UNIVERSITY OF CALIFORNIA, SAN DIEGO

The distribution and evolution of protein kinase and phosphatase families in the three superkingdoms of life

A Dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics

by

Kristine Mary Briedis

Committee in charge:

Professor Philip Bourne, Chair

Professor Vineet Bafna

Professor Alexander Hoffmann

Professor William Loomis

Professor Wei Wang

2008

The Dissertation of Kristine Mary Briedis is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2008

To my loving family.

I thank God every day for all of you.

I don’t know how I got so lucky.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people to whom I am indebted for their support and guidance during my graduate studies. I profusely thank my advisor, Dr. Phil Bourne. His direction and encouragement have been invaluable. This research would not have been nearly as complete without his insightful critiques and comments. I am grateful to have been part of his lab.

I also thank the other members of my committee: Dr. Bill Loomis, Dr. Vineet Bafna, Dr. Alex Hoffmann and Dr. Wei Wang. Their comments and suggestions during both my proposition exam and individual conversations have helped shape this research. I very much appreciate their generous donation of time and attention.

I must thank the other members of the Bourne lab, past and present. Specifically, the idea for this project grew out of conversations with Dr. Eric Scheeff and Dr. Song Yang. Numerous other lab members have also provided valuable advice and asked probing questions that oftentimes provoked me to delve more deeply into an area than I might otherwise have. It is a wonderful experience to be around such great minds on a daily basis.

I would be remiss if I did not thank the other students of the Bioinformatics program for their friendship and scholarly discussions. I am especially grateful to Leah Barrera, Chris Benner, Michele Day, Kasey Hutt, Eugene Ke and Coleman

Mosley. My life is richer and graduate school has been a much more enjoyable experience for having known them.

My deepest gratitude and love goes out to my family. They were always quick with words of encouragement and support when needed. I thank my parents for instilling in me the belief that I can accomplish whatever I set my mind to doing. And I thank my little brother for always making me laugh amid his constant reminders that somehow he was the first of us to get a "real job". I'm not really sure if that should be considered "winning". And finally, a special thank you to my loving, late Grandma Briedis. The courage she showed in life is all the inspiration I will ever need. Tev bija taisniba.

Chapter 2 contains the complete reprint of the paper Briedis KM, Starr A, and Bourne PE (2008) Analysis of the human kinome using methods including fold recognition reveals two novel kinases. *PLoS One* 13(2): e1597. I was a primary author of this paper.

VITA

2001	Bachelor of Science, Genetics with Distinction Iowa State University, Ames, IA
2003	Teaching Assistant, University of California, San Diego "Using Internet Resources in Molecular Biology", Pharmacology/Biology 207
2004	Teaching Assistant, University of California, San Diego "Biological Data Representation and Analysis", Pharmacology 201/Bioinformatics I
2008	Doctor of Philosophy, Bioinformatics University of California, San Diego

PUBLICATIONS

Articles

· Briedis KM, Starr A and Bourne PE (2007) Analysis of the Human Kinome Using Methods Including Fold Recognition Reveals Two Novel Kinases. PLoS One 3:1597

· Briedis KM and Bourne PE (In Press) Protein properties: the structural view. Bioinformatics for Protemics, Humana Press

Conference Poster

· The Encyclopedia of Life: a new web resource for domain-based protein annotation data. Conference on Research in Computation Molecular Biology (RECOMB) 2004

ABSTRACT OF THE DISSERTATION

The distribution and evolution of protein kinase and phosphatase families in the three superkingdoms of life

by

Kristine Mary Briedis

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2008

Professor Philip Bourne, Chair

Protein phosphorylation and dephosphorylation plays a critical role in the regulation of many important cellular processes. The protein families responsible for

this, the kinases and phosphatases, have been the focus of enormous amounts of research. However, our knowledge of these families is in many respects still incomplete, as prior studies have oftentimes focused only on humans and other higher eukaryotes. The advent of the genome sequencing era now allows us to examine these protein families on a more global scale. I present here a study of protein kinase and phosphatase families in 115 completely sequenced genomes. This is an important contribution towards understanding not only which families are present in different lineages, but also how the evolution of these families relates to each other.

In chapter 2, I define the human kinome using a method called iGAP. This method combines sequence similarity and fold recognition methods to annotate proteins. I searched the human proteome for members of the eukaryotic protein kinase-like superfamily and identified two novel putative kinases. In subsequent chapters, I extend this focus to include phosphatases and other genomes in the Eukarya, Bacteria and Archaea superkingdoms.

Chapter 3 is centered on phosphatases. I built profile hidden Markov models of known phosphatase families and searched 115 complete proteomes for the presence or absence of these families. I define which genomes and lineages contain particular families and discuss what we can learn about the evolution of the phosphatase families.

In chapter 4, I present a similar study of the kinases. I built models for microbial and eukaryotic kinase families and searched the same 115 proteomes for the presence or absence of the kinase families. I report here the results and discuss the evolutionary implications, incorporating past sequence and structure-based research of the evolution of the protein kinase-like superfamily.

Chapter 5 compares and contrasts the evolutionary patterns of protein kinase and phosphatase families that target either the same substrate or each other. I report the presence or absence of these families in the aforementioned species. I then compare the phylogenetic profiles of these families and discuss how the evolution of each family relates to the other.

1 Introduction

1.1 Protein phosphorylation

Protein phosphorylation and dephosphorylation plays an essential role in regulating a multitude of cellular processes such as gene transcription, cell growth, cell shape and differentiation [1-3]. Protein kinases are responsible for phosphorylating proteins. Their action is countered by protein phosphatases, which dephosphorylate those same protein targets.

Together, protein kinases and phosphatases comprise a significant proportion of genomes. Manning *et al.* [4] found 518 protein kinases in humans, encompassing nearly 2% of the human genome. It is estimated that roughly 150 phosphatases exist in humans [2,3,5].

Historically, more research has focused on kinases than phosphatases. However, the importance of phosphatases to critical cellular functions has grown increasingly clear. Consequently, phosphatases have received more attention in recent years [2].

Initially, kinase and phosphatase research was limited to small-scale, localized studies of specific proteins. However, the more recent explosion of

genome sequencing presents us with the ability to perform large-scale analyses of entire kinomes and phosphatomes.

Following the release of the human genome draft in 2000, attention turned to large-scale sequencing efforts of other organisms in all three superkingdoms of life—Archaea, Bacteria and Eukarya. The Genomes On Line Database (GOLD) [6] is an online resource that tracks current genome sequencing projects. In 2001, the database contained information on 350 sequencing projects [7]. By 2005, this number had grown to 1575 [8]. As of September 2007, the database reported 2905 sequencing projects worldwide [6]. This number is expected to continue quickly growing as advances in genome sequencing technology result in increased efficiency and decreased cost.

As a result of the enormous amount of data being generated from these sequencing projects, many scientific efforts have focused on developing methods to analyze, classify and predict the function of protein sequences. Such tools are vital to bioinformatics researchers and their ability to interpret the raw sequencing data and turn it into useful scientific knowledge.

The theory that proteins with a similar sequence or structure are related and can be used to infer function of unknown proteins has led to the development of sequence-based homology detection tools. A number of techniques have been proposed, including the use of pairwise sequence comparison (e.g. BLAST [9]), profile-based searches to detect more remote homology (e.g. PSI-BLAST [9]),

profile hidden Markov models [10], neural networks [11] and support vector machines [12].

The confluence of an increased knowledgebase of protein kinase and phosphatase families, a large number of sequenced genomes and the development of powerful bioinformatics tools to analyze this data presents us with the unique opportunity to study protein kinases and phosphatases in the context of the tree of life. This allows us to compare and contrast which kinase and phosphatase families are present in the three superkingdoms, and consider the evolutionary implications of such.

To this end, I have utilized several of the aforementioned bioinformatics methods in an effort to classify the presence or absence of protein kinase and phosphatase families in 115 organisms spread across the Eukarya, Bacteria and Archaea superkingdoms. Here, I give a brief introduction to these methods and discuss their relative performance as background to the research discussed in this dissertation.

1.2 BLAST

Originally described in 1990, BLAST (Basic Local Alignment Search Tool) is now a commonly used method to search nucleotide and protein databases for sequences with regions of high similarity to a query sequence [13]. BLAST initially

attempts to find "word" matches of a given length between the query sequence and the sequence database. Matches that score above a given threshold are then extended and reported if the extended alignment meets the user-specified cutoff scoring value. A predefined substitution matrix is generally used in the scoring, allowing for greater consideration to be given to more likely "conservative" substitutions than biologically unlikely replacements of amino acids. This algorithm allows a user to search large databases for similar sequences in a relatively short amount of time [9,13].

1.3 PSI-BLAST

PSI-BLAST (Position Specific Iterated BLAST) was published in 1997 [9]. PSI-BLAST is an extension of the previously described BLAST algorithm. During a PSI-BLAST search, a position specific scoring matrix (PSSM) is constructed from a multiple sequence alignment of the hits above a certain threshold returned from an initial BLAST search. The PSSM (aka "profile") is then used to query the database in a second BLAST search. A user-defined number of BLAST iterations are performed and the profile is refined after each subsequent search [9].

PSI-BLAST allows the database search to favor sequences with highly conserved residues at particular positions, while allowing other positions that are not commonly conserved in a protein family to match a wider range of amino acids.

This type of search is useful in allowing a user to search for more distantly related sequences that may not be found by a single simple BLAST iteration [14].

1.4 Profile hidden Markov models

Profile hidden Markov models (HMMs) are another method that utilizes profiles built from multiple sequence alignments to search for distantly related sequences. Profile HMMs are probabilistic models that can be used to score how similar a sequence is to a given family [15].

A number of programs aimed at applying HMM theory to biological problems have been developed, including SAM [16] and HMMER [10]. The research presented here uses HMMER to construct profile HMMs and search for protein kinase and phosphatase protein families.

A profile HMM uses a multiple sequence alignment to model the distribution of amino acids and the probability that an insertion or deletion may occur at a particular sequence position. A query sequence can be aligned and scored against this model. Profile HMMs provide a more realistic representation of a protein or nucleotide sequence family, as probabilities are determined from the sequences of actual family members. This derivation of such parameters as the transition probability to or from a gap and the probability of the emission of a specific amino acid or nucleotide at a particular position allows for a more biologically relevant

model. For example, profile HMMs can take into account that amino acid insertions may be more likely to occur in surface loops of protein structures and thus are more likely to be hydrophilic residues [10]. Techniques such as the use of sequence weighting, pseudocounts and mixture Dirichlet priors have also been developed as a way to introduce variability into a model to minimize the risk of "over-fitting" or "overtraining" the model to a sequence alignment of highly similar sequences [10,17,18]. Thus, profile hidden Markov models allow researchers to perform sensitive large-scale searches and classifications of proteins and genes.

1.5 Performance of these methods

Several benchmark tests have compared the performance of traditional BLAST to that of profile HMMs in biological sequence classification. One such study compared the ability of PSI-BLAST, SAM (profile HMM method) and BLAST to correctly identify related sequences of known protein structures whose sequence identity to each other was 40% or less. It was found that at a 0.00002% false positive rate, the profile HMMs (using SAM) identified approximately 35% of the related protein sequences and PSI-BLAST found 30%. This was roughly twice the success rate of BLAST, which identified 15% of the known protein evolutionary relationships [19]. Thus, while the traditional pairwise method of BLAST is useful to find clearly related sequences, profile HMMs (and to a slightly lesser extent PSI-BLAST) are better able to identify more distantly related sequences. Given this

success, profile HMMs have been used in the construction of a number of commonly used community resources to classify large datasets of protein sequences, including Pfam [20], Superfamily [21] and PANTHER [22].

I have utilized the aforementioned methods to search completed (though not necessarily final drafts) of 115 organisms of the Eukarya, Bacteria and Archaea superkingdoms for the presence and absence of protein kinase and phosphatase families. I present here the results and a discussion of the evolutionary implications of this research.

2 Analysis of the Human Kinome Using Methods Including Fold Recognition Reveals Two Novel Kinases

Protein sequence similarity is a commonly used criterion for inferring the unknown function of a protein from a protein of known function. However, proteins can diverge significantly over time such that sequence similarity is difficult, if not impossible, to find. In some cases, a structural similarity remains over long evolutionary time scales and once detected can be used to predict function.

Here we employed a high-throughput approach to assign structural and functional annotation to the human proteome, focusing on the collection of human protein kinases, the human kinome. We compared human protein sequences to a library of domains from known structures using WU-BLAST, PSI-BLAST, and 123D. This approach utilized both sequence comparison and fold recognition methods. The resulting set of potential protein kinases was cross-checked against previously identified human protein kinases, and analyzed for conserved kinase motifs.

We demonstrate that our structure-based method can be used to identify both typical and atypical human protein kinases. We also identify two potentially novel kinases that contain an interesting combination of kinase and acyl-CoA dehydrogenase domains.

2.1 Introduction

Most proteome-wide functional annotation focuses on sequence similarity, however, this ignores valuable information that protein structure can provide--an important consideration in the era of structural genomics when many more protein structures are becoming available [23]. In some cases, the sequence between two proteins has diverged too far to find any significant sequence similarity with current methods, but a structural similarity can still be seen [24-26]. For example, Hon *et al.* crystallized the aminoglycoside phosphotransferase APH(3')-IIIa and found a surprising homology to eukaryotic protein kinases (ePKs) [27]. About half of the sequence folded into a structure typical of ePKs, despite a very low sequence identity. The major structural differences were found in the area of the protein that determined substrate specificity [27]. Likewise, Holm and Sander found two glucosyltransferases that shared less than 10% sequence identity, but still contained strong structural similarities that indicated evolutionary relatedness [28]. These two examples illustrate that the structures of proteins can reveal surprising similarities that are undetected by sequence identity alone. Notwithstanding, one must be cautious in assigning relatedness based on structural similarity alone. It is possible for two proteins with a similar structure to function in different ways. For example, lysozyme and α-lactalbumin have similar structures and a 40% sequence identity, but differ in function [29]. It is also possible for proteins to arrive at a similar structure through convergent rather than divergent evolution. Subtilisin and chymotrypsin are

serine endopeptidases that share a catalytic triad, but no other sequence or fold similarity [29].

We have established a high-throughput approach to provide accurate structure and functional annotation termed the Encyclopedia of Life (EOL) [30], based on the desire to annotate a large number of sequenced proteomes. EOL uses a pipeline approach termed the integrated Genome Annotation Pipeline (iGAP), which we have applied in examining the set of human kinases, the human kinome, in an attempt to uncover distant homologs not previously seen.

2.2 Integrated genome annotation pipeline (iGAP)

iGAP (Figure 2.1**)** compares already identified protein sequences from whole proteomes against a comprehensive structure fold library (FOLDLIB). The fold library was built from a combination of Protein Data Bank (PDB) protein chains [31] and protein domains defined by SCOP [32] and PDP [33]. SCOP domain sequences were filtered at 90% identity. Since there is a delay between protein structures being added to the PDB and classified by SCOP, PDB chains were clustered at 90% identity, parsed with PDP, and added to the SCOP domains to generate a more complete library. The collection of SCOP, PDP and PDB sequences were then clustered at 90% identity to determine the final FOLDLIB composition [30].

The core of the pipeline consists of tools that search for sequence and fold similarity, including the sequence comparison programs WU-BLAST [34] and PSI-BLAST [14], and the threading program 123D [30,35]. Protein sequences from completed proteomes were first compared to FOLDLIB using WU-BLAST. Then, PSI-BLAST profiles were generated for each input protein sequence using three iterations and a default H-value of 1e-06. Lastly, the protein sequences were compared to FOLDLIB using the fold recognition program 123D [30].

The result is a set of putative structure and function assignments including a novel statistical measure of reliability (Shindyalov *et al.* unpublished). Reliability is defined using a consensus approach with SCOP as a benchmark. Using a test set of non-redundant SCOP folds, Shindyalov *et al.* counted the number of consistently and inconsistently predicted assignments by WU-BLAST for each target sequence. The hits were binned by E-value and the specificity was averaged over all values in the bin, resulting in a reliability assignment. Reliability is defined as the number of positions with consistent predictions divided by the total number of positions having two or more hits to the same SCOP fold.

Using this method, it is found that the probability of traditional E-value assignments being correct varies between proteomes since they are not random, and indeed are not random in different ways. For example, using WU-BLAST to assign SCOP folds to proteomes, to reach a level of 1 error per 1000 annotations, one must use an E-value cutoff of $1x10^{-8}$ for *Arabidopsis thaliana* but only $1x10^{-2}$ for

Caenorhabditis elegans. EOL individually benchmarks every genome and assigns a reliability index that can be used to compare different genomes. The reliability index is set by determining the E-values required for a sequence to be consistently identified with a fold and binning the hits by E-value. The resultant reliability index is termed A through E and corresponds to 99.9%, 99%, 90%, 50%, and 10% specificity, respectively [30].

2.3 The human kinome

We utilized this pipeline to characterize the collection of human protein kinases. Eukaryotic protein kinases (ePKs) regulate signal transduction reactions in the cell, influencing many processes including metabolism, apoptosis and transcription [4].

The collection of kinases has previously been defined by several groups including Cheek *et al*. [36] and Manning *et al*. [4]. Cheek *et al.* searched multiple species for all enzymes that catalyze the transfer of an ATP terminal phosphate group, while Manning *et al*. focused on both typical and atypical protein kinases in humans. Atypical protein kinases (aPKs) were defined by Manning *et al.* as proteins that have weak sequence similarity to the ePKs, but still have protein kinase activity.

Since our study focuses on the human protein kinase superfamily, we compared our results with that of Manning *et al.* [4]. They published the "complete"

human kinome paper in 2002 based on homologies detected using Hidden Markov Models (HMMs). HMMs were developed by Manning *et al.* for the ePK family and the PIKK, RIO, ABC1, PDK, and alpha kinase atypical families. The HMMs were used to search against Genbank, SwissProt, dbEST, Celera human genome, Incyte LifeSeqGold, and internal SUGEN and Pharmacia sequence databases. Full-length gene predictions were determined for putative kinase hits, and confirmed in most cases by cDNA cloning [4].

Our approach differs in several ways. By including the threading program 123D, we incorporate fold recognition along with sequence similarity, possibly leading to the identification of more distant homologs. We also searched Ensembl's [37] draft assembly 34 v19.34.a.1 of the human genome, which presumably differs from the genome draft used by Manning *et al.* in 2001-2002.

Utilizing Hidden Markov Models (HMMs) along with EST and cDNA data, Manning *et al.* found 518 human protein kinases. This accounts for almost 2% of all human genes, and makes protein kinases one of the largest eukaryotic gene families [4]. Most human kinases contain a eukaryotic protein kinase (ePK) catalytic domain. This catalytic domain shows remarkable conservation, specifically with respect to critical residues and motifs, as previously described by Hanks and Hunter [38]. However, the HMM method employed by Manning *et al.* is only one approach to identifying specific protein families across a whole proteome. We thus compared the human kinome as classified by the EOL pipeline to the Manning set. We determined

that our method performs well in classifying the kinome and we present here two putative novel kinases.

2.4 Methods

Assembly 34 v19.34.a.1 of the Ensembl [37] human genome draft was run through iGAP [30], including WU-BLAST [34], PSI-BLAST [14], and the threading program 123D [35]. The subset predicted to contain the protein kinase superfamily was selected for further study. Protein kinase domains are generally 250-300 amino acids in length [38]. Thus, our set of candidate proteins was filtered to exclude near-identical sequences and those shorter than 200 amino acids to exclude proteins that despite a short sequence or structural similarity cannot contain a full, active kinase domain. Since it was unknown at the beginning of the study how sensitive iGAP would be in identifying full kinase domains, we selected proteins with a predicted kinase domain of 120 amino acids (roughly half the length of a typical protein kinase domain) or greater for further study. To ensure we didn't miss any abnormally short kinases, we also included any proteins that did not meet the above criteria, but appeared to contain at least two conserved subdomains from Hanks and Hunter's ePK domain analysis [38].

The proteins found were mapped to the Manning *et al.'s* human kinome using BLAST at a 90% sequence identity cutoff point. This strict threshold was set so

proteins were not erroneously mapped to each other. However, it was done with the understanding that given human genome draft changes, some proteins may fall below this identity threshold that should be considered equivalent to each other.

Of the remaining 324 potentially unique proteins, 234 were selected that matched to a kinase domain by 123D, in hopes of exploiting any distant structural similarities that would be overlooked when considering sequence alone. Many of these predictions were at a lower reliability and were deemed false positives. These false positives likely share some structural, but not functional, similarity to the kinase fold. Sequences of "A" or "B" reliability were analyzed for conserved kinase domain motifs and blasted against NCBI's NR database [39]. Including the aforementioned sequences that contained Hanks and Hunter ePK subdomains [38], our final data set consisted of 153 sequences (Table S1).

2.5 Analysis of the human kinome using iGAP

Overall, the human kinase set identified here by EOL agreed with the set of kinases found by Manning *et al.* In addition, we analyzed 153 potential novel protein kinase sequences (selected as described in the Methods section) using Pfam [40] and found 44 contained an assignment for either an ePK or atypical kinase domain. Based on these Pfam results, our sequences were classified into the following groups (followed by the sequence count in parentheses):

choline/ethanolamine kinase (5), fructosamine kinase (2), protein kinase (20), PI3_PI4 kinase (17) and not kinase (109) (See Supplementary Tables S1 and S2 for data).

Most of the differences between our human kinome and that found by Manning *et al.* can be attributed to analyzing a different draft of the human genome. Only one kinase exists in both Manning *et al.'s* human kinome and our Ensembl human genome draft that EOL did not identify (LRRK2 UniProt:Q5S2007 [41]). Upon further investigation, it was discovered that the Ensembl LRRK2 protein was only 400 amino acids long in our draft, and was missing the protein kinase domain. Ensembl lengthened the LRRK2 sequence in a subsequent draft to 2527 amino acids, including the protein kinase domain. Ten other protein kinases from the Manning *et al.* kinome match proteins in our set at a lower score than our cutoff for mapping, probably due to using slightly different gene predictions and data sets. These ten proteins, upon closer inspection, were manually mapped to the Manning *et al.* kinome. For example, ENSP00000330379 has a 98% local sequence identity to EphA10 in Manning *et al.'s* human kinome, but is 462 amino acids shorter. It is annotated in Ensembl as EphA10 precursor. The ten proteins, along with reasons for their poor mapping, are described in further detail in Supplementary Table S3.

Some of the kinases identified by EOL are from protein families that are part of the protein kinase-like SCOP superfamily (d.144.1), but are not classified in the "protein kinases, catalytic subunit" family. This includes the atypical kinase families

actin-fragmin kinases, MHCK/EF2 kinases, phosphoinositide 3-kinases, choline kinases, aminoglycoside phosphotransferases, and the RIO1-like kinases [32]. Some of these EOL kinases were not present in Manning *et al.'s* set, but were already deposited and identified in NCBI's [39] Non-Redundant database (NR) as kinases. In an effort to pinpoint the source of differences, we looked at the methods used by Manning *et al.* to classify the sequences. Manning *et al.'s* paper states that they developed Hidden Markov Models (HMMs) for some of the atypical families, including PIKK, RIO, ABC1, PDK, and alpha kinase [4]. In comparison, the EOL search included only those atypical kinases present in the ePK superfamily, as defined by SCOP (d.144.1) [32]. Thus, it is not surprising that the EOL human kinome contains a different set of atypical kinases than Manning *et al.'s* kinome. For example, to the best of our knowledge Manning's group did not build an HMM to look for choline/ethanolamine kinases. EOL's human kinome, however, correctly classified five such proteins (SCOP family d.144.1.8) in the human proteome.

