony Chung Tung Lo, Cynthia J. Meyer, Steven Stanford and Krista S. Zamora.

METHODS

P-type ATPases from fully sequenced prokaryotic genomes were initially retrieved from TransportDB (Ren et al, 2007). The Transporter Classification Database (TCDB; www.tcdb.org; Saier et al, 2006, 2009) was then used to screen each of the selected genomes for additional members of this superfamily and to verify assignments of these proteins as alpha-subunit homologues. To ensure that all possible P-type ATPase alpha-subunit sequences had been retrieved, one protein from each organism was blasted as a query sequence in NCBI PSI-BLAST searches, screening the proteomes of all organisms under study (Altschul et al, 1998, 2005; Thever et al, 2009). Redundant sequences were eliminated, and truncated sequences were either reconstituted to their full length when possible, or eliminated manually. Short sequences were not always artifactual, resulting from the presence of pseudogenes or from inaccurate sequencing or initiation codon assignment. Many short sequences were haloacid dehydrogenase (HAD) domain proteins of ~230 amino acyl residues (aas) and some other dehydrogenases of larger sizes.

Multiple alignments and phylogenetic trees were generated using the CLUSTAL X (Neighbor Joining) and TreeView programs, respectively, with default settings (Thompson et al, 1997; Zhai et al, 2002). The WHAT (Zhai et al, 2001) and AveHas (Zhai et al, 2001) programs were used to perform topological analyses of single protein sequences and multiply aligned sequences, respectively. The latter program is based on the CLAUSTALX program generated multiple alignment. Other programs such as the TMHMM and HMMTOP

programs (Ikeda et al, 2001; Kall et al, 2002; Krogh et al, 2001; Moller et al, 2001; Tusnady et al, 2001) were used for confirmatory purposes. Modified GAP, IC (InterCompare) and GS (Getscore) programs were used to compare sequences to provide statistically significant evidence for homology (Dayhoff et al, 1983; Devereux et al, 1984; Saier et al, 1994, 2009; Yen et al, 2009). Comparison scores, expressed in standard deviations (S.D.) were based on the GAP and IC programs (Saier et al, 1994, 2009; Wang et al, 2009). A value of 10 S.D. corresponds to a probability of 10^{-24} that the observed degree of sequence similarity could have occurred by chance (Dayhoff et al, 1983). This value was used as the criterion to establish homology in binary comparisons.

In order to reconstruct full-length ATPase sequences from partial sequences, the DNA in front of or behind the recognized fragmentary sequence was determined in the three co-directional reading frames using the BCM Search Launcher program (Smith et al, 1996). If the translated amino acid sequence exhibited clear similarity to a close homologue, it was manually fused to the fragment in a way that minimized the numbers and sizes of gaps in the binary alignment. At the N-termini, this most frequently resulted from incorrect initiation codon selection, while in the C-terminal region, it most frequently resulted from a sequencing error (often a frameshift) leading to premature termination. In such cases, the remaining sequence was similarly identified in one of the three open reading frames.

Searches for conserved motifs in addition to the nine currently recognized motifs (Moller et al, 1996), either in P-type ATPases from a specific group of

organisms, or for specific P-type ATPase families, was performed with the MEME program (Bailey et al, 1996). Default settings were used although the setting “any number of repetitions” was selected for predictions of how a single motif was distributed among the homologues. The locations of the motifs were determined relative to the positions of the TMSs, the later using the WHAT or HMTMM program (see above). ATPase families were examined for conserved motifs using the same approach.

For the integrated analysis of P-type ATPase data outside of the Cyanobacteria and the miscellaneous bacterial phyla, further data compiled by the Saier Lab was used. The same approach was taken to gather the outside data, thus it made sense in an integrated sense. I have used their data along with mine to conduct an integrated analysis of the P-type ATPase proteins across all the major kingdoms of life.

Parts of the Methods section are being prepared for publication. The thesis author will be a co-investigator and co-author of this paper. Co-authors include: Milton H. Saier, Henry Chan, Charmy Gandhi, Kunal Hak, Danielle Harake, Kris Kumar, Perry Lee, Tze T. Li, Hao Yi Liu, Tony Chung Tung Lo, Cynthia J. Meyer, Steven Stanford and Krista S. Zamora.

RESULTS

1. Cyanobacterial P-type ATPases

1.A. P-type ATPases Encoded Within the Genomes of Nine Cyanobacteria

Nine cyanobacteria, belonging to nine different genuses, all with fully sequenced genomes, were analyzed for genes encoding P-type ATPase catalytic subunits. Fifty-nine such homologues were identified. These organisms, their genome sizes, the number of recognized ORFs in each genome, and the representation of P-type ATPases in nine families of these enzymes are tabulated in Table 1. Genome sizes vary from 2.6 Mbp for *Thermosynechococcus elongatus* BP-1 to 8.3 Mbp for *Acaryochloris marina* MBIC11017. The diversity of this group of organisms therefore correlates with their varied genome sizes. Surprisingly, there is substantial variation reported with respect to the numbers of ORFs per unit of genome size. Thus, the lowest gene density is recorded for *Trichodesmium erythraeum* IMS101 with 571 genes per million base pairs while the highest gene density is observed for *Prochlorococcus marinus* str. MIT 9303 with 1110 genes per million base pairs. Whether these differences represent actual differences in genome construction or reflect different annotation methods has yet to be ascertained.

Table 1 summarizes the distributions of cyanobacterial P-type ATPases affiliated with nine different families of these enzymes. It can readily be seen that two of these families (Family 2 (Ca^{2+}) and Family 5 (Copper)) are overrepresented relative to all other families. In fact, they both have more than twice as many members (19 and 17, members, respectively) as do any other

families represented among these cyanobacteria. In third place, we find, Family 6 (heavy metals), Family 7 (Kdp-type K^+) and FUPA32 (unknown specificity) where each of these families have six to seven members. All other families have only one or two members represented within the nine organisms studied. These poorly represented families include Family 1 (Na^+,K^+), Family 4 (Mg^{2+}), FUPA23 and FUPA30. Families 3, 8 and 9, present predominantly in eukaryotes, are lacking, as are all of the eukaryotic FUPA families and most of the prokaryotic FUPA families.

The individual proteins found in the nine cyanobacteria examined are listed in Table 2 according to family. This table also presents the organismal sources of these proteins as well as their sizes and GenBank Index (gi) numbers. It can be seen that different families have substantially different average molecular sizes. The largest ATPases are within the (Na^+,K^+) ATPase Family 1 with an average size of 971±1 aas. These are followed by the Family 2 (Ca^{2+}) ATPases with an average size of 929±21 aas. The third largest ATPases are the FUPA23 enzymes with an average size of 825±3 aas. The Copper ATPases come next with 771±34 aas, and the Family 6 HM ATPases are about 50 residues smaller (721±88 aas) with much greater size variation. In fact, the largest of these enzymes is 238 aas larger than the smallest. The differences between these enzymes is in the N-terminal region where Nsp2 has an N-terminal extension of 30 residues relative to Ava4 and an additional ~75 residues at positions 131 – 205. Examination of this repeat element proved that it is the heavy metal binding domain present in both Copper and HM ATPases in variable

numbers.

1.B. Phylogenetic Analyses of Cyanobacterial P-type ATPases

The phylogenetic tree, showing all 59 P-type ATPase alpha-subunits, without standards from TCDB, is shown in Fig. 3. Proteins, in general, cluster according to both typological type and family. Thus, all Type I proteins (Families 5, 6, and 32) cluster at the bottom of the tree, all Type II proteins (Families 1, 2, 4, 23 and 30) cluster at the top of the tree, and the Type III proteins (Family 7) cluster tightly together at the upper left hand side of the tree. It is apparent that proteins assigned to Family 2 (Ca^{2+}) exhibit widely divergent sequences falling into five different clusters, one of which (Ter4) actually appears to be more distant from the other Ca^{2+}-ATPases than are the Family 1 (Na^{+},K^{+}) and the FUPA 30 family ATPases, based on the tree shown in Fig. 3. Nevertheless, TC BLAST searches suggested that Ter4 is most closely related to the Ca^{2+}-ATPases in TCDB than to members of the other Type II families. The FUPA 23 and FUPA 30 family proteins cluster somewhat more distantly from the other Type II ATPases. Further analyses of Type II ATPases will be presented below.

1.C. Type I ATPases

Among the Type I ATPases, the Copper ATPases fall into two deeply branching clusters or subfamilies, and the same is true of the HM ATPases. Deep branching is also seen for the clusters of six FUPA 32 proteins.

A 16S rRNA tree was derived for the nine organisms studied (Fig. 4). *Nostoc* and *Anabaena* cluster tightly together as expected with *Trichodesmium* and *Synechocystis* clustering more loosely with them at the bottom of the tree.

Synechococcus and *Gloeobacter* cluster together on a branch that also includes *Thermosynechococcus* and *Prochlorococcus*.

Family 5 (Copper) includes 17 proteins with from one to four members per organism. In fact, Family 5 is the only family that has at least one representative in each of the nine organisms studied. The most copper-transporting ATPases are observed for *Nostoc*, with four such proteins. Surprisingly, its close relative, *Anabaena,* only has two. Close orthologous relationships are observed for Ava6 and Nsp12, and also for Ava16 and Nsp4. On two occasions, *Anabaena* has apparently lost its *Nostoc* ortholog. This can be seen in the cases of Nsp1 and Nsp3. Nsp1 shows distant relationships to the other proteins and could have arisen by either vertical or horizontal transmission. Neither cluster 1 nor cluster 2 shows relationships that suggest orthology. It appears that genes encoding Copper ATPases have undergone deletion and frequent horizontal transfer.

Family 6 (HM) follows a much clearer pattern. The organisms that contain the HM proteins cluster together, as can be seen in the 16S rRNA tree (Fig. 4). They include *Anabaena, Nostoc, and Synechocystis w*hich show close evolutionary relationships. The relationships of both clusters 1 and 2 are consistent with orthology. However, considering the closeness of *A. variabilis* to *Nostoc* sp. PCC 7120, Ava4 and Nsp2 are sufficiently distant in sequence to suggest that they might not be orthologous.

The only organisms containing FUPA32 enzymes are in the *Nostoc, Anabaena* and *Thermosynechococcus* genuses (Figs. 3 and 4). There are two sets of *Anabaena/Nostoc* orthologs, sufficiently closely related to each other to

suggest that they arose by an early gene duplication event, long before the separation of these two species. The two *Thermosynechococcus* paralogues are too distant from each other to suggest gene duplication as the origin of these proteins.

1.D. Type II ATPases

Family 2 (Ca^{2+}-ATPases) is a family of highly sequence-divergent proteins. Just as in Family 5 (Copper), a broad spectrum of organismal representation is observed. At least one Family 2 ATPase is present in every organism studied except for *Prochlorococcus marinus* str. MIT9303. Fig. 3 shows the variations observed in Family 2 as explained below. Families 1, 4, 23 and 30 are all distantly related to the Ca^{2+}-ATPases (upper half of Fig. 3) which occur in multiple subfamilies. We therefore conducted more detailed analyses of Type II ATPases including all such members from TCDB plus additional proteins identified by conducting NCBI BLAST searches with representative cyanobacterial orphan proteins. The proteins included in this study are presented in Table 3, and the phylogenetic tree including these 68 proteins is shown in Fig. 5.

In contrast to all previous studies, we could identify nine, distinct, deeply branching clusters or subfamilies of Ca^{2+}-ATPases. These are labeled Ca1 to Ca9 in Fig. 5. The tree also shows the positions of Families 1 (Na^+,K^+), 3 (H^+ or Mn^{2+}/Cd^{2+}), 4 (Mg^{2+}) and 9 (Na^+ or K^+), so labeled. The previously recognized eukaryotic families of Ca^{2+}-ATPases are Ca1 (golgi), Ca4 (endoplasmic reticulum (ER)/sarcoplasmic reticulum (SR)) and Ca9 (plasma membrane (PM)). Of these

three well-characterized families from animals, only one family, including the Family 9 (PM) enzymes is represented in one of the cyanobacteria examined here. The tree reveals that cyanobacteria include putative Ca^{2+}-ATPases belonging to four subfamilies, three of which are not yet found in eukaryotes and one of which (Ter4) belongs to the same family as the plasma membrane (PM) enzymes of eukaryotes. It is also interesting to note that three Ca^{2+}-ATPases, classified in TCDB, one from *Aquifex aeolicus* (Ca6 = 3.A.3.2.24), one from *Plasmodium falciparum* (Ca7 = 3.A.3.2.8) and one from *Clostridium acetobutylicum* (Ca8 = 3.A.3.2.20), each comprises its own deep-rooting cluster or subfamily.

One of the proteins included in this study, Ter4, from *Trichodesmium erythraeum* IMS101, proved to have a topology distinct from any ATPase yet characterized. The hydropathy plot for this protein, shown in Fig. 6, reveals several characteristics of typical Type II ATPases. Thus, peaks 1 through 10, as labeled in Fig. 6, are all present as expected for a Type II ATPase. Surprisingly, however, two strong peaks of hydrophobicity were identified in this protein between TMSs 3 and 4. We label these two peaks A and B in Fig. 6.

The 75 aas, which comprise the inserted region of Ter4, were BLASTed against the NCBI NR protein database. This region proved to be present in putative Ca^{2+}- ATPases from several cyanobacteria including *Arthrospira maxima* (gi 209525516)*, Nostoc punctiforme* (gi 186686018)*, Lyngbya* sp*.* PCC 8106 (gi 119485128), and *Cyanothece* sp. ATCC 51142 (gi 172035065).

The cyanobacterial 16S ribosomal RNA tree shown in Fig. 7 reveals the

phylogenetic relationships of the represented cyanobacteria to each other. The organisms shown to have the two TMS insert between TMSs 3 and 4 are presented in bold print. These organisms are *N. punctiforme, Lyngbya* sp. PCC 8106, and *Cyanothece* sp. ATCC 51142; the 16S rRNA for *Arthrospira maxima* could not be found. Examination of the tree shown in Fig. 7 revealed that these three organisms cluster in the upper half of the tree, together with four organisms, which apparently lack these unusual Ca^{2+}-ATPases. None of the organisms shown in the lower half of the tree contains such a protein**.**

1.E. Type III ATPases

Type III Kdp ATPase □ (B)-subunits (8, 12) cluster tightly together. The surprising feature of the Kdp ATPases is that although there are seven such proteins, three of them are derived from *A. variabilis* and two are from *Nostoc* sp. PCC 7120. Thus, only four of the nine cyanobacteria examined possess these enzymes. To the best of our knowledge, this is the first time three Kdp-type ATPases have been identified in a single organism.

Comparing the 16S rRNA tree in Fig. 4 with the members of Family 7 in Fig. 3, we see that the four organisms represented in Fig. 4 show the same relative phylogenetic relationships as the corresponding 16S rRNAs. That is, *Nostoc* and *Anabaena* are very closely related, *Synechocystis* is substantially more distant, while *Gloeobacter* is most distant of all. The three *Anabaena* paralogues appear to have risen from two distinct gene duplication events, the earlier one giving rise to Ava5 and the precursor of Ava2 and Ava12; the second one gave rise to Ava2 and Ava12. These events occurred long before the

divergence of *Nostoc* from *Anabaena*. Only in the case of Ava2 was the *Nostoc* orthologue lost.

1.F. Conserved Motifs

The nine most conserved motifs in P-type ATPases for the cyanobacterial enzymes are compiled in Table 4. For two of the families, Families 4 (Ava8) and 30 (Ava15), where only a single member was identified in the above described studies, additional members of these families were identified by conducting NCBI BLAST searches. For Family 4 proteins, the GenBank Index numbers (gi#s) were: 22695030, 170760778, 168179304, and 91203403. For Family 30, they were: 85857966, 116754767, 237745459, and 171058719. The ClustalX alignments of proteins from each of these families were used to identify the nine motifs. The characteristics of these motifs are summarized in Table 4.

Motif 1, PGD, is fully conserved only in Families 2 and 4, but in Family 1, this motif is RGD, and in Families 7 and 30, the dominant residue at position 1 is also R. The G is fully conserved in all families except Family 30. The D is fully conserved only in Families 1, 2, 4 and 30, but it is the dominant residue in all other families except Family 6. PAD is motif 2, which is fully conserved only in Families 2 and 4. However the D is fully conserved in all of the families.

Motif 3, TGES, is well conserved in all of the nine TCDB families examined. Families 1, 6, 7, 23, and 32 have this motif fully conserved. In fact, TGES is the consensus motif in all families except Family 2, where this motif is TGAA. The G is fully conserved in all families, and the T is fully conserved in all families except FUPA30.

Motif 4, PEGL, shows tremendous variability. Only Families 1 and 2 have perfect conservation of the PEGL motif. Variations on the PEGL theme are observed for the remaining families, but it is interesting to note that Family 7 has a fully conserved motif, PTTI.

Motif 5, DKTGTLT, the phosphorylation site motif, is fully conserved in most families, except in Family 2, where in one protein, Ava5, the D is substituted by a G. Only in Family 6 is the L substituted by I in two proteins, Ava6 and Nos1, and in Family 7, this residue (I) is fully conserved.

Motif 6, KGAPE, is fully conserved only in Families 1, 4, and 30. The only residue showing perfect conservation is the G, which appears in every one of the families studied. The poorest conservation is observed for Families 5 and 6 where only the G is conserved.

Motif 7, DPPR, is fully conserved only in Family 1; however, Family 4 exhibits the motif DPPK. In all other families this motif is imperfectly conserved. The D is fully conserved in all families except FUPA32, where a single protein, Tel5, has an N replacing the D. In the last position, R and K are found exclusively.

Motif 8, MVTGD, shows good conservation. In Families 1, 2, 4, 7 and 30, the residues are fully conserved; although only in Family 1 do we observe the established motif. The TGD sub-motif is fully conserved in all families except FUPA23 where the fully conserved sub-motif is SGD. Position 2 shows full conservation in most families, but it can be L, I, or V, with decreasing frequencies in this order.

Motif 9, VAVTGDGVNDSPALKKADIGVAM, shows moderate conservation, not surprising in view of the size of this motif. In general, the FUPA families, with the possible exception of FUPA32, show motif conservation comparable to that of the other families, suggesting functionality.

1.G. Hydropathy Profiles

As noted above, all families identified in cyanobacteria belong to topological Type I, II or III, based on hydropathy plots generated with both the WHAT and AveHas programs. Families 5, 6 and 32 exhibit the Type I topology; Families 1, 2, 4, 23, and 30 exhibit the typical Type II topology, and Family 7 exhibits the Type III topology.

2. P-type ATPases in Miscellaneous Bacterial Phyla

2.A. Organisms and P-type ATPases Included in This Study

Six organismal phyla including fourteen organisms with fully sequenced genomes were examined for the presence of the catalytic subunits of P-type ATPases. A total of 31 such proteins were identified. The different organismal phyla examined were: *Aquificae*, *Chlamydiae*, *Chlorobi, Chloroflexi, Deinococcus-Thermus*, and *Thermotogae*. The organisms and their respective genome sizes, numbers of ORFs and numbers of P-type ATPases are tabulated in Table 5.

One organism, *Aquifex aeolicus* VF5, represents the *Aquificae.* It contains three P-type ATPases. The four Chlamydial species examined were *Chlamydia muridarum* Nigg, *Chlamydophila abortus* S26/3, *Chlamydophila caviae* GPIC, and *Chlamydophila pneumoniae* TW-183. Each of these organisms contains only one P-type ATPase. These organisms also show very similar genome sizes and ORF numbers, in agreement with this meager P-type ATPase representation. Three *Chlorobi* species were examined: *Chlorobium chlorochromatii* CaD3, *Chlorobium tepidum* TLS, and *Pelodictyon luteolum* DSM 273. Each of these organisms contains three P-type ATPases, and they have similar genome sizes and numbers of ORFs (Table 5). The phylum *Chloroflexi* includes *Dehalococcoides sp.* CBDB1, *Dehalococcoides ethenogenes* 195, and *Thermomicrobium roseum* DSM 5159. Both *Dehalococcoides sp.* CBDB1 and *Dehalococcoides ethenogenes* 195 were found to have one P-type ATPase while *Thermomicrobium roseum* DSM 5159 was found to have 4 such proteins. *T.*

roseum DSM 5159 has a genome size of 2 Mbps as opposed to 1.4-1.5 Mbps for the other two organisms. The number of reported ORFs is also higher for *T. roseum* DSM 5159 (1922) than for *Dehalococcoides sp.* CBDB1 and *D. ethenogenes* 195 (1458, 1580, respectively). Representing the *Deinococcus-Thermus* group are *Deinococcus radiodurans* R1 and *Thermus thermophilus* HB27. Both bacteria have four P-type ATPases and the expected numbers of ORFs considering their genome sizes (Table 5). The last phylum examined, *Thermotogae*, is represented by a single organism, *Thermotoga maritima* MSB8, which contains only one P-type ATPase. Thus, significant variation in the numbers of ORFs per unit of genome size was not observed for any of the organisms studied, and no organism proved to contain more than four P-type ATPases.