2.6 Identification of two novel kinases

Here we focus on two particularly interesting potential kinases that were classified by Ensembl as acyl-CoA dehydrogenase family members. A BLAST search against NR showed these proteins to be ACAD10 [UniProt: Q6JQN1; Ensembl: ENSP00000325137] and ACAD11 [UniProt:Q709F0; Ensembl:ENSP00000264990].

ACAD10 has been previously identified as being involved in the β-oxidation of fatty acids [42]. EOL recognized the acyl-CoA dehydrogenase domain, but also assigned a kinase domain as part of the sequence. 123D produced the strongest kinase hit, with a weaker hit from PSI-BLAST. Figure 2.2 shows an alignment of the protein to common kinase motifs. Clearly, the nucleotide position loop and Brenner's phosphotransferase motif [43] are well conserved. Less well conserved is the choline kinase motif. It is interesting to note, however, that some of the most critical functional residues of choline kinases as identified by Yuan *et al.* are conserved [44].

ACAD11 is 279 amino acids shorter than ACAD10, and has a similar arrangement of acyl-CoA dehydrogenase and kinase domains. The difference in length is mostly attributable to a hydrolase domain that is present in ACAD10, but not ACAD11. A BLAST alignment between ACAD10 and ACAD11 shared a 46% sequence identity overall (excluding the hydrolase domain), and a 48% sequence identity in the kinase domain. At the time of our initial study, the protein corresponding to ACAD11 in Ensembl did not contain a kinase domain. However, it has since been lengthened in a subsequent release and appears to contain a kinase domain with similar features to ACAD10, as shown in Figure 2.2 (see Supplementary Figure S1 for a longer alignment).

The kinase domains of ACAD10 and ACAD11 appear to be most similar to a choline kinase or an aminoglycoside phosphotransferase (APH) domain. The

similarity between the APH and choline kinase families was previously noted by Scheeff and Bourne [45] in a study of the structural evolution of the protein kinase-like superfamily. Structural analysis revealed conservation in their C-terminal subdomains that was not observed to exist in other kinase families [45]. EOL, Pfam, and Superfamily [21] annotate the protein kinase domains of ACAD10 and ACAD11 as an APH domain with higher confidence than a choline kinase domain, however, the aforementioned similarities to the choline kinase motif are intriguing.

Choline kinases phosphorylate choline to produce phosphocholine [46]. This pathway eventually produces phosphatidylcholine, a component of cell membranes. Choline kinase is a particularly important atypical kinase as it has been shown to play a role in several types of cancer. Over-activity of choline kinase and increased concentrations of phosphocholine have been identified in breast cancer cells [47]. Increased phosphocholine levels have also been reported in prostate and brain tumors [48].

Aminoglycoside phosphotransferases (APHs) are also an interesting atypical kinase family, present in bacteria. As previously mentioned, Hon *et al.* revealed a surprising structural similarity between APH and eukaryotic protein kinases (ePK) [27]. APHs have been implicated in antibiotic resistance. They phosphorylate aminoglycoside hydroxyl groups. In bacteria this can result in inactivation of aminoglycoside antibiotics such as kanamycin and gentamicin. However, APHs have also been shown to phosphorylate some ePK substrates. Daigle *et al.*

demonstrated that two APHs had the ability to phosphorylate some Ser/Thr protein kinase substrates, though at a slower rate than aminoglycoside phosphorylation [49]. This could perhaps offer an explanation as to how a kinase domain with similarities to APHs would function in eukaryotes.

The domain arrangement of ACAD10 shown in Figure 2.3 was the only human protein identified as such in the Superfamily database [21]. Superfamily and Pfam found proteins with the same domain structure in *Mus musculus* (mouse), *Caenorhabditis elegans* (worm), *Caenorhabditis briggsae* (worm), *Bos taurus* (cow), *Ciona intestinalis* (sea squirt), *Macaca mulatta* (rhesus monkey), *Monodelphis domestica* (opossum) and *Pan troglodytes* (chimp) [21,40]. Similar proteins also exist in bacteria [50].

2.7 Conclusion

In conclusion, we have utilized both sequence and structure-based tools to annotate the human kinome. We were successful in identifying both ePK and atypical kinases. We were particularly intrigued by ACAD10 and ACAD11, which contain acyl-CoA dehydrogenase and apparent kinase domains. The cellular function of such a combination of domains and the level of kinase activity for these proteins remains to be determined.

Acknowledgements

Chapter 2 contains the complete reprint of the paper Briedis KM, Starr A, and Bourne PE (2008) Analysis of the human kinome using methods including fold recognition reveals two novel kinases. *PLoS One* 13(2): e1597. I was a primary author of this paper.

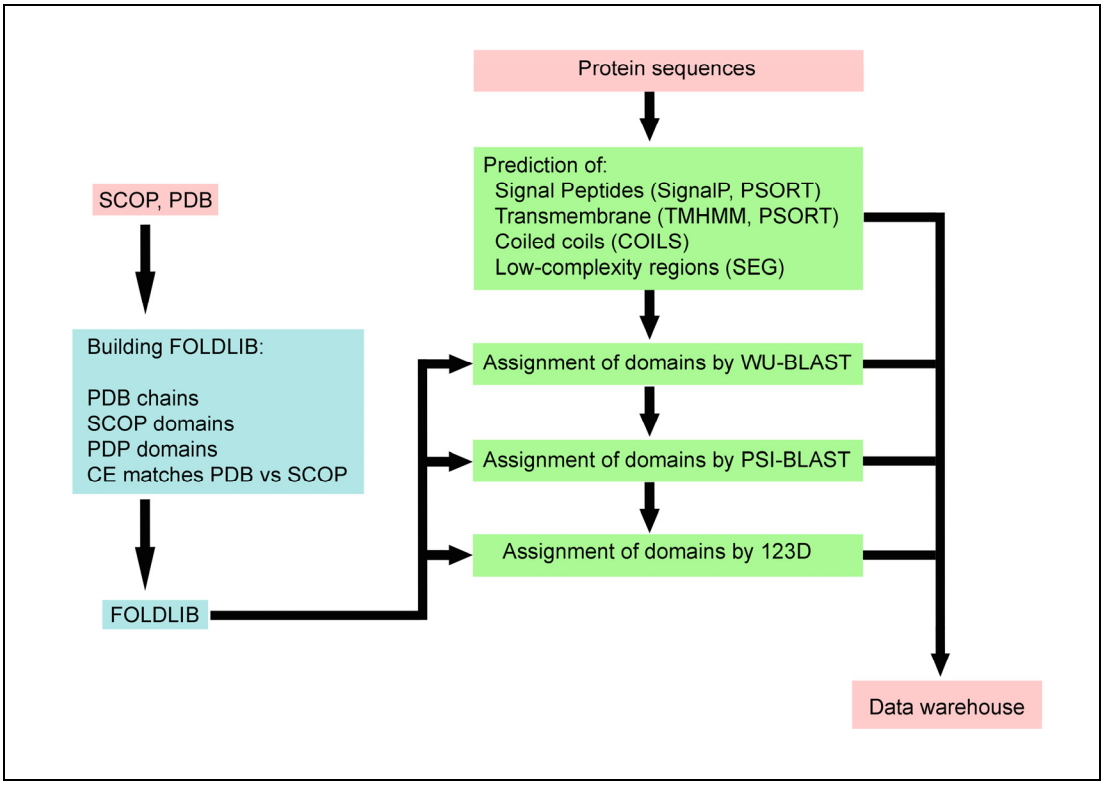

Figure 2.1 Diagram of the iGAP pipeline.

Protein sequences are compared to a domain library using WU-BLAST, PSI-BLAST, and 123D.

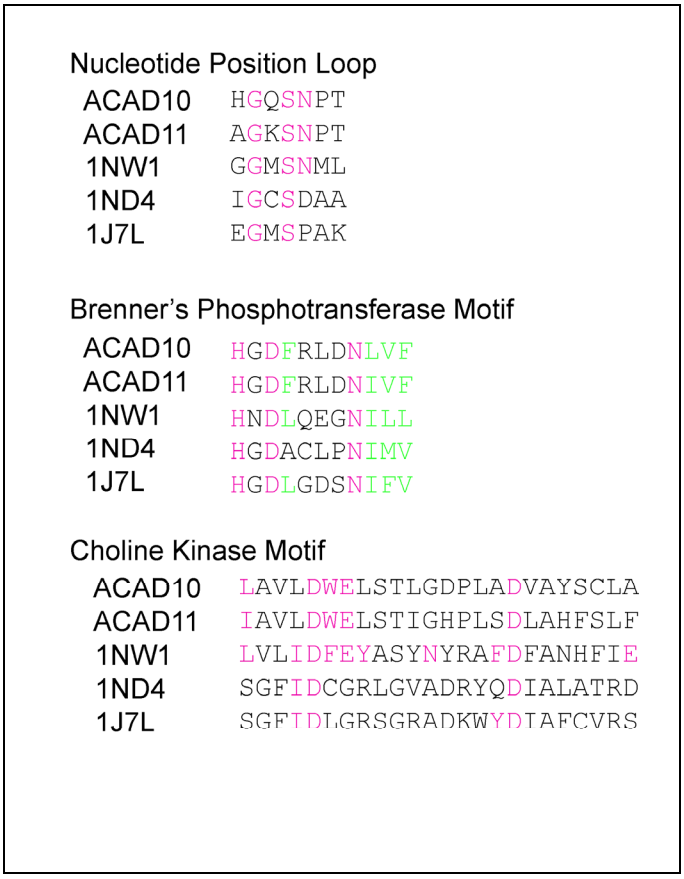

Figure 2.2 Conserved kinase motifs in ACAD10 and ACAD11.

ACAD10 and ACAD11 contain conserved kinase motifs such as the nucleotide position loop, a phosphotransferase motif, and part of a choline kinase motif. Residues in pink are highly conserved; residues in green are commonly large hydrophobic amino acids. ACAD10 and ACAD11 are aligned with the choline kinase 1NW1 and aminoglycoside phosphotransferases 1ND4 and 1J7L for comparison.

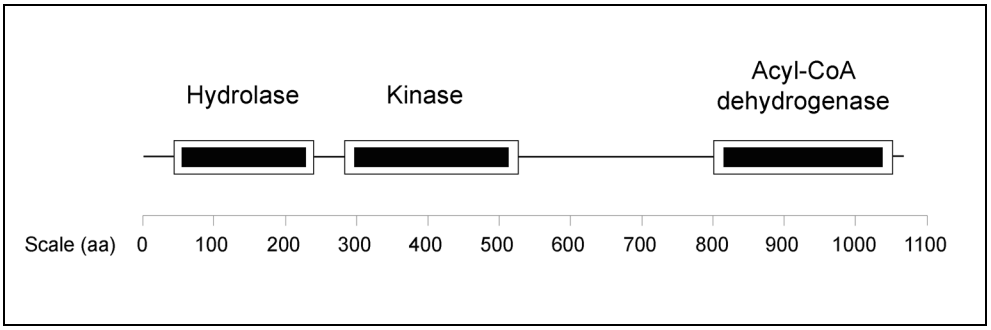

Figure 2.3 Domains identified in ACAD10.

iGAP identified hydrolase, kinase, and acyl-CoA dehydrogenase domains in ACAD10.

3 Protein Phosphatases in the Eukarya, Bacteria, and Archaea Superkingdoms

3.1 Introduction

Protein phosphorylation and dephosphorylation are essential components of cellular activity [51]. Protein kinases function by phosphorylating tyrosine, serine and threonine residues. Their counterparts, protein phosphatases, are responsible for the opposite dephosphorylation activity. This action is critical to regulating a number of cellular processes including cellular proliferation, differentiation and metabolism [52]. Historically, protein phosphatases have not received as much attention as protein kinases. However, in recent years, this important group of proteins has begun to be the focus of increased research efforts [2].

Protein phosphatases can be generally classified by substrate specificity into two main groups: protein tyrosine phosphatases (PTPs) and serine/threonine phosphatases [52]. They can be further subdivided based on sequence and structure similarities, as described elsewhere in this chapter.

Previous phosphatase evolutionary studies have been focused on sequence and, in some cases, structural differences [53-55]. The advent of complete genome sequencing provides us with a unique opportunity to study the evolution of protein

phosphatase families from a global phylogenetic perspective. To this end, I present here a study that defines the presence or absence of phosphatase families in completed genomes across the Eukarya, Archaea and Bacteria superkingdoms and discusses the evolutionary implications of such.

3.2 Genome-wide phosphatase searches

Proteome drafts of completely sequenced eukaryotic (listed in Appendix A), bacterial (listed in Appendix B) and archaeal (listed in Appendix C) organisms were collected and searched for the presence of protein phosphatase families. Putative phosphatases were identified through literature review as well as searches conducted using both NCBI BLAST [9] and HMMER [10]. BLAST is a pairwise alignment tool that finds areas of local similarity between sequences, and is commonly used to identity protein family homologs [9]. HMMER is a program that builds profile Hidden Markov Models (HMMs), which are probabilistic models built from profiles of multiple sequence alignments of a protein family [10]. In benchmark tests, profile HMMs have been shown to be adept at identifying protein family members [16].

HMMs representing phosphatase families were collected from Pfam [20] and PANTHER [22,56], community resources that manually curate and provide models for different protein families. In addition, profile HMMs were constructed for smaller subfamilies that were not represented in the Pfam or PANTHER databases.

Protein sequences known to belong to the subfamilies were collected via literature and NCBI BLAST database searches. As much as possible, initial seed sequences were confined to those phosphatases with published experimental evidence. When necessary, sequences with clear conservation of critical residues and domains were also included to expand the training data set. Multiple sequence alignments were built using ClustalW2 [57] and manually inspected and adjusted for known conserved sequence motifs. HMMs were built and calibrated with the HMMER package version 2.3.2 [10]. Models were run against a negative control data set of phosphatase sequences from other families to help estimate the scoring threshold below which false positives were likely to occur.

The top scoring hit in each genome was manually evaluated for conservation of known sequence motifs and critical residues, and was BLASTed against NCBI's nonredundant (NR) database. In cases where a putative family member was not found, lower scoring matches were also manually evaluated, up to an E-value of 2.0. The resulting presence or absence of families was plotted on eukaryotic, bacterial and archaeal phylogenetic trees derived from NCBI's taxonomy database [58].

3.3 Protein tyrosine phosphatases in Archaea and Bacteria

Originally, protein phosphorylation and dephosphorylation was thought to be an exclusively eukaryotic function. It was hypothesized that this functionality

evolved to meet the needs of more complex multicellular organisms. However, in the late 1980s and early 1990s, both serine/threonine and tyrosine phosphatases were identified in viruses and Bacteria [59].

Several Archaea groupings contained putative PTPs (Figure 3.1). Four of the five Crenarchaeota species appeared to have PTPs. The Euryarchaeota phylum was more mixed, with two of three *Methanosarcina* genomes having possible PTPs, as well as the related three *Pyrococcus* and *Thermococcus kodakarensis* organisms. The lone Nanoarchaeota did not appear to contain any PTPs.

PTP-like proteins were scattered throughout the bacterial genomes as well (Figure 3.2). Two of the four Cyanobacteria, four Proteobacteria, two Firmicutes and the *Deinococcus radiodurans* species contained putative PTPs. There were also a few weak hits in other genomes (not shown), where the PTP active site has been conserved. However, the rest of the protein sequences were too divergent to be able to conclude whether the proteins were truly protein tyrosine phosphatases.

3.4 Eukaryotic PTP groups

Eukaryotic protein tyrosine phosphatases can be grouped based on substrate specificity and sequence and structure similarity into four classes: the Class I PTPs, which include the classical PTPs and the dual-specificity phosphatases; the Class II low-molecular weight PTPs; the Class III CDC25-like phosphatases; and the Asp-

based PTPs (Table 3.1) [2]. The class I, II and III PTPs all use a similar catalytic mechanism in which a cysteine residue plays a critical role in a nucleophilic attack on the phosphate, leading to its hydrolysis and the formation of a phosphocysteine intermediate (Figure 3.3) [52]. It has been suggested that this catalytic mechanism is an example of convergent evolution among these three classes [60,61]. The Asp-based PTPs, conversely, utilize an aspartate as the nucleophile [62]. The classical PTPs specifically dephosphorylate phosphotyrosine residues, while the dual-specificity phosphatases are able to target both tyrosine and serine/threonine phosphorylated amino acids [2].

3.5 Class I Cys-based protein tyrosine phosphatases

3.5.1 Classical protein tyrosine phosphatases

The so-called "classical" PTPs can be subdivided into two groups: receptor-like protein tyrosine phosphatases (RPTPs) and non-receptor protein tyrosine phosphatases (NRPTPs). Both are considered to have evolved from a common ancestor and are structurally classified in the same SCOP family (c.45.1.2), along with the dual-specificity phosphatases [2,32]. All The RPTPs contain a transmembrane region and are represented in humans by at least 21 genes. 17 human genes encode the intracellular NRPTPs [2].

3.5.2 RPTPs

RPTPs consist of an extracellular domain, a transmembrane region, and a PTP domain [2]. The extracellular domains of RPTPs vary in length and functional domain content, with some proteins containing fibronectin III-like repeats, immunoglobulin-like domains and/or glycosylation sites [63]. In addition, five human RPTP subgroups (R1/R6, R2A, R2B, R4 and R5) contain two protein tyrosine phosphatase domains [2]. The membrane-proximal PTP domain is named D1 and the membrane-distal domain is termed D2. In almost all cases, the D2 domain is missing critical residues and is not catalytically active. However, the D2 domain still serves an important role by contributing to the stability and specificity of the RPTP and is involved in RPTP dimerization [64,65].

RPTP dimerization and regulation is still not completely understood. Representatives of several RPTP subgroups have been shown to have the ability to dimerize, but it is not yet certain if this is true of every RPTP [66]. There is evidence that RPTP dimerization can affect activity (both active and inactive conformations have been identified) and ligand binding properties of the enzymes [66-68].

The D1 and D2 domains of the RPTPs LAR and CD45 have shown fairly high structural and sequential similarity to each other [69,70]. A BLAST alignment of the D1 and D2 LAR domains has 46% sequence identity. Furthermore, the PTPase activity of the normally catalytically inactive D2 domain of LAR was experimentally restored by mutating only two residues [69].

RPTPs were found only in metazoan genomes (Figure 3.4), consistent with previous RPTP discoveries [71]. In addition, I analyzed the PTP domain arrangement of the putative RPTPs. All metazoans were found to contain at least one RPTP with both a single and a tandem PTP domain arrangement.

3.5.3 NRPTPs

NRPTPs contain very similar catalytic phosphatase domains to the D1 domain of RPTPs, but lack the transmembrane and extracellular regions [72]. In fact, several RPTPs have even been shown to exist in an alternative isoform as cytoplasmic PTPs [53]. Many NRPTPs also contain additional domains that determine cellular location [72].

NRPTPs were found in almost every eukaryotic genome (Figure 3.4). However, no classical PTPs were located in a grouping of parasitic genomes (*Plasmodium falciparum, Plasmodium yoelii, Theileria parva* and *Theileria annulata*) and the diatom *Thalassiosira pseudonana*.

3.5.4 Dual-specificity phosphatases

The dual-specificity phosphatases (DSPs) differ from the classical PTPs in their ability to dephosphorylate not only tyrosine residues, but also serine/threonine

residues. This additional ability is reflected in their protein structure. The catalytic cleft of DSPs is shallower than that of tyrosine-only PTPs. This allows the shorter phosphoserine and phosphothreonine residues to access the catalytic cysteine. The deeper pocket of classical PTPs restricts access to only the longer phosphotyrosine [54]. Human DSPs can be classified into the following seven groups: classical DSPs (aka MAPK phosphatase or MKP), slingshots (SSHs), phosphatase of regenerating liver (PRL), cell division cycle 14 (CDC14), phosphatase and tensin homolog (PTEN), myotubularins and the atypical DSPs [2]. They target a wide variety of substrates, some of which are described in the following sections.

3.5.5 Classical dual-specificity phosphatases

The "classical" dual-specificity phosphatase family is comprised of phosphatases that dephosphorylate and inactivate mitogen-activated protein kinases (MAPK) [2]. MAPK signal transduction pathways are crucial to cellular function. Disruption of proper MAPK signaling can lead to the development of cancer or other diseases [73]. MAPKs contain both tyrosine and threonine regulatory residues that undergo phosphorylation. Dual-specificity phosphatases have been identified that target both phospho-residues. In addition, both tyrosine and serine/threonine-specific phosphatases have been located that can target the phosphorylation sites individually [74]. The dephosphorylation of either residue inhibits MAPK activity [54].

The classical DSPs differ from other DSPs (often termed the "atypical" DSPs) in that the classical DSPs contain an inactive rhodanese-like domain which assists the phosphatase in targeting MAPKs. Based on analysis of sequence divergence, it is thought that the classical DSPs represent a younger family that arose after an atypical DSP acquired a rhodanese domain, perhaps in response to an increased need for efficient MAPK regulation [54]. Classical DSPs were present in every metazoan genome (Figure 3.4). All proteins contained both a DSP domain and a rhodanese-like domain.

3.5.6 CDC14 phosphatases

CDC14s are involved in cell cycle regulation in eukaryotes [75]. It is essential for mitotic exit in *Saccharomyces cerevisiae*, and is involved in regulating cytokinesis in *Schizosaccharomyces pombe* [76]. CDC14 targets a number of substrates, including many cyclin-dependent kinase (CDK) substrates [75].

Putative CDC14s were found in all fungi and animals except the reduced genome of *Encephalitozoon cuniculi* (Figure 3.4). In addition, no strong candidates for CDC14s were found in *Entamoeba histolytica*, *Dictyostelium discoideum* or the Apicomplexa parasites: *Plasmodium falciparum*, *Plasmodium yoelii*, *Theileria annulata* and *Theileria parva*. A candidate CDC14 was found in the related ciliate *Tetrahymena thermophila*. It contains 10 of 11 conserved, critical CDC14 residues

as described by Gray *et al.* [76]. A BLAST search against the NR database revealed a 46% sequence identity to human CDC14.

The plant pathogens *Phytophthora ramorum* and *Phytophthora sojae* both contain a putative CDC14, with 84% and 86% sequence identity respectively to the known CDC14 that is expressed during sporulation in *Phytophthora infestans* [77]. There is also a protein in *Thalassiosira pseudonana* with a 42% BLAST sequence identity to *P. infestans* CDC14 but it contains only 6 of the 11 commonly conserved CDC14 residues. The catalytic PTP domain, however, is well conserved, indicating it is likely some kind of dual-specific phosphatase.

Interestingly, a CDC14-like sequence was found in the green algae *Chlamydomonas reinhardtii* but not the higher plants *Arabidopsis thaliana* or *Oryza sativa*. The *C. reinhardtii* protein contains 10 of the 11 aforementioned critical CDC14 residues. It also has a 52% BLAST sequence identity with mouse CDC14. In February 2008, Kerk *et al.* also noted this apparent CDC14 in the *C. reinhardtii* genome [78]. They agree that the protein contains many hallmarks of a CDC14.

3.5.7 Slingshot phosphatases

The slingshot, or SSH, family of phosphatases was first identified in 2002 [79]. Three SSH genes have been found in humans (SSH-1, SSH-2, and SSH-3) [80]. *Drosophila*, however, contains only one known SSH gene. SSHs contain three

highly conserved N-terminal protein domains, termed A, B, and P (catalytic phosphatase) domain [81]. The phosphatase domain is thought to be distantly related to mitogen-activated protein kinase phosphatases [79]. The C-terminal domain of SSHs is more varied, with the exception of a serine-rich motif present in human and mouse but not *Drosophila* [81].