Table 6 summarizes the P-type ATPases identified in these bacteria. Family 5 (Copper) appears in every group studied with the exception of *Chlamydia*, in which only Family 6 (heavy metal) is represented. The organismal group with the largest number of P-type ATPase family represetation belongs to the phylum *Chloroflexi*. Members of a total of six families, including a single functionally uncharacterized P-type ATPase (Tro5), belonging to family FUPA24, were encoded. The single *Thermotogae* species contains only one P-type ATPase (Tma1) belonging to Family 5.

Two proteins (Tro4 and Dra4) appeared to be truncated, based on initial inspection. They were reconstructed based on the DNA sequences deposited in the NCBI nucleic acid database. The reported sequence of Tro4 (*T. roseum)* was

658 aas long, but upon inspecting its alignment with other Family 4 proteins it was discovered that an expected N-terminal portion of the protein was missing. It was found that the wrong initiation codon had been selected, preventing the initial part of the protein from appearing in the NCBI sequence. After reconstructing the protein, its length was increased to 879 aas. The newly introduced N-terminus showed good alignment with other P-type ATPases. In the case of Dra4 (*D. radiodurans)*, a stop codon in the C-terminal portion of the gene was introduced, probably because of a sequencing error. Upon inspection of the DNA sequence and its potential translation products, the protein could be extended from the original 538 aas to 748 aas. Just as in the case of Tro4, the reconstructed protein aligned well with other proteins of its family (Family 6). Conserved motifs in these proteins were identified in the reconstructed regions, providing evidence that the reconstruction was performed correctly and that these proteins are probably functional. The FUPA24 protein (Tro5) was further examined due to its large size (1607 aas). It was found to contain fused structure, as will be discussed in the section entitled “Type II ATPases” below. Otherwise, there were no anomalous proteins of unexpected sizes or topologies (Table 6).

2.B. Phylogenetic Analyses

The phylogenetic tree showing the 31 proteins studied, along with family assignments, is presented in Fig. 8. The proteins generally cluster as expected according to family and typological type. Type I proteins (Families 5 and 6) cluster at the left side of the tree; the Type II proteins (Families 1, 2, 3, 4 and 24) cluster at the bottom of the tree, and the Type III protein (Dra3; Family 7)

branches to the right of the tree.

2.C. Type I ATPases

As was stated earlier, all but one of the small bacterial groups studied (*Chlamydiae*) contain at least one Copper ATPase. The highest number of Copper ATPases was in the *Chlorobi*, where each of the three organisms studied contained at least one and sometimes two Copper ATPases. From the 16S RNA tree (Fig. 9) it can be seen that *Chlorobium chlorochromatii, Chlorobium tepidum and Pelodictyon luteolum* (*Chlorobi*) are close relatives, and this relationship can be observed in the case of the Plu1, Cch1, and Cte3-ATPases (Family 5). It should be noted that Cch2 is also in Family 5, but it clusters further away from the three above-mentioned orthologues. *Aquifex aeolicus* VF5 (*Aquificae*) has two Family 5 proteins (Aae2 and Aae3). Another close orthologous relationship is observed in the case of *Dehalococcoides sp.* and *Dehalococcoides ethenogenes* (Fig. 9). This relationship within the *Chloroflexi* group is observed for the proteins Dsp1 and Det1. It should also be noted that the third protein from *Chloroflexi* that belongs to Family 5 (Tro1) clusters further away from these two orthologues. Close orthology is seen in the Chlamydial group, where all of the organisms studied appear to be closely related (Fig. 8). The *Deinococcus-Thermus* group includes three Family 5 ATPases, two from *Deinococcus radiodurans* R1 and one from *Thermus thermophilus* HB27.

Family 6 (HM) has representatives in three of the six small groups of bacteria studied. In fact, all of the P-type ATPases present in *Chlamydia* are from Family 6 and are probably orthologous. The *Deinococcus-Thermus* group

contains three such P-type ATPases. The reconstructed protein (Dra4) exhibits the nine conserved motifs, as will be shown below. No other anomalies were observed in the sizes or distributions of the proteins in Family 6. *Chloroflexi* proved to contain two Family 6 paralogues (Tro2 and Tro3) from *Thermomicrobium roseum* DSM 5159.

2.D. Type II ATPases

Only one protein (Tro7) was found phylogenetically to belong to Family 1 (Na^+, K^+)-ATPases; it was identified in *Thermomicrobium roseum* DSM 5159, which belongs to the *Chloroflexi*. Family 2 (Ca^{2+}-ATPases) has representation in four of the six groups of bacteria studied. The largest number is seen in *Chlorobi*, with three such proteins. *Pelodictyon luteolum* DSM 273 has 2 such paralogues (Plu2 and Plu3), and *Chlorobium chlorochromatii* CaD3 contains one (Cch3). An orthologous relationship is suggested for Plu2 and Cch3. *Chloroflexi* have one representative in Family 2, Tro6 in *T. roseum* DSM 5159. *Aquifex aeolicus* VF5 (*Aquificae*) similarly contains a Family 2 P-type ATPase (Aae1) as does the *Deinococcus-Thermus* group, with a single Family 2 protein (Tth1) present in *Thermus thermophilus* HB27.

The rest of the Type II proteins are seen in the *Chlorobi* and *Chloroflexi*. The *Chlorobi* phylum contains one Family 3 (H^+; Mn^{2+}) protein (Cte2) and one Family 4 (Mg^{2+}) protein (Cte1), both from *Chlorobium tepidum* TLS. The only other Family 4 protein is observed in *Chloroflexi*, more specifically in the organism *Thermomicrobium roseum* DSM 5159. In fact, this organism contains all of the Type II families observed for the *Chloroflexi* group.

Members of the FUPA24 family are also of Type II (Fig. 10). Although it is functionally uncharacterized, it is known that within this family, “repeat” ATPase sequences are always present. Such fusions can be seen in the only FUPA24 protein identified in these groups of bacteria, Tro5, from *Thermomicrobium roseum* DSM 5159 (*Chloroflexi*). This protein is 1607 aas in length and probably arose by a fusional event, as appears to be true of other FUPA24 members. The first part of the protein (1-765 aas) is probably inactive based on the fact that conserved motifs that are crucial to P-type ATPase function are absent. The second part of the protein, however, is probably active and displays all well conserved motifs, thus suggesting functionality. These characteristics of Tro5 are characteristic of other FUPA24 proteins. The hydropathy profile of Tro5 is in agreement with these conclusions and reveals a basic Type II topology (Fig. 10).

2.E. Type III ATPases

The only Type III protein (K^+; Dra3) was found in *Deinococcus radiodurans* R1. The size and hydropathy plot of the protein were as expected, and this protein was not examined further.

2.F. Conserved Motifs

Analyses of the nine conserved motifs in the various groups of bacteria considered in this section are presented in Table 7. Examining motifs within the single *Chloroflexi* member of Family 1, and also the corresponding motifs from members of Family 2 from *Chlorobi and Chloroflexi,* we see that all of these

proteins have the consensus motifs for all of the first eight motifs. Even the long hinge motif (Motif 9) shows remarkable conservation among these proteins. This is particularly surprising when it is considered that these motifs were defined on the basis of eukaryotic proteins, primarily the Family 1 and Family 2 proteins found in animals. The Family 2 member from *Deinococcus-Thermus,* the Family 3 enzyme from *Chlorobi,* and the Family 4 ATPase from *Chlorobi* also show excellent conservation of all motifs with only a few exceptions. The exceptions include: (1) motif 8 for the *Chlorobi* and *Chloroflexi* Family 2 proteins where an I is substituted for a V at position 2, (2) motifs 4 and 6 for the *Chlorobi* Family 3 enzyme where motif 4 is PVAL and motif 6 corresponds to the consensus except that a Q (terminal position) replaces the E, (3) in the *Chlorobi* Family 4 protein, motif 4 is PEML, deviating at a single position, and motif 8 corresponds to the consensus, except that an L replaces the V at position 2. Each of these aforementioned proteins shows variation in the long motif 9.

Turning to the Copper and Heavy Metal transporters (Families 5 and 6 respectively), we find much more substantial motif variation. Thus, in all of these enzymes, motifs 1, 2, 4 and 6 through 9 show deviation from the consensus motifs (see Table 7). However, there are different consensus motifs that appear to apply to these two families. In motif 1 (PGD), the terminal D is often replaced by E, while in motif 2 (PAD) we see that the second position is almost always substituted by a more hydrophobic residue, V, L or T. Motif 3 (TGES) is largely conserved, although the terminal S can be replaced by other residues. Motif 4 (PEGL), involved in ion binding, is most frequently PCAL, although in the *Aquifex*

Family 5 proteins the consensus motif is PHAL and in the *Deincococcus-Thermus* homologue, it is PCAM. While motif 5 (DKTGTLT) is almost always fully conserved, motif 6 (KGAPE) shows huge variation in these proteins. Surprisingly, this motif is different for the proteins in each organismal group, and also different for the proteins within each of these two families. The same is true of motif 7 (DPPR), but the differences are more extreme. Only the D at position 1 is fully conserved, and the R at position 4 can only be substituted by K. Motif 8 (MVTGD) only shows variation at position 2 where L, I or V can be present. Motif 9 shows variation as expected, although, the GDG-NDAPAL sub motif is conserved among these enzymes.

A Family 7 ATPase is present only in *Deinococcus.* The first three motifs correspond precisely to the consensus motifs, but all others show variation (see Table 7). Finally, the single FUPA24 protein from *Chloroflexi* exhibits twice the normal size with poor motif conservation in the first half, but good motif conservation in the second half as noted above (Table 7). In fact, motifs 1 through 5 and 7 are perfectly conserved, corresponding to the consensus motifs. Further, each of motifs 6 and 8 show variation only at a single position. Motif 9 is also well conserved.

2.G. Hydropathy Profiles

All families identified in these miscellaneous small groups of bacteria belong to topological Type I, II or III, based on hydropathy plots generated with the WHAT and AveHas programs. Families 5 and 6 exhibit the Type I topology; Families 1, 2, 3, 4 and 24 (c-terminal half) exhibit the Type II topology except that the FUPA24 homologue exhibits the characteristic fusion, and Family 7 exhibits the typical Type III topology.

3. Integration of P-type ATPase Data

3A. Distribution of P-type ATPases in the major kingdoms within the three domains of life

P-type ATPases were assigned to families for each of the major kingdoms examined in this integration. The subtotals, as well as percentages, for each of the domains according to family were tabulated in Table 8 in bold print at the bottom of each category and the grand totals, as well as percentages, for all organisms combined are summarized at the bottom of the table. 529 P-type ATPases, from bacteria, a total of 505 ATPases from eukaryotes and 200 enzymes from archaea were included in our studies. The total number of P-type ATPases analyzed is therefore 1234. Among the families, the two with greatest representation are families 5 with 22.6% and 2 with 20% of the P-type ATPases. With about half as many proteins, second place goes to families 6 with 11.1% and 8 with 12.2%. Families 3 and 7 are in third place with about 5% each. Family 4 is in 4th place with 3.5%. Families 9, 10 and 27 have about 2%, while families 20, 23, 25 and 32 each has about 1%. All others are present in smaller amounts.

Family 1, Na/K-ATPases were identified in all three domains but with different percentages. Eukaryotes have 7.5% of their P-type ATPases derived from Family 1. Archaea have 3.5% and bacteria have 1.1%. The same order is observed for Family 2, with 24.2% for eukaryotes, 20.5% for archaea and 15.9% for bacteria. Family 3 has about 9% for both eukaryotes and archaea, but only 0.6% for bacteria. Family 4 shows the reverse of this pattern with highest representation in bacteria (6.8%), lower in archaea (2.5%) and very low in

eukaryotes (0.4%). The dominant Family 5, surprisingly, is over-represented in archaea with 40.5%. It is also the dominant family in bacteria with 29.5%, but eukaryotes have only 8.3%. Family 6 has equal representation in bacteria and archaea (~18%), but low representation in eukaryotes (1.4%). Family 7, potassium ATPases, are well represented in bacteria (10.8%), and moderately represented in archaea (4.5%), but completely absent in eukaryotes. By contrast, Family 8 is present exclusively in eukaryotes (30%) and is not represented in prokaryotes. Thus, in eukaryotes, Family 8 has the largest representation, with Family 2 in second place, and Family 3 and 5 in third place (slight over 8%). Family 9 ATPases are similarly represented only in eukaryotes with 4.8%. Finally, Family 10 with 5.2% (one per eukaryotic organism examined) is also lacking in prokaryotes.

FUPA families 11 - 22 are similarly found only in eukaryotes and their distributions vary from 0.2% for 6 of these families, corresponding to a single member, all the way up to 2.2% for Family 20 with 11 members.

Families 23-32 are represented only in prokaryotes. Only one of these families, Family 32, is represented in archaea. Within the bacteria, one FUPA family predominates. This is Family 27 with 5.3% of the total P-type ATPases. Other families with good representation in bacteria and archaea are families 23 and 32.

Eukaryotic ATPases have already been analyzed with respect to organismal distribution. It was reported, for example, that while families 2, 5 and 8 are nearly ubiquitous, several others are restricted to certain eukaryotic

kingdoms. For example, Na/K ATPases are lacking in plants and most unicellular eukaryotes, but are present in large numbers in animals and ciliates, with low numbers in fungi. Just the opposite is observed for proton transporting ATPases, which predominate in plants, fungi and unicellular eukaryotes. It is interesting to note that while both of these families are well represented in archaea, their representation in bacteria is poor.

Family 5, Copper ATPases, and Family 6, Heavy Metal ATPases, are well represented in prokaryotes; both bacteria and archaea, to roughly 3 to 15 fold greater relative amounts than in eukaryotes. In fact, both families are lacking in ciliates, and Family 6 is lacking in all eukaryotic kingdoms except plants. By contrast, almost every prokaryotic phylum includes members of these two families. Family 7, lacking in eukaryotes, represents 4.5% of the archaeal ATPases and 10.8% of the bacterial ATPases

In summary, families 1, 3, 8, 9 and 10 predominate in eukaryotes, while families 5, 6 and 7 predominate in prokaryotes. Family 2 is present in all three domains of life with similar probabilities.

The archaea also have over 50% of their ATPases within families 5 and 6. This contrasts with eukaryotes, which have less than 10% of their ATPases in these two families. In almost all organisms, the copper ATPases outnumber the Heavy Metal ATPases, with just a few exceptions. Particularly noteworthy are fungi, unicellular eukaryotes and some of the small bacterial phyla, where Heavy Metal ATPases are lacking. Exceptions include firmicutes and *Chlamydia* where the Heavy Metal ATPases predominate. In fact, in

Chlamydia the only ATPases present are within the Heavy Metal family. It should be noted, however, that families 5 and 6 overlap substantially in specificity.

It is noteworthy that in eukaryotes, type II ATPases predominate over type I enzymes, while in prokaryotes the opposite is true. Correlating with this observation is the interesting fact that all eukaryotic FUPA families are derived from type II, while most prokaryotic FUPA families are of or are derived from type I. This observation may explain why the fact that the FUPA families of eukaryotes are exclusively of type II (sometimes with N-terminal modifications), while the FUPA families of prokaryotes are usually of type I.

3B. Size Comparisons Among P-type ATPases

ATPases were analyzed according to size, both by family and by organismal type (Table 9). For bacteria, the sizes of Family 1 ATPases are the largest (946 aas), and then in order of size follow Family 2 (892 aas), Family 3 (866 aas), Family 4 (863 aas), Family 5 (765 aas), Family 6 (695 aas) and Family 7 (684 aas). The same size order is observed for the archaea except that Family 4 ATPases are substantially larger than the Family 3 ATPases (by 51 residues). It is interesting to note, that for all type II ATPases the bacterial enzymes are larger than the archaeal enzymes, but that for types I and III ATPases the archaeal proteins are larger. Examining the same families in eukaryotes, the same pattern is observed except that Family 5 ATPases are the largest among families 1-6. This is due to the increased number of N-terminal regulatory Heavy Metal binding domains (up to seven per enzyme). Among the remaining eukaryotic-specific

families there is substantial size variation, but all of these ATPases have sizes in excess of a 1000 residues. The largest of these families are families 12 (1998 aas), 19 (1807 aas), 22 (1541 aas), 18 (1491 aas), 14 (1390 aas), 16 (1388 aas), 21 (1372 aas), 8 (1304 aas), 15 (1292 aas), 10 (1249 aas), 13 (1212 aas), 20 (1187 aas), 11 (1146 aas), 17 (1096 aas) and Family 9 (1074 aas). Thus, the total range for all eukaryotic FUPA families is 1074-1998 aas, while the range for all eukaryotic ATPases is 942 (Family 4) through 1998 (Family 12). It should be noted that size variation within these families is sometimes very substantial (as great as 50% of the average size, as for Family 10).

Comparing the eukaryotic protein sizes with the bacterial protein sizes (families 1-6), we see that in every case, the eukaryotic proteins are larger. However, the variation is substantial. The percent increases are as follows: Family 1 – 20%, Family 2 – 25%, Family 3 – 10%, Family 4 – 9%, Family 5 – 50% and Family 6 – 35%.

The prokaryotic FUPA families, in general, show less size variation within families and also between families, with the sole exception of Family 24. As noted above, this family represents a fusion between a type II and a type I ATPase. Arranging these families according to size, we see that they are ranked in the order: Family 24 (1468 aas), 26 (898 aas), 31 (860 aas), 28 (849 aas), 30 (846aas), 23 (815 aas), 27 (800 aas), 29 (795 aas), 32 (715aas) and Family 25 (666 aas). Thus, excluding Family 24 with a fusion of two ATPases, the largest proteins are in Family 26, while the smallest proteins are in Family 25.

3C. Motif Analyses

Motif 1 (PGD) The nine conserved motifs were analyzed both by family and by organismal types (see tables 10 and 11). In Family 1, motif 1 (PGD), only position 1 shows variation (Table 10). Examination of Table 29 reveals that only in proteobacteria and spirochetes is the first residue a P. In the cyanobacteria, this residue is an R. In all other organismal types (all archaea and eukaryotes) the preferred residue is a V. In Family 2 only the final D is fully conserved. At position 1, the consensus sequence for each group of organisms can only be a P or a V, while at position 2, although not fully conserved, the consensus sequence is always a G. In families 3 and 4 the consensus motif is always PGD, but in Family 3 none of the residues is fully conserved, although the last two residues in Family 4 are conserved. In families 5 and 6 no residue is fully conserved, but the consensus sequence for Family 5 is PGE, while that for Family 6 is PGD. In the consensus sequences of these two families, the third residue can be either D or E with only one exception in each family. The G at position 2 is the consensus residue in every group of organisms. However, at position 1, substantial variation is observed, where six different residues appear in consensus sequences. In Family 7 no residue is fully conserved, and again, substantial variation occurs in position 1. In Family 8, the dominant motif is VGD except for ciliates where this motif is VGH. Family 9 has this motif, PGD, fully conserved.

In most of the eukaryotic FUPA families motif 1 is PGD, sometimes fully conserved, but instead of a P at position 1, an I occurs in FUPA 17, and an N occurs at position 3 in FUPA 11. In the prokaryotic FUPA families the dominant

motif is also PGD, although variations occur. However, these variations are seldom conserved among the family members.

Motif 2 (PAD) Motif 2 is somewhat better conserved than motif 1. A D at position 3 is present in all consensus sequences within all families. The A at position 2 can be replaced by a V (families 5, 17, 21, 23, 25 and 26) or a C (families 10-16, 18-20 and 22). A Q in Family 13 and a V in Family 23 replace the P at position 1.

Motif 3 (TGES) Motif 3 is well conserved between families with relatively little variation. The G at position 2 is fully conserved among all families. The greatest variation occurs in residue 4, where the S can be substituted with T, A, R or P. The T at position 1 is found in the consensus sequence for all families except Family 8 where a D replaces the T. The E is also well conserved in almost all consensus sequences, except in Family 26 where a fully conserved H exists and in Family 28 where a poorly conserved A is found.

Motif 4 (PEGL) This motif is relatively poorly conserved among most families. However, the P at position 1 is the best conserved of the four residues, although an N occurs at that position in Family 11, and an S occurs in the consensus sequence for Family 32. At position 2, a C is found in families 5, 6, 25, 27-29 and 32. This residue in families 5 and 6 homologues is believed to be involved in Heavy Metal binding (Wu et al, 2006). Therefore, in FUPA families possessing a C at this position one might suggest that Heavy Metals are the substrates. Also, variations in this position may be dictated by the ions transported. Since position 2 appears to be important for substrate recognition, it

is of interest to note that in families 1, 2, 4, 9, 23, 24 and 30, an E is present, consistent with the known cation specificities for the families of known function (1, 2, 4 and 9). However, it should be noted that neither the C nor the E is particularly well conserved in most of these families.

The best conservation for motif 4 is observed in families 12-18 where the fully conserved motif PPAL is observed. This motif is found in no other family. These families are all of the type V topology, which includes families 11-21. It should be noted that families 13-16 comprise a sub-super-family present in animals, fungi, slime molds and ciliates, respectively. The remaining families having this motif (families 12, 17 and 18) are represented by single proteins, each from a unicellular eukaryote, one of them being a fungus. This common motif suggests that all of these ATPases may serve a common function as suggested previously for families 13-16 (Thever and Saier, 2009).

Motif 5 (DKTGTLT) Motif 5 is by far the best-conserved motif within the P-type ATPases. It includes the aspartyl residue at position 1 that is the site of phosphorylation. Surprisingly, it is not fully conserved in all families. For example, in both families 1 and 2, although the D is present in the consensus sequences, it is not found in all members of these two families. In Family 1 the D is fully conserved in all organisms except the ciliates, while in Family 2 this aspartyl residue is not fully conserved in the cyanobacteria. The few exceptions where D is not observed may be non-functional pseudogenes. However, another exception is that of Family 26 where a fully conserved asparagine residue (N) replaces the D.

In most of the families, variation within motif 5 occurs primarily at position 6 where other hydrophobic residues can replace the L. However, more extensive divergence occurs in a few families (families 1, 2, 5, 24, 26, 27 and 31). In a few of these families, it appears possible that this lack of conservation is due to the presence of pseudogenes. This seems likely in families 1, 5, 26 and 31. The poor conservation of several residues in families 1 and 31 supports this conclusion. For example, in Family 1 most proteins exhibit well conserved motifs, but in two proteins from the ciliate *Tetrahymena thermophila,* we observed very poor conservation of all motifs leading to the conclusion that these two genes are pseudogenes. Similarly, in Family 31 one of the three proteins in this family exhibits very poor motif conservation. This protein from *Methylococcus capsulatus* (GI-53804058; 1068 aas) proved, not only to lack the conserved motifs, but also to possess a 250 residue C-terminal domain that is homologous to many bacterial proteins, primarily from firmicutes, but also from other bacterial kingdoms. The function of this domain is unknown. We tentatively propose that the inactive ATPase serves as an anchor for the C-terminal catalytic domain.

Motif 6 (KGAPE) Motif 6 is the second least well-conserved motif among the nine analyzed. Only eight of the 32 families exhibit this motif in the consensus sequence, and only one family with more than one member, Family 14, has KGAPE fully conserved. However, families 13 and 15 have fully conserved KGSPE, and families 16, 18 and 20 show this motif as the consensus sequence. The best-conserved residues in motif 6 are the first two positions (KG).

Motif 7 (DPPR) Motif 7 is even less well conserved than motif 6. Only in

one family, Family 9, is DPPR fully conserved. However, in four other families a different motif is fully conserved. These four families are: families 14 and 16 where NKLK is fully conserved, Family 26 where the fully conserved motif is DRGV, and Family 28 where the conserved motif is DPLR. In fact, only in four families does DPPR occur as the consensus sequence. These families are 1-3 and 9. The most conserved residue in motif 7 is at position 4, where all families but two have either an R or a K. The D at position 1 is seldom replaced by another residue, except that an N appears in 11 of the families. Fourteen families have a proline at position 2, and this residue is particularly common in the functionally characterized families. An R or a K occurs as the consensus residue at position 2 in 10 families. A P occurs in position three for only 7 of the families, most notably in families 1-4 and 9. An aliphatic hydrophobic residue, L, V or I appears fifteen times at this position.

Motif 8 (MVTGD) Motif 8 consists of two hydrophobic residues at positions 1 and 2, followed by the well conserved motif (T/S)GD. Only one family, Family 15, exhibits the MVTGD motif fully conserved. However, Family 26 has MLSRD fully conserved at this position. The results show that while M is the primary residue at position 1 (21 families), aliphatic hydrophobic residues can replace it. In position 2, the L, V or I can be replaced by semi-polar residues such as C, S or A.

Motif 9 (Hinge Motif) In almost all families, the sub-motif GDG-ND is conserved. In fact, these five residues are the most conserved residues within this motif.

3D. Topological Types: Eukaryotes and Prokaryotes

In eukaryotes, topological Type II predominates. The standard Type II topology is present in families 1-4, 8 and 9. The standard Type I topology are found only in families 5 and 6. Novel eukaryotic topologies (Types IV, V and VI) are all based on the Type II topology. It is interesting to note that in eukaryotes Type II ATPases predominate over Type I ATPases, and all of the FUPA families found in eukaryotes, including the three novel topological types described below, are all based on the Type II topology (Thever and Saier, 2009). By contrast, in prokaryotes the Type I topological type predominates over Type II ATPases, and all but two of the prokaryotic FUPA families (families 24 and 30) exhibit, or are based on, the Type I topology.

Table 11 summarizes the topological types of the prokaryotic FUPA families as well as their organismal distributions. Each of these FUPA families exhibits a unique distribution. Family 23 derives from actinobacteria, cyanobacteria and firmicutes, while Family 24 derives from actinobacteria, chlorobi and gamma- and delta-proteobacteria. Interestingly, only in the chlorobi protein was the family 24 N-terminal half easily recognizable as Type I. In both the actinobacteria and the proteobacteria, sequence divergence was so great in this half of the protein that its origin was difficult to ascertain.

Family 25 ATPases are derived from six different bacterial kingdoms including both Gram-positive and Gram-negative phyla (Table 11). Family 26 was identified only in actinobacteria, three orders of proteobacteria and the spirochaetes. Families 27, 28 and 31 are found only in proteobacteria, 27 being

present in both beta and gamma orders, while families 28 and 31 were identified only in gamma-proteobacteria. Family 29 was also limited, being present only in delta-proteobacteria and flavobacteria. Family 30, in addition to being present in alpha-, beta- and delta-proteobacteria, was represented in *bacteroides*, cyanobacteria and *spirochaetes*. Finally, Family 32 is the most widely distributed family having members in all five orders of proteobacteria, six additional bacterial phyla and euryarchaeota. It is worth noting that of the novel FUPA families, only Family 32 was identified in archaea.

Three new predicted topological types were identified in eukaryotic ATPases. All of them are based on the type II topology with variations at the N-termini. These three new types (types IV-VI) differ in the numbers of TMSs preceding TMS1. Type IV has 12 transmembrane segments and is found in Family 10. The extra 2 TMSs are designated A' and B' (Thever and Saier, 2009). Topological Type V, observed for families 11-21, is predicted to have 11 transmembrane segments, the usual 10 observed for Type II ATPases plus an extra N-terminal TMS, which we call TMS C (Thever and Saier, 2009). The presence of such a TMS, if correct, would bring the N-termini of these proteins to the external surface of the membrane. Topological Type VI is predicted to have 13 TMSs. As for types IV and V, the extra TMSs are N-terminal. These three TMSs are designated D, E and F. As for type V, this putative topology implies that the N-terminus is localized externally.

In prokaryotes we discovered only one new topological type (Type VII) that was found in Family 24. Additionally, the topology of Family 25 was not

determined due to the fact that different members exhibited different characteristics.

Parts of the Results section are being prepared for publication. The thesis author will be a co-investigator and co-author of this paper. Co-authors include: Milton H. Saier, Henry Chan, Charmy Gandhi, Kunal Hak, Danielle Harake, Kris Kumar, Perry Lee, Tze T. Li, Hao Yi Liu, Tony Chung Tung Lo, Cynthia J. Meyer, Steven Stanford and Krista S. Zamora.

Discussion

P-type ATPases were identified in nine different genuses of Cyanobacteria. These include: *Acaryochloris, Anabaena, Gloeobacter, Nostoc, Prochlorococcus, Synechococcus, Synechocystis, Thermosynechococcus and Trichodesmium.* Of these the highest number of the ATPases was found in *Anabaena variabilis* with 17 members and the second highest was found in *Nostoc sp.* with 12 members. Analyzing the distribution of the P-type ATPase proteins in these eight genuses showed that the variation of distribution of these enzymes within members of this phylum was tremendous. The genome size of the members did not prove to be indicative of the number of ATPases to be found within the genuses. Thus, for example, *Acaryochloris marina* MBIC11017 with the largest genome has only four P-type ATPases, while *Anabaena variabilis* ATCC 29413, with a genome size 15% smaller, has 17 such enzymes. Even more strikingly, *Synechocystis* sp. PCC 6803, with a genome size less than half that of *A. marina,* has nine such enzymes; more than twice that observed for *A. marina*. The families of functionally characterized P-type ATPases found within these cyanobacteria include family 1, 2 and 4-7 and 3 families that are functionally uncharacterized (FUPA). Along with the Cyanobacteria P-type ATPases were also identified in six miscellaneous phyla of bacteria. These phyla include: Aquificae, Chlorobi, Chloroflexi, Chlamydia, Deinococcus-Thermus and Thermotogae. The most number of ATPases was found in the pylum Chlorobi, with a total of 9 ATPases. Just as in the Cyanobacteria there was no indication that genome size was indicative of the number of P-type ATPases identified.

The Family 5, or Copper, ATPases are present in each one of the cyanobacterial genuses. Chlamydia, which is comprised solely of Family 6 ATPases, is the only miscellaneous phylum that does not contain these Family 5 proteins. This is interesting to note due to the importance of these proteins. Mutations in two such enzymes, the human copper-transporting P-type ATPases, ATP7A and ATP7B, lead to Menkes disease and Wilson disease, respectively (Barnes et al, 2005). Both ATP7A (TC# 3.A.3.5.6) and ATP7B (TC# 3.A.3.5.3) regulate intracellular copper concentrations and participate in the biosynthesis of copper-dependent secretory enzymes; both transport copper from the cytosol into the endomembrane system, ATP7B in the liver and ATP7A in other cells. ATP7A directs copper within the transgolgi network to the enzymes, dopamine beta-monooxgenase, peptidylglycine alpha-amidating monooxygenase, lysyl oxidase, and tyrosinase, depending on the cell type (Prohaska et al, 2004). ATP7B delivers copper to a plasma ferroxidase, ceruloplasmin, in the liver (Barnes et al, 2005). Inactivation of ATP7A or ATP7B leads to copper accumulation in the cytosol. Other Copper ATPases can function in either uptake or efflux, and with Cu^{+} and/or Cu^{2+} as the primary transported metal ion(s), although some catalyze transport of other monovalent and divalent cations as well (Odermatt et al, 1993; Solioz et al, 2003). Their functions, regulation and biogenesis have been studied extensively (Lutsenko et al, 2007; Veldhuis et al, 2009).

The copper and heavy metal ATPases contain metal binding domains. An example of the presence and number of these domains is seen in a protein from

Anabaena (Ava4) and *Nostoc* (Nsp2). These domains are also found in Mercuric (Hg^{2+}) reductases, extracytoplasmic Hg^{2+} binding proteins and heavy metal chaperone proteins (Yamaguchi et al, 2007). While Ava4 has just one such repeat, Nsp2 has two. Moreover, a HM homologue from *Desulfitobacterium hafniense* DCB-2 (gi 219669744), a firmicute of the clostridial group, proved to have four such repeats, and a Copper ATPase from the animal, *Branchiostoma floridae* (gi 219406998), proved to have seven such repeats. Thus, the difference in size between Ava4 and Nsp2, and the size variations observed for the HM and Copper ATPases, appears to be largely due to the variable numbers of N-terminal heavy metal binding domains. Within all other families of ATPases, size variation was minimal. This size variation was also observed in the miscellaneous phylas where the Copper ATPases identified showed varying Family 2 and Family 6 ATPase sizes.

Calcium ATPases are present in all of the Cyanobacteria studied besides for *Prochlorococcus*. These calcium ATPases are also present in all but 2 of the miscellaneous phylas examined (Chlamydia and Thermatogae). Ca^{2+}-ATPases of Family 2 may catalyze Ca^{2+}/K^{+} or Ca^{2+}/H^{+} antiport. A single organism often possesses multiple isoforms of these enzymes. Olesen et al. (Olsen et al, 2007) have described functional studies and three crystal structures of the rabbit skeletal muscle Ca^{2+}-ATPase. These represent the phosphoenzyme intermediates associated with (1) Ca^{2+} binding, (2) Ca^{2+} translocation and (3) dephosphorylation. They are based on complexes with a functional ATP analogue, beryllium fluoride or aluminum fluoride. The structures complete the

cycle of nucleotide binding and cation transport of the Ca^{2+}-ATPase. Phosphorylation of the enzyme triggers a conformational change that leads to opening of a luminal exit pathway defined by the transmembrane segments, TMS 1 through TMS 6. TMSs 1-6 represent the canonical membrane domain of Type II P-type pumps. Ca^{2+} release is promoted by translocation of the TMS 4 helix, exposing Glu 309, Glu 771 and Asn 796 to the lumen. The mechanism explains how P-type ATPases are able to form the steep electrochemical gradients required for key functions in eukaryotic cells (Olsen et al, 2007).

The Calcium family is shown to be composed to numerous subfamilies. No Cyanobacterial protein showed resemblance to the Golgi or Sarco/Endoplasmic type of Calcium ATPases. The protein Ter4 from *Trichodesmium erythraeum* is the only Calcium ATPase that was part of a previously identified group of plasma membrane eukaryotic Calcium ATPases. It can be seen that Ter4 branches further from the rest of the Family 2 proteins and this can be explained by the fact that the PM type ATPases compose a separate group of Ca2+ ATPases. These enzymes eject Ca2+ out of the cell are documented in all animal and plant cells (Carafoli, 1994). They are known to share catalytic properties of other ion-motive P-type ATPases, but still retains its unique regulatory properties (Carafoli, 1994). This protein, Ter4, was found to contain two extra TMS segments. These two segments were also identified in other putative Calcium ATPases in cyanobacteria. From this observation, we suggest that the insertion that gave rise to these two extra putative TMSs between TMSs 3 and 4 arose during the early evolution of the cyanobacterial

phylum. Judging from the 16S RNA tree of the Cyanobacteria including the organisms that contain this interesting insertion it can be seen that they all group in the upper portion of the tree. The proteins in the bottom half of the tree do not contain this insertion. Thus, we suggest that the insertion occurred before segregation of the organisms represented in the top half of the tree, but after they diverged from the organisms in the bottom half of the tree. If so, then these proteins were lost from the remaining four organisms represented in the top half of the tree that do not contain a protein with the 2 TMS insertion.

From Table 3, it can be seen that all of the eukaryotic proteins in subfamily Ca9 (PM) are substantially larger than the prokaryotic members of this subfamily. This is in accordance with expectation since eukaryotic homologues of prokaryotic transport proteins are on the average 40% larger than those from bacteria (Chung et al, 2001). In this case, the average size of the eukaryotic homologues was 21% larger than the average size of the bacterial homologues.

Orthologous relationships are observed for the *Anabaena* and *Nostoc* proteins. This can be seen in the case of Ava9 and Nsp10 or Ava10 and Nsp11. In some cases this orthologous relationship is not observed as certain proteins must have lost their orthologues through time. The 16S RNA tree also reveals this close relationship of *Anabaena* and *Nostoc*. There are two observable paralogues in Family 32 (Tel2 and Tel5), but they are too distant to suggest that gene duplication occurred to yield these proteins. The fact that various functionally uncharacterized and characterized families contain proteins that do not exhibit these orthologous and paralogous relationships, points to the

possibility that horizontal gene transfer (HGT) occurred. Recent studies by Martinez et al**,** (2006) reported detectable horizontal gene transfer of genes encoding bacterial P-type ATPases in various bacteria isolated from a deep subsurface environment that is free of heavy-metal contamination, which points to the possibility of underestimating the extent of HGT among closely related bacterial lineages (Anderson et al, 2003; Coombs et al, 2004).

All of the proteins except for Ava7 and Tel5 found in this study are expected to be functional based on the motif analysis. Since these motifs are located in the region of the proteins that is necessary for function, their conservation is a considerable sign of their functionality. The first three of these motifs are located in the small cytosolic Region B loop between TMS 2 and TMS 3 (Fig. 2). The residues within this loop, particularly those in Motif 3, may enhance the stability of this region, thereby providing more favorable reaction kinetics for the transition between the E1 and E2 states. Motif 3, TGES, is conserved in all of the cyanobacterial and miscellaneous protein families. The importance of this motif is observed from crystal structures (Felsenstein et al, 1997) that have shown that the conserved TGES loop of the Ca^{2+}-ATPase, isolated in the Ca_2E1 state, becomes inserted in the catalytic site in the *E*2 states. Motif 4, PEGL, shows poor conservation but this is expected since this motif serves as part of the ion binding site and therefore determines substrate specificity (Thever et al, 2009). In two Cyanobacterial cases (Families 5 and 6) the residues PCAL, are observed. The cysteine residue plays a role in heavy metal ion binding (Liu et al, 2006). In contrast to Family 5, the A is not fully

conserved in Family 6 proteins. Motif 5, DKTGTLT, contains the transiently phosphorylated aspartyl (D) residue and exhibits nearly full conservation in most P-type ATPases (Okkeri et al, 2002). The remaining residues of motif 5 may play roles in catalysis and help maintain the structure of this region of the C-domain. Conservation of the motif is good except for the substitution of the D for a G in Cyanobacterial protein Ava7. If this residue truly replaces the D, and this residue assignment is not due to a sequencing error, this enzyme must be inactive due to the importance of the D residue for phosphorylation. Resequencing of this region is advised.

In the case of another important motif, DPPR, the D is conserved fully except for one protein in Cyanobacterial FUPA32 protein Tel5, where the D is replaced by an N. Due to the importance of this residue to the phosphorylation activity of the enzyme the fact that there is substitution points to the non-functionality of Tel5. Motif 9, VAVTGDGVNDSPALKKADIGVAM, forms part of a flexible hinge region that joins Domain C with the C-terminal transmembrane domain (Moller et al, 1996). This “hinge” region-J” helps provide the flexibility needed for conformational changes that occur during ion translocation (Moller et al, 1996). This motif shows good conservation in the areas that are documented to be submotifs. The GDG and ND submotifs are fully conserved in all but one family (Family 32), in which the first D is substituted with an E in one protein, Tel5, pointing again to the question of its functionality. These submotifs are characteristic of all P-type ATPases of eukaryotes (Thever et al, 2009) and prokaryotes (this study). The remaining motifs show varying levels of

conservation. In the miscellaneous phyla a great level of conservation is observed for the important motifs 3-5 and 9, thus, all the proteins studied are thought to be functional.