Slingshot phosphatases play an important role in cellular actin rearrangement. Cofilin/ADF (actin depolymerizing factor) proteins sever actin filaments and increase the rate at which actin monomers leave the pointed end of an actin filament [82]. When cofilin/ADF is phosphorylated by a LIM or TES kinase, it is unable to bind actin, and thus unable to depolymerize actin. The dephosphorylation of cofilin/ADF allows it to resume actin-depolymerization activities [83]. Cofilin/ADF is dephosphorylated by both slingshot phosphatases [79] and chronophin [84]. There is also evidence that SSH phosphatases dephosphorylate LIM kinases, resulting in downregulation of LIM kinase activity towards cofilin/ADF [80]. While all three SSH proteins dephosphorylate cofilin, their cellular location and expression patterns differ, suggesting they fill different roles in cellular and developmental functions [81].

Slingshot phosphatases were found in only higher eukaryotes, as shown in Figure 3.4. No slingshot genes were found in other eukaryotes, including *Arabidopsis thaliana, Caenorhabditis elegans, Caenorhabditis briggsae* and *Saccharomyces cerevisiae*, indicating that other mechanisms such as chronophin

(discussed in section 5.6.5) bear responsibility for cofilin/ADF dephosphorylation activity in these species.

3.5.8 PRL phosphatases

The phosphatase of regenerating liver (PRL) family has three known members in humans: PRL-1, PRL-2 and PRL-3 [85]. The amino acid sequences of PRLs have been fairly well conserved between family members. Human PRL-1 and PRL-2 share the highest sequence identity, 86%. PRL-1 and PRL-3 have a 78% sequence identity, and PRL-2 and PRL-3 are least similar at 75%. PRLs also show high sequence identity between mammals. The human and mouse PRL-3 sequences share a 96% sequence identity, while PRL-1 and PRL-2 sequences are 100% identical [85]. However, while similar in amino acid sequence [86], the three PRLs localize in different tissues. In mouse and rat studies, PRL-1 was found mostly in the brain and muscle, PRL-2 localized to skeletal muscle, and PRL-3 was expressed in both skeletal and cardiac muscle [87].

The specific function of PRLs has proved difficult to ascertain. Among phosphatases, PRLs show greatest sequence similarity to the CDC14 (20.3%) and PTEN (16.9%) proteins [86]. PRLs contain a C-terminal prenylation motif (CAAX, where “A” stands for an aliphatic or aromatic amino acid), unique among other protein tyrosine phosphatases [88,89], and can be modified by farnesylation [90].

They also contain an alanine instead of the usual serine or threonine next to the catalytic arginine in the P loop of the phosphatase domain [89]. PRLs, particularly PRL-3, have been implicated in different human cancers. PRL-3 expression has been found to be up-regulated in colorectal and breast carcinomas. PRL-3 expression levels also appear to be a useful biomarker in predicting whether cancer has metastasized to other locations in the body [88]. There have been fewer studies regarding the expression levels of PRL-1 and PRL-2 in relation to cancers. However, Wang *et al.* reported similar expression levels of the three PRLs in colorectal carcinoma samples, indicating PRL-1 and PRL-2 may also play a role in cancer progression [91].

Putative PRLs were found in several eukaryotic groups (Figure 3.4). All metazoans, including fly, human, mouse, chicken, frog, zebrafish and worm contained at least one PRL protein. In addition, hits were found in a number of protists, including *Dictyostelium discoideum, Trypanosoma brucei, Plasmodium falciparum* and *Phytophthora ramorum*. The marine diatom *Thalassiosira pseudonana* sequence hit did not have a CAAX box at the C-terminal. However, a subsequent search against a more recent draft of the genome (JGI v3.0) successfully found an updated sequence that did contain the prenylation site. A putative *Entamoeba histolytica* PRL was also identified. It had strong BLAST hits to PRLs in other species, although it contained only one aliphatic/aromatic amino acid in the CAAX box, making the assignment slightly questionable. Recently, a single copy of PRL was found in the malaria parasite *Plasmodium falciparum* [89]. My results

confirm this as the only PRL protein in the current draft of the genome, and also find PRLs present in related Apicomplexa species. No PRL protein was found in the ciliate *Tetrahymena thermophila*. In addition, no PRL proteins were found in fungal genomes.

3.5.9 PTENs

Phosphatase and tensin homolog (PTEN) was originally discovered in humans and classified as a tumor suppressor gene [92]. It has since been heavily studied and found to preferentially target phosphatidylinositol 3,4,5-trisphosphate (PIP_3) [92,93]. PIP_3 is produced when a phosphatidylinositol 3-kinase phosphorylates phosphoinositide-4,5-bisphosphate PIP_2. PIP_3 then continues as part of a cascade that ultimately affects cell growth, migration, survival and metabolism [94]. Loss of PTEN function has been linked to human cancer. Without proper regulation by PTEN, PIP_3 accumulates in the cell and leads to increased activation of its downstream signals, causing oncogenic cell growth [93]. PTEN contains a phosphatase domain, a PIP_2 binding domain, two PEST homology domains, and a PDZ binding domain [94]. An interesting feature of the PTEN catalytic phosphatase domain is the conservation of two lysines. Most PTENS include the consensus active site motif of HCKAGKGRTG [95,96].

Alonso *et al.* also group the phosphatases transmembrane phosphatase with tensin homology (TPTE), TPTE and PTEN homologous inositol lipid phosphatase (TPIP), the tensins (tensin , tensin2, and tensin3), and C1-TEN in the PTEN-related PTP subgroup [2]. TPTE is similar to PTEN in its phosphatase and C2 domains. However, it appears to be an inactive phosphatase. Leslie *et al.* were able to restore phosphatase activity to TPTE by mutating two residues in the phosphatase domain [97]. TPIP, published in 2001, is closely related to TPTE. However, at least one splice form of TPIP has been shown to have phosphoinositide phosphatase activity [98].

C1-TEN and tensin have high sequence identity and are similar in domain organization. Both protein families contain a phosphatase domain, a Src homology 2 (SH2) domain and a phosphotyrosine binding (PTB) domain [99]. However, C1-TEN also contains a N-terminal cysteine-rich C1 domain, and tensin has a N-terminal region that interacts with actin [99,100]. Similarly to PTEN, C1-TEN negatively regulates the Akt/PKB signaling cascade, resulting in inhibition of cell survival, migration and proliferation. Tensin is missing an active site cysteine residue that is critical for phosphatase activity [99]. A study of tensin knockout mice suggests that tensin is critical to proper renal function, muscle regeneration and cell migration [101].

The PTEN family is well represented throughout the eukaryotes. Most genomes had strong hits to the PTEN models and were supported by analysis of

sequence motifs and domain content (Figure 3.4). Two variations were identified in fungal genomes on the above mentioned active site consensus motif. Most of the Saccharomycetes contained a PTEN sequence with a methionine residue in place of alanine. In addition, on a wider scale, almost all of the Fungi studied also had a serine substituted for threonine. These substitutions were previously known to exist in the *Saccharomyces cerevisiae* PTEN homolog [102]. My results suggest this slight sequence divergence took place early in Fungi evolution, as they appear to be conserved in a number of species.

Curiously, a possible PTEN hit in the fungi *Phanerochaete chrysosporium* (sequence jgi|3655) had an apparent insertion in the middle of the active site. A multiple sequence alignment of PTEN family proteins revealed a 17 amino acid insertion between the first lysine residue and alanine. Otherwise, the motif was well conserved in this sequence. If this insertion truly exists in the protein catalytic site, it brings into question whether the protein can function normally. Alternatively, it might simply result from an inaccurate protein or gene prediction.

No PTEN family proteins were detected in the fungi *Encephalitozoon cuniculi* or in a cluster of the following four species: the malarial parasites *Plasmodium falciparum* and *Plasmodium yoelii*, and the bovine parasites *Theileria annulata* and *Theileria parva*.

3.5.10 Myotubularin phosphatases

The myotubularin family contains lipid phosphatases. They dephosphorylate the D3 position of PI(3)P and PI(3,5)P_2 [103]. The first myotubularin gene (*MTM1*) was identified in yeast in 1996 [104]. Currently, the human genome is known to contain 14 myotubularin genes. However, six of these genes code for catalytically inactive proteins. These inactive myotubularins have been shown to associate with, and possibly regulate, catalytically active myotubularins [105]. Myotubularins contain multiple protein domains, including PTP and pleckstrin homology (PH)-GRAM domains [106]. Myotubularins generally conform to the consensus motif of VHCSDGWDRT, though the inactive subgroups deviate somewhat [96,107].

Despite having very similar substrate specificity, catalytically active myotubularins have been shown to have different roles and are not functionally redundant with each other [103]. Thus, when mutated, myotubularins can lead to multiple diseases, including Charcot-Marie-Tooth disease and X-linked myotubular myopathy (XLMTM) [105,108].

Most eukaryotic genomes were found to contain at least one myotubularin protein (Figure 3.4). The majority conserved the VHCSDGWDRT motif, with a few minor differences. Roughly half of the fungal species substituted isoleucine for valine, and one species (*Candida glabrata*) had a leucine residue instead of valine. Two other eukaryotes also had an isoleucine instead of a valine (*Caenorhabditis elegans* and *Tetrahymena thermophila*), and one had a leucine substituted for the

valine (*Thalassiosira pseudonana*). The grouping of *Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania major* all contain a threonine in place of the serine.

There were a few genomes in which no putative myotubularins were identified: *Aspergillus nidulans*, *Neurospora crassa* and *Chlamydomonas reinhardtii*. These three species all have related organisms in the eukaryotic tree that contain putative myotubularins, and it is unclear whether they have lost the gene or if it is the result of an incomplete proteome prediction or a missed assignment. As with several other phosphatase families mentioned in this study, *Encephalitozoon cuniculi* has no apparent myotubularin protein. There is also no apparent myotubularin protein present in a cluster of four protist genomes: *Plasmodium falciparum*, *Plasmodium yoelii*, *Theileria annulata* and *Theileria parva*. It is possible that the myotubularin gene was lost in the common ancestor of these four species.

3.6 Class II Cys-based protein tyrosine phosphatases

3.6.1 Low molecular weight protein tyrosine phosphatases (LMPTP)

Low molecular weight protein tyrosine phosphatases (LMPTP) are one of the most conserved families of PTPs [2,109]. Bacterial LMPTPs have been isolated from several genomes, including *Acinetobacter johnsonii* and *Escherichia coli.* These sequences share roughly 30% sequence identity with human LMPTPs [109].

LMPTPs are encoded by the polymorphic *ACP1* gene [2]. While some genomes seem to encode a single LMPTP, Rudbeck *et al.* found two isoforms of LMPTP proteins present in cow, pig, and five species of fish (plaice, cod, eel, ray and shark) in 2003 [110]. Allelic variants of LMPTP have been linked to a number of diseases, including diabetes, Alzheimer's disease and rheumatoid arthritis [2].

There are two motifs commonly found in LMPTPs. The general CX_5R PTP motif is CXGNXCR in LMPTPs. There are also two conserved residues, DP, that are generally followed by one, and oftentimes two, tyrosines. These motifs have been found to be conserved in both bacterial and eukaryotic LMPTPs [109].

LMPTPs were present in a majority of eukaryotic genomes, specifically in the fungi, animal and plant groupings (Figure 3.5). Initially no hits to the *Ciona intestinalis* genome were found, but a subsequent search against an updated genome assembly revealed the presence of a LMPTP containing the signature motifs mentioned above.

The *Caenorhabditis briggsae* genome had a weak hit that contained the DP motif, but only partially matched the CXGNXCR motif. The *C. briggsae* sequence CBP03401 was FPGNICR, substituting a phenylalanine in place of the catalytically critical cysteine. Thus, it is doubtful that this protein is a catalytically active phosphatase. The *Caenorhabditis elegans* genome did not contain any LMPTPs. An updated version of the genome draft has a sequence that has an 81% sequence identity to the *C. briggsae* protein, but is 51 amino acids shorter and missing the

region that the CXGNXCR motif would occur in. It seems likely that these two proteins are homologous, and the *C. elegans* protein prediction may be incorrectly truncated.

The *Danio rerio*, *Cryptococcus neoformans*, *Debaromyces hansenii*, *Candida albicans* and *Canis familiaris* genomes had weak hits to the LMPTP model. Upon further analysis, it was observed that these sequences appeared to be truncated protein predictions. The dog protein shared an 89% BLAST sequence identity with human LMPTP, but was only 83 amino acids long (compared to 158 amino acids for human LMPTP) and did not contain the CXGNXCR motif. Likewise, the zebrafish protein was 120 amino acids long and had a 77% BLAST sequence identity to human LMPTP. It too was missing the CXGNXCR motif. The *C. neoformans, D. hansenii* and *C. albicans* proteins contained the DP motif, but were predicted to begin with a methionine residue present just after where the CXGNXCR motif should have ended. Low-molecular weight phosphatase models from other sources (Pfam [20] and SMART [111]) also classified these proteins as LMPTPs. Thus, it is suspected these uncertain cases may be due to an artifact of the genome sequencing and protein prediction process rather than a loss of LMPTP.

Aspergillus nidulans and *Yarrowia lipolytica* also had weak hits to the LMPTP model. The sequences both clearly contained the DP motif followed by two tyrosines, but did not contain the CX_5R motif. Low-molecular weight phosphatase models from Pfam and SMART also hit these sequences, but given the lack of a

commonly conserved catalytic motif, it is unlikely that they are truly functioning LMPTPs.

Several genomes had no hits to the LMPTP model. *Encephalitozoon cuniculi* did not appear to contain a LMPTP. No LMPTPs were found in some of the protists, including the Euglenozoa branch, the Alveolates branch, or the genome of *Thalassiosira pseudonana*. A recent phosphatase study of *Leishmania major, Trypanosoma cruzi and Trypanosoma brucei* (the Euglenozoa grouping) also found no LMPTPs in these genomes, supporting my results [112].

In the bacterial genomes, there were some clear LMPTPs as well as weaker hits to other proteins that appeared to more likely function as arsenate reductases (Figure 3.2). All Cyanobacteria genomes studied included putative LMPTPs. In addition, roughly half of the organisms in the Firmicutes and Proteobacteria phylums contained LMPTPs. Putative LMPTPs were also found in two archaeal species: *Thermococcus kodakarensis* and *Pyrococcus furiosus* (Figure 3.1). It is also worth noting that there were additional, weaker hits in several organisms. These proteins appear to be arsenate reductases, which are thought to be related to the LMPTPs [113]. This apparent evolutionary relationship is discussed further in section 3.10.

3.7 Class III Cys-based protein tyrosine phosphatases

3.7.1 CDC25 phosphatases

CDC25 is the lone subgroup present in the Class III cysteine-based protein tyrosine phosphatases [2]. Humans contain three CDC25 genes: CDC25A, CDC25B and CDC25C. CDC25s are critical for cell cycle progression. They are responsible for dephosphorylating, and thus activating, cyclin-dependent kinases (CDK) [114]. All three human isoforms are currently thought to be involved in the G1-S and G2-M cell cycle transitions [115,116]. CDC25s also play a role in cellular response to DNA damage [115]. Overexpression of all three human CDC25 isoforms has been observed in various cancers, including pancreatic cancer, colorectal cancer, breast cancer and non-Hodgkin lymphoma [115,117].

The sequence of the N-terminal region shows low conservation in CDC25s. Phosphatase activity, protein expression levels and association with other proteins is controlled by phosphorylation and ubiquitination sites present in this region. The C-terminal is more highly conserved and includes the catalytic domain. While CDC25s contain the (HCX_5R) motif common, the rest of the sequence shows little homology to other PTPs [118]. CDC25s target phosphotyrosine and phosphothreonine residues [2].

Putative CDC25s were found in all fungi and animal genomes (Figure 3.5), but not in the Archaea or Bacteria superkingdoms. Initially, the CDC25 models did

not locate a CDC25 in either *Takifugu rubripes* or *Tetraodon nigroviridis*. However, after rerunning the models against updated genome drafts, putative CDC25s were found in both. CDC25s were also identified in *Dictyostelium discoideum* and *Entamoeba histolytica*.

The presence or absence of CDC25 homologs in plant genomes has been a matter of much debate [119]. A CDC25-like protein has previously been identified in *Arabidopsis* and shown to bind CDKs *in vitro*. However, experiments by Dhankher *et al.* suggest that the protein actually functions as an arsenate reductase [120]. A recent study of two similar CDC25-like protein in rice was also inconclusive, with experiments showing *in vitro* phosphatase activity, but *in vivo* arsenate reductase function [121]. Further experimental study is needed to clarify these discrepancies.

Interestingly, a possible CDC25 was identified in the plant pathogen *Phytophthora sojae*. The protein contains the conserved CDC25 motifs DCR and CE(Y/F)SXXR [122,123]. A similar CX_5R motif is present in arsenate reductases, thus it is possible that this protein is simply an arsenate reductase. However, it is worth noting that a BLAST search against NCBI's Non-Redundant (NR) database showed the highest similarity to other CDC25s, including a 41% sequence identity with mosquito CDC25. In addition, the sequence was run against arsenate reductase HMMs curated by Pfam [20] and PANTHER [124]. Neither matched the *P. sojae* protein, while PANTHER's HMM specific to CDC25 did with an e-value of 2.5e-51.

3.8 Asp-based protein tyrosine phosphatases

3.8.1 Eyes absent phosphatases

The Eyes Absent (Eya) phosphatase family was originally studied as a necessary protein for *Drosophila* eye development. It has since been shown to function not only as a transcription factor, but also as a phosphatase. Humans contain four Eya genes, while *Drosophila* have only one [62].

Eya is a member of the haloacid dehydrogenase (HAD) superfamily. Although Eya has been shown to have protein tyrosine phosphatase (PTP) activity, it uses a different catalytic mechanism. Most PTPs use a catalytic cysteine as a nucleophile, but HAD phosphatases use an aspartate instead [125]. There has also been speculation that Eya could additionally act as a serine-threonine phosphatase, but this requires further study [62,126].

Eya phosphatases contain two conserved domains. The C-terminal Eya domain (ED) is 271 amino acids long and is highly conserved in the Eya family. The N-terminal contains a somewhat less conserved Eya domain 2 (ED2), present in the middle of a heavily proline-serine-threonine area [62]. Eya also contains two MAPK phosphorylation sites. Eya activity in *Drosophila* is positively regulated by phosphorylation of these sites [127].

Strong hits were found to the Eya models in higher eukaryotes, including worms, vertebrates, insects and plants (Figure 3.5). No putative Eya phosphatases

were identified in Archaea or Bacteria. Eya homologs were previously published in *Arabidopsis thaliana* and rice [128,129]. The Eya models found these homologs, but only produced a very weak hit to *Chlamydomonas reinhardtii*, a green algae. A February 2008 study by Kerk *et al.* also failed to identify any Eya homologs in *C. reinhardtii* [78]. The protein found does not contain the commonly conserved catalytic site, suggesting it is not an active phosphatase, but does seem to be somewhat similar to other Eya proteins. A BLAST search against NR revealed that a portion of the C-terminal in the *C. reinhardtii* protein matches other Eya phosphatases with a roughly 30% sequence identity. This same 100 amino acid stretch of the *C. reinhardtii* protein shares a 35% sequence identity with the *Arabidopsis* homolog, but shows no apparent similarity to the rest of the protein.

Another interesting result of this search was fairly strong hits to the Eya models in both *Phytophthora ramorum* and *Phytophthora sojae*, plant pathogens. An alignment of these protein sequences to known Eya proteins supports this finding, as several motifs commonly conserved in Eya proteins were found in the *Phytophthora* sequences (Figure 3.6). A BLAST search revealed a 33% and 32% sequence identity between the *P. sojae* and *P. ramorum* proteins and a rat Eya protein, respectively. To the best of our knowledge, possible Eya homologs have not previously been reported outside of metazoan or plant genomes.

3.9 Serine/Threonine phosphatases

Serine/threonine phosphatases dephosphorylate phosphoserine and phosphothreonine residues. They can be broadly classified into three main groups—the phosphoprotein phosphatase (PPP) family, the metal-ion dependent protein phosphatase (PPM) family, and the more recently identified FCP family [3,51,130]. Thus far, the only known function of the FCPs is to dephosphorylate RNA polymerase II [3]. This study concentrated on the PPM and PPP families. The PPP family is the larger of the two families in eukaryotes and can be further subdivided, as described below [51]. The two groups are thought to have evolved separately [131] and are structurally classified into different superfamilies by SCOP (Structural Classification of Proteins) (Table 3.2) [32].

3.9.1 PPP family

While members of the PPP family have been found in all three superkingdoms, PPP family members have different characteristics in eukaryotes and bacteria. Bacterial PPPs tend to have broader substrate specificity and have not been found in some completed genomes [59,132]. There is a group of PPPs in eukaryotes that appears to be more closely related to these bacterial PPPs than to conventional eukaryotic PPPs. Andreeva *et al.* [132] studied this group and discovered that all bacterial PPPs and "bacterial-like" eukaryotic PPPs contain the consensus motif

(I/L/V)D(S/T)G, which is not present in other eukaryotic and archaeal PPPs. Likewise, in a 2001 study, Kennelly [51] found an average of 27-31% sequence identity between the catalytic cores of conventional eukaryotic and archaeal PPPs, and only 17-19% between eukaryotic/archaeal PPPs and bacterial PPPs. PPP function is generally regulated through the addition of varying targeting and regulatory domains to the catalytic core [51].

The eukaryotic PPP phosphatases can be grouped into several distinct subfamilies: PPP1, PPP2A (aka PPP2), PPP2B (aka calcineurin or PPP3) and "non-conventional" or "novel" phosphatases that have more recently been identified (eg-PP4, PP5, PP6 and PP7) [130]. These newer phosphatase groups can in some cases be classified as subfamilies based on sequence similarity [51,52]. This study focuses on the more well-known groupings of PPP1, PPP2A, PPP2B and their associated subfamilies as classified by Kennelly (Table 3.3) [51].

As previously mentioned, the specificity and function of PPPs is largely dependent on its interaction with other subunits. For example, the catalytic domains of PP1 and PP2A have high phosphatase activity and low specificity. PP1 alone has over 50 regulatory proteins that it can associate with [133]. This allows PPPs to be involved in a number of different cellular functions.

My research supports the prevailing thought that PPPs exist in all eukaryotic genomes, but not necessarily in all bacterial or archaeal genomes [51,132]. Strong hits were found to putative PPP1 and PP2A phosphatase families in all eukaryotic

genomes studied (Figure 3.7). PP2Bs were found to be in most, but not all, eukaryotic genomes (Figure 3.7). No PP2B proteins were detected in the plants (*Arabidopsis thailiana*, *Oryza sativa* and *Chlamydomonas reinhardtii*). This is consistent with several previous studies that have also failed to find any PP2B homologs in plants [78,134]. PP2B proteins were also not found in cow parasites *Theileria parva* and *Theileria annulata*. The other two parasites in the same phylum, Apicomplexa, have apparently conserved their PP2Bs, as they showed strong hits to putative PP2Bs. No putative PP2Bs were found in the marine diatom *Thalassiosira pseudonana*. Lastly, while PP1 and PP2A were found in *Encephalitozoon cuniculi*, it did not contain any PP2Bs.

Evidence of putative PPPs was found in some, but not all, archaeal genomes (Figure 3.1). Each of the five species examined in the phylum Crenarchaeota appeared to contain a PPP. In the Euryarchaeota phylum, results were mixed, with possible PPPs present in seven genomes. In the bacteria, putative PPPs were detected in three of the four cyanobacteria, only one in the Firmicutes phylum (*Bacillus anthracis*), *Deinococcus radiodurans*, *Thermotoga maritima*, and roughly half of the Proteobacteria genomes (Figure 3.2).

3.9.2 PPM family

The PPM phosphatases contains the PP2C/PPM1 family [52]. Structural studies have indicated that the PPM and PPP phosphatases have evolved separately [135]. In contrast to PPPs, PPMs are monomeric enzymes [136].

With the exception of *Encephalitozoon cuniculi*, putative PP2Cs were found in every eukaryotic genome studied (Figure 3.7). Conversely, there was only one possible instance of a PPM protein in an archaeal genome, *Thermoplasma volcanium* (Figure 3.1). The ORF for this putative PPM was noted by Kennelly to be the only known PPM in the nine archaeal genomes published as of 2003 [137].

In Bacteria, hits to possible PPM phosphatases were mixed (Figure 3.2). Genomes in the Actinobacteria, Firmicutes, Spirochaetes and Cyanobacteria phylums contained putative PP2Cs. The hits in the Proteobacteria phylum were more scattered. In 2002, Kennelly [59] reported on several Bacteria genomes that contained no apparent PPM ORFs. However, my study identified possible PPMs in a few of these species. For example, a *Mycobacterium leprae* sequence (gi 15826883) contains 11 of 14 commonly conserved residues found in eukaryotes, including four aspartic acid residues critical to binding Mg^{2+} or Mn^{2+} necessary for catalysis [3]. A putative PPM was also found in *Deinococcus radiodurans* (gi 15807498) with 11 of 14 conserved amino acids, including the important aspartic acids. Similarly, a putative PP2C was identified in *Mycoplasma genitalium* that

Kennelly had not found (gi 12044960). This protein also had 11 of 14 conserved residues.

3.10 Overall evolution of the phosphatases

A number of evolutionary observations can be made from these findings, both on a global and a more local scale. Here, I first consider the serine/threonine phosphatases. Then, I discuss the tyrosine phosphatases followed by an examination of several specific lineages and species.

Serine/threonine phosphorylation is clearly an ancient trait, but differences can be noted between the PPP and PPM families. Given the abundant presence of PP2C phosphatases in eukaryotes and the virtual absence in Archaea, it seems clear that PPM phosphatases emerged after the divergence of Archaea. As previously noted, the lone putative PP2C in Archaea is present in *Thermoplasma volcanium.* This singular Archaea PP2C was previously noted by Kennelly in a study of nine archaeal genomes [137]. My analysis of an additional 14 Archaea species (23 in total) supports the theory that *T. volcanium* most likely acquired this protein through a horizontal gene transfer. The other serine/threonine phosphatase class, the PPPs, appear to be older than the PPMs. PPPs were found in all eukaryotes and many archaeal and bacterial genomes studied. Based on the widespread presence of PPPs in all three superkingdoms and in light of past sequence homology studies [132,138],

it can be suggested that the PPP ancestor may have even been present in the last common ancestor of Eukarya, Bacteria and Archaea. A sequence homology study including the more newly identified PPP subgroups would help shed light on this possibility.

The protein tyrosine phosphatases, while likely older than PPMs, are harder to date with respect to the PPPs but they do show different characteristics among the distinct classes of conventional PTPs, LMPTP, CDC25s, and Eya phosphatases. Eya and CDC25, believed to have arisen separately from the conventional PTPs [60,61], appear to be the youngest tyrosine phosphatases, as they are present only in eukaryotes.

There has been debate over the evolution of conventional PTPs and LMPTPs. PTPs and LMPTPs share a common catalytic motif of CX_5R and use a similar catalytic structural mechanism, but have no other sequence homology [139]. It has been alternatively suggested that the two groups have either evolved through convergent evolution or from a very distant ancestor through circular permutation [137,140-144]. In my study, both groups were found in all three superkingdoms, suggesting they may be of relatively similar age. However, there is another evolutionary question surrounding the LMPTPs that involves their relation to arsenate reductases.

As noted previously, the LMPTP models had weaker hits to arsenate reductase proteins in several archaeal and bacterial genomes. Bennett *et al.* [113]

previously noticed similarities between a *Bacillus subtilis* arsenate reductase and the LMPTPs. The arsenate reductase had roughly 18% sequence identity with a mammalian LMPTP, shared some structural similarities and contained the CX_5R motif. In addition, the *B. subtilis* arsenate reductase showed a very limited amount of PTPase activity *in vitro*. *In vivo*, arsenate reductase reduces arsenate to arsenite [145].

The bacterial arsenate reductases aligned in Bennett *et al.* share the CXGNXCR motif found in LMPTPs, but do not contain the LMPTP DPYY motif [113]. Given the aforementioned similarities and the propensity of the LMPTP models to weakly hit arsenate reductases, these results seem to support the theory that arsenate reductases and LMPTPs are likely related. However, further experimental characterization of the putative LMPTPs and arsenate reductases, particularly in the lesser studied Archaea, is required to fully characterize these families. Such study is also needed to determine whether arsenate reductases may have low levels of phosphatase activity *in vivo.*

On a subfamily level, the PRLs and slingshot phosphatases both show an interesting history. PRLs were present in almost all eukaryotic genomes, with the notable exception of plant and fungi genomes. The complete absence of this protein family in both lineages suggests it was likely lost by their respective common ancestors, as opposed to lost individually in every genome.

While this study generally focused only on the presence or absence of phosphatase families in genomes, I did delve further into the total number of proteins present in several phosphatase and kinase families, including the slingshot phosphatases. Multiple slingshot genes were found in all vertebrates studied, but only one slingshot gene was present in insects and *Ciona intestinalis*. The timing of this apparent gene duplication suggests the possibility that the multiple copies of slingshot directly resulted from the proposed genome duplications thought to have taken place sometime in early vertebrate divergence [148]. This idea is further discussed in section 5.6.6.

On a more local scale, several eukaryotic genomes appear to have lost multiple phosphatase families. The Apicomplexa species (*Plasmodium falciparum*, *Plasmodium yoelii*, *Theileria parva* and *Theileria annulata*) had no detectable classical PTP, PTEN, CDC14 or myotubularin family members. Interestingly, the related ciliate *Tetrahymena thermophila* did contain representatives of these families. Thus, it appears that the Apicomplexa genomes may have lost a significant number of phosphatases. Perhaps the most likely explanation for such a loss would be the parasitic nature of the Apicomplexa. All four genomes are obligate parasites, which in general have been shown to endure significant gene loss through evolution.

The Microsporidian *Encephalitozoon cuniculi* was also missing a number of phosphatase families. Similar to the aforementioned cases, *E. cuniculi* is a parasite and contains a very compacted genome [146]. It is unknown whether the genome

has truly lost all of these families or if some of the sequences have simply diverged too far for the models to detect. However, given that *E. cuniculi* has the smallest known eukaryotic genome [147], it is not unreasonable to conclude that it likely contains a much more streamlined collection of phosphatases than other eukaryotes.

3.11 Conclusion

This work has presented a comprehensive analysis of protein phosphatase families present in 115 genomes spanning the Eukarya, Bacteria and Archaea superkingdoms. The study included 12 known protein tyrosine phosphatase families and 4 serine/threonine phosphatase families. I also compared the phylogenetic patterns of evolution of the families and discussed what they may indicate in terms of the overall evolutionary history of the phosphatases. This is an example of what we can learn as a greater number of genomes are sequenced. In the future, we will be presented with continued opportunities to study and refine our knowledge of the protein phosphatases and their evolution.

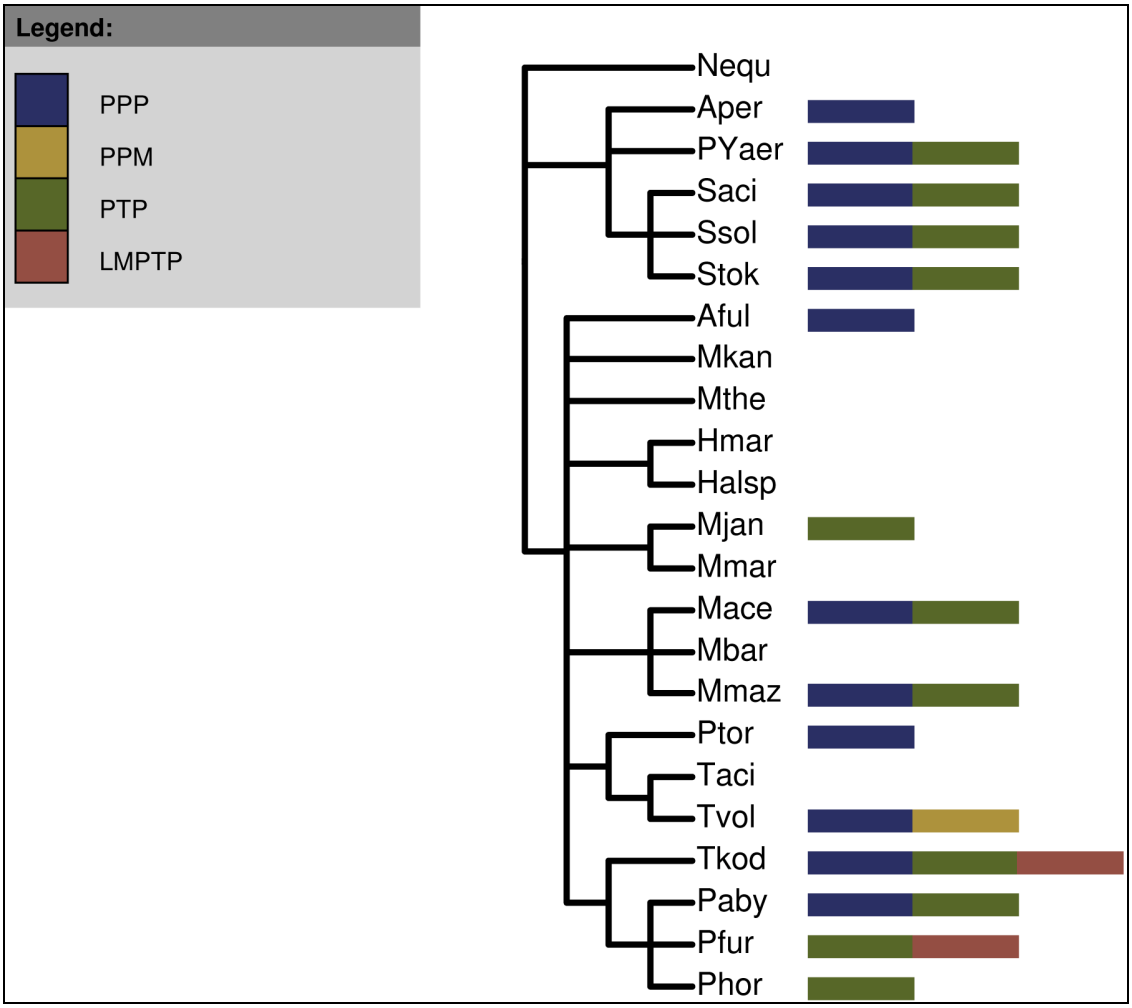

Figure 3.1 Archaea phosphatases.

PPP, PPM, PTP and LMPTP protein phosphatase families present in the Archaea species. Blue bars represent the PPP family, gold bars the PPM family, green bars the PTP family, and red bars the LMPTP family.

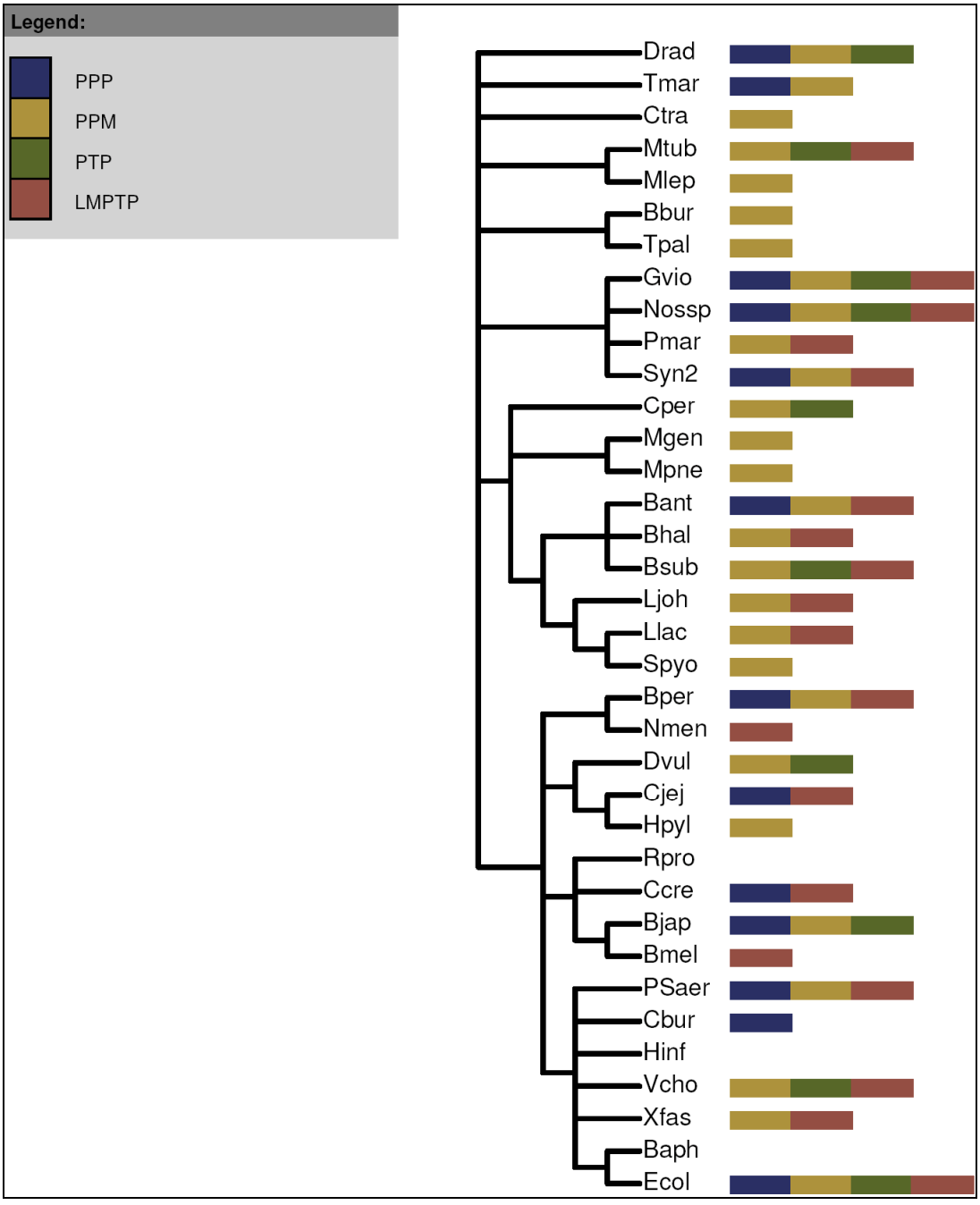

Figure 3.2 Bacteria phosphatases.

PPP, PPM, PTP and LMPTP protein phosphatase families present in the Bacteria species. Blue bars represent the PPP family, gold bars the PPM family, green bars the PTP family, and red bars the LMPTP family.

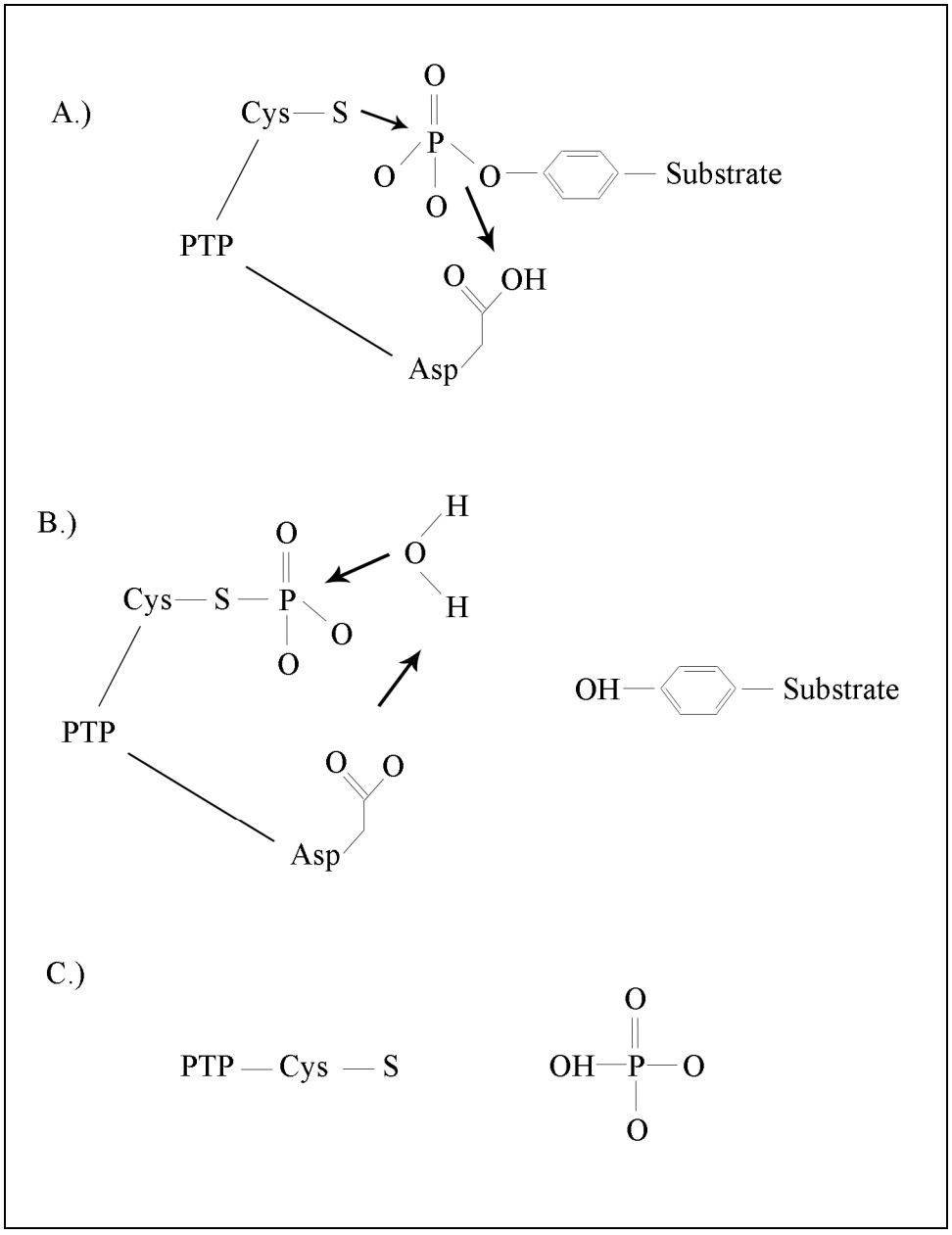

Figure 3.3 Protein tyrosine phosphatase mechanism.

A.) Nucleophilic attack by sulfur atom of the catalytic cysteine residue on phosphate group of phospho-tyrosine substrate.
B.) Hydrolysis of phospho-cysteine intermediate.
C.) Phosphate group is freed and PTP is ready for future dephosphorylation reactions.

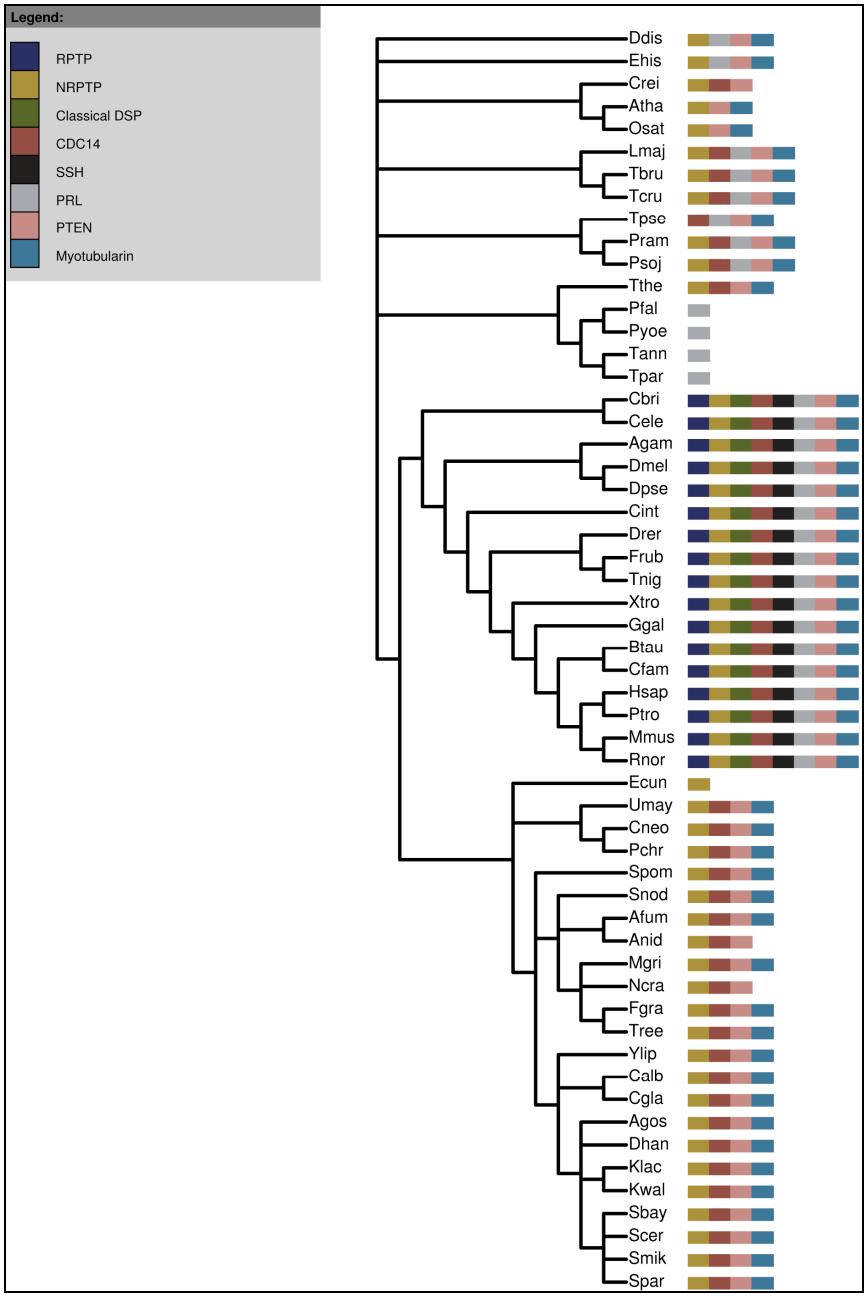

Figure 3.4 Eukaryotic Class I Protein Tyrosine Phosphatases

Class I PTPs in the Eukarya superkingdom. Dark blue bars represent the RPTPs, gold bars the NRPTPs, green bars the classical DSPs, red bars the CDC14s, black bars the SSHs, gray bars the PRLs, pink bars the PTENs, and light blue bars the myotubularins.