The only functionally uncharacterized family found in the miscellaneous group of bacteria is Family 24. It was identified in *Thermomicrobium roseum (*Tro5) from the phylum Chloroflexi. This protein is about twice as long as the average bacterial ATPase alpha subunit (1607aas). Upon further investigation it was discerned that this protein resulted possibly by a fusion of a topological type I and II ATPases. This type of ATPase was also found in Actinobacteria. The C-terminal half of these proteins exhibit Type II topology, and display good motif conservation. The N-terminal half, on the other hand, shows no motif conservation, and seems to resemble a topological Type I protein based on the hydropathy plot. This N-terminal half thus seems not needed for the functioning of this enzyme. In Cyanobacteria, the functionally uncharacterized families are FUPA 23, 30 and 32. The topological types for these were Type II for 23 and 30 and Type I for 32. Using single amino acid mutations in the ion binding sites of Na+ and K+ P-type ATPases, Gadsby et al (2006) have found four amino acids that are crucial for determining ion specificities of these ATPases. These four important amino acids were found to be E321, G337, G805 and T806 (Gadsby et al, 2006). Examining these specific amino acids in the proteins from all the families examined showed that these amino acids were mostly neutral. Thus, it would be impossible to predict whether the FUPA families transport a cation, an anion, or even more specifically, what unique type of ion.

The integrated P-type ATPase data proved to be very revealing. A total of 1234 P-type ATPases were analyzed from bacteria, archaea and eukaryotes. Just as expected, the highest representation was observed in families 5 and 2. Certain of families (8, 9, 11-22) were found only in eukaryotes, yet families 23-32 are represented only in prokaryotes. Due to the observed differences in the representation of eukaryotic Na/K ATPases and proton transporting ATPases, we propose that these two types of enzymes perform comparable functions in maintaining monovalent cation concentration differences (sodium vs. protons) in the different organism that contain the proteins.

The Family 7 ATPases are not found in Eukaryotic organisms but show good representation in the prokaryotes. Interestingly, experimental evidence suggests that these ATPases function primarily as scavengers when external potassium concentrations are exceptionally low. They may also be important in osmotic stress adaptation. It should also be noted that Family 2 ATPases are represented in all three domains of life with similar probabilities, as mentioned above, pointing to their importance in the proper functioning of cells.

As noted above, eukaryotic P-type ATPase distribution is kingdom specific. The same is true in prokaryotes. For example, calcium ATPases are prevalent in cyanobacteria, firmicutes and euryarchaeota, but with appreciable representation in proteobacteria, bacteroides and actinobacteria. Interestingly, all of the organisms with high calcium ATPase representation are free-living organisms. Obligate pathogens or symbionts tend to lack these enzymes. The same is true for copper and Heavy Metal ATPases that are represented in high numbers in

free-living organisms, while pathogents and symbionts appear to have far fewer. For example, small phyla such as *Aquificae*, *Thermatogae*, *Chloroflexi*, *Chlamydia* and *Deinococcus-Thermus* possess between 50 to 100% of their ATPases in these two families; actinobacteria, *Bacteroides*, *Chlorobi* and fusobacteria have almost 50% of their total ATPases occurring within these two families.

The integration also shed more light on the existing idea that eukaryotic ATPases are larger than the prokaryotic ones. The results were in agreement with a previous study showing that on the average, eukaryotic proteins are 40% larger than bacterial proteins (Chung et al, 2001). As noted above, this increase in size was in part due to the increased numbers of heavy metal binding doimains observed, for example, in family 5 and 6 proteins.

Comparing the integrated motif analysis we can see interesting results. In the case of Family 26, a fully conserved asparagines residue (N) replaces the D in the crucial motif DKTGTLT. These enzymes may either be inactive, or they may be subject to deamidation. Spontaneous deamidation in proteins has been observed (Kosky et al, 2009) but deamidases are also known (Vinogradov et al, 1976). The large variation in conservation of the motifs across the major kingdoms of life is expected due to the evolutionary divergence that contributes to unique characteristics of each organism in our biosphere.

Topological analysis of the P-type ATPases across eukaryotes and prokaryotes was performed and used to check for the presence of unique topological types. In the past, in isolated studies of this sort, errors were made in

assigning a family of P-type ATPases to a novel topological type, due to the lack of integrated data that can provide clear evidence of novelty. In prokaryotes the only new topological type (Type VII) was that of Family 24. These fusion proteins are indeed unique and are represented in numerous prokaryotes. However, the first half of these proteins exhibits none of the conserved motifs and exhibits a topology that is very difficult to interpret. This first half has diverged in sequence to such an extent that it can no longer be an active P-type ATPase. We presume that this domain has evolved a distinctive function such as, protein-protein interaction, or as an anchor for macromolecular assembly (Charbit *et al*, 1996).

Family 25, which was previously thought to be a novel topological type, but the integrated analysis proved that it would be difficult to assign this family to a uniform topological type. The AveHas plot with all family members included appeared similar to the Type I topology. However, deviation from this topology appears possible, and the different members of the family may exhibit somewhat different topologies. Further work would have to be performed to thoroughly analyze this family of ATPases to ascertain which existing topological type it belongs to, or if it is in-fact a unique type.

In eukaryotes, three novel topological types were found, as mentioned above (types IV – VI). Thever and Saier (2009) have previously described these new types and our integrated data proves to show that indeed these are novel types of ATPases, which exhibit altogether new topologies. The fact that all the other families fit into existing categories further solidified existing knowledge and research. The grouping of these proteins into families, subfamilies and

topological types is crucial to our understanding of their function and structure.

Parts of the Discussion section are being prepared for publication. The thesis author will be a co-investigator and co-author of this paper. Co-authors include: Milton H. Saier, Henry Chan, Charmy Gandhi, Kunal Hak, Danielle Harake, Kris Kumar, Perry Lee, Tze T. Li, Hao Yi Liu, Tony Chung Tung Lo, Cynthia J. Meyer, Steven Stanford and Krista S. Zamora.

Appendix

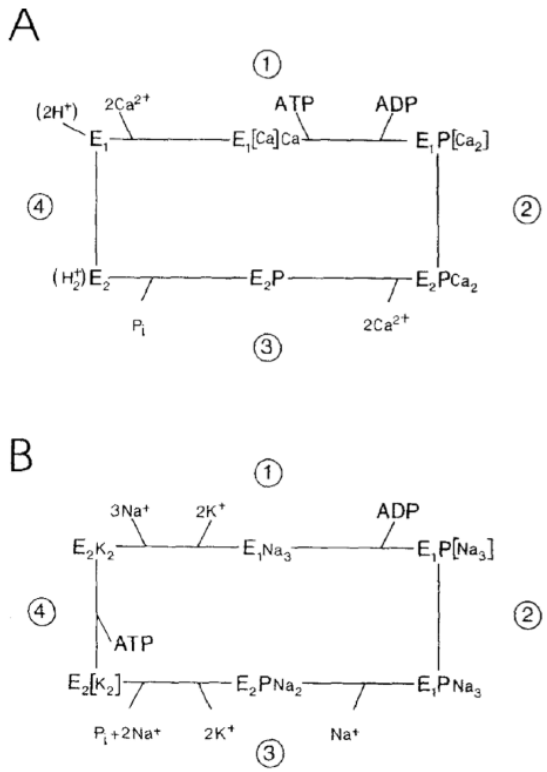

Figure 1: Kinetic schemes for (**A**) Ca^{2+} transport by the SERCA ATPase and (B) Na^+,K^+-exchange by the Na^+,K^+-ATPase. The schemes are based on the assumption of two major conformational states, E_1 and E_2, which are characterized by their ability to become phosphorylated by ATP and P_i, respectively. For the SERCA ATPase (A), Ca^{2+}-dependent phosphorylation of the enzyme results at step (1) in the formation of $E_1P[Ca_2]$, with occluded Ca^{2+}, from E_I[Ca]Ca, with half-occluded Ca^{2+}. After conversion of $E_1P[Ca_2]$ to E_2PCa_2, Ca^{2+} is de-occluded and quickly released towards the extracytosolic side in step (2). This is followed by dephosphorylation to E_2 in step (3), and re-formation of E_1 in step (4), with or without transfer of protons (1 or 2) to the cytosol. For the Na^+,K^+-ATPase (**B**), $E_1P[Na_3]$, with occluded Na^+, is formed in step (1) by Na^+-dependent phosphorylation from ATP. Translocation in step (2) gives rise to the formation of E_1PNa_3, an intermediate that is in rapid equilibrium with the previous intermediate and which therefore can be rapidly dephosphorylated by ADP. Dephosphorylation in step (3) begins by formation of E_2P, which is characterized by the binding of 2 cations ($2K^+$ or $2Na^+$), instead of the $3Na^+$ that are bound in the E_1 state. Step (3) can be considered as consisting of three steps: (i) release of one Na^+ to form $E_2PNa^+_2$, which then (ii) exchanges with 2 K^+ and (iii) dephosphorylates. Concomitant with dephosphorylation, the bound 2 K^+ are occluded, and in step (4), the occluded K^+ ions transported and released on the cytosolic side in a process that is accelerated by ATP. The cycle is then complete and can repeat (Møller et al, 1996).

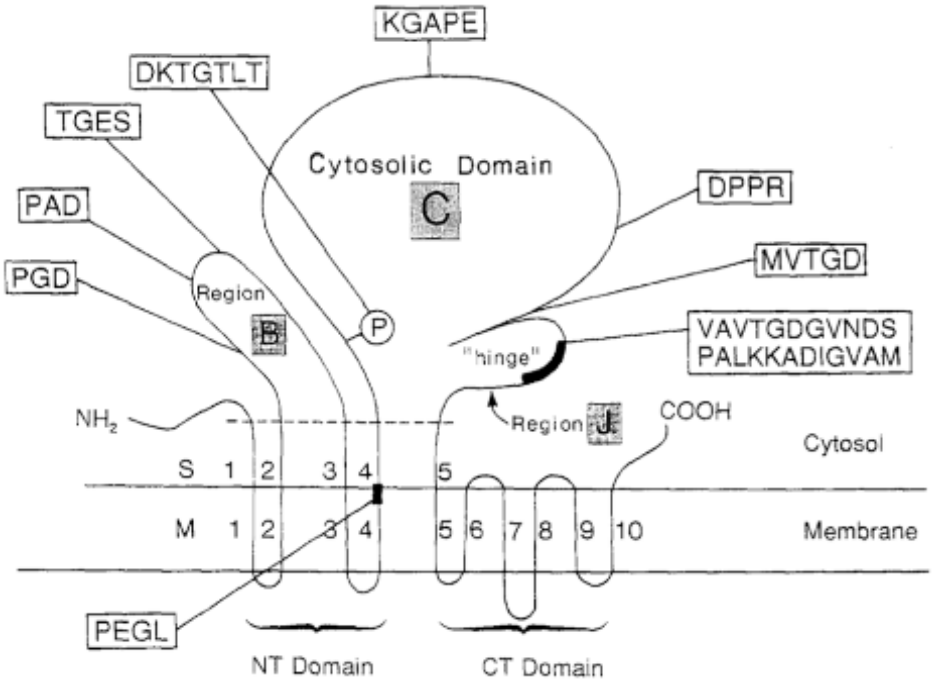

Figure 2: Characteristic sequence motifs of Type II ATPases, represented by those present in the Na^+,K^+ -ATPase. Most of these motifs are also seen in Type I and Type III ATPases, but sometimes in modified forms. Conservative substitutions include the last residue (Ser) in TGES which is replaced by Asn in *S. cerevisiae* and by Ala in the putative Ca^{2+}-ATPase of the cyanobacterium *Synechococcus* PCC 7942; the glutamate (E) residue in the PEGL motif which is replaced by Val or Ile in plant and fungal proton ATPase and by Cys or His in Type I ATPases forming part of a putative heavy metal binding CPC or CPH motif; the leucine (L) in the phosphorylation motif (DKTGTLT) which is replaced by Isoleucine (I) in some Type I ATPases. In the 'hinge' motif, few conservative mutations occur, except for the two lysine (K) residues which in some ATPase types are replaced by one *(S. cerevisiae* Ca^{2+}-ATPase and bacterial Cu^+-transporting ATPases) or atwo (bacterial Cd^{2+}-ATPases and Cu^{2+}-ATPases of man) aliphatic (A, L or M) amino acyl residues. In the PGD and PAD motif of Region *B*, only the glycine and aspartate residue, respectively, are absolutely conserved in Type I ATPases [10]. Furthermore, the KGAP and DPPR motifs, characteristic of Type II ATPases, are not always present in readily recognizable forms in Type I ATPases (Møller et al, 1996).

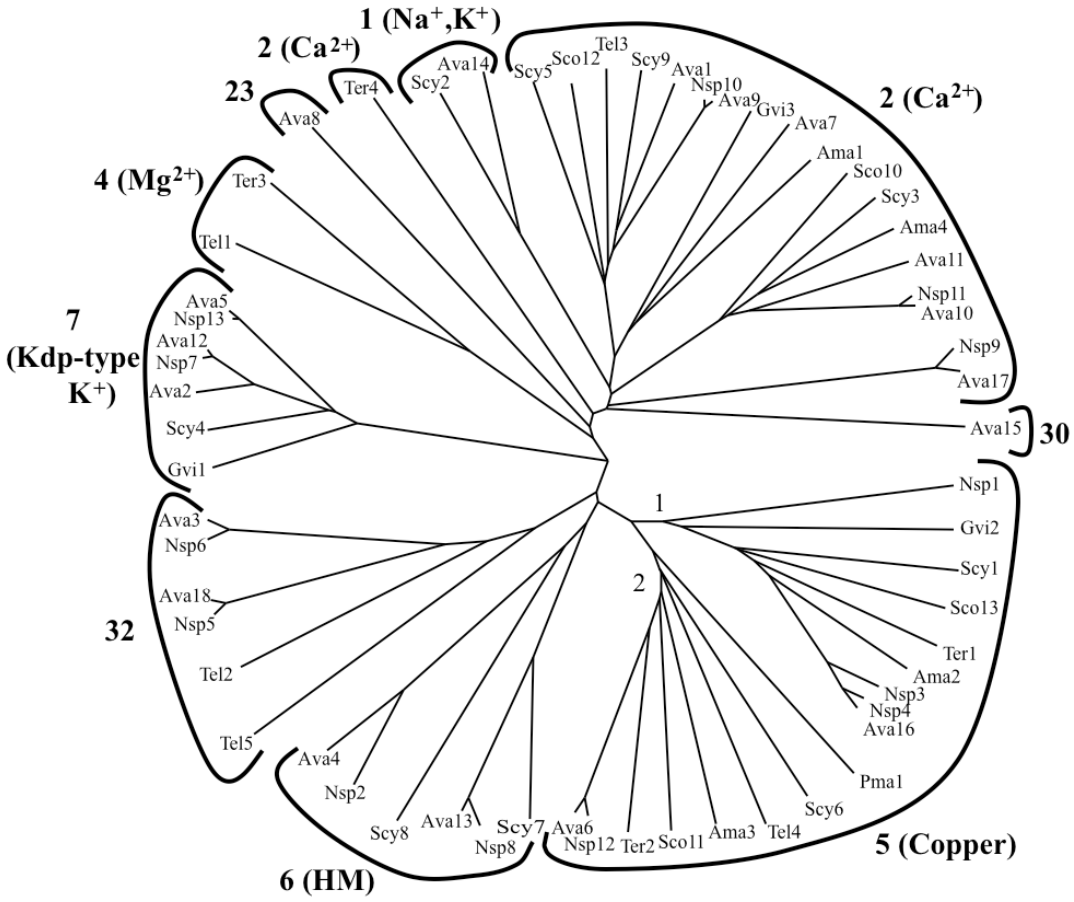

Figure 3: Phylogenetic tree including fifty-nine cyanobacterial P-type ATPases representing six functionally characterized families and three functionally uncharacterized families.

Figure 4: 16S rRNA tree of the nine cyanobacterial genuses included in this study.

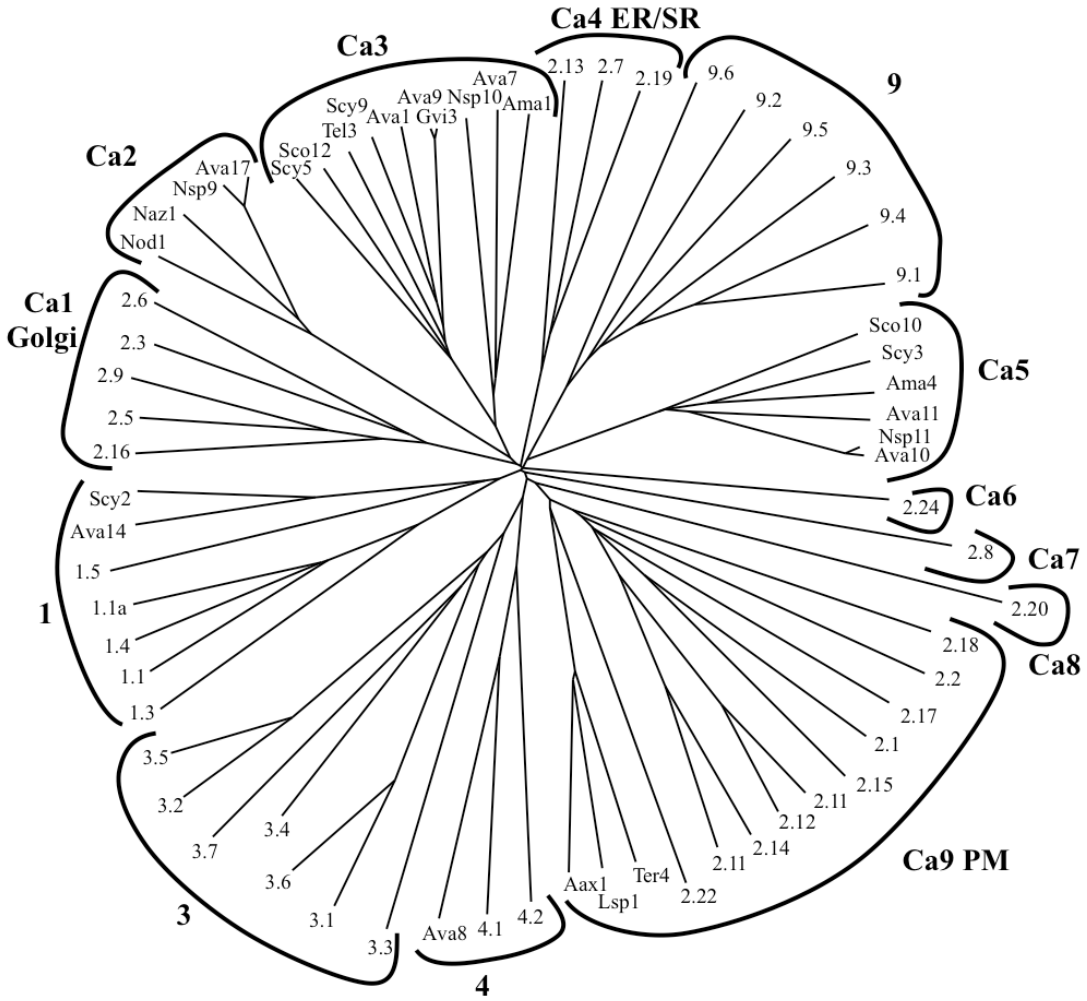

Figure 5: Phylogenetic tree of Type II ATPases included in this study along with standards from TCDB; indicated by the last two digits of the TC number. Standards for Golgi, ER/SR and Plasma Membrane (PM) Ca^{2+}-ATPase types are indicated.