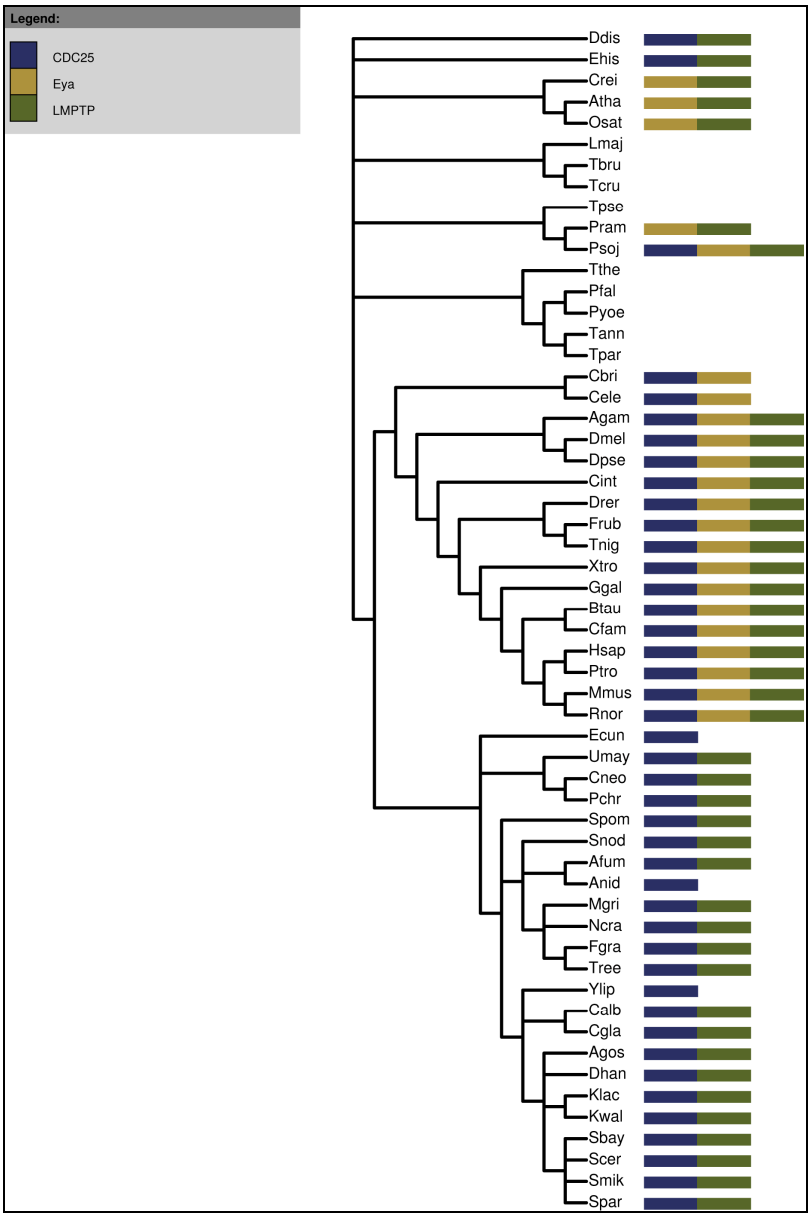

Figure 3.5 CDC25, Eya, and LMPTP phosphatases in Eukarya superkingdom.

Representation of the presence and absence of CDC25, Eya, and LMPTP phosphatases in eukaryotes. Blue bars are CDC25, gold bars are Eya, and green bars are LMPTP.

	Motif I:	Motif II:	Motif III:
	(DXDXT/V)	(hhhT/S)	(K…GDGXXD/E)
Human:	DLDET	ILVT	K...GDGVEE
Mouse:	DLDET	ILVT	K...GDGVEE
Worm:	DIDDI	VVLS	K...TSG-DT
Fly:	DLDET	VLVT	H...GDGNEE
P. sojae:	DLDET	VLVT	K...GDGLEE
P. ramorum:	DLDET	VLVT	K...GDGLEE

Figure 3.6 Alignment of known and putative Eya conserved motifs.

Alignment of putative Eya phosphatases in *Phytophthora sojae and Phytophthora ramorum* with known Eya phosphatases. Blue residues in motifs I, II, and III are commonly conserved in Eya proteins.

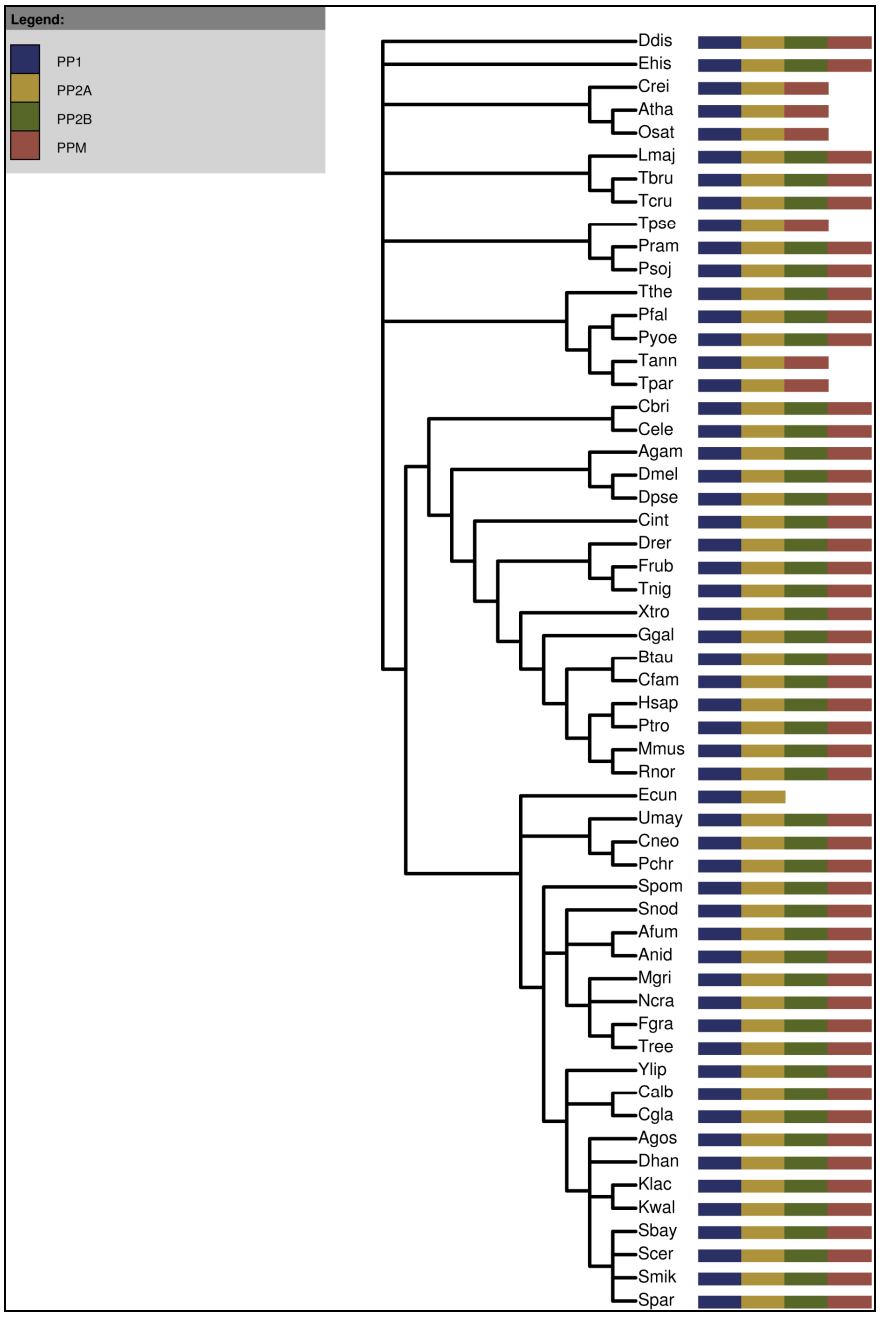

Figure 3.7 Serine/Threonine phosphatases in eukaryotes.

This tree shows the presence or absence of the PPP (PP1, PP2A, PP2B) and PPM families of serine/threonine phosphatases. Blue bars represent PP1, gold bars are PP2A, green bars are PP2B, and red bars are PPM.

Table 3.1 Protein tyrosine phosphatase families.

Protein Tyrosine Phosphatases
Class I: Classical PTPs
Receptor Protein Tyrosine Phosphatases (RPTP)
Non-receptor Protein Tyrosine Phosphatases (NRPTP)
Class I: Dual-Specificity Phosphatases
Classical DSPs (aka MAPK Phosphatases, MKP)
Slingshot (SSH)
Phosphatase of Regenerating Liver (PRL)
Cell Division Cycle 14 (CDC14)
Phosphatase and Tensin Homolog (PTEN)
Myotubularin
Atypical Dual-Specificity Phosphatases
Class II
Low Molecular Weight Protein Tyrosine Phosphatases (LMPTP)
Class III
Cell Division Cycle 25 (CDC25)
Asp-based PTP
Eyes Absent Phosphatases (Eya)

Table 3.2 SCOP classification of phosphatase groups.

Family	SCOP Superfamily Name	SCOP SCCS ID
PPP	Metallo-dependent phosphatases	d.159.1
PPM	PP2C-like	d.219.1
CDC25	Rhodanese/Cell cycle control phosphatase	c.46.1
PTP/DSP	Phosphotyrosine protein phosphatases II	c.45.1
LMPTP	Phosphotyrosine protein phosphatases I	c.44.1

Table 3.3 Serine/Threonine phosphatase families.

Serine/Threonine Phosphatases
PPP Family
PP1
PP2A
PP2B
PPM Family
PP2C

4 Protein Kinases in the Eukarya, Bacteria, and Archaea Superkingdoms

4.1 Introduction

Protein kinases are responsible for the phosphorylation of proteins in the cell. They have historically attracted much attention, for good reason. Many signal transduction processes are controlled by protein kinases. They are intimately involved in a number of critical cell activities such as transcription, cell cycle control, cytoskeletal rearrangement and cellular growth and differentiation [4]. Disruption of normal protein kinase activity can have drastic consequences. Protein kinases have been implicated in numerous diseases and cancers, including breast cancer, leukemia, lung cancer and colon cancer [149,150]. Protein kinases also comprise one of the largest protein superfamilies in humans, constituting almost 2% of the human genome [4].

Eukaryotic protein kinases have received the most attention, with complete kinome studies being published at an increasingly rapid rate as more and more genomes are completely sequenced. Kinomes have been published for a wide-ranging collection of eukaryotes including human [4], mouse [151], fly (*Drosophila*) [152], worm (*Caenorhabditis elegans*) [153,154], yeast (*Saccharomyces cerevisiae*)

[155], *Dictyostelium discoideum* [156], *Tetrahymena thermophila* [157], *Entamoeba histolytica* [158], the parasite *Plasmodium falciparum* [159] and three kinetoplastids (*Leishmania major, Trypanosoma brucei* and *Trypanosoma cruzi*) [160]. These studies have enabled an extensive classification of protein kinase families and subfamilies [4,38]. Conversely, the collection of protein kinases present in Bacteria and Archaea has not been as well studied, until recently.

The growing field of metagenomics, as well as increased sequencing efforts of bacterial and archaeal genomes affords us the opportunity to delve deeper into the evolution of protein kinases. For example, a major project was recently undertaken by the Venter Institute to sequence the contents of ocean water samples taken from around the globe [161]. This effort resulted in over 17 million predicted amino acid sequences from bacterial, archaeal and viral genomes [162]. Following that study, Kannan *et al.* [163] utilized metagenomic data to further define protein kinase families as they exist in microorganisms. They mined protein sequences from the GOS dataset and other public databases for eukaryotic protein kinase-like kinases (ELKs), identifying a total of 27,677 eukaryotic protein kinases (ePKs) and 18,699 ELKs. These kinases were then classified into 20 distinct families (Table 4.1). This work, combined with the ever-increasing quantity of sequenced genomes, allows us to take a deeper look at prokaryotic kinase evolution.

I present here a study defining the presence and absence of the aforementioned 20 microbial protein kinase families and eight groups of eukaryotic

protein kinases in 115 species. These organisms consist of completely sequenced genomes from the three superkingdoms of Eukarya, Bacteria and Archaea. The results shed light on which kinase families are present in both eukaryotic and microbial lineages. I also incorporate these findings with those of past sequence and structural studies of the protein kinase-like superfamily and discuss what this combined knowledge suggests about the evolution of different kinase families.

4.2 Methods

As described in the previous chapter, proteome drafts of completely sequenced organisms were collected and systematically searched for the presence or absence of protein kinase families. Putative kinases were found through both literature and *in silico* searches using BLAST [9] and HMMER [10] methods.

Hidden Markov Models (HMMs) representing the previously mentioned 20 prokaryotic protein kinase families were provided by the authors of the Kannan *et al.* paper [163]. Profile HMMs for the ePK groups were built locally. Sequences from the defined kinomes of human [4], yeast [155], worm [154], *Dictyostelium discoideum* [156], fly [152] and *Tetrahymena thermophila* [157] were collected. Multiple sequence alignments were constructed using ClustalW2 [57] and manually inspected and adjusted for alignment of known kinase motifs. Profile HMMs were built and calibrated using the HMMER package, version 2.3.2 [10]. Models were

then tested against the previously classified mouse [151] and sea urchin [1] kinomes to verify performance and gauge false positive scoring thresholds.

Putative kinase classification was verified in several ways. Sequences were manually inspected for commonly conserved domain motifs. Proteins were also BLASTed against NCBI's non-redundant database [58] and KinBase, a database of protein sequences from previously characterized kinomes [152] to further check for similarity to known kinase domains. Additionally, sequence classification was checked using the outside community resources of Pfam [20] and PANTHER [22,56] in cases where appropriate kinase family models were available.

4.3 Microbial kinases

4.3.1 BLRK

The BLRK family is uncharacterized, but contains ePK-like kinases with leucine-rich repeats [163]. Members of this family were found in only one bacterial genome--*Pseudomonas aeruginosa* (Figure 4.1). BLRKs were also not found in any of the Archaea species.

There were a few putative BLRKs in eukaryotes (Figure 4.2). A clear homolog is present in the fungi *Neurospora crassa*. A possible BLRK relative was also identified in the algae *Chlamydomonas reinhardtii*. This protein had a weaker hit than the *N. crassa* kinase (1.1e-36 vs. 2.8e-151). Although not a perfect match to

the BLRK consensus sequence motifs in Kannan *et al.* [163], it is worth noting it does contain 19 out of 26 residues present in all previously known BLRKs, including several unique to BLRKs. In addition, it had a strong hit to the PANTHER "leucine-rich repeat containing protein" HMM [22], providing further support for this grouping.

4.3.2 Bub1

Bub1 is thought to function in the checkpoint regulating proper mitotic spindle assembly [163]. It has been found to phosphorylate Cdc20, which leads to inhibition of the anaphase-promoting complex *in vitro* [164]. Bub1 is mainly present in eukaryotes, though Kannan *et al.* did find 9 possible homologs in the GOS data [163].

As expected, this kinase family was sparsely distributed in Bacteria. Possible Bub1 proteins were found in *Bacillus subtilis* and *Bacillus halodurans*. The *B. subtilis* protein shows a greater similarity to the Bub1 consensus motifs than the *B. halodurans* sequence. The proteins share a 48% sequence identity, as found by BLAST [9]. It is unclear if these proteins truly belong to the Bub1 family, but they should be considered candidates for further experimental characterization, particularly in the case of the *B. subtilis* kinase. No Bub1 kinases were found in archaeal genomes.

As previously mentioned, the Bub1 family is widespread in eukaryotes (Figure 4.2). Putative Bub1s were found in all fungi except the reduced genome of *Encephalitozoon cuniculi*. Bub1 homologs were also identified in all Metazoa with the exception of *Ciona intestinalis* and the zebrafish *Danio rerio*. The *D. rerio* genome contains what appears to be an incomplete prediction of a potential Bub1 kinase. The protein matches well with the N-terminal Bub1 domain consensus sequence motifs, but the *D. rerio* protein prediction ends shortly after the HXDXXXXN catalytic motif.

Although a putative Bub1 was located in the rice genome (*Oryza sativa*), there were no Bub1 kinases found in *Arabidopsis thaliana*. A previous search for Bub1 in *Arabidopsis* by another group also failed [165]. Bub1s were also not found in the algae *Chlamydomonas reinhardtii*. Additionally, none of the kinetoplastids, stramenopiles or apicomplexa genomes in this study contained Bub1s.

4.3.3 PknB

Based on sequence and structural similarities, bacterial PknB kinases are thought to be related to the ePK family [166,167]. PknB has been experimentally shown to be essential for proper cellular shape and growth of *Mycobacterium tuberculosis* [168,169]. Possible substrates for this family include Wag31, an

ortholog of a cell division protein, and RV1422, a protein of unknown function. This family has also shown the ability to undergo autophosphorylation [169].

Putative PknBs were found in only one Archaea (*Picrophilus torridus*) (Figure 4.3), and almost half the Bacteria genomes (Figure 4.1). Bacterial PknBs were mostly clustered in the Actinobacteria, Cyanobacteria and Firmicutes phylums. Only three of sixteen Proteobacteria species studied contained a potential PknB kinase: the alpha-proteobacterium *Bradyrhizobium japonicum*, the delta-proteobacterium *Desulfovibrio vulgaris* and the gamma-proteobacterium *Pseudomonas aeruginosa*. Neither of the Spirochaetes genomes in this study contained PknBs.

While no PknBs were found in eukaryotes, it was noticed that eukaryotic ePKs tended to also have weaker hits to the PknB model. This observation and comparison of their conserved sequence motifs as found by Kannan *et al.* [163] supports their putative relationship mentioned above. The evolutionary implications of this observation are discussed later in this dissertation.

4.3.4 HRK

The HRK family consists of eukaryotic haspin protein kinases and two related kinases mostly found in viral genomes [163]. A mouse haspin has been implicated in cell cycle progression [170]. Human haspin has been shown to

phosphorylate histone H3 during mitosis, although the physiological consequences of this are unclear as haspin siRNA did not prevent chromosome condensation [171,172].

Kannan *et al.* found 259 family members in the GOS dataset [163]. However, no HRKs were present in any of the Archaea or Bacteria genomes in this study. This is consistent with previous findings of haspin kinases only in eukaryotes, and the other two subgroups classified as viral-specific [163,170].

Putative HRKs were found in most plant, metazoan and fungal species (Figure 4.2). No HRK proteins were located in the fungi *Debaromyces hansenii* and *Candida albicans*, the fly *Drosophila pseudoobscura* and the frog *Xenopus tropicalis*. All of these organisms have related species that contain putative haspins, thus it is quite possible that the family has not been lost but rather is simply missing from the genome drafts. No HRKs were found in the groupings of Apicomplexa and *Tetrahymena*, the kinetoplastids, or the stramenopiles.

4.3.5 GLK

Kannan *et al.* identified a previously unannotated family, termed the glycosylase-linked kinases (GLKs). It appears to be a relatively small family, with only 38 members in the GOS dataset. Kannan *et al.* found some of these kinases to be either fused to, or neighbors of, a DNA glycosylase domain [163].

GLK family members were found in a number of Archaea, including all Crenarchaeota species, *Nanoarchaeum equitans*, all four Thermococci organisms studied, both Methanococci species and *Archaeoglobus fulgidus* (Figure 4.3). A possible GLK was found in only one bacterium, *Clostridium perfringens*. Among the kinase models, this protein sequence produced the best hit to the GLK family, but it was not a very strong score. While some kinase motifs are clearly present, other commonly conserved residues are not contained in the protein sequence and it is doubtful that the protein truly belongs to the GLKs.

Interestingly, putative GLKs were also found in the related eukaryotes *Phytophthora ramorum* and *Phytophthora sojae* (Figure 4.2). The strongest kinase model matched by these proteins was clearly the GLK family, and the alignments of the proteins show good conservation of GLK consensus sequence motifs. These two species were the only eukaryotes found to contain GLKs.

4.3.6 Bud32

The Bud32 family (also known as the piD261 kinases), have previously been found only in Eukarya and Archaea [173,174]. The Bud32 kinase is essential for correct cellular growth in yeast, and has been shown to phosphorylate the tumor suppressor protein p53 in humans [175].

Initially, no Bud32 kinases were found in the Archaea species *Halobacterium sp. NRC-1*, *Haloarcula marismortui* and *Methanosarcina barkeri*. However, subsequent searches against updated genome drafts revealed strong hits to putative Bud32 kinases in all three organisms. Thus, members of the Bud32 family were identified in all Archaea studied (Figure 4.4).

Additionally, Bud32 family kinases were found in virtually all eukaryotes (Figure 4.5). *Stagonospora nodorum* contained a possible Bud32, but the protein sequence had an apparent deletion in the middle of the kinase domain, eliminating several important kinase motifs. If this is not simply a protein prediction error, it is uncertain that such a protein could actually be an active kinase. It is also unclear if the algae *Chlamydomonas reinhardtii* contains a Bud32 kinase. A weak hit was found to a protein that contains many, but not all, of the commonly conserved Bud32 motifs. It does, however, match the Bud32 model the strongest out of all the kinase models. No Bud32 kinases were located in the *Ciona intestinalis* genomes. There were also no Bud32 kinases found in any of the bacterial species studied.

4.3.7 RIO

RIO kinases were originally discovered in 2002 in yeast and were shown to be essential for cell cycle progression and proper ribosome biogenesis [176]. They have been shown to phosphorylate serine *in vitro*. A structure of RIO2 also revealed

surprising similarity to DNA binding proteins, suggesting RIO2 may bind DNA [177]. RIO1 does not, however, contain the same DNA-binding domain [178]. They have previously been found in Eukarya, Archaea and a small number of Bacteria [163].

RIO family members were initially found in all Archaea studied, with the same three exceptions as the aforementioned Bud32 results (*Halobacterium sp. NRC-1*, *Haloarcula marismortui* and *Methanosarcina barkeri*). As above, a search against NCBI's updated genomes resulted in the successful location of RIO kinases in all three species (Figure 4.4). Further emphasizing the critical role that RIO kinases apparently play in both eukaryotes and Archaea, putative RIOs were found in every eukaryotic genome included in this study (Figure 4.5).

RIO kinases were identified in only two bacterial genomes: *Pseudomonas aeruginosa* and *Deinococcus radiodurans* (Figure 4.6). This finding is consistent with a previous study that also found RIO kinases in only these two bacterial genomes [179].

4.3.8 KdoK

The Kdo kinase family contains sugar kinases that phosphorylate lipopolysaccharides (LPSs) [180,181]. LPSs are major components of the outer membrane in Gram-negative bacteria [182]. Thus, as expected, putative Kdo kinases

were confined to Gram-negative bacteria: *Bordetella pertussis*, *Escherichia coli*, *Haemophilus influenza*e, *Pseudomonas aeruginosa*, *Vibrio cholerae* and *Xylella fastidiosa* (Figure 4.6). No Kdo kinase family members were found in Archaea or eukaryotic genomes.

4.3.9 CAK

The CAK family includes choline kinases, ethanolamine kinases and aminoglycoside phosphotransferases [163]. Eukaryotic choline kinases phosphorylate choline to create phosphorylcholine as part of the CDP-choline pathway [183]. This pathway results in the formation of phosphatidylcholine, a major component of the membrane. A choline kinase homolog is also present in some Bacteria (often termed LicA), where it is involved in the incorporation of phosphocholine on lipopolysaccharides [184,185]. Ethanolamine kinases fulfill a similar responsibility, phosphorylating ethanolamine to eventually form phosphatidylethanolamine, a membrane phospholipid. Phosphatidylethanolamine can also be methylated and turned into phosphatidylcholine [186]. Aminoglycoside phosphotransferases (APHs) are bacterial kinases that target aminoglycoside antibiotics, providing Bacteria with antibiotic resistance [27].

CAK family members were found in most eukaryotic genomes (Figure 4.7). There was a somewhat questionable hit in the *Saccharomyces mikatae* genome. The

protein identified contains many commonly conserved CAK motifs, and has a 75% BLAST sequence similarity to a CAK in *Saccharomyces cerevisiae*. However, the protein prediction appears to be shortened and is missing part of the N-terminal region. No CAKs were located in the compacted genome of *Encephalitozoon cuniculi*. Weaker hits were also found to the three kinetoplastids (*Leishmania major, Trypanosoma cruzi* and *Trypanosoma brucei*). While these protein sequences do vary somewhat from the CAK consensus sequence, they contain a majority of the highly conserved CAK residues and motifs. Additionally, a BLAST search against NCBI's non-redundant database returned top hits to CAK family members.

Only two Archaea groupings contained possible CAK family members (Figure 4.8). Both genomes from the Halobacteria class (*Halobacterium sp. NRC-1* and *Haloarcula marismortui*) contained apparent CAKs. Additionally, two of the three Archaea in the Methanomicrobia class had proteins with conserved CAK sequence motifs: *Methanosarcina acetivorans* and *Methanosarcina mazei*.

CAKs were found in a majority of bacterial genomes (Figure 4.9). All nine firmicutes contained CAK family members. Three cyanobacteria (*Gloebacter violaceus, Nostoc* and *Synechocystis*) had putative CAKs, while the other cyanobacterium in this study (*Prochlorococcus marinus*) had a very weak hit that could not be satisfactorily confirmed based on sequence and BLAST analysis alone. CAKs were also identified in *Deinococcus radiodurans*, *Treponema pallidum,* one of the two actinobacteria (*Mycobacterium tuberculosis*) and over half of the

Proteobacteria species. The Proteobacteria hits included three of four alpha-proteobacteria, both beta-proteobacteria and six of seven gamma-proteobacteria organisms.

There were two other bacterial genomes that had uncertain hits to possible CAK proteins. The delta-proteobacterium *Desulfovibrio vulgaris* contained a protein that conserves most of the CAK sequence motifs. A BLAST search against NCBI's non-redundant database also returned strong hits to APHs in other species but only the asparagine of the commonly conserved HXDXXXXN catalytic motif is present. Additionally, there was a protein fragment prediction in the epsilon-proteobacterium *Campylobacter jejuni* with a strong match to the last three-quarters of the CAK domain. However, without the full protein sequence, this prospective kinase cannot be definitively classified as a CAK family member.

4.3.10 HSK2

The HSK2 grouping contains a family of homoserine kinases (HSKs) distantly related to the protein kinase-like family [36,163,187]. Homoserine kinases are critical to the first of a two-step process in threonine biosynthesis. HSKs phosphorylate homoserine, producing phosphohomoserine. The phosphohomoserine is then isomerized and dephosphorylated, resulting in the production of threonine

[187,188]. Another group of homoserine kinases catalyzes the same process but are unrelated to the HSK2s and protein kinase-like family [36].

Putative members of the HSK2 family were identified in six Bacteria (Figure 4.9). Three of the four alpha-proteobacteria species in this study contained HSKs: *Caulobacter crescentus*, *Bradyrhizobium japonicum* and *Brucella melitensis*. In addition, both beta-proteobacteria analyzed contained HSK2s (*Bordetella pertussis* and *Neisseria meningitidis*). Only one of the gamma-proteobacteria, *Pseudomonas aeruginosa*, had a homoserine kinase in the HSK2 family. *Bacillus anthracis* contained a weak HSK2 hit, but this protein could not be definitively classified based only on BLAST and sequence analysis. No HSK2s were found in any archaeal or eukaryotic genomes studied.

4.3.11 FruK

The FruK family contains fructosamine kinase homologs [163]. Mammalian fructosamine kinases phosphorylate fructosamines, destabilizing them and leading to their detachment from proteins. This process of deglycation has been identified as a protein repair mechanism [189,190]. Bacterial homologs of fructosamine 3-kinases were recently characterized by Gemayel *et al.* [191]. These bacterial proteins, along with some mammalian homologs, were demonstrated to act as ribulosamine/erythrulosamine 3-kinases [191,192]. This finding led Gemayel *et al.*

to conclude that the ability of mammalian and avian fructosamine kinases to phosphorylate fructosamines was likely a more recent evolutionary event [191]. Plant FruK family members have also been shown to be ribulosamine/erythrulosamine 3-kinases, likely acting in a similar deglycation protein repair functional role [189].

FruK family members were found in only one archaeal genome—*Haloarcula marismortui* (Figure 4.8). Our Bacteria hits were confined to cyanobacteria and gamma-proteobacteria organisms (Figure 4.9). All four cyanobacteria genomes analyzed and the gamma-proteobacteria species *Escherichia coli* and *Vibrio cholerae* contained FruK proteins.

In the eukaryotes, FruK proteins were found in all metazoans, with the exception of insects (Figure 4.7). FruK has previously been found in *Arabidopsis thaliana* [189]. I successfully identified this kinase, as well as FruKs in rice and the algae *Chlamydomonas reinhardtii*. Putative FruKs were also located in all fungi studied from the Pezizomycotina subphylum (*Aspergillus nidulans, Aspergillus fumigatus, Fusarium graminearum, Trichoderma reesei, Magnaporthe grisea, Neurospora crassa* and *Stagonospora nodorum*). There were also apparent FruKs in all three fungi in the phylum Basidiomycota (*Ustilago maydis, Cryptococcus neoformans* and *Phanerochaete chrysosporium*). Additionally, FruKs were found in one kinetoplastid (*Trypanosoma cruzi*) and one stramenopile—the diatom

Thalassiosira pseudonana. Related species in these two groupings did not contain any apparent FruKs.

4.3.12 MTRK

Methylthioribose (MTR) kinases comprise the MTRK family [163]. MTR kinases are involved in methionine salvage. Specifically, MTR kinase phosphorylates MTR into MTR-1-phosphate, which can later be converted into methionine. This process permits organisms to grow in non-methionine sulfur environments [193]. Expression of MTR kinase has been linked to both environmental methionine levels and conditions of sulfur, nitrogen and carbon starvation [193,194]. While MTR kinases have been identified in plants (*Arabidopsis* and rice) and a few Bacteria, mammals use a different methionine salvage pathway that does not require MTR kinase [195,196]. Given this difference between parasites and humans, the MTR pathway has generated interest as a potential drug target [197].

MTRKs were found in not only the higher plants of *Arabidopsis thaliana* and *Oryza sativa*, but also the algae *Chlamydomonas reinhardtii* (Figure 4.7). MTRK proteins were also found in *Tetrahymena thermophila* and possibly in the marine diatom *Thalassiosira pseudonana*. The *T. pseudonana* protein shows fairly good MTRK sequence motif conservation, but is shorter than the MTRK consensus

sequence and ends prior to several unique C-terminal residues commonly conserved in MTRKs.

No MTRKs were found in Archaea organisms, and only a limited number were identified in Bacteria (Figure 4.9). The proteobacterium *Bradyrhizobium japonicum* and two *Bacillus* species (*B. anthracis* and *B. subtilis*) both contained strong MTRK hits with well conserved sequence motifs.

4.3.13 UbiB

Kinases in the UbiB family are required for coenzyme Q (aka ubiquinone) biosynthesis [198]. Ubiquinone is a component of the electron transport chain in the eukaryotic mitochondrial membrane and the prokaryotic plasma membrane [198,199]. UbiBs have previously been found in eukaryotes (where they are also known as the ABC1 family) and Bacteria [163].

UbiB kinases were present in roughly half the archaeal genomes studied (Figure 4.10). All three *Sulfolobus* species contained putative UbiBs. UbiBs were also present in four of eight Euryarchaeota classes: Halobacteria, Methanobacteria, Methanomicrobia and Thermoplasmata. No UbiBs were located in *Nanoarchaeum equitans*.

Putative UbiBs were found in 19 Bacteria (Figure 4.11). This included all Cyanobacteria, Actinobacteria, Alpha-proteobacteria and Beta-proteobacteria species

included in this study. Additionally, UbiB family members were found in five of seven Gamma-proteobacteria studied, but only two of nine Firmicutes (*Bacillus halodurans* and *Clostridium perfringens*). No UbiB kinases were apparent in *Chlamydia trachomatis, Deinococcus radiodurans, Thermotoga maritima* or either of the Spirochaetes (*Borrelia burgdorferi* and *Treponema pallidum*).

UbiBs were located in almost every eukaryotic genome (Figure 4.12). This family was present in every kinetoplastid, stramenopile, Alveolata and plant genome studied. It was also present in *Dictyostelium discoideum* as well as most Metazoa and Fungi species. However, no UbiB family members were located in *Entamoeba histolytica* or the compacted *Encephalitozoon cuniculi* genome. A questionable UbiB was also found in *Ciona intestinalis*. The protein sequence includes some conserved UbiB motifs, but is shorter than the UbiB consensus sequence, resulting in the absence of several commonly conserved regions of the protein. This may be due simply to a gene or protein prediction error. Conversely, if the protein sequence is correct, it seems unlikely that the protein would contain a functional kinase domain.

4.3.14 MalK

Maltose kinases comprise the MalK family [163]. Maltose kinase was originally identified in 1996 in an *Actinoplanes* bacterium [200]. It was experimentally characterized and found to phosphorylate the disaccharide maltose,

resulting in maltose 1-phosphate [201]. A maltose kinase was also later experimentally demonstrated to be present in *Streptomyces coelicolor* [202].

A fairly limited number of species were found with maltose kinases. Only two Archaea, *Aeropyrum pernix* and *Picrophilus torridus,* contained putative MalKs (Figure 4.10). Of the bacterial genomes, maltose kinases were identified in four organisms (Figure 4.11). One actinobacterium, *Mycobacterium tuberculosis* had a MalK protein. The remaining three MalK containing species were all Protebacteria: *Bradyrhizobium japonicum, Bordetella pertussis* and *Pseudomonas aeruginosa*. No MalKs were found in any eukaryotic proteomes.

4.3.15 RevK

Kannan *et al.* characterized a novel kinase family termed reverse kinase, or RevK. The function of these kinases is unknown. They do not include the N-terminal glycine-rich ATP-binding loop common to protein kinases. However, the C-terminal often does contain an ATPase domain with the ATP-binding GXXGXG sequence motif [163].

The results from this study indicate that among the three superkingdoms of life, RevK is restricted to only Bacteria (Figure 4.11). RevKs were found in *Mycobacterium tuberculosis*, two cyanobacteria (*Synechocystis sp. PCC 6803* and *Nostoc sp. PCC 7120*), and three proteobacteria (*Caulobacter crescentus,*

Bradyrhizobium japonicum and *Pseudomonas aeruginosa*). Without any functional annotation, it is difficult to speculate why this kinase is not present in Archaea or Eukarya. Future experimental characterization of this kinase family is needed.

4.3.16 CapK

Another novel kinase group published by Kannan *et al.* is the capsule kinase (CapK) family. The function of CapKs is unclear. However, their chromosomal location may offer a clue, as some GOS Bacteria members were located near genes involved in capsule synthesis [163]. This potential functional linkage is of interest, as much research has been devoted to characterizing capsules in the context of vaccine research and development [203,204].

No CapKs were found in the archaeal or eukaryotic species. Of the bacterial genomes included in this study, only two contained capsule kinases: *Escherichia coli K12* and *Vibrio cholerae O1 biovar* (Figure 4.11). Of these, *E. coli* is known to produce a capsule, while this particular strain of *V. cholerae* does not [205,206]. CapKs present an interesting kinase target for experimental characterization to determine their true function.

4.3.17 PI3K

A small number of PI3Ks were found in the GOS data set. This family includes the lipid kinases phosphoinositide 3-kinase (PI3K) and phosphoinositide 4-kinase (PI4K), as well as the phosphoinositide 3-kinase related kinases (PIKKs) [163]. PI3 and PI4 kinases phosphorylate phosphoinositides at their 3- and 4-hydroxyl, respectively [207,208]. These phosphorylated lipids can then act as second messengers, meaning PI3Ks and PI4Ks play roles in a multitude of cellular processes including cell motility, adhesion, proliferation, apoptosis and cytoskeletal organization [208,209].

The catalytic kinase domain of PIKKs shares many commonalities with that of PI3Ks and PI4Ks. However, the PIKKs have been shown to phosphorylate not lipids, but proteins. At least five PIKKs have been characterized as Ser/Thr kinases, while a sixth member, TRRAP, lacks critical catalytic residues and may not be an active kinase [210,211]. The PIKKs are involved in detecting DNA damage [212].

The results from this study support past studies that have classified the PI3K family as eukaryotic kinases [163,211]. PI3K family members were found in all eukaryotic genomes (Figure 4.13). PI3Ks were absent in the Bacteria and Archaea genomes.

4.3.18 AlphaK

Like the PI3Ks, the Alpha kinases are a eukaryotic kinase family [163]. They share very little sequence homology with typical protein kinases [213]. However, structural comparisons of their kinase domain with that of ePKs have shown surprising similarity between the two families and have led to the classification of the alpha kinases as part of the "atypical" branch of the eukaryotic protein kinase-like superfamily [214]. The alpha kinase family includes several distinct kinases, including the myosin heavy chain kinases, the elongation factor 2 kinases and the channel kinases [4,213].

None of the Archaea or Bacteria genomes in this study contained alpha kinases. Putative alpha kinases were found in a number of eukaryotic organisms (Figure 4.13). All vertebrates had strong hits to alpha kinases. Alpha kinases were also identified in three invertebrates (*Caenorhabditis briggsae, Caenorhabditis elegans* and *Ciona intestinalis*). Alpha kinases were not, however, apparent in the three insects analyzed (*Anopheles gambiae, Drosophila melanogaster* and *Drosophila pseudoobscura*). *Dictyostelium discoideum,* the three kinetoplastids (*Leishmania major, Trypanosoma cruzi* and *Trypanosoma brucei*), and the three stramenopiles (the plant parasites *Phytophthora ramorum* and *Phytophthora sojae* and the marine diatom *Thalassiosira pseudonana*) all contained putative alpha kinases.

Additionally, the presence of alpha kinases was found in *Tetrahymena thermophila*, but not in the related phylum of animal parasites, Apicomplexa. Very few Fungi were found to have alpha kinases. *Neurospora crassa* contained a very strong hit that contained most commonly conserved alpha kinase residues and sequence motifs. Weaker hits were found in the related fungi *Trichoderma reesei* and *Magnaporthe grisea*. Sequence alignments of these possible alpha kinases revealed conservation of most alpha kinase sequence motifs. A BLAST search against NCBI's non-redundant database further supported their inclusion in this family, with a number of hits to alpha kinases in other species. The *T. reesei* and *M. grisea* proteins shared a 33% and 28% sequence identity respectively with a *T. thermophila* alpha kinase.

Lastly, although no alpha kinases were found in the higher plant genomes of *Arabidopsis thaliana* and *Oryza sativa* (rice), an interesting alpha kinase candidate was found in the algae *Chlamydomonas reinhardtii*. Manual sequence analysis showed very good conservation of commonly conserved residues and motifs in the first half of the alpha kinase domain. BLAST analysis also showed it to have 58% sequence identity to a mouse alpha kinase. However, the *C. reinhardtii* protein sequence prediction abruptly ends in the middle of the kinase domain. If this is simply a protein or gene prediction error, it seems likely that this protein is truly an alpha kinase.

4.3.19 IDHK

The IDHK family is comprised of isocitrate dehydrogenase kinases [163]. Isocitrate dehydrogenase kinase phosphorylates the Krebs cycle enzyme isocitrate dehydrogenase (IDH) [215]. IDH is involved in controlling the level of isocitrate entering the Krebs cycle versus the glyoxylate bypass. The glyoxylate bypass allows the cell to use acetate, ethanol or fatty acids as a carbon source, permitting cellular growth in these environments [216]. Phosphorylation inactivates IDH, resulting in an increased level of isocitrate being directed to the glyoxylate bypass [215]. Interestingly, *E. coli* isocitrate dehydrogenase kinase also contains apparent phosphatase activity and ATPase activity [216]. While originally little similarity was detected between IDHK and conventional protein kinases, in-depth structure and sequence analysis has since revealed similarities in critical catalytic residues. Thus, an evolutionary link has been proposed between IDHKs and ePKs [217].

The analysis from this study found IDHKs present in a very limited number of genomes. IDHKs were pinpointed in the gamma-proteobacteria *Escherichia coli* and *Pseudomonas aeruginosa* (Figure 4.14). No IDHKs were identified in any archaeal or eukaryotic genome in this study.

4.3.20 Eukaryotic protein kinase-like

The ePK family is comprised of kinases with the "typical" eukaryotic protein kinase-like fold. Kannan *et al.* found 2753 ePKs present in the GOS dataset [163]. As expected, ePKs were present in every eukaryotic genome (Figure 4.2). This kinase family is further broken down in eukaryotes later in this dissertation. A smaller number of hits were also found in archaeal (Figure 4.3) and bacterial (Figure 4.1) genomes. Most potential archaeal ePKs were located in the Crenarchaeota phylum. Four of these five genomes contained proteins with strong hits to the ePK model. Overall, they showed good conservation of common ePK motifs as well. The one genome with no apparent ePK was *Aeropyrum pernix*. Outside of the Crenarchaeota, putative ePKs were found in *Pyrococcus furiosus* and *Picrophilus torridus*. Initially, ePKs were not found in *Haloarcula marismortui*, but a subsequent search against an updated genome draft located a candidate ePK in this species as well.

Putative bacterial ePKs were found in two cyanobacteria: *Synechocystis sp. PCC 6803* and *Nostoc sp. PCC 7120*. Additionally, the cyanobacterium *Gloebacter violaceus* had a kinase that could not be definitively classified. It shows similarity to both the ePK and pknB consensus sequences and scores almost evenly between the two HMMs. Experimental characterization is required to clarify the function of this protein.

ePKs were also located in *Deinococcus radiodurans, Coxiella burnetii* and both *Mycoplasma* genomes (*Mycoplasma genitalium* and *Mycoplasma pneumoniae*). A few possible but not definitive ePKs were found in other Bacteria. *Bacillus anthracis* and *Bacillus subtilis* contained possible ePKs, although the APE motif was difficult to locate and apparently at least partially not conserved in either protein. The rest of the protein sequences showed clear kinase motifs. Another uncertain hit belonged to the *Pseudomonas aeruginosa* genome. It contained a protein with clear kinase motifs in the C-terminal half of the domain, but the glycine-rich loop and VAIK motifs were not clear. No other bacterial genomes contained apparent ePKs.

4.4 ePK groups in eukaryotes

Analyses of eukaryotic protein kinases have further defined eight main groups (Table 4.2) [4,38,147]. I discuss here the presence or absence of these groups in 56 eukaryotic genomes. I also delve deeper into selected subfamilies in the next chapter.

4.4.1 Ubiquitous ePK groups

Four of the eight groups were present in every eukaryotic genome studied: AGC, CAMK, CK1 and CMGC (Figure 4.15). Their presence throughout the

superkingdom emphasizes the critical roles these kinases play in eukaryotic cell growth and survival, and are described below.

The AGC group of serine/threonine protein kinases was originally named for cAMP-dependent protein kinase A, protein kinase G, and protein kinase C [218]. It contains mainly cyclic-nucleotide and calcium-phospholipid-dependent kinases [38]. AGC kinases are involved in a variety of essential cellular processes such as cell growth and differentiation [219].

The CAMK group contains Ca^{2+}/calmodulin-dependent protein kinases [156]. In the human kinome 74 CAMKs have been identified, far more than that in the yeast (21), worm (46) or fly kinomes (32) [4]. CAMKs have a number of functional roles effecting, among others, protein synthesis, myosin activation and calcium levels in the heart [220].

Casein kinase 1 and its close relatives comprise the CK1 group [4]. Casein kinases have been implicated in regulating DNA repair pathways and cell morphogenesis [221]. They have also been connected to control of mammalian circadian rhythm, and may even play a role in Alzheimer's disease [221,222].

CMGCs include the cyclin-dependent kinases (CDKs), the mitogen-activated protein kinases (MAPKs), glycogen synthase (GSK) and the Clk (aka CDK-like) kinases [38]. CMGCs are involved in cell development and proliferation [159].

This study showed the other kinase groups (RGC, STE, TK and TKL) to be present in only a proportion of eukaryotic genomes. I present these findings below.

4.4.2 RGC group

The receptor guanylate cyclase (RGC) group was first declared a distinct protein kinase grouping by Manning *et al.* [4]. These proteins show some similarity with the tyrosine kinase domain, but are missing a highly conserved aspartic acid critical to catalytic activity and are generally thought to lack kinase activity, with the possible exception of one such protein [4,223,224]. RGCs induce guanylate cyclase activity when ligand bound, catalyzing the formation of cGMP from GTP [224,225]. The analysis of the RGC family in this study found these proteins confined to only metazoan genomes (Figure 4.15). This supports previous kinome analyses that have found no RGCs present in the *Saccharomyces cerevisiae*, *Dictyostelium discoideum* or *Tetrahymena thermophila* genomes [152,156,157].

4.4.3 STE group

STE kinases were originally named after the "sterile" yeast mutants in which they were first identified [159]. This group includes kinases involved in MAPK cascades (although MAP kinases themselves are part of the CMGC group, not the STE group) [4]. Putative STEs were present in almost every eukaryotic genome

with a few notable exceptions (Figure 4.15). None were identified in the reduced microsporidia genome of the parasite *Encephalitozoon cuniculi*. These findings support a study published in September 2007 that also was unable to locate STEs in *E. cuniculi* [147].

The results from my study suggest that STEs may have been lost in some Apicomplexa parasites. No STEs were found in three of the four Apicomplexa species included in this study (*Plasmodium yoelii, Theileria parva* and *Theileria annulata*). A STE was, however, identified in the related *Tetrahymena thermophila* genome. STEs in *T. thermophila* have previously also been noted by Eisen *et al.* [157].

Two past studies have conflicted on the presence of STEs in *Plasmodium falciparum,* with a 2004 study finding the *P. falciparum* kinome lacking STEs [159] and a 2007 automated classification of several eukaryotic kinomes listing STE as present in the parasite [226]. However, the 2007 study does not discuss this finding, nor list a sequence identification number for the putative STE.

The STE models found what may be a possible STE kinase in *P. falciparum*. The protein in question contains a kinase domain that shows similarity to known STE domains. A BLAST search of the kinase domain against KinBase (a database of kinases from five known kinomes, administrated by the Manning group [152,156]) finds a 36% sequence identity with a *Dictyostelium* STE kinase.

Additionally, the top ten hits returned are all to STE kinases from *Dictyostelium,* human, fly and worm.

Searching NCBI's non-redundant database by BLASTing the entire *P. falciparum* protein only found matches to other generically annotated proteins labeled as either putative kinases or hypothetical proteins. However, a specific BLAST search against NR using only the kinase domain returned hits against a number of STE kinases, including a 36% local sequence identity match to a *T. Thermophila* STE. Additionally, a search against the PANTHER [22] database of protein family HMMs produced a strong hit (5e-64) to a MAPKK model. A search of the full protein using InterProScan [227] found no other significant hits to protein domains other than the kinase domain. Thus, this protein may be an interesting target to pursue experimentally, in hopes of further clarifying its function as a potential STE and determining whether STEs have indeed been completely, or nearly completely, lost in the Apicomplexa ancestor sometime after the divergence of *T. thermophila*.