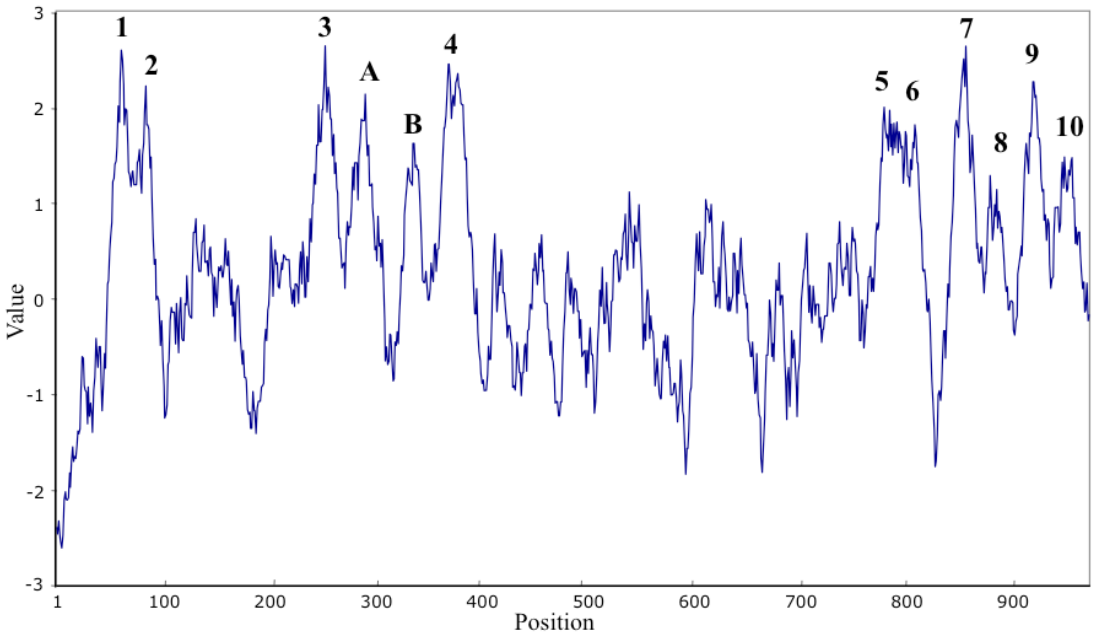

Figure 6: Hydropathy plot of a Type II protein, Ter4, showing two extra putative TMSs between peaks 3 and 4 (labeled A and B, respectively).

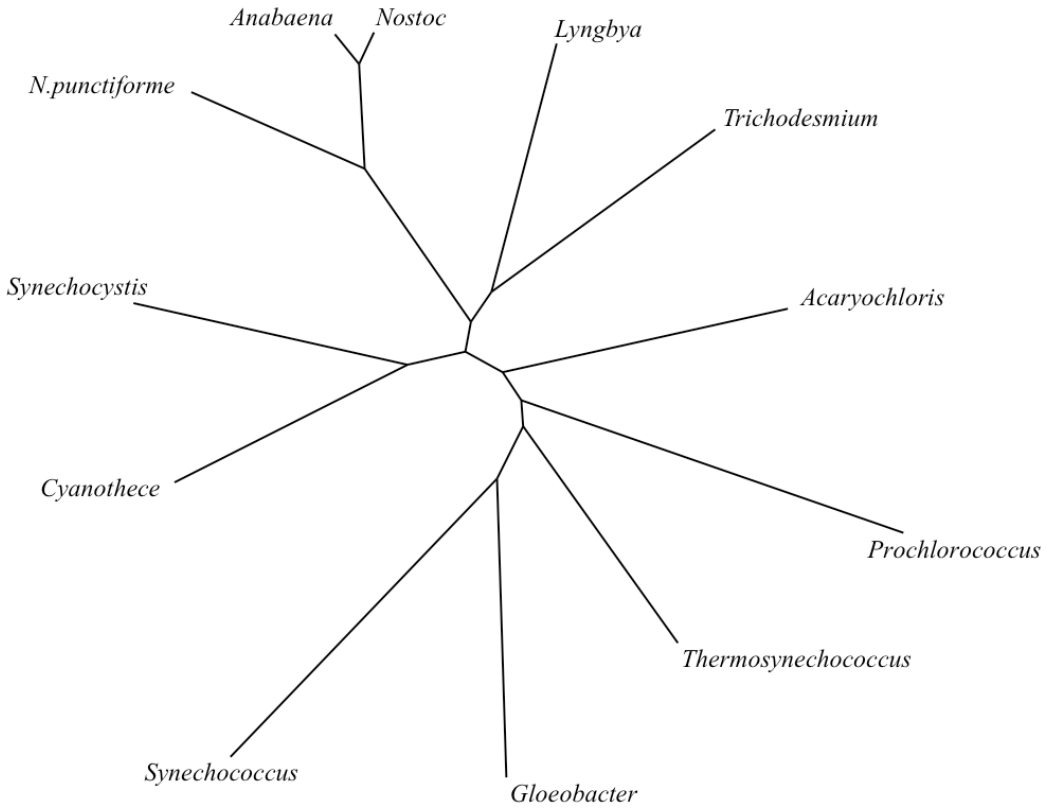

Figure 7: 16S rRNA tree for the nine studied cyanobacteria, along with three of the four additional cyanobacteria showing the two extra TMS in Ter4. (marked in bold).

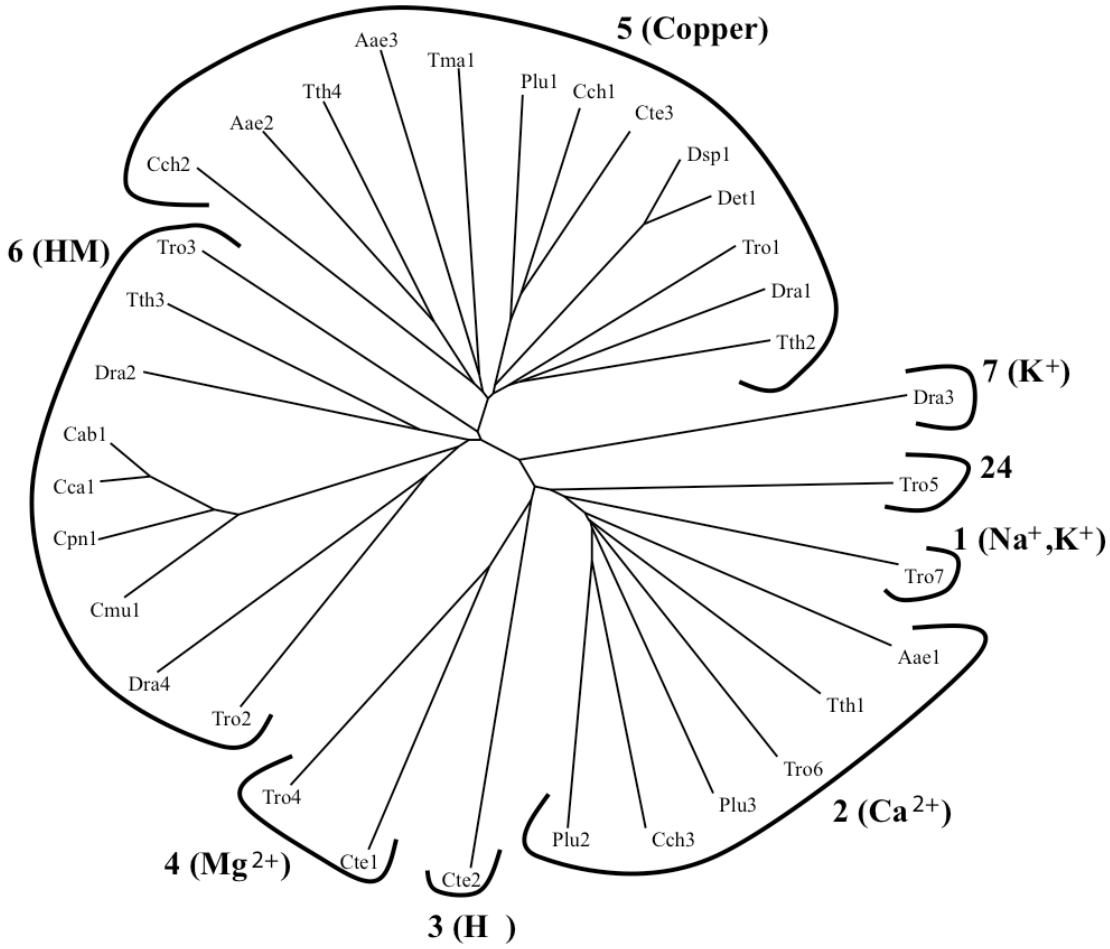

Figure 8: Phylogenetic tree of P-type ATPases from miscellaneous small bacterial phyla.

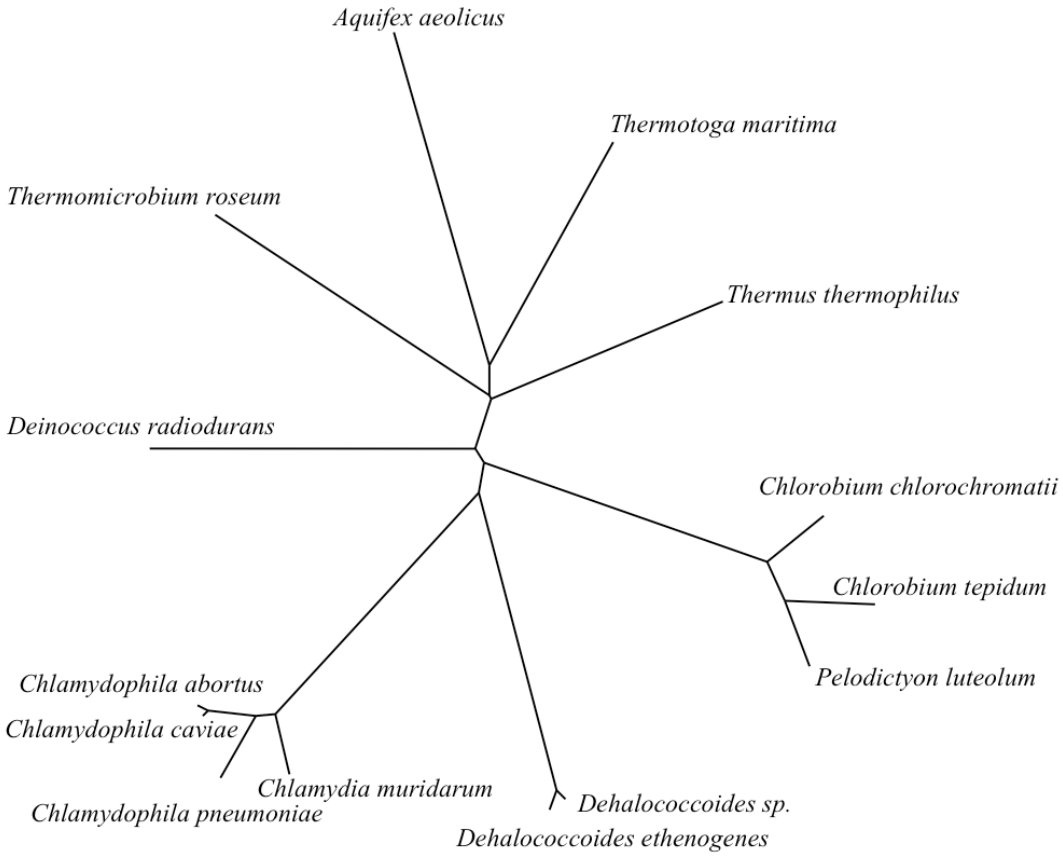

Figure 9: Phylogenetic tree of 16S rRNAs from the bacterial species represented in the small miscellaneous phyla included in this study.

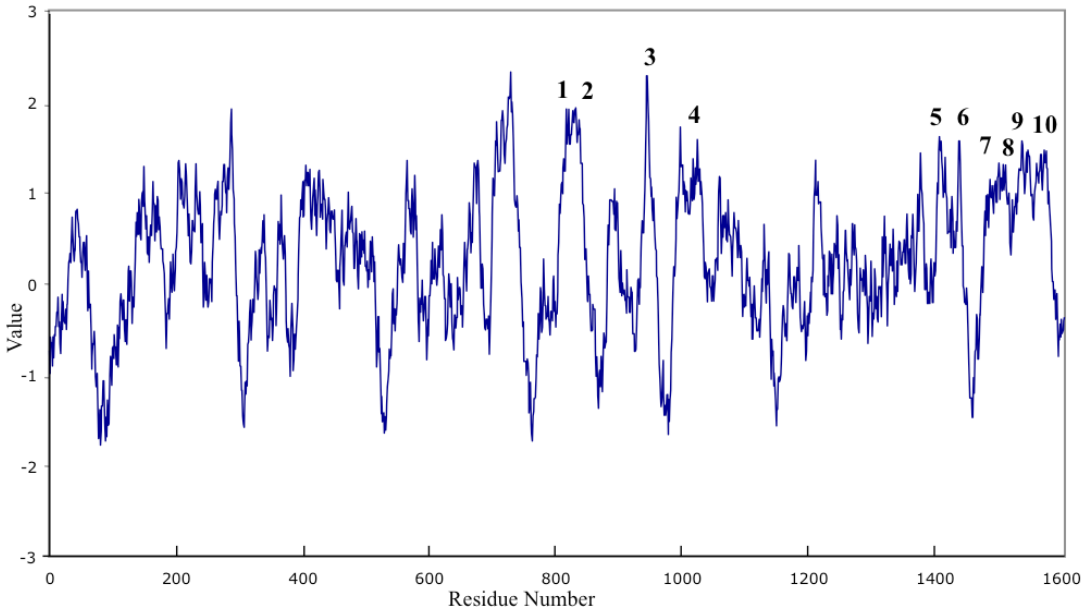

Figure 10: Topology (WHAT program) of the Tro5 protein, a representation of FUPA Family 24.

Table 1 : Numbers of P-type ATPases in the families represented in nine Cyanobacteria.
A total of fifty-nine P-type ATPases, present in nine Cyanobacteria were identified. The complete organismal names with along with the corresponding abbreviations are provided.

Organisms	Genome Size	# of ORFs	Family 1 (Na+, K+)	Family 2 (Ca2+)	Family 4 (Mg2+)	Family 5 (copper)	Family 6 (HM)	Family 7 (Kdp)	FUPA 23	FUPA 30	FUPA 32	Total
Acaryochloris marina MBIC11017 (Ama)	8.3	6254	0	2	0	2	0	0	0	0	0	**4**
Anabaena variabilis ATCC 29413 (Ava)	7.1	5043	1	5	1	2	2	3	0	1	2	**17**
Gloeobacter violaceus PCC 7421 (Gvi)	4.7	4430	0	1	0	1	0	1	0	0	0	**3**
Nostoc sp. PCC 7120 (Nsp)	7.2	5366	0	2	0	4	2	2	0	0	2	**12**
Prochlorococcus marinus str. MIT 9303 (Pma)	2.7	2997	0	0	0	1	0	0	0	0	0	**1**
Synechococcus sp. JA-2-3B'a(2-13) (Sco)	3.1	2862	0	2	0	2	0	0	0	0	0	**4**
Synechocystis sp. PCC 6803 (Scy)	3.9	3172	1	3	0	2	2	1	0	0	0	**9**
Thermosynechococcus elongatus BP-1 (Tel)	2.6	2476	0	1	0	1	0	0	1	0	2	**5**
Trichodesmium erythraeum IMS101 (Ter)	7.8	4451	0	1	0	2	0	0	1	0	0	**4**
Total			**2**	**17**	**1**	**17**	**6**	**7**	**2**	**1**	**6**	**59**

Table 2: P-type ATPases Identified in nine Cyanobacteria.
Fifity-nine cyanobacterial P-type ATPases are categorized here according to TCDB Family. Included are the functionally characterized and functionally uncharacterized families as described in the text.
Notes:
HM is the abbreviation for Heavy Metal
Gene Bank Index Number is referred to as the GI number in this paper.

Abbreviation	Organism	Protein Size (#aas)	GenBank Index# (GI)
Family 1 (Na^+,K^+)			
Ava14	*Anabaena variabilis ATCC 29413*	971	75910286
Scy2	*Synechocystis sp. PCC 6803*	972	16329893
Average Size ± S.D.		**971 ± 1**	
Family 2 (Ca^{2+})			
Scy5	*Synechocystis sp. PCC 6803*	953	16330730
Scy9	*Synechocystis sp. PCC 6803*	945	16331945
Ava1	*Anabaena variabilis ATCC 29413*	946	75812388
Nsp10	*Nostoc sp. PCC 7120*	957	17230867
Ava9	*Anabaena variabilis ATCC 29413*	953	75909598
Tel3	*Thermosynechococcus elongatus BP-1*	941	22294948
Sco12	*Synechococcus sp. JA-2-3B'a(2-13)*	929	86609103
Gvi3	*Gloeobacter violaceus PCC 7421*	921	37523784
Ava7	*Anabaena variabilis ATCC 29413*	914	75908122
Ama1	*Acaryochloris marina MBIC11017*	910	158334135
Scy3	*Synechocystis sp. PCC 6803*	905	16330489
Ama4	*Acaryochloris marina MBIC11017*	910	158339407
Nsp11	*Nostoc sp. PCC 7120*	911	17231215
Ava10	*Anabaena variabilis ATCC 29413*	912	75909762
Ava11	*Anabaena variabilis ATCC 29413*	915	75909770
Sco10	*Synechococcus sp. JA-2-3B'a(2-13)*	921	86608249
Ter4	*Trichodesmium erythraeum IMS101*	974	113478169
Nsp9	*Nostoc sp. PCC 7120*	995	17230737
Ava17	*Anabaena variabilis ATCC 29413*	1002	75911096
Average Size ± S.D.		**929 ± 21**	
Family 4 (Mg^{2+})			
Ava8	*Anabaena variabilis ATCC 29413*	912	75908125
Family 5 (Copper)			
Scy1	*Synechocystis sp. PCC 6803*	745	16329860
Nsp3	*Nostoc sp. PCC 7120*	753	17158771
Nsp4	*Nostoc sp. PCC 7120*	753	17229119

Table 2 continued

Abbreviation	Organism	Protein Size (#aas)	GenBank Index# (GI)
Ava16	*Anabaena variabilis ATCC 29413*	753	75910433
Ama2	*Acaryochloris marina MBIC11017*	754	158334138
Ter1	*Trichodesmium erythraeum IMS101*	758	113475045
Sco13	*Synechococcus sp. JA-2-3B'a(2-13)*	771	86609786
Gvi2	*Gloeobacter violaceus PCC 7421*	747	37523616
Nsp1	*Nostoc sp. PCC 7120*	724	17158728
Scy6	*Synechocystis sp. PCC 6803*	780	16331210
Nsp12	*Nostoc sp. PCC 7120*	815	17231274
Ava6	*Anabaena variabilis ATCC 29413*	813	75907770
Ter2	*Trichodesmium erythraeum IMS101*	773	113475254
Sco11	*Synechococcus sp. JA-2-3B'a(2-13)*	864	86608948
Ama3	*Acaryochloris marina MBIC11017*	794	158334993
Tel4	*Thermosynechococcus elongatus BP-1*	745	22295646
Pma1	*Prochlorococcus marinus str. MIT 9303*	774	124024338
Average Size ± S.D.		**771 ± 34**	
Family 6 (HM)			
Ava4	*Anabaena variabilis ATCC 29413*	751	75907348
Nsp2	*Nostoc sp. PCC 7120*	879	17158758
Scy8	*Synechocystis sp. PCC 6803*	721	16331908
Ava13	*Anabaena variabilis ATCC 29413*	641	75910063
Nsp8	*Nostoc sp. PCC 7120*	694	17230653
Scy7	*Synechocystis sp. PCC 6803*	642	16331905
Average Size ± S.D.		**721 ± 88**	
Family 7 (Kdp)			
Gvi1	*Gloeobacter violaceus PCC 7421*	694	37520143
Ava5	*Anabaena variabilis ATCC 29413*	725	75907417
Nsp13	*Nostoc sp. PCC 7120*	701	17231737
Ava2	*Anabaena variabilis ATCC 29413*	715	75812488
Ava12	*Anabaena variabilis ATCC 29413*	708	75910052
Nsp7	*Nostoc sp. PCC 7120*	708	17230645
Scy4	*Synechocystis sp. PCC 6803*	690	16330521
Average Size ± S.D.		**706 ± 12**	
Family 23			
Ter3	*Trichodesmium erythraeum IMS101*	831	113477109
Tel1	*Thermosynechococcus elongatus BP-1*	826	22293874
Average Size ± S.D.		**825 ± 3**	
Family 30			
Ava15	*Anabaena variabilis ATCC 29413*	867	75910290
Family 32			

Table 2 continued

Abbreviation	Organism	Protein Size (#aas)	GenBank Index# (GI)
Tel5	*Thermosynechococcus elongatus BP-1*	769	22295937
Tel2	*Thermosynechococcus elongatus BP-1*	769	22294378
Ava3	*Anabaena variabilis ATCC 29413*	737	75907215
Nsp6	*Nostoc sp. PCC 7120*	735	17230400
Ava18	*Anabaena variabilis ATCC 29413*	770	75911260
Nsp5	*Nostoc sp. PCC 7120*	771	17229496
Average Size ± S.D.		**758 ± 44**	

Table 3: Type 2 (Ca^{2+}) P-Type ATPases from cyanobacteria as well as standards from TCDB (6/14/09).
Notes:
*'s indicate added proteins
Numbers (such as 3.3, 3.4, etc.) refer to the TCDB family and sub-family numbers.