4.4.4 TKL group

The tyrosine kinase-like (TKL) group consists of kinases that bear some resemblance to tyrosine kinases, yet generally act as serine/threonine kinases [159]. These include the LIS kinases (LISK), the interleukin-1 receptor-associated kinase

(IRAK), MLK, receptor-interacting protein kinase (RIPK), activin and TGF-β receptors (STRK), and Raf kinases [4].

TKLs were present in a number of eukaryotes (Figure 4.15). Putative TKLs exist in all metazoans, *Dictyostelium discoideum, Entamoeba histolytica* and all three plant genomes (*Arabidopsis thaliana, Oryza sativa* and *Chlamydomonas reinhardtii*). Additionally, the presence of TKLs was noted in the diatom *Thalassiosira pseudonana* and the related parasitic stramenopiles *Phytophthora ramorum* and *Phytophthora sojae*. No TKLs were found in the kinetoplastids (*Leishmania major, Trypanosoma cruzi* and *Trypanosoma brucei*).

A previous study noted the presence of TKLs in two Apicomplexa genomes: *Plasmodium falciparum* and *Plasmodium yoelii* [226]. My study confirmed these findings, but did not locate any TKLs in two other Apicomplexa species. Neither *Theileria annulata* nor *Theileria parva* appeared to contain TKLs. This suggests the possibility that TKLs may have been lost in the common *Theileria* ancestor. TKLs were found to be present in the more distantly related alveolate *Tetrahymena thermophila,* agreeing with the findings in the recent publication of the *T. thermophila* genome [157].

The aforementioned study also found TKLs present in two of nine fungi tested [226]. The presence of TKLs was confirmed in *Cryptococcus neoformans* and *Phanerochaete chrysosporium*. Additionally, a putative TKL was found in the related fungi *Ustilago maydis*. These results confirm the absence of TKLs in both

the seven other previously studied fungi species and the 13 additional fungi genomes included in this study.

4.4.5 TK group

The tyrosine kinases (TKs) phosphorylate tyrosine residues. They exist in both receptor and cytoplasmic forms and are involved in intercellular and intracellular communication [156,226]. Historically, the tyrosine kinase group has been thought of as a metazoan kinase development [152]. The presence of phosphotyrosine has, however, been noted in species not thought to contain true TKs [156,157]. This seeming contradiction can be partially explained by the presence of some dual-specificity kinases that have shown the ability to phosphorylate not only serine/threonine residues, but also tyrosine residues [156,228].

Recently, the theory that TKs are confined to metazoan genomes has been the subject of some debate. Several kinome studies of non-metazoan genomes have been unable to locate members of the TK group [156,157,160]. However, a few research groups have put forth suggestions that TKs may exist in *Entamoeba histolytica* and plant genomes [158,226,229]. A separate study concluded that no tyrosine kinases were present in *Arabidopsis* [228].

My analysis found no tyrosine kinases present in *Tetrahymena thermophila, Dictyostelium discoideum* or the kinetoplastids (*Leishmania major, Trypanosoma*

cruzi and *Trypanosoma brucei*), consistent with previous studies (Figure 4.15) [156,157,160]. There were also no tyrosine kinases present in the three stramenopile species (*Thalassiosira pseudonana, Phytophthora ramorum* and *Phytophthora sojae*) or any fungi genome. Tyrosine kinases were found in all metazoans.

The classification of the plant genomes was less clear. All candidate TKs were stronger matches to the TKL models. BLAST checks against known TKs and TKLs in KinBase were inconclusive, returning hits for a few proteins to both TK and TKL domains with near equal scores. Analysis of the plant sequences did not show very good conservation of common tyrosine kinase motifs.

However, strong hits were found to the TK models in the *Entamoeba histolytica* genome. An *E. histolytica* protein shared a 36% sequence identity to a worm tyrosine kinase domain and returned all top hits to tyrosine kinases from other fly and human genomes. A BLAST search against NCBI's non-redundant database also found strong hits against known tyrosine kinase family members. Additionally, both Pfam and PANTHER databases classified this protein as a tyrosine kinase family member with high confidence and the protein matches well to known tyrosine kinase motifs.

These results seem to correspond with Shiu *et al.* [229], who found that several *E. histolytica,* but not *Arabidopsis,* sequences clustered with known metazoan tyrosine kinases. Recently, putative tyrosine kinases have been identified in the unicellular choanoflagellate *Monosiga brevicollis*, a close relative of

metazoans [230]. A more in-depth paper regarding this organism's kinome is soon to be published [231]. The question of non-metazoan tyrosine kinases will hopefully continue to be an area of focus for future experimental studies to clarify the functions and substrate specificities of these putative kinases.

4.5 Evolution of the protein kinases

A number of observations can be made from the phylogenetic distribution of kinase families in this research. Combined with previous studies, the data provides insight into the evolutionary history of protein kinases.

It has been proposed, based on similarity to the ePKs, that the pknB family arose from an early horizontal transfer from Eukarya into Bacteria [173]. The Kannan *et al.* [163] study also supported this apparent sequence relationship. Based on profile-profile alignments of the 20 microbial families found in the GOS dataset, they grouped the families into five general clusters (Table 4.1). The ePKs grouped together with the pknB, Bub1, BLRK, HRK and GLK families.

This study supports the notion that the pknB family appeared after the divergence of the three superkingdoms, as pknBs were found in a number of Bacteria but no eukaryotes and only one Archaea (*Picrophilus torridus*). This would suggest that *P. torridus* acquired pknB by horizontal transfer from a bacterium.

Combining the phylogenetic distribution results from this study with the Leonard *et al.* [173] and Kannan *et al.* [163] observations of ePK sequence similarity, the theory of an early ePK horizontal transfer into Bacteria and subsequent creation of the pknB family is possible though it is also feasible that an ancestral protein was present before the superkingdoms diverged. These results also suggest that pknBs have been lost by a majority of the proteobacteria, as only three of the proteobacteria genomes in this study contained putative pknBs. Conversely, the presence of pknBs remains strongly clustered in the Cyanobacteria, Firmicutes and Actinobacteria species included in this study. Interestingly, these results show putative ePKs in Bacteria follow a somewhat similar distribution to that of the pknBs. They seem to be more prevalent in the Cyanobacteria and Firmicutes, while only one ePK candidate was found in the proteobacteria.

The Kannan *et al.* study included the BLRKs in the same cluster as ePKs and pknBs. Given the dearth of BLRKs in eukaryotes and the complete absence of BLRKs in the archaeal genomes, it is likely this family also emerged following the divergence of the Bacteria from the other two superkingdoms. Furthermore, it is possible the family is present mostly in gamma-proteobacteria. Of the Bacteria genome set in this study, BLRKs were found in only one species (a gamma-proteobacterium). Out of curiosity, I performed a subsequent search against an additional 81 complete bacterial genomes. This search found only three additional microorganisms that contained BLRK: *Pseudomonas syringae*, *Shewanella oneidensis* and *Vibrio parahaemolyticus*. Interestingly, these three are all gamma-

proteobacteria species, suggesting BLRKs may be confined, or mostly confined, to this branch of Bacteria. It will be interesting to see if this trend holds true as more Bacteria are sequenced.

It is possible that given the aforementioned proposed relationship to ePKs and pknBs, the BLRK family may have followed a similar evolutionary pattern as pknBs. That is, perhaps the BLRKs arose from a horizontal transfer and subsequent specialization of a eukaryotic kinase in the proteobacteria ancestor. Alternatively, they may have branched off in Bacteria from an ancestral protein or even the pknBs themselves. This seems a more likely explanation than the scenario of BLRKs being present before Bacteria divergence and subsequently being lost in all Archaea and/or almost all eukaryotes. As more family members are identified, it will be interesting to pursue this theory further by utilizing sequence-based phylogenetic trees and comparing structures as they become available.

A similar scenario is seen in the GLK family. GLKs were found almost exclusively in Archaea, but clustered with the ePK group. However, there were two putative GLKs present in eukaryotes in the related plant parasites *Phytophthora ramorum* and *Phytophthora sojae*. This would seem to suggest that the GLKs evolved from after the Archaea divergence and were horizontally transferred to *P. ramorum and P. sojae*.

The other families that clustered with ePKs, pknBs and BLRKs are Bub1 and HRK [163]. Bub1s and HRKs were found to be mainly present in eukaryotes. No

Bub1s were present in the archaeal genomes, and there were only two questionable hits in the Bacteria. Similarly, no HRKs were apparent in the archaeal and bacterial genomes. Thus, it can be hypothesized that the HRK and Bub1 families arose after Eukarya diverged from Archaea and Bacteria.

Kannan *et al.* also found the RIO and Bud32 kinase families to cluster together, though not as strongly as the aforementioned ePK grouping [163]. According to my study, they also show an almost identical phylogenetic pattern. While no putative Bud32s were located in the *Ciona intestinalis* genome and questionable Bud32s were found in *Stagonospora nodorum* and *Chlamydomonas reinhardtii*, both RIO and Bud32 appeared to be present in all other eukaryotes and Archaea, but very few Bacteria in this study. This would suggest that these kinase families likely emerged early in eukaryotic and Archaea evolution.

One other kinase family was also included in the Bud32/RIO cluster: KdoK. KdoK shows a very different pattern of evolution than Bud32 and RIO. KdoKs were not found in any eukaryotes or Archaea. Even in Bacteria, the presence of KdoK was rather sparse and confined to proteobacteria. Given their suggested similarity to Bud32/RIO and their limited distribution in Bacteria, it can be speculated that this family may have arisen from a horizontal transfer of a Bud32/RIO-related kinase (or their ancestor protein) from an Archaea or eukaryote into the proteobacteria ancestor. Further phylogenetic study of these proteins would shed light on this question of a common ancestor for these three families.

A third cluster of kinases was seen between the CAK, FruK, MTRK and HSK2 families [163]. According to the analysis from this study, these four families show different phylogenetic patterns. It can be hypothesized that the CAK family may be oldest amongst this group, as it is the most widely distributed. It is present in most Eukarya, Bacteria and the Halobacteria and Methanomicrobia archaeal genomes. Thus, it is possible it emerged before the three superkingdoms diverged and was subsequently lost in other Archaea lineages. Alternatively, it may have emerged in either Eukarya or Bacteria and then was horizontally transferred early in their evolution. This possible ancient origin is further supported by a study in which Scheeff *et al.* [45] created a phylogenetic tree based on an extensive structural comparison of kinase structures. They found strong evidence that the so-called atypical kinases, which include CAKs, diverged very early from the typical ePKs.

The emergence of the related HSK2s, conversely, appears to be a much more recent evolutionary event. No evidence of HSK2s was found in eukaryotes or Archaea. Instead, the results indicate HSK2s are limited to the proteobacteria.

The remaining two families, MTRK and FruK, are more puzzling. MTRKs were found in a very small number of species. They were most prevalent in the eukaryotes, with putative kinases present in the plant genomes, *Tetrahymena thermophila* and possibly *Thalassiosira pseudonana*. MTRKs were only found in three Bacteria (two firmicutes and one proteobacteria), and no Archaea. Given this limited sampling of MTRKs, it is difficult to pinpoint their evolutionary history. It

seems likely that they developed more recently, after the three superkingdoms diverged. Whether they first emerged in eukaryotes (perhaps in the plants) and were later transferred to the few Bacteria genomes, or vice versa, is unclear.

FruKs were present in many eukaryotes, a few Bacteria and only one Archaea. Thus, it is possible they emerged early in eukaryotic evolution and were later transferred to the *Haloarcula marismortui* archaeum. This scenario seems to fit with the Bacteria occurrences as well, as the hits were clustered in the cyanobacteria except for two putative gamma-proteobacteria FruKs. Thus a FruK may have been transferred at some point from eukaryotes to a cyanobacteria ancestor. It will be interesting to see if this pattern holds true as more Bacteria are sequenced and analyzed.

The remaining families clustered separately in Kannan *et al*, although the CapK, RevK, MalK and UbiB kinases showed more similarity to the aforementioned three groupings than the PI3Ks, AlphaKs and IDHKs [163]. Of the former kinase families, the results suggest UbiB is the most ancient. UbiB kinases were present in almost every eukaryote, as well as spread throughout many Archaea and Bacteria. This would indicate that UbiB may have been present before the three superkingdoms diverged.

MalK, RevK and CapK were present in far fewer genomes. MalKs were found in two Archaea (one creanarchaeota and one euryarchaeota), and four Bacteria. Three of the four Bacteria were proteobacteria, but they were in different

classes. Thus, although this family appears to have emerged after eukaryotes diverged, it is difficult to pinpoint its exact point of origin. The ancestry of the two novel groupings of RevK and CapK is also hard to ascertain, though both families appear to have arisen after the Bacteria diverged. CapKs were identified in only two Bacteria species; both were gamma-proteobacteria. RevKs were also found to be present only in Bacteria, though the six species were scattered throughout different phylums. The results suggest IDHKs have also appeared after the superkingdoms diverged. Putative IDHKs were found in only two of the bacterial genomes, both of which are gamma-proteobacteria.

The two remaining families, AlphaKs and PI3Ks, show very little sequence similarity to the other kinases [163]. Their evolutionary relationship to the protein kinase-like superfamily has been established through detailed structural comparison [45]. Scheeff *et al.* showed that the closest structural relative to the alpha kinases may even be the PI3K family. An evolutionary relationship between alpha kinases and the typical ePKs was also noted by Drennan *et al.* [213].

Both of these atypical families appear to have emerged more recently than the previously described atypical kinase family CAK. AlphaKs and PI3Ks are present throughout different eukaryotic phylums, but they were not found in the Archaea and Bacteria species. Alpha kinases do not appear to be as universally-distributed as the PI3Ks. While the PI3Ks were present in all eukaryotic genomes and thus likely emerged early in eukaryotic evolution, the alpha kinases were most

prevalent in the metazoa, kinetoplastid and stramenopile genomes. Alpha kinases are scattered throughout some other eukaryotic organisms, but on a much smaller scale. If alpha kinases emerged early in eukaryotic evolution, these results suggest a number of genomes have since lost this family. An alternative explanation is that the family arose later in eukaryotic evolution and subsequently spread to select organisms in other lineages.

In an unrooted sequence-based phylogenetic tree construction of human kinases, the CAMK and AGC groups cluster to one end of the tree and the RGCs, TKs and TKLs are present at the opposite end, with CMGC, CK1 and STE falling in between (Figure 4.16) [4]. The Scheeff *et al.* [45] study produced a similar phylogenetic tree from their structure-based alignment of kinases (Figure 4.17). The main difference between the two trees was the placement by Scheeff *et al.* of the CK1 group closer than the STEs to the TK and TKL groups. In the Manning *et al.* tree, the STEs were found to occupy this position.

My findings can be correlated with these ePK phylogenetic trees quite nicely. The results suggest that the AGC, CAMK, CK1 and CMGC are the oldest families. These four groups were present in all of the eukaryotes. The STEs also likely arose early in eukaryotic evolution, though they have apparently since been lost in a few eukaryotes (*Encephalitozoon cuniculi* and at least some apicomplexa species).

The remaining kinase families present at the “other” end of the phylogenetic tree appear to be younger groups, particularly TK and RGC. The RGCs were

confined to the metazoan genomes in this study, suggesting they arose later in eukaryote evolution after the metazoans diverged. Likewise, the TKs were mainly found in metazoans, with additional possibilities in *Entamoeba histolytica*. Given the previously mentioned apparent presence of TKs in a choanoflagellate, it is possible that the TKs emerged in the choanoflagellate and metazoan ancestor and were then horizontally transferred to *E. histolytica* [230,231]. The TKLs were likely produced earlier in the eukaryotic tree than TKs and RGCs, as putative TKLs were found in many, but not all eukaryotic genomes. However, the exact emergence of TKL is more difficult to pinpoint. Perhaps the simplest explanation is that they diverged from the other ePKs early in eukaryotic evolution and were then later lost by the kinetoplastids, most fungi, and the *Theileria* Apicomplexa ancestor.

4.6 Conclusion

In conclusion, I have presented a comprehensive study of the presence and absence of both microbial and eukaryotic kinase families in over 100 genomes. I have traced the phylogenetic patterns of evolution of these families through the Archaea, Bacteria and Eukarya superkingdoms of life. I have also attempted to interpret and correlate my results with those of sequence and structure-based kinase studies previously undertaken. I look forward to future advancements in this area, and eagerly anticipate an even deeper understanding of kinase evolution as more and more diverse species are sequenced and made available for study.

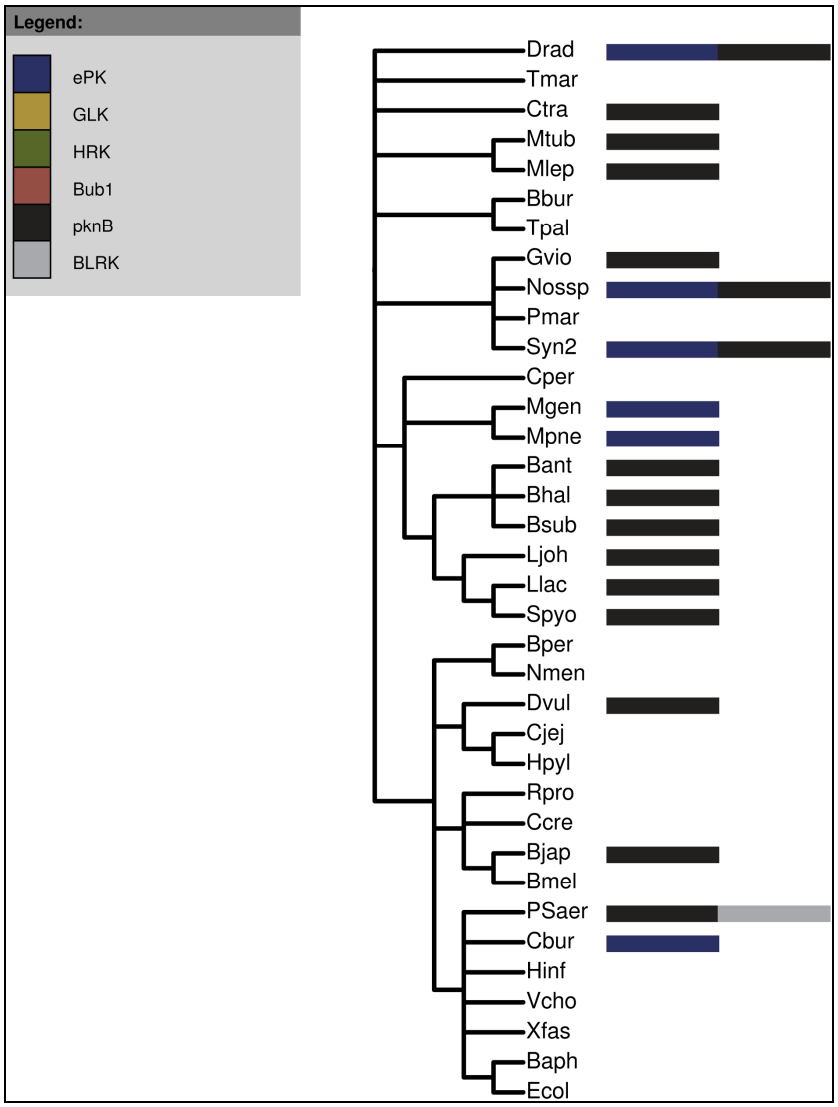

Figure 4.1 Bacterial microbial kinases, group 1.

Group 1 microbial kinase families present in Bacteria. The blue bars represent the ePK family, the black bars represent the pknB family, and the gray bars represent the BLRK family.

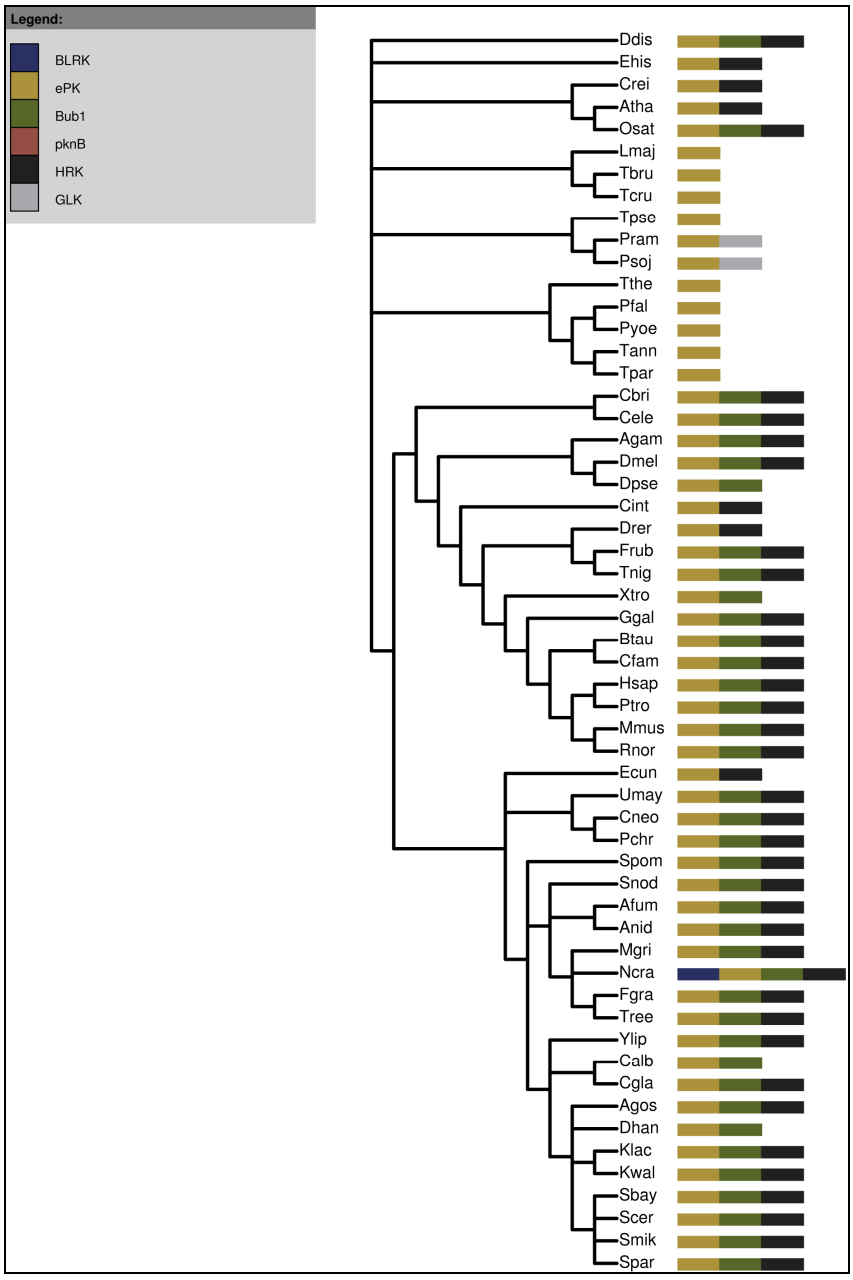

Figure 4.2 Eukaryotic microbial kinases, group 1.

Group 1 microbial kinase families present in Eukarya. The blue bars represent the BLRK family, the gold bars are the ePK family, the green bars are the Bub1 family, the black bars are the HRK family, and the gray bars are the GLK family.