Abbreviation	Organism	Protein Size (# aas)	GenBank Index#
Ca1 (Golgi)			
2.16	*Caenorhabditis elegans*	901	75028081
2.5	*Homo sapiens*	919	68068024
2.9	*Homo sapiens*	946	218511924
2.3	*Saccharomyces cerevisiae*	950	114301
2.6	*Neurospora crassa*	1025	74698463
		948 +/- 47	
Ca2			
Nsp9	*Nostoc sp. PCC 7120*	995	17230737
Ava17	*Anabaena variabilis ATCC 29413*	1002	75911096
Nod1[2]	*Nodularia spumigena CCY9414*	1002	119512071
Naz1[2]	*Nostoc azollae 0708*	999	225520741
		999 +/- 3	
Ca3			
Scy5	*Synechocystis sp. PCC 6803*	953	16330730
Sco12	*Synechococcus sp. JA-2-3B'a(2-13)*	929	86609103
Tel3	*Thermosynechococcus elongatus BP-1*	941	22294948
Scy9	*Synechocystis sp. PCC 6803*	945	16331945
Ava1	*Anabaena variabilis ATCC 29413*	946	75812388
Nsp10	*Nostoc sp. PCC 7120*	957	17230867
Ava9	*Anabaena variabilis ATCC 29413*	953	75909598
Gvi3	*Gloeobacter violaceus PCC 7421*	921	37523784
Ava7	*Anabaena variabilis ATCC 29413*	914	75908122
Ama1	*Acaryochloris marina MBIC11017*	910	158334135
		937 +/- 17	
Ca4 (ER/SR)			
2.13	*Arabidopsis thaliana*	1061	12643704
2.7	*Homo sapiens*	1042	114312

Table 3 continued

Abbreviation	Organism	Protein Size (# aas)	GenBank Index#
2.19	*Arabidopsis thaliana*	998	122229987
		1034 +/- 32	
Family 9			
9.6	*Ustilago maydis*	1125	74704927
9.2	*Schizosaccharomyces pombe*	1037	114303
9.5	*Ustilago maydis*	1100	195540542
9.3	*Debaryomyces occidentalis*	1082	74675873
9.4	*Zygosaccharomyces rouxii*	1048	74676231
9.1	*Saccharomyces cerevisiae*	1091	114302
		1080 +/- 33	
Ca5			
Sco10	*Synechococcus sp. JA-2-3B'a(2-13)*	921	86608249
Scy3	*Synechocystis sp. PCC 6803*	905	16330489
Ama4	*Acaryochloris marina MBIC11017*	910	158339407
Ava11	*Anabaena variabilis ATCC 29413*	915	75909770
Nsp11	*Nostoc sp. PCC 7120*	911	17231215
Ava10	*Anabaena variabilis ATCC 29413*	912	75909762
		912 +/- 5	
Ca6			
2.24	*Aquifex aeolicus*	835	81343521
Ca7			
2.8	*Plasmodium falciparum*	1228	74967369
Ca8			
2.2	*Clostridium acetobutylicum*	847	81530583
Ca9 (PM)			
2.18	*Toxoplasma gondii*	1405	75023636
2.2	*Saccharomyces cerevisiae*	1173	728904
2.17	*Dictyostelium discoideum*	1115	166203130
2.1	*Homo sapiens*	1241	14286105
2.15	*Caenorhabditis elegans*	1228	74763859
2.11	*Arabidopsis thaliana*	1020	30316378
2.12	*Arabidopsis thaliana*	1014	12229639
2.14	*Arabidopsis thaliana*	1086	150421517
2.11	*Arabidopsis thaliana*	1074	12643246
2.22	*Streptococcus thermophilus CNRZ1066*	878	81559531
Ter4	*Trichodesmium erythraeum IMS101*	974	113478169
Lsp1[2]	*Lyngbya sp. PCC 8106*	976	119485128
Aax1[2]	*Arthrospira maxima CS-328*	972	209525516

Table 3 continued

Abbreviation	Organism	Protein Size (# aas)	GenBank Index#
		1088 +/- 142	
Family 4			
4.2	*Leptospira interrogans*	843	81830773
4.1	*Salmonella typhimurium*	903	543864
Ava8	*Anabaena variabilis ATCC 29413*	912	75908125
		886 +/- 37	
Family 3			
3.3	*Lactobacillus plantarum*	758	81325414
3.1	*Neurospora crassa*	920	114347
3.6	*Saccharomyces cerevisiae*	918	1168544
3.4	*Methanocaldococcus jannaschii*	805	47606650
3.7	*Arabidopsis thaliana*	949	12644156
3.2	*Leishmania donovani*	974	20981683
3.5	*Trypanosoma brucei*	912	75013788
		890 +/- 79	
Family 1			
1.3	*Heterosigma akashiwo*	1330	75213257
1.1	*Homo sapiens*	1035	148877240
1.4	*Rattus norvegicus*	1036	1703464
1.1	*Homo sapiens*	1023	114374
1.5	*Leptospira biflexa serovar Patoc strain 'Patoc 1 (Paris)'*	1046	183222107
Ava14	*Anabaena variabilis ATCC 29413*	971	75910286
Scy2	*Synechocystis sp. PCC 6803*	972	16329893
		1059 +/- 123	

Table 4: Analysis of Nine Conserved Motifs in P-type ATPases of Cyanobacteria
The nine conserved motifs of the functionally characterized and uncharacterized families were identified based on the multiple alignments of the appropriate proteins. Only the most common motif residues are presented in this table along with the degree of conservation.
Notes:
* is the symbol on the multiple alignments for an identity.
: is the symbol on the multiple alignments representing a close similarity.
. is the symbol on the multiple alignments signifying a more distant similarity as determined by the CLUSTAL X program.
_ represents a space in the multiple alignments.

Consensus:	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Family 1 (Na^+,K^+)									
Seq. Motif	RGD	SAD	TGES	PEGL	DKTGTLT	KGAPL	DPPR	MVTGD	VAVTGDGVNDAPALRAAHIGVAM
Conservation	***	***	****	****	*******	*****	****	*****	**********:**:**
Family 2 (Ca^{2+})									
Seq. Motif	PGD	PAD	TGAA	PEGL	DKTGTLT	KGRIK	DPAR	MITGD	VAMTGDGANDAPALKKADIGIAM
Conservation	***	***	**_:	****	.******	**__.	**.*	*****	*******_*******:****:**
Family 4 (Mg^{2+})									
Seq. Motif	PGD	PAD	TGES	PEML	DKTGTLT	KGAVE	DPPK	ILTGD	VGFLGDGINDAAALREADVGISV
Conservation	***	***	***:	****	*******	*****	****	*****	**::*******.***:*******
Family 5 (Copper)									
Seq. Motif	PGD	PVD	TGES	PCAL	DKTGTLT	LGNLD	DTLR	LLTGD	VAMVGDGINDAPALAQADVGISL
Conservation	.*:	*.*	***.	****	*******	_*__.	*__*	:*:**	:.*****:******:_*::**.:
Family 6 (HM)									
Seq. Motif	PGE	PLD	TGES	PCAL	DKTGTLT	VGNRR	DGIR	MLTGD	VAMVGDGINDAPALAAADVGIAM
Conservation	**_	._*	****	**.*	*****:*	.*:__	*_::	*::**	:.:*******:*.:*_*.:.:*:
Family 7 (Kdp)									
Seq. Motif	RGD	PAD	TGES	PTTI	DKTGTIT	KGAVD	DIVK	MLTGD	VAMTGDGTNDAPALAQANVGVAM
Conservation	:.*	*.*	****	****	*******	****.	**:*	*****	**********:**:**
FUPA23									
Seq. Motif	PGD	VVD	TGES	PVGL	DKTGTLT	LGAPE	DRLR	IISGD	VAMIGDGVNDVLSLKQANVGIAM
Conservation	.**	*.*	****	*_**	*******	****:	*.**	:****	**********:**:.***
FUPA30									
Seq. Motif	RED	PAD	TGES	PNEI	DKTGTLT	KGAPE	DPVR	MITGD	VAMTGDGVNDAPALKAANIGIAM
Conservation	*__	.**	:***	*:*:	*******	*****	**:*	*****	**********:*::*:*****
FUPA32									
Seq. Motif	PGD	PVD	TGES	GTGI	DKTGTLT	VGSER	DPLR	MLTGD	VAFVGDGINDSPALAYADVSVSF
Conservation	.*.	*:*	****	_:**	*******	**:_:	:_:*	::***	**:_*:*:**:_*:****::::::

Table 5: Properties of organisms from miscellaneous small bacterial phyla. A total of thirty-one P-type ATPases, present in six miscellaneous bacterial phyla were identified. The complete organismal names with along with the corresponding abbreviations are provided.

Organisms	Genome Size	# ORFs	# P-Type ATPases
Chlorobi			
Chlorobium chlorochromatii CaD3 (Cch)	2.5	2002	**3**
Chlorobium tepidum TLS (Cte)	2.1	2245	**3**
Pelodictyon luteolum DSM 273 (Plu)	2.4	2083	**3**
Aquificae			
Aquifex aeolicus VF5 (Aae)	1.5	1529	**3**
Thermotogae			
Thermotoga maritima MSB8 (Tma)	1.8	1858	**1**
Chloroflexi			
Dehalococcoides sp. CBDB1 (Dsp)	1.4	1458	**1**
Dehalococcoides ethenogenes 195 (Det)	1.5	1580	**1**
Thermomicrobium roseum DSM 5159 (Tro)	2	1922	**7**
Chlamydia			
Chlamydia muridarum Nigg (Cmu)	1.1	904	**1**
Chlamydophila abortus S26/3 (Cab)	1.1	932	**1**
Chlamydophila caviae GPIC (Chl)	1.2	998	**1**
Chlamydophila pneumoniae TW-183 (Cpn)	1.2	1113	**1**
Deinococcus-Thermus			
Deinococcus radiodurans R1 (Dra)	2.6	2629	**4**
Thermus thermophilus HB27 (Tth)	1.9	1982	**4**
Total			**31**

Table 6: P-type ATPases Identified in six bacterial phyla.
Thirty-one P-type ATPases from miscellaneous bacterial phyla are categorized here according to TCDB Family. Included are the functionally characterized and functionally uncharacterized families as described in the text.
Notes:
HM is the abbreviation for Heavy Metal
Gene Bank Index Number is referred to as the GI number in this paper.

Abbreviation	Organism	Protein Size	GenBank Index (GI) #
Chlorobi			
Family 2 (Ca^{2+})			
Cch3	*Chlorobium chlorochromatii* CaD3	912	78189109
Plu2	*Pelodictyon luteolum* DSM 273	898	78186809
Plu3	*Pelodictyon luteolum* DSM 273	891	78187096
Average protein size ± S.D.		**900 +/- 10**	
Family 3 (H^+/K^+)			
Cte2	*Chlorobium tepidum* TLS	869	21674501
Family 4 (Mg^{2+})			
Cte1	*Chlorobium tepidum* TLS	886	21673463
Family 5 (Copper)			
Cch1	*Chlorobium chlorochromatii* CaD3	761	78188927
Cch2	*Chlorobium chlorochromatii* CaD3	814	78189332
Plu1	*Pelodictyon luteolum* DSM 273	743	78187092
Cte3	*Chlorobium tepidum* TLS	758	21673644
Average protein size ± S.D.		**769 +/- 31**	
Aquificae			
Family 2 (Ca^{2+})			
Aae1	*Aquifex aeolicus* VF5	835	15606121
Family 5 (Copper)			
Aae2	*Aquifex aeolicus* VF5	664	15606387
Aae3	*Aquifex aeolicus* VF5	679	15606617
Average protein size ± S.D.		**671 +/- 10**	
Thermotogae			
Family 5 (Copper)			
Tma1	*Thermotoga maritima* MSB8	726	15643086

Table 6 continued

Abbreviation	Organism	Protein Size	GenBank Index (GI) #
Chloroflexi			
Family 1 (Na^+,K^+)			
Tro7	*Thermomicrobium roseum* DSM 5159	873	221632607
Family 2 (Ca^{2+})			
Tro6	*Thermomicrobium roseum* DSM 5159	926	221635716
Family 4 (Mg^{2+})			
Tro4	*Thermomicrobium roseum* DSM 5159	658	221633428
Family 5 (Copper)			
Dsp1	*Dehalococcoides* sp. CBDB1	828	73748721
Det1	*Dehalococcoides ethenogenes* 195	828	57234243
Tro1	*Thermomicrobium roseum* DSM 5159	842	221633394
Average protein size ± S.D.		**832 +/- 8**	
Family 6 (HM)			
Tro2	*Thermomicrobium roseum* DSM 5159	846	221633749
Tro3	*Thermomicrobium roseum* DSM 5159	699	221633506
		772 +/- 104	
FUPA24			
Tro5	*Thermomicrobium roseum* DSM 5159	1607	221635885
Chlamydia			
Family 6 (HM)			
Cmu1	*Chlamydia muridarum* Nigg	659	15834725
Chl1	*Chlamydophila caviae* GPIC	657	29840655
Cpn1	*Chlamydophila pneumoniae* TW-183	658	33242228
Cab1	*Chlamydophila abortus* S26/3	657	62185469
Average protein size ± S.D.		**657 +/- 1**	
Deinococcus-Thermus			
Family 2 (Ca^{2+})			

Table 6 continued

Abbreviation	Organism	Protein Size	GenBank Index (GI) #
Tth1	*Thermus thermophilus* HB27	809	46199082
Family 5 (Copper)			
Dra1	*Deinococcus radiodurans* R1	847	15807440
Tth2	*Thermus thermophilus* HB27	798	46199660
Tth4	*Thermus thermophilus* HB27	687	46199673
Average protein size ± S.D.		**777 +/- 82**	
Family 6 (HM)			
Dra2	*Deinococcus radiodurans* R1	728	15807741
Dra4	*Deinococcus radiodurans* R1	748	15807638
Tth3	*Thermus thermophilus* HB27	683	46198662
Average protein size ± S.D.		**719 +/- 33**	
Family 7 (Kdp)			
Dra3	*Deinococcus radiodurans* R1	675	10957402

Table 7: Analysis of Nine Conserved Motifs in P-type ATPases of Miscellaneous Bacterial Phyla

The nine conserved motifs of the functionally characterized and uncharacterized families were identified based on the multiple alignments of the appropriate proteins. Only the most common motif residues are presented in this table along with the degree of conservation.

Notes:

* is the symbol on the multiple alignments for an identity.

: is the symbol on the multiple alignments representing a close similarity.

. is the symbol on the multiple alignments signifying a more distant similarity as determined by the CLUSTAL X program.

_ represents a space in the multiple alignments.

Consensus:	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Family 1 (Na^+,K^+)									
Chloroflexi									
Seq. Motif	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MVTGD	VAMTGDGVNDAPALRQADVGVAM
Family 2 (Ca^{2+})									
Chlorobi									
Seq. Motif	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MITGD	VVAMTGDGVNDAPALKRADVGIA
Conservation	.**	***	****	**.*	*******	*****	****	*****	**********:**::::*::
Chloroflexi									
Seq. Motif	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MITGD	VAVTGDGVNDAPALRAAEIGVAM
Deinococcus-Thermus									
Seq. Motif	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MVTGD	VAMTGDGVNDAPALKRADVGVAM
Family 3 (H^+/K^+)									
Chlorobi									
Seq. Motif	PGD	PAD	TGES	PVAL	DKTGTLT	KGAPQ	DPPR	MVTGD	VSMTGDGVNDAPALKKADCGIAV
Family 4 (Mg^{2+})									
Chlorobi									
Seq. Motif	PGD	PAD	TGES	PEML	DKTGTLT	KGAPE	DPPR	MLTGD	VAFLGDGINDAPALREADVGISV
Family 5 (Copper)									
Aquificae									
Seq. Motif	PGE	PTD	TGES	PHAL	DKTGTLT	VGTLE	DRIK	MITGD	VAMVGDGVNDAPALIQADVGIAI
Conservation	.**	_**	***.	*_*:	*******	.:___	*_::	::***	*.***********_*:*:***:
Chlorobi									
Seq. Motif	RPD	PVD	TGES	PCAL	DKTGTIT	PLAQA	DTIK	MLTGD	VAMAGDGINDAPALALADVGIAM
Conservation	_.*	*:*	:**.	****	*****:*	*_*:*	*_**	:::**	*_*.**********:_*_.::::
Chloroflexi									
Seq. Motif	PGE	PVD	TGES	PCAL	DKTGTLT	PGAGL	DILK	MLTGD	VAMVGDGINDAPALAKADVGIAI
Conservation	***	***	****	***:	*******	**_*:	*__:	:****	*.*************:******:

Table 7 continued

Consensus:	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Deinococcus-Thermus									
Seq. Motif	PGD	PVD	TGES	PCAM	DKTGTLT	LGLPL	DPIR	MVTGD	VAFVGDGINDAPALAGADVGVAIG
Conservation	_**	*.*	***.	*_*:	*****:*	**:_:	*__*	::***	..*********.**__**:*:*:*
Family 6 (HM)									
Chlamydia									
Seq. Motif	SGE	PLD	TGEK	PCAL	DKTGTLT	KGVRG	DTPR	MLTGD	ILMVGDGINDAPALAQATVGVAMG
Conservation	***	***	****	****	*******	:**:*	*_**	*****	*:******************:***
Chloroflexi									
Seq. Motif	PGD	PVD	TGES	PCAL	DKTGTLT	RGVRA	DVVR	MLTGD	VAMVGDGVNDAPALAAADVGIAM
Conservation	**:	***	***.	.**:	*******	_****	*_:*	:****	*.*****:*******_*******
Deinococcus-Thermus									
Seq. Motif	PGE	PAD	TGES	PCAL	DKTGTLT	KGVAE	DEPR	MLTGD	VAMVGDGINDAPALAAADVGIAM
Conservation	**	.**	***.	***:	*******	_.:__	*__*	:****	:.:*********.::*_.:****
Family 7 (Kdp)									
Deinococcus-Thermus									
Seq. Motif	RGD	PAD	TGES	PTTI	DKTGTIT	KGAVD	DIVR	MITGD	VAMMGDGTNDAPALAQADVGLAM
FUPA24									
Chloroflexi									
Seq. Motif	PGD	PAD	TGES	PEGL	DKTGTLT	KGSPE	DPPR	MITGD	VAMTGDGVNDSAALRLADVGMAM

Table 8: Integrated Analysis of Family Representation of P-Type ATPases
The table is organized by organismal type and and family, showing the numbers of P-type ATPases present in each family from each organism and showing percentages taken from the total numbers within each family and the totals are given at the bottom of the table.

Family	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Organismal Type											
Eukaryotes											
Animal	11	18	0	0	5	0	0	25	0	4	0
Plant	0	28	21	0	10	7	0	21	0	2	0
Fungi	4	33	16	1	20	0	0	35	19	8	0
1-Cell Euk	2	32	8	1	7	0	0	47	5	11	1
Ciliate	21	11	0	0	0	0	0	23	0	1	0
Total	**38**	**122**	**45**	**2**	**42**	**7**	**0**	**151**	**24**	**26**	**1**
Percent	**7.52 %**	**24.16 %**	**8.91 %**	**0.40 %**	**8.32%**	**1.39%**	**0.00 %**	**29.90%**	**4.75 %**	**5.15 %**	**0.20 %**
Archaea											
Crenarchaeota	0	2	4	0	14	2	1	0	0	0	0
Euryarchaeota	7	39	14	5	66	34	8	0	0	0	0
Korarchaeota	0	0	1	0	1	0	0	0	0	0	0
Total	**7**	**41**	**19**	**5**	**81**	**36**	**9**	**0**	**0**	**0**	**0**
Percent	**3.50 %**	**20.50 %**	**9.50 %**	**2.50 %**	**40.50 %**	**18.00 %**	**4.50 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**
Bacteria											
Spirochaetes	1	1	0	2	7	3	1	0	0	0	0
Actinobacteria	0	5	0	1	23	12	6	0	0	0	0
Firmicutes	0	23	0	9	16	18	9	0	0	0	0
Proteobacteria											
a	0	2	0	4	14	9	6	0	0	0	0
b	0	3	0	2	9	3	3	0	0	0	0
g	1	11	0	12	33	19	13	0	0	0	0
d	1	6	2	0	6	1	2	0	0	0	0
e	0	0	0	0	8	3	1	0	0	0	0
Bacteroides	0	6	0	3	7	6	6	0	0	0	0
Flavobacteria	0	0	0	0	2	2	2	0	0	0	0
Fusobacteria	0	2	0	0	2	3	0	0	0	0	0
Chlorobi	0	3	1	1	4	0	0	0	0	0	0
Aquificae	0	1	0	0	2	0	0	0	0	0	0
Thermotogae	0	0	0	0	1	0	0	0	0	0	0
Chloroflexi	1	1	0	1	3	2	0	0	0	0	0
Chlamydia	0	0	0	0	0	4	0	0	0	0	0
Deinococcus Thermus	0	1	0	0	3	3	1	0	0	0	0
Cyanobacteria	2	19	0	1	17	6	7	0	0	0	0
Total	**6**	**84**	**3**	**36**	**157**	**94**	**57**	**0**	**0**	**0**	**0**
Percent	**1.13 %**	**15.88 %**	**0.57 %**	**6.81 %**	**29.68 %**	**17.77 %**	**10.78 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**
Total	**51**	**247**	**67**	**43**	**280**	**137**	**66**	**151**	**24**	**26**	**1**
Percent	**4.13 %**	**20.02 %**	**5.43 %**	**3.48 %**	**22.69 %**	**11.10 %**	**5.35 %**	**12.24%**	**1.94 %**	**2.11 %**	**0.08 %**

Table 8 continued

Family	12	13	14	15	16	17	18	19	20	21	22
Organismal Type											
Eukaryotes											
Animal	0	5	0	0	0	0	0	0	0	0	0
Plant	0	0	0	0	0	0	0	0	0	0	0
Fungi	0	0	6	0	0	1	0	0	0	0	0
1-Cell Euk	1	0	0	3	0	0	1	0	0	1	6
Ciliate	0	0	0	0	11	0	0	1	11	0	0
Total	**1**	**5**	**6**	**3**	**11**	**1**	**1**	**1**	**11**	**1**	**6**
Percent	**0.20 %**	**0.99 %**	**1.19 %**	**0.59 %**	**2.18 %**	**0.20 %**	**0.20 %**	**0.20 %**	**2.18 %**	**0.20 %**	**1.19 %**
Archaea											
Crenarchaeota	0	0	0	0	0	0	0	0	0	0	0
Euryarchaeota	0	0	0	0	0	0	0	0	0	0	0
Korarchaeota	0	0	0	0	0	0	0	0	0	0	0
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Percent	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**
Bacteria											
Spirochaetes	0	0	0	0	0	0	0	0	0	0	0
Actinobacteria	0	0	0	0	0	0	0	0	0	0	0
Firmicutes	0	0	0	0	0	0	0	0	0	0	0
Proteobacteria											
a	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0	0
Bacteroides	0	0	0	0	0	0	0	0	0	0	0
Flavobacteria	0	0	0	0	0	0	0	0	0	0	0
Fusobacteria	0	0	0	0	0	0	0	0	0	0	0
Chlorobi	0	0	0	0	0	0	0	0	0	0	0
Aquificae	0	0	0	0	0	0	0	0	0	0	0
Thermotogae	0	0	0	0	0	0	0	0	0	0	0
Chloroflexi	0	0	0	0	0	0	0	0	0	0	0
Chlamydia	0	0	0	0	0	0	0	0	0	0	0
Deinococcus Thermus	0	0	0	0	0	0	0	0	0	0	0
Cyanobacteria	0	0	0	0	0	0	0	0	0	0	0
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Percent	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**
Total	**1**	**5**	**6**	**3**	**11**	**1**	**1**	**1**	**11**	**1**	**6**
Percent	**0.08 %**	**0.41 %**	**0.49 %**	**0.24 %**	**0.89 %**	**0.08 %**	**0.08 %**	**0.08 %**	**0.89 %**	**0.08 %**	**0.49 %**

Table 8 continued

Family	23	24	25	26	27	28	29	30	31	32	Total
Organismal Type											
Eukaryotes											
Animal	0	0	0	0	0	0	0	0	0	0	**68**
Plant	0	0	0	0	0	0	0	0	0	0	**89**
Fungi	0	0	0	0	0	0	0	0	0	0	**143**
1-Cell Euk	0	0	0	0	0	0	0	0	0	0	**126**
Ciliate	0	0	0	0	0	0	0	0	0	0	**79**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**505**
Percent	**0.00 %**	**0.00 %**	**0.00 %**	**0.00%**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	
Archaea											
Crenarchaeota	0	0	0	0	0	0	0	0	0	0	**23**
Euryarchaeota	0	0	0	0	0	0	0	0	0	2	**175**
Korarchaeota	0	0	0	0	0	0	0	0	0	0	**2**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**200**
Percent	**0.00 %**	**0.00 %**	**0.00 %**	**0.00%**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**0.00 %**	**1.00 %**	
Bacteria											
Spirochaetes	0	0	0	0	2	0	0	1	0	1	**19**
Actinobacteria	8	6	4	3	0	0	0	0	0	0	**68**
Firmicutes	5	0	2	0	0	0	0	0	0	0	**82**
Proteobacteria											
a	0	0	2	0	10	0	0	2	0	1	**50**
b	0	0	0	0	3	0	0	1	0	1	**25**
g	0	0	1	0	13	2	0	0	3	0	**108**
d	0	0	1	0	0	0	1	1	0	1	**22**
e	0	0	0	0	0	0	0	0	0	1	**13**
Bacteroides	0	0	0	0	0	0	0	0	0	0	**28**
Flavobacteria	0	0	0	0	0	0	2	1	0	0	**9**
Fusobacteria	0	0	0	0	0	0	0	0	0	3	**10**
Chlorobi	0	0	0	0	0	0	0	0	0	0	**9**
Aquificae	0	0	0	0	0	0	0	0	0	0	**3**
Thermotogae	0	0	0	0	0	0	0	0	0	0	**1**
Chloroflexi	0	1	0	0	0	0	0	0	0	0	**9**
Chlamydia	0	0	0	0	0	0	0	0	0	0	**4**
Deinococcus Thermus	0	0	0	0	0	0	0	0	0	0	**8**
Cyanobacteria	2	0	0	0	0	0	0	1	0	6	**61**
Total	**15**	**7**	**10**	**3**	**28**	**2**	**3**	**7**	**3**	**14**	**529**
Percent	**2.84 %**	**1.32 %**	**1.89 %**	**0.57%**	**5.29 %**	**0.38 %**	**0.57 %**	**1.32 %**	**0.57 %**	**2.65 %**	
Total	**15**	**7**	**10**	**3**	**28**	**2**	**3**	**7**	**3**	**16**	**1234**
Percent	**1.22 %**	**0.57 %**	**0.81 %**	**0.24 %**	**2.27 %**	**0.16 %**	**0.24 %**	**0.57 %**	**0.24 %**	**1.30 %**	

Table 9: Size distribution of P-type ATPases, categorized by family and organismal type.
The integrated analysis of the size of P-type ATPases across the major kingdoms of life is presented in this table. The average sizes are given according to number of amino acids within each member of a family and the standard deviation is also presented.

Table 9

Family	1	2	3	4	5	6	7	8	9	10	11
Organismal Type											
Bacteria											
Actinobacteria	-	939 +/- 45	-	892	752 +/- 33	672 +/- 50	667 +/- 80	-	-	-	-
Firmicute	-	899 ± 39	-	895 ± 24	766 ± 80	682 ± 41	681 ± 7	-	-	-	-
Proteobacteria	894±3	915±49	863±7	906±27	798±70	736±61	689±30	-	-	-	-
Fuso.Flavo.Bacter.	-	886±17	-	883	780±49	635±28	679±3	-	-	-	-
Cyanobacteria	971 ± 1	929 ± 21	-	912	771 ± 34	721 ± 88	706 ± 12	-	-	-	-
Chlorobi	-	900 +/- 10	869	886	769 +/- 31	-	-	-	-	-	-
Aquificae	-	835	-	-	671 +/- 10	-	-	-	-	-	-
Thermotogae	-	-	-	-	726	-	-	-	-	-	-
Chloroflexi	873	926	-	658	832 +/- 8	772 +/- 104	-	-	-	-	-
Chlamydia	-	-	-	-	-	657 +/- 1	-	-	-	-	-
Deinococcus Thermus	-	809	-	-	777 +/- 82	719 +/- 33	675	-	-	-	-
Spirochaetes	1046	878	-	873 +/- 43	776 +/- 55	653 +/- 16	692	-	-	-	-
All Bacteria	**946 +/- 79**	**892 +/- 42**	**866 +/- 4**	**863 +/- 84**	**765 +/- 41**	**694 +/- 45**	**684 +/- 13**	**-**	**-**	**-**	**-**
Archaea											
Crenarchaeota	-	891 +/- 4	795 +/- 18	-	758 +/- 46	592 +/- 10	701	-	-	-	-
Euryarchaeota	924 +/- 28	848 +/- 37	802 +/- 23	851 +/- 10	800 +/- 83	719 +/- 90	690 +/- 24	-	-	-	-
All Archaea	**924 +/- 28**	**884 +/- 36**	**800 +/- 21**	**851 +/- 10**	**792 +/- 80**	**712 +/- 92**	**691 +/- 22**	**-**	**-**	**-**	**-**
Archaea/Bacteria	**-2.30%**	**-1.40%**	**-7.60%**	**-1.40%**	**3.50%**	**2.60%**	**1.00%**				
Eukaryotes											
Ciliate	1190 +/- 146	1258 +/- 353	-	-	-	-	-	1428 +/- 477	-	1423 +/- 380	-
Fungi	1058 +/- 32	1134 +/- 159	976 +/- 46	929	1160 +/- 100	-	-	1432 +/- 178	1099 +/- 83	1315 +/- 132	-
Animals	1031 +/- 18	1067 +/- 108	-	-	1295 +/- 176	-	-	1239 +/- 174	-	1153 +/- 57	-
Plants	-	1045 +/- 27	946 +/- 34	-	964 +/- 53	937	-	1181 +/- 66	-	1238 +/- 84	-
1-Cell Euk	1269 +/- 52	1060 +/- 64	931 +/- 67	954	1164 +/- 203	-	-	1240 +/- 274	1049 +/- 84	1117 +/- 147	1146
All Eukaryotes	**1137 +/- 112**	**1113 +/- 88**	**951 +/- 23**	**942 +/- 18**	**1148 +/- 136**	**937**	-	**1304 +/- 117**	**1074 +/- 35**	**1249 +/- 124**	**1146**
Eukaryotes/Bacteria	**20.00%**	**25.00%**	**10.00%**	**9.20%**	**50.00%**	**35.00%**					
Overall Average	**1028 +/-133**	**960 +/- 120**	**897 +/- 65**	**876 +/- 77**	**862 +/- 183**	**717 +/- 83**	**685 +/- 12**	**1304 +/- 117**	**1074 +/- 35**	**1249 +/- 124**	**1146**

Table 9 continued

Family	12	13	14	15	16	17	18	19	20	21	22
Organismal Type											
Bacteria											
Actinobacteria	-	-	-	-	-	-	-	-	-	-	-
Firmicute	-	-	-	-	-	-	-	-	-	-	-
Proteobacteria	-	-	-	-	-	-	-	-	-	-	-
Fuso.Flavo.Bacter.	-	-	-	-	-	-	-	-	-	-	-
Cyanobacteria	-	-	-	-	-	-	-	-	-	-	-
Chlorobi	-	-	-	-	-	-	-	-	-	-	-
Aquificae	-	-	-	-	-	-	-	-	-	-	-
Thermotogae	-	-	-	-	-	-	-	-	-	-	-
Chloroflexi	-	-	-	-	-	-	-	-	-	-	-
Chlamydia	-	-	-	-	-	-	-	-	-	-	-
Deinococcus Thermus	-	-	-	-	-	-	-	-	-	-	-
Spirochaetes	-	-	-	-	-	-	-	-	-	-	-
All Bacteria	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
Archaea											
Crenarchaeota	-	-	-	-	-	-	-	-	-	-	-
Euryarchaeota	-	-	-	-	-	-	-	-	-	-	-
All Archaea	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
Archaea/Bacteria											
Eukaryotes											
Ciliate	-	-	-	-	1388 +/- 281	-	1491	1807	1187 +/- 220	-	1541 +/- 430
Fungi	-	-	1390 +/- 120	-	-	1096	-	-	-	-	-
Animals	-	1212 +/- 63	-	-	-	-	-	-	-	-	-
Plants	-	-	-	-	-	-	-	-	-	-	-
1-Cell Euk	1998	-	-	1292 +/- 209	-	-	-	-	-	1372	-
All Eukaryotes	**1998**	**1212**	**1390**	**1292 +/- 209**	**1388 +/- 281**	**1096**	**1491**	**1807**	**1187 +/- 220**	**1372**	**1541 +/- 430**
Eukaryotes/Bacteria											
Overall Average	**1998**	**1212**	**1390**	**1292 +/- 209**	**1388 +/- 281**	**1096**	**1491**	**1807**	**1187 +/- 220**	**1372**	**1541 +/- 430**

Table 9 continued

Family	23	24	er25	26	27	28	29	30	31	32	Average
Organismal Type											
Bacteria											
Actinobacteria	772±61	1329±415	698±62	898±70	-	-	-	-	-	-	**858 +/- 192**
Firmicute	849 ± 104	-	612 ± 11	-	-	-	-	-	-	-	**769 +/- 115**
Proteobacteria	-	-	689±74	-	785±40	849±3	798	841±13	860±198	678±49	**807 +/- 82**
Fuso.Flavo.Bacter.	-	-	-	-	-	-	793.5±2	838	-	735	**778 +/- 91**
Cyanobacteria	825 ± 3	-	-	-	-	-	-	867	-	758 ± 44	**829 +/- 96**
Chlorobi	-	-	-	-	-	-	-	-	-	-	**856 +/- 59**
Aquificae	-	-	-	-	-	-	-	-	-	-	**753 +/- 116**
Thermotogae	-	-	-	-	-	-	-	-	-	-	**726**
Chloroflexi	-	1607	-	-	-	-	-	-	-	-	**945 +/- 337**
Chlamydia	-	-	-	-	-	-	-	-	-	-	**657**
Deinococcus Thermus	-	-	-	-	-	-	-	-	-	-	**745 +/- 60**
Spirochaetes	-	-	-	-	816 +/- 5	-	-	837	-	699	**808 +/- 121**
All Bacteria	**815 +/- 39**	**1468 +/- 196**	**666 +/- 47**	**898**	**800 +/- 22**	**849**	**795 +/- 4**	**846 +/- 14**	**860**	**715 +/- 31**	**850 +/- 177**
Archaea											
Crenarchaeota	-	-	-	-	-	-	-	-	-	-	**747 +/- 111**
Euryarchaeota	-	-	-	-	-	-	-	-	-	706 +/- 1	**805 +/- 80**
All Archaea	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**706 +/- 1**	**795 +/- 87**
Archaea/Bacteria										**-1.30%**	**-6.50%**
Eukaryotes											
Ciliate	-	-	-	-	-	-	-	-	-	-	**1412 +/- 195**
Fungi	-	-	-	-	-	-	-	-	-	-	**1159 +/- 169**
Animals	-	-	-	-	-	-	-	-	-	-	**1166 +/- 102**
Plants	-	-	-	-	-	-	-	-	-	-	**1052 +/- 129**
1-Cell Euk	-	-	-	-	-	-	-	-	-	-	**1132 +/- 424**
All Eukaryotes	-	-	-	-	-	-	-	-	-	-	**1265 +/- 274**
Eukaryotes/Bacteria											**49.00%**
Overall Average	**815 +/- 39**	**1468 +/- 196**	**666 +/- 47**	**898**	**800 +/- 22**	**849**	**795 +/- 4**	**846 +/- 14**	**860**	**715 +/- 31**	

Table 10: Integrated motif analysis
The nine conserved motifs of the functionally characterized and uncharacterized families were identified based on the multiple alignments of the appropriate proteins. Only the most common motif residues are presented in this table along with the degree of conservation.
Notes:
* is the symbol on the multiple alignments for an identity.
: is the symbol on the multiple alignments representing a close similarity.
. is the symbol on the multiple alignments signifying a more distant similarity as determined by the CLUSTAL X program.
_ represents a space in the multiple alignments.

Consensus:	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Family 1									
Seq. Motif:	VGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MVTGD	VAVTGDGVNDAPALRAAHIGVAM
Conservation	_**	.**	.*._	*__:	:_...:*	**___	:___	::*__	____*.___:___::_..:*___
Family 2									
Seq. Motif:	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DPPR	MITGD	VAMTGDGANDAPALKKADIGIAM
Conservation	_.*	__:	.*__	*___	.****:*	.*___	*__:	::***	___.***_**___:__:_:._.:
Family 3									
Seq. Motif:	PGD	PAD	TGES	PVAL	DKTGTLT	KGAPQ	DPPR	MVTGD	VAMTGDGANDAPALKKADIGIAV
Conservation	_.:	_.*	***_	*_*:	*****:*	**:__	____	::___	_.:_***::*.*:*:_:._*_**
Family 4									
Seq. Motif:	PGD	PAD	TGES	PEML	DKTGTLT	KGAVE	DPPK	ILTGD	VGFMGDGINDAPALRKADVGISV
Conservation	_**	*.*	.**.	*:_*	*****:*	**:__	*___	:::**	*.::***:**__.::_:::.:..
Family 5									
Seq. Motif:	PGE	PVD	TGES	PCAL	DKTGTLT	KGLEL	DPIK	MLTGD	VAMVGDGINDSQALAQADVGIAM
Conservation	_._	__*	.*__	:___	**:*_:*	_____	____	:_:**	_____
Family 6									
Seq. Motif:	PGD	PAD	TGES	PCAL	DKTGTLT	KGVRA	DELR	MLTGD	VAMVGDGVNDAPALAAADVGIAM
Conservation	___	___	.**_	._..	*****:*	_____	:___	:_:**	___:*:*_:*___:____..___
Family 7									
Seq. Motif:	RGD	PAD	TGES	PTTI	DKTGTIT	KGAVD	DIVK	MITGD	VAMMGDGTNDAPALAQANVGVAM

Table 10 continued

Consensus:	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Conservation	.__	__:	****	____	*****:*	:_:__	*___	_____	..__:._:._.:.:.:_:_____
Family 8									
Seq. Motif:	VGD	PAD	DGET	PISL	DKTGTLT	KGADS	DGLR	VLTGD	TLAIGDGANDVSMIQMADVGVGI
Conservation	_*_	___	:*._	____	*****:*	_____	____	:_***	___:**._**__::__._:*:*:
Family 9									
Seq. Motif:	PGD	PAD	TGES	PEGL	DKTGTLT	KGAVE	DPPR	MVTGD	SSMTGDGVNDAPAIKSSNMGIAM
Conservation	***	***	***_	*:.*	*******	***_:	****	::***	_:**********::_:::*:**
Family 10									
Seq. Motif:	PGD	PCD	TGES	PPEL	DKTGTLT	KGSPE	CPLK	MITGD	TLMCGDGTNDVGALKQAHVGVAL
Conservation	__*	__*	:**.	*.::	*****:*	**:**	.__:	::***	_:*.******:.***_:._*:::
Family 11									
Seq. Motif:	PGN	PCD	TGET	NPAI	DKTGTLT	KGSPE	NKLK	MCTGD	TMFCGDGANDSGALSSADVGLAL
Conservation									
Family 12									
Seq. Motif:	PGD	PCD	TGEA	PPAL	DKTGTLT	KGAPE	NRLK	MVTGD	VGMCGDGANDAGALREADVGVGL
Conservation									
Family 13									
Seq. Motif:	PGD	QCD	TGES	PPAL	DKTGTLT	KGSPE	NRLK	MVTGD	VAMCGDGANDCGALKAAHAGISL
Conservation	*_*	_**	****	****	*******	*****	*_:*	*:***	*.*********.***.*..****
Family 14									
Seq. Motif:	PGD	PCD	TGES	PPAL	DKTGTLT	KGAPE	NKLK	MCTGD	CGFCGDGANDCGALKAADVGISL
Conservation	_**	*.*	****	****	*******	*****	****	*_***	_.******************:**
Family 15									
Seq. Motif:	PGD	PCD	TGES	PPAL	DKTGTLT	KGSPE	NRIK	MVTGD	VGMCGDGANDCGALKAAHVGISL
Conservation	***	***	****	****	*******	*****	*:*	*****	**********************
Family 16									

Table 10 continued

Consensus:	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Seq. Motif:	PGD	PCD	TGES	PPAL	DKTGTLT	KGSPE	NKLK	MVTGD	CGMCGDGANDCGALKTADMGISL
Conservation	_**	_**	****	****	*******	**:**	****	*:***	_.*********_***_**:*:**
Family 17									
Seq. Motif:	IGD	PVD	TGES	PPAL	DKTGTLT	KGAPE	SPLR	ICSGD	VGFCGDGANDCIALKQADVGVSL
Conservation									
Family 18									
Seq. Motif:	PGD	PCD	TGES	PPAL	DKTGTLT	KGSPE	NKLR	MSTGD	VGMCGDGNNDILAIQSANIGIAI
Conservation									
Family 19									
Seq. Motif:	PGD	PCD	TGES	PPSL	DKTGTLT	KGAPE	NPLK	MATGD	__MVGDGANDCGALKQADIGLAL
Conservation									
Family 20									
Seq. Motif:	PGD	PCD	TGES	PPFL	DKTGTLT	KGSPE	NKLK	IISGD	VGMIGDGANDCSAIKQADIGISF
Conservation	_**	*_*	***_	*.__	*******	**:**	*.**	::***	:_*:***:**._:*::.*:****
Family 21									
Seq. Motif:	PGD	PVD	TGES	PLDL	DKTGTLT	KGAPT	AAMR	MLTGD	VLVCGDGVNDIAAMREADVSVAM
Conservation									
Family 22									
Seq. Motif:	PGD	PCD	TGES	PVWT	DKTGTLT	KGAFE	NQIR	ILTGD	TGMVGDGTNDCGALKISHAGISL
Conservation	___	*.*	***:	*.:_	*******	:.___	._::	::***	__*_***_**__*.:_:._.:_.
Family 23	PGD	VVD	TGES	PEGL	DKTGTLT	LGAPE	DPLR	VISGD	VAMIGDGVNDVLSLKQANVGIAM
Seq. Motif:	.__	.._	***:	*_*:	*****:*	:***:	:.__	::***	.*********.:*::_:._.:.*
Conservation									
Family 24	PGD	PAD	TGES	PEGL	DKTGTLT	KGAPE	DTAR	LITGD	TAMVGDGANDAAAIRMADVGIGV
Seq. Motif:	*._	***	****	***:	******:	**.**	*..*	::***	.*.***:*******_*_:*:*:_
Conservation									

Table 10 continued

Consensus: Family 25	**PGD**	**PAD**	**TGES**	**PEGL**	**DKTGTLT**	**KGAPE**	**DPPR**	**MVTGD**	**VAVTGDGVNDSPALKKADIGVAM**
Seq. Motif:	PGD	PVD	TGES	PCPL	DKTGTLT	KAFVD	DEPR	MLTGD	VMMVGDGVNDAPALAAADVGVAM
Conservation	_*_	***	.**.	**.:	*****:*	_.___	*_:_	:_:**	._::***:****.:__*_:**__
Family 26									
Seq. Motif:	PGV	PVD	TGHR	PVAL	NRVGTLT	GEDPE	DRGV	MLSRD	VAMVGDESVMDCLRVADVGILMD
Conservation	**	*.*	****	*::*	**:*:*:	**:*:	****	*****	:*::**.*: * ::**:**:*:.
Family 27									
Seq. Motif:	PGD	PAD	TGES	PCAL	DKTGTLT	IGQPA	DRIR	LLSGD	VLMVGDGINDAPVLAAAHVSVAM
Conservation	___	..*	.**_	***:	****.**	_____	:__:	:_:**	.___***:**_._:__*._*_:_
Family 28									
Seq. Motif:	PGE	PVD	TGAP	PCAL	DLNGTLT	IIGNK	DPLR	ICTGA	VAMVGDAANDATALAASDIGIAV
Conservation	_*_	***	_*__	**:*	*******	_**_*	****	:*:**	***:*******.*:*_**:****
Family 29									
Seq. Motif:	KGD	PVD	TGEA	PCAL	DKTGTIT	IKIGS	NQYR	ILSGD	VMMVGDGLNDAGALAQSNVGISI
Conservation	**:	***	****	***:	*******	*****	*.**	:****	**********:**********
Family 30									
Seq. Motif:	PGD	PAD	TGES	PEEL	DKTGTLT	KGAPE	DPPK	MITGD	VAMTGDGVNDAPALKAAHIGIAM
Conservation	___	.**	:***	*:*:	*****:*	**:**	**_:	:****	**:*******.*:*::*.:****
Family 31									
Seq. Motif:	AGD	PVD	TGES	PTLV	DKTGTLT	KGAEI	DPEI	VVIDD	VALVGDGVNDSPALKKADVSVSL
Conservation	_**	*_*	_*_.	*::_	*__.___	.._.:	____	:__._	.__***__:::____.*.*::.:
Family 32									
Seq. Motif:	AGD	PVD	TGES	SCAL	DKTGTLT	KGEEL	DPVR	MLTGD	VIMVGDGINDSPALSAADVSVAM
Conservation	_..	*.*	***.	_:::	*****:*	_____	:_:*	::***	:_:_*:*:**:_*:__*._::_:

Table 11: Topological types of ten novel prokaryotic families (FUPA23-32) and the range of organisms in which these proteins have been identified.
Note:
[1] I(N): The N-terminal sequence divergent half appears to be derived from a primordial Type I ATPase. II(C): The C-terminal well-conserved half is of the type II typology.

TC#	Type	Organisms
3.A.3.23	II	Actinobacteria Cyanobacteria Firmicutes
3.A.3.24	I(N) II (C)[1] (VII)	Actinobacteria Chlorobi γ, δ-Proteobacteria
3.A.3.25	I	Acidobacteria Actinobacteria Firmicutes Planctomycetes α, β, γ-Proteobacteria Verrucomicrobia
3.A.3.26	I	Actinobacteria α, β, γ-Proteobacteria Spirochaetes
3.A.3.27	I	β, γ-Proteobacteria
3.A.3.28	I	γ-Proteobacteria
3.A.3.29	I	Flavobacteria δ-Proteobacteria
3.A.3.30	II	Bacteroidetes Cyanobacteria α, β, δ-Proteobacteria Spirochaetes
3.A.3.31	I	γ-Proteobacteria
3.A.3.32	I	Actinobacteria Cyanobacteria Euryarchaeota Firmicutes Fusobacteria α β γ δ & ε-Proteobacteria Spirochaetes Verrucomicrobia

REFERENCES

Altschul, S. F., and E. V. Koonin. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci **23:**444-7.

Altschul, S. F., J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schaffer, and Y. K. Yu. 2005. Protein database searches using compositionally adjusted substitution matrices. FEBS J **272:**5101-9.

Anderson, R. T., H. A. Vrionis, I. Ortiz-Bernad, C. T. Resch, P. E. Long, R. Dayvault, K. Karp, S. Marutzky, D. R. Metzler, A. Peacock, D. C. White, M. Lowe, and D. R. Lovley. 2003. Stimulating the in situ activity of *Geobacter* species to remove uranium from the groundwater of a uranium-contaminated aquifer. Appl Environ Microbiol **69:**5884-91.

Artigas, P., and D. C. Gadsby. 2003. Na^+/K^+-pump ligands modulate gating of palytoxin-induced ion channels. Proc Natl Acad Sci U S A **100:**501-5.

Bailey, T. L., and C. Elkan. 1995. The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol **3:**21-9.

Barnes, N., R. Tsivkovskii, N. Tsivkovskaia, and S. Lutsenko. 2005. The copper-transporting ATPases, menkes and wilson disease proteins, have distinct roles in adult and developing cerebellum. J Biol Chem **280:**9640-5.

Blasi F., F. Denti, M. Erba, R. Cosentini, R. Raccanelli, A. Rinaldi, L. Fagetti, F. Esposito, U. Ruberti and L. Allegra. 1996. Detection of *Chlamydia pneumonia* but not *Helicobacter pylori* in Atherosclerotic Plaques of Aortic Aneurysms. J. of Clinical Microbiology **11:** 2766 – 2769.

Carafoli E. 1994. Biogenesis: Plasma membrane calcium ATPase: 15 years of work on the purified enzyme. FASEB J. **13**: 993 – 1002.

Charbit A., J. Reizer and M.H Saier. 1996. Function of the duplicated IIB domain and oligometric structure of the fructose permease of *Escherichia coli.* J Biol Chem. **271:** 9997 – 10003.

Chung, Y. J., C. Krueger, D. Metzgar, and M. H. Saier, Jr. 2001. Size comparisons among integral membrane transport protein homologues in bacteria, Archaea, and Eucarya. J Bacteriol **183:**1012-21.

Clarke, D. M., T. W. Loo, and D. H. MacLennan. 1990. Functional consequences of alterations to amino acids located in the nucleotide binding domain of the Ca^{2+}-ATPase of sarcoplasmic reticulum. J Biol

Chem **265:**22223-7.

Collier J. L. and A. R. Grossman. 1992. Chlorosis induced by nutrient deprivation in *Synechococcus* sp. strain PCC 7941: not all bleaning is the same. Journal of Bacteriology **174**: 4718 – 4726.

Coombs, J. M., and T. Barkay. 2004. Molecular evidence for the evolution of metal homeostasis genes by lateral gene transfer in bacteria from the deep terrestrial subsurface. Appl Environ Microbiol **70:**1698-707.

Dayhoff, M. O., W. C. Barker, and L. T. Hunt. 1983. Establishing homologies in protein sequences. Methods Enzymol **91:**524-45.

Deckert G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay, R. Huber, R.A. Feldman, J.M. Short, G.J. Olsen, R.V. Swanson. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* 1998. Nature **392:** 353 – 358.

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

Dodds W. K., D. A. Gudder and D. Mollenhauer. 1995. The ecology of *Nostoc*. J. Phycol. **31**: 2 – 18.

Dufresne A., M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. Tandeau de Marsac, J. Weissenbach, P. Wincker, Y. I. Wolf and W. R. Hess. 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. PNAS **100**: 10020 – 10025.

Felsenstein, J. 1997. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol **46:**101-11.

Frikha-Gargouri O., R. Gdoura, A. Znazen, N.B. Arab, J. Gargouri, M.B. Jemaa and A. Hammami. 2008. Evaluation and optimization of a commercial enzyme linked immunosorbent assay for detection of *Chlamydophila pneumonia* IgA antibodies. BMC Infectious Diseases **8**: 1471 – 2334.

Fulda S., F. Huang, F. Nilsson, M. Hagemann and B. Norling. 2000. Proteomics of *Synechocystis* sp. strain PCC6803: identification of periplasmic proteins in cells grown at low and high salt concentrations.

Eur. J. Biochem **267**: 5900 – 5907.

Gadsby, D. C. 2007. Structural biology: ion pumps made crystal clear. Nature **450:**957-9.

Garcia-Pichel F. 1998. Solar ultraviolet and the evolutionary history of cyanobacteria. Origins of Life and Evolution of the Biosphere **28**: 321 – 347.

Gerencser, G. A., and J. Zhang. 2003. Existence and nature of the chloride pump. Biochim Biophys Acta **1618:**133-9.

Griffiths E. and R. S Gupta. 2007. Identification of signature proteins that are distinctive of the *Deinococcus-Thermus* phylum. Int. Microbiology **10:** 201 – 208.

Haupt, M., M. Bramkamp, M. Coles, H. Kessler, and K. Altendorf. 2005. Prokaryotic Kdp-ATPase: recent insights into the structure and function of KdpB. J Mol Microbiol Biotechnol **10:**120-31.

Ikeda, M., M. Arai, D. M. Lao, and T. Shimizu. 2002. Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. In Silico Biol **2:**19-33.

Kall, L., and E. L. Sonnhammer. 2002. Reliability of transmembrane predictions in whole-genome data. FEBS Lett **532:**415-8.

Kosky A.A., V. Dharmavaram, G. Ratnaswamy, M.C. Manning. 2009. Multivariate analysis of the sequence dependence of asparagines deamidation rates in peptides. Pharm Res. Epub. Ahead of Print.

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

Kube M., A. Beck, S.H. Zinder, H. Kuhl, R. Reinhardt and L. Adrian. 2005. Genome sequence of the chlorinated compound-respiring bacterium *Dehalococcoides* species strain CBDB1. Nature Biotechnology **23:** 1269 – 1273.

Kühlbrant, W. 2004. Biology, structure and mechanism of P-type ATPases. Nat Rev Mol Cell Biol **5:**282-95.

Liu, J., S. J. Dutta, A. J. Stemmler, and B. Mitra. 2006. Metal-binding affinity of

the transmembrane site in ZntA: implications for metal selectivity. Biochemistry **45:**763-72.

Lutsenko, S., N. L. Barnes, M. Y. Bartee, and O. Y. Dmitriev. 2007. Function and regulation of human copper-transporting ATPases. Physiol Rev **87:**1011-46.

Martinez, R. J., Y. Wang, M. A. Raimondo, J. M. Coombs, T. Barkay, and P. A. Sobecky. 2006. Horizontal gene transfer of PIB-type ATPases among bacteria isolated from radionuclide- and metal-contaminated subsurface soils. Appl Environ Microbiol **72:**3111-8.

Miyashita H., H. Ikemoto, N. Kurano, S. Miyachi and M. Chihara. 2003. *Acaryochloris marina* gen. et sp. nov. (cyanobacteria), an oxygenic photosynthetic prokaryote containing Chl D as a major pigment. J. Phycol. **39**: 1247 – 1253.

Møller, J. V., B. Juul, and M. le Maire. 1996. Structural organization, ion transport, and energy transduction of P-type ATPases. Biochim Biophys Acta **1286:**1-51.

Möller, S., M. D. Croning, and R. Apweiler. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics **17:**646-53.

Nakamura Y., T. Kaneko, S. Sato, M. Ikeuchi, H. Katoh, S. Sasamoto, A. Watanabe, M. Iriguchi, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, N. Nakazaki, S. Shimpo, M. Sugimoto, C. Takeuchi, M. Yamada and S. Tabata. 2002. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. DNA Research **9**: 123 – 130.

Nakamura Y., T. Kaneko, S. Sato, M. Mimuro, H. Miyashita, T. Tsuchiya, S. Sasamoto, A. Watanabe, K. Kawashima, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, N. Nakazaki, S. Shimpo, C. Takeuchi, M. Yamada and S. Tabata. 2003. Complete genome Structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. DNA Research **10**: 137 – 145.

Nelson K.E., R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G Sutton, R.D. Fleischmann, J.A. Eisen, O. White, S.L. Salzberg, H.O. Smith, J.C. Venter and C.M. Fraser. 1999. Evidence for lateral gene transfer between Arcahea and bacteria from genome sequence of

Thermotoga maritima. Nature **399:** 323 – 329.

Odermatt, A., H. Suter, R. Krapf, and M. Solioz. 1993. Primary structure of two P-type ATPases involved in copper homeostasis in Enterococcus hirae. J Biol Chem **268:**12775-9.

Okkeri, J., E. Bencomo, M. Pietila, and T. Haltia. 2002. Introducing Wilson disease mutations into the zinc-transporting P-type ATPase of *Escherichia coli*. The mutation P634L in the 'hinge' motif (GDGXNDXP) perturbs the formation of the E2P state. Eur J Biochem **269:**1579-86.

Olesen, C., M. Picard, A. M. Winther, C. Gyrup, J. P. Morth, C. Oxvig, J. V. Møller, and P. Nissen. 2007. The structural basis of calcium transport by the calcium pump. Nature **450:**1036-42.

Overmann J. and H. van Gemerden. 2000. Microbial interactions involving sulfur bacteria: implications for the ecology and evolution of bacterial communities. FEMS Microbiol. Lett. **24:** 591 – 599.

Overmann, J. and K. Schubert. 2002. Phototrophic consortia: model systems for symbiotic interrelations between prokaryotes. Arch. Microbiol. **177:** 201 – 208.

Poulsen, L. R., R. L. López-Marqués, and M. G. Palmgren. 2008. Flippases: still more questions than answers. Cell Mol Life Sci **65:**3119-25.

Prohaska, J. R., and A. A. Gybina. 2004. Intracellular copper transport in mammals. J Nutr **134:**1003-6.

Ren, Q., K. Chen, and I. T. Paulsen. 2007. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res **35:**D274-9.

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., C. V. Tran, and R. D. Barabote. 2006. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res **34:**D181-6.

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

Saier, M. H., Jr., M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan. 2009. The

Transporter Classification Database: recent advances. Nucleic Acids Res **37:**D274-8.

Sato N. and N. Murata. 1980. Temperature shift-induced responses in lipids in the blue-green alga, *Anabaena variabilis.* The central role of diacylmonogalactosylglycerol in thermo-adaptation. Biochimica et Bipysica Acta **619:** 353 – 366.

Smith, R. F., B. A. Wiese, M. K. Wojzynski, D. B. Davison, and K. C. Worley. 1996. BCM Search Launcher--an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. Genome Res **6:**454-62.

Solioz, M., and J. V. Stoyanov. 2003. Copper homeostasis in Enterococcus hirae. FEMS Microbiol Rev **27:**183-95.

Tandeau de Marsac, N. and J. Houmard. 1993. Adaptation of cyanobacteria to environmental stimuli: new steps towards molecular mechanisms. FEMS Microbiology Reviews **104**:119 – 190

Thever, M. D., and M. H. Saier, Jr. 2009. Bioinformatic Characterization of P-Type ATPases Encoded Within the Fully Sequenced Genomes of 26 Eukaryotes. J Membr Biol.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25:**4876-82.

Tuschak C., J Glaeser and J. Overmann. 1999. Specific detection of green sulfur bacteria by in situ hybridization with a fluorescently labeled oligonucleotide probe. Arch. Microbiol. **171:** 265 – 272.

Tusnady, G. E., and I. Simon. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics **17:**849-50.

Van Baalen C. and R. M. Brown. 1969. The ultrastructure of the marine blue green alga, *Trichodesmium erythraeum,* with special reference to the cell wall, gas vacuoles, and cylindrical bodies. Arch. Mikrobiol. **69:** 79 – 91.

Veldhuis, N. A., A. P. Gaeth, R. B. Pearson, K. Gabriel, and J. Camakaris. 2009. The multi-layered regulation of copper translocating P-type ATPases. Biometals **22:**177-90.

Vinogradov B.D., A.G. Chigaleichik, S.S. Rylkin and M.F. Merkulov. 1976.

Purification and properties of L-glutamine and L-asparagine deaminase from *Pseudomonas aurantiaca* IBPM-14. Prikl Biokhim Mikrobiol. **12:** 704 – 708.

Wang, B., M. Dukarevich, E. I. Sun, M. R. Yen, and M. H. Saier Jr. 2009. Membrane Porters of ATP-binding Cassette (ABC) Transport Systems are Polyphyletic. J Membrane Biol **in press**.

Yamaguchi, A., D. Tamang, and M. Saier. 2007. Mercury Transport in Bacteria. Water, Air, & Soil Pollution **182:**219-234.

Yen, M. R., J. Choi, and M. H. Saier Jr. 2009. Bioinformatic Analyses of Transmembrane transport: Novel Software for Deducing Protein Phylogeny, Topology, and Evolution. J Mol Microb Biotech **17:**163-176.

Zhai, Y., and M. H. Saier, Jr. 2001. 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., and M. H. Saier, Jr. 2001. 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., and M. H. Saier, Jr. 2002. A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins. J Mol Microbiol Biotechnol **4:**375-7.