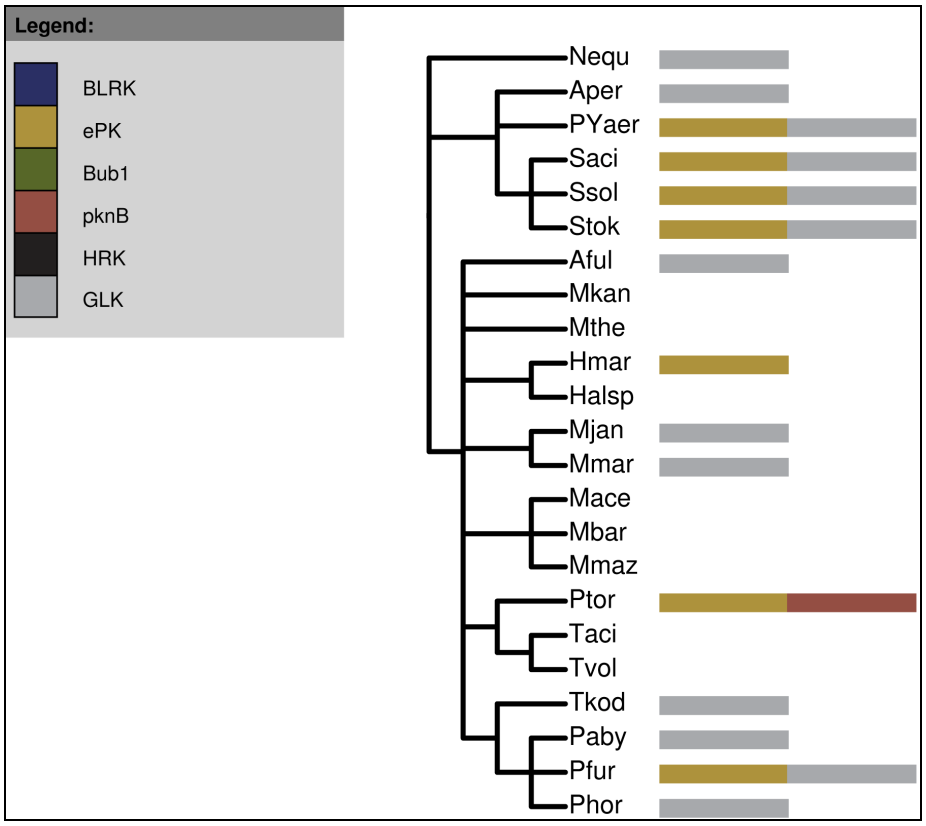

Figure 4.3 Archaeal microbial kinases, group 1.

Group 1 microbial kinase families present in Archaea. The gold bars are the ePK family, the red bars are the pknB family, and the gray bars are the GLK family.

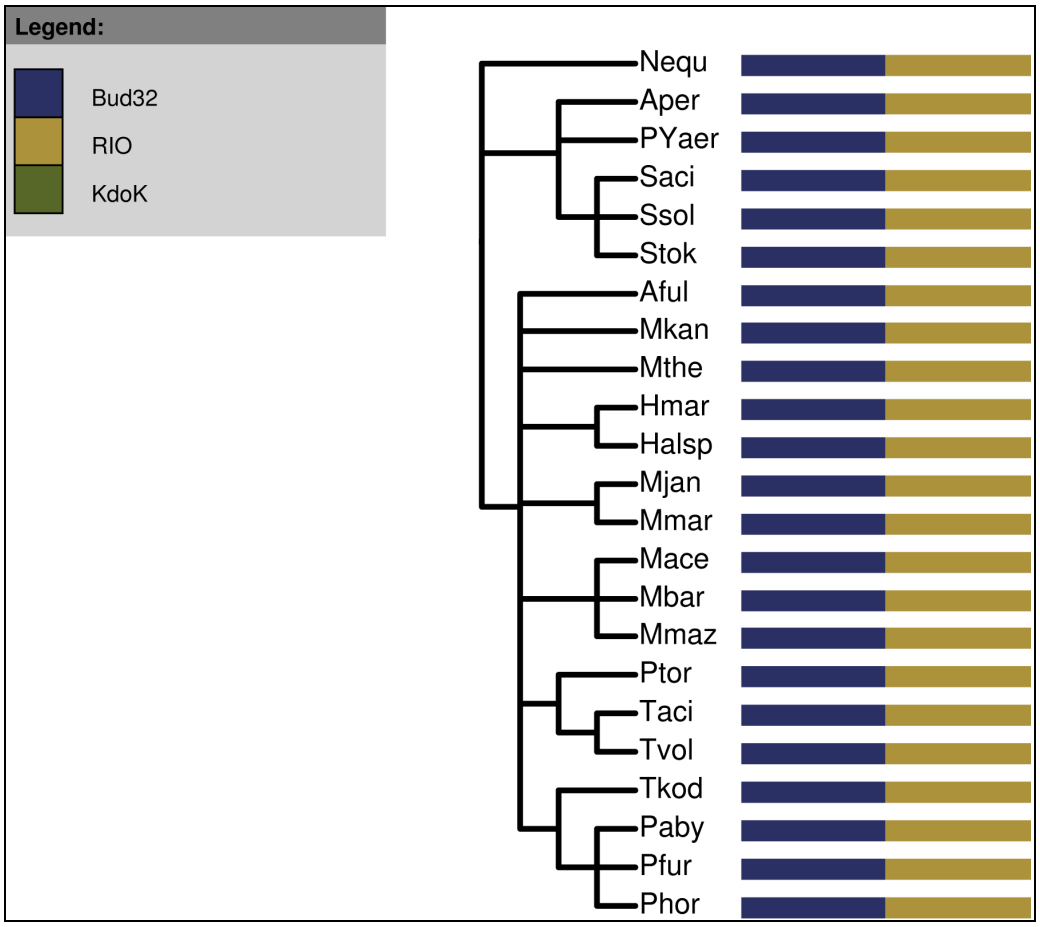

Figure 4.4 Archaeal microbial kinases, group 2.

Group 2 microbial kinase families present in Archaea. The blue bars are the Bud32 family and the gold bars are the RIO family.

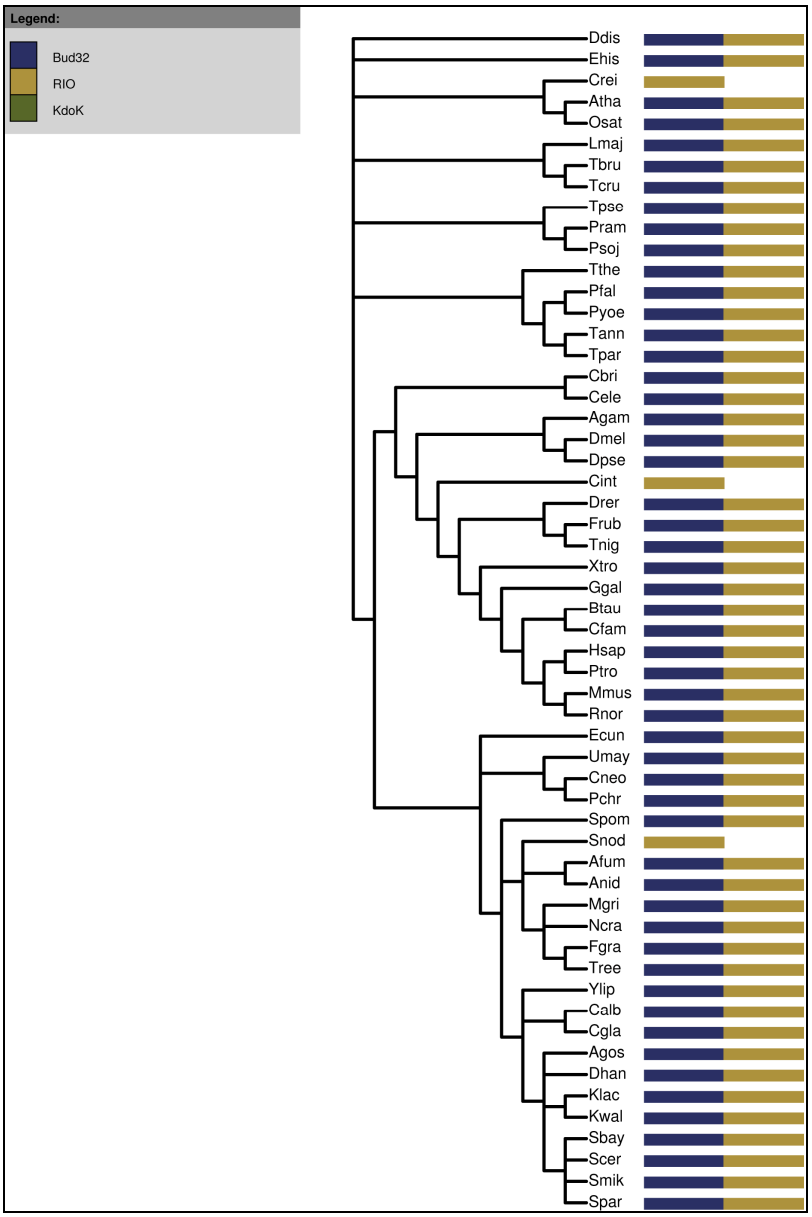

Figure 4.5 Eukaryotic microbial kinases, group 2.

Group 2 microbial kinase families present in Eukarya. The blue bars are the Bud32 family and the gold bars are the RIO family.

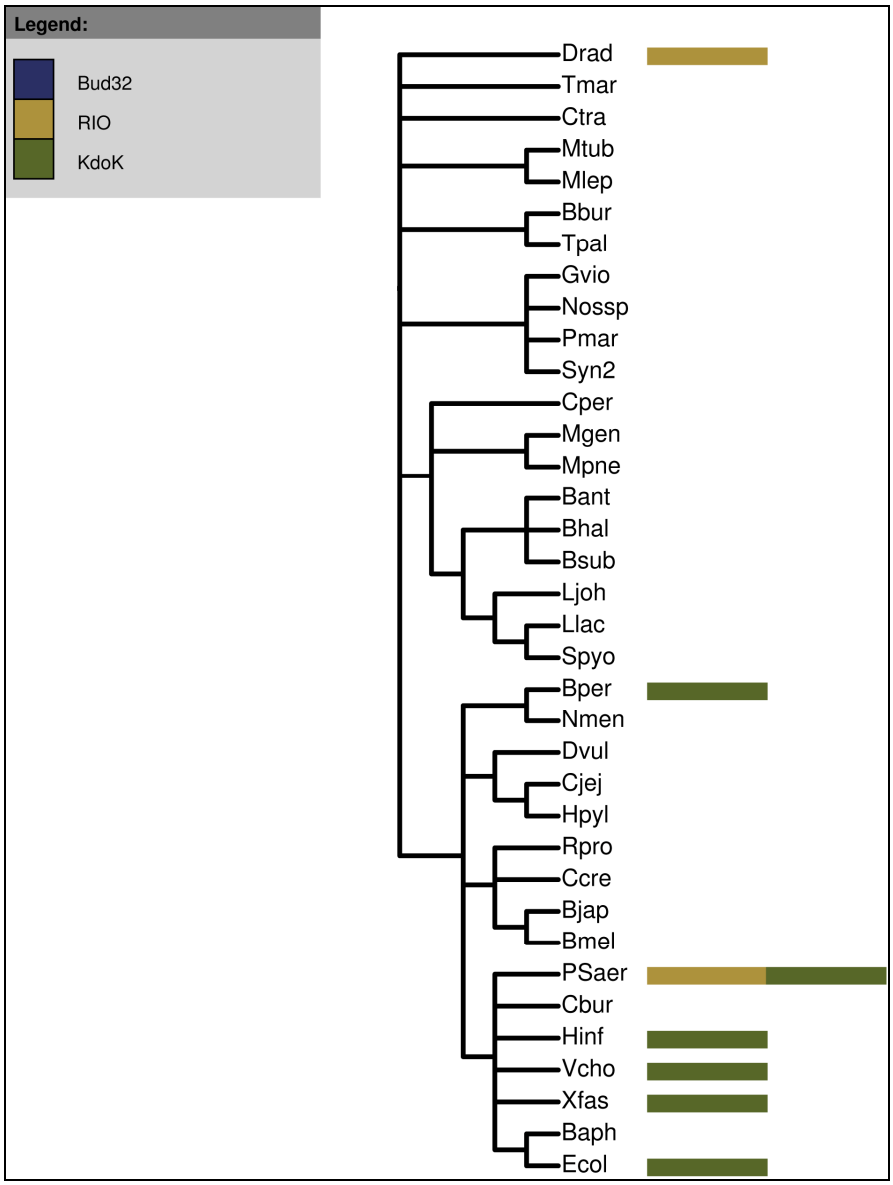

Figure 4.6 Bacterial microbial kinases, group 2.

Group 2 microbial kinase families present in Bacteria. The gold bars are the RIO family and the green bars are the KdoK family.

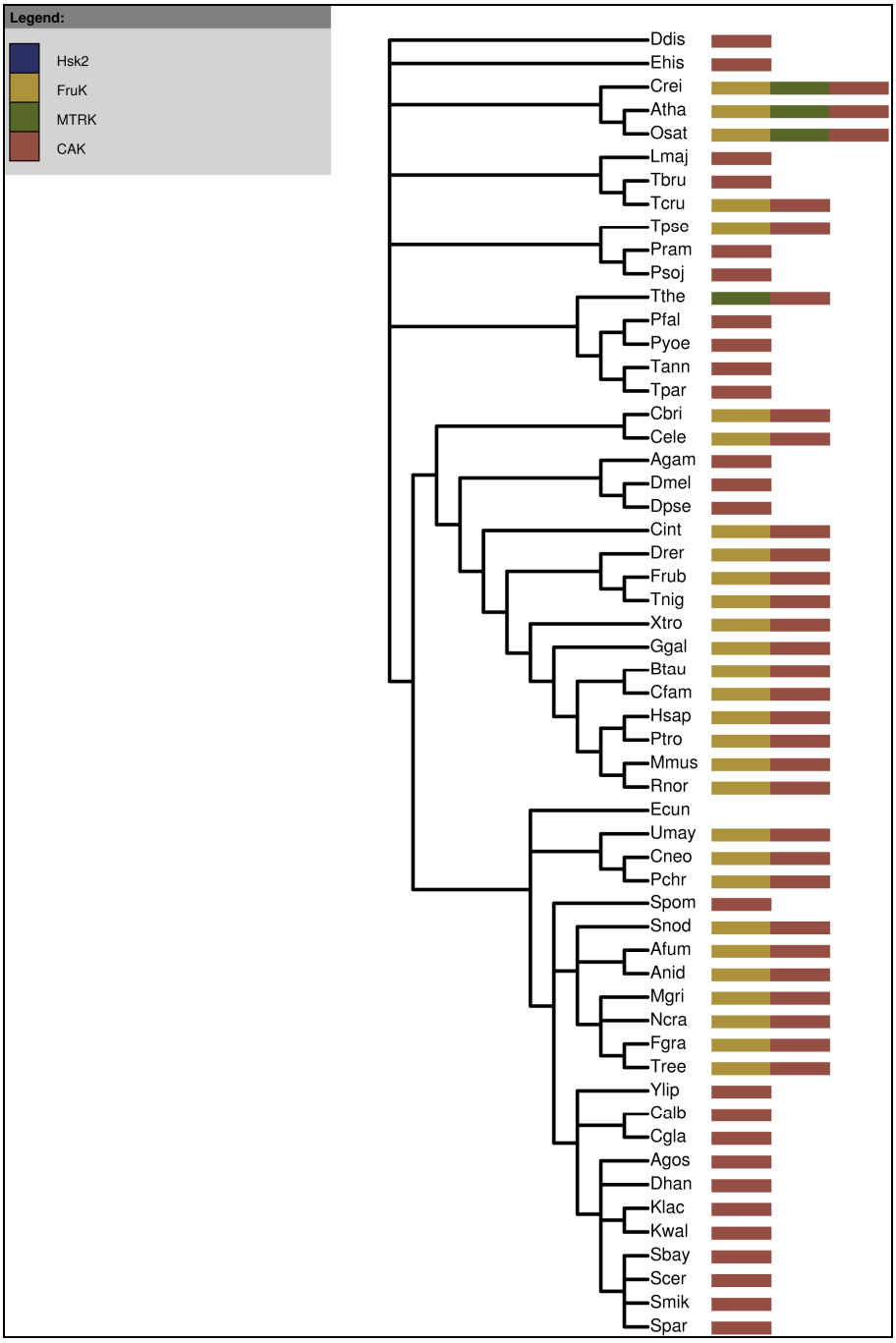

Figure 4.7 Eukaryotic microbial kinases, group 3.

Group 3 microbial kinase families present in Eukarya. The gold bars are the FruK family, the green bars are the MTRK family, and the red bars are the CAK family.

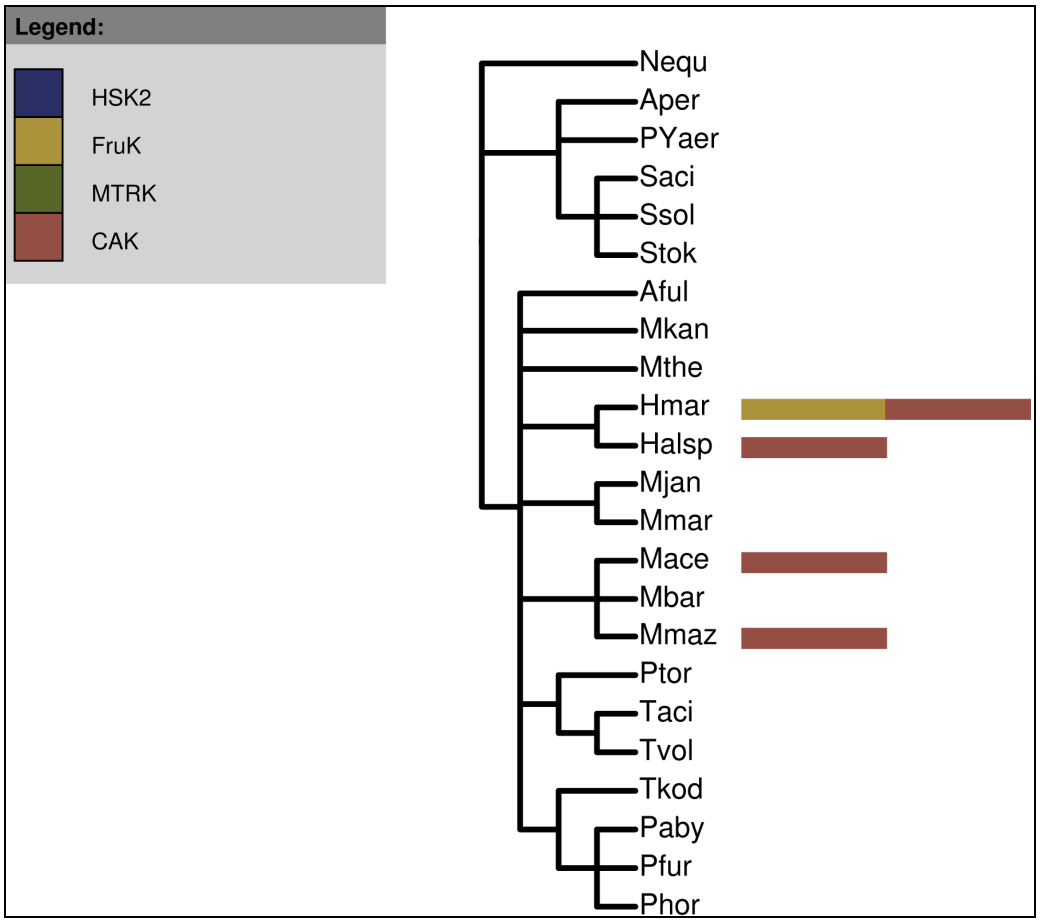

Figure 4.8 Archaeal microbial kinases, group 3.

Group 3 microbial kinase families present in Archaea. The gold bars are the FruK family and the red bars are the CAK family.

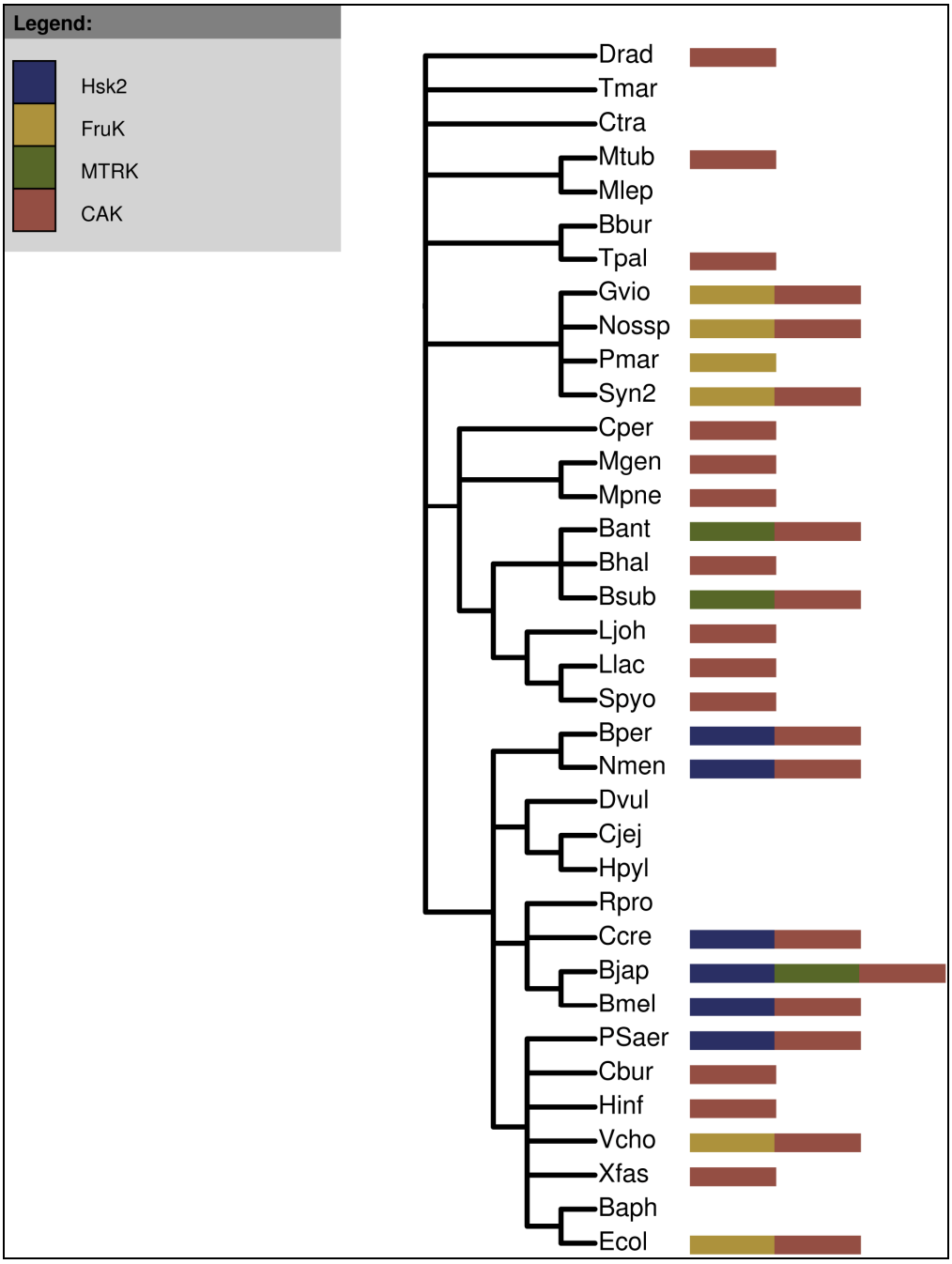

Figure 4.9 Bacterial microbial kinases, group 3.

Group 3 microbial kinase families present in Bacteria. The blue bars are the Hsk2 family, gold bars are the FruK family, the green bars are the MTRK family, and the red bars are the CAK family.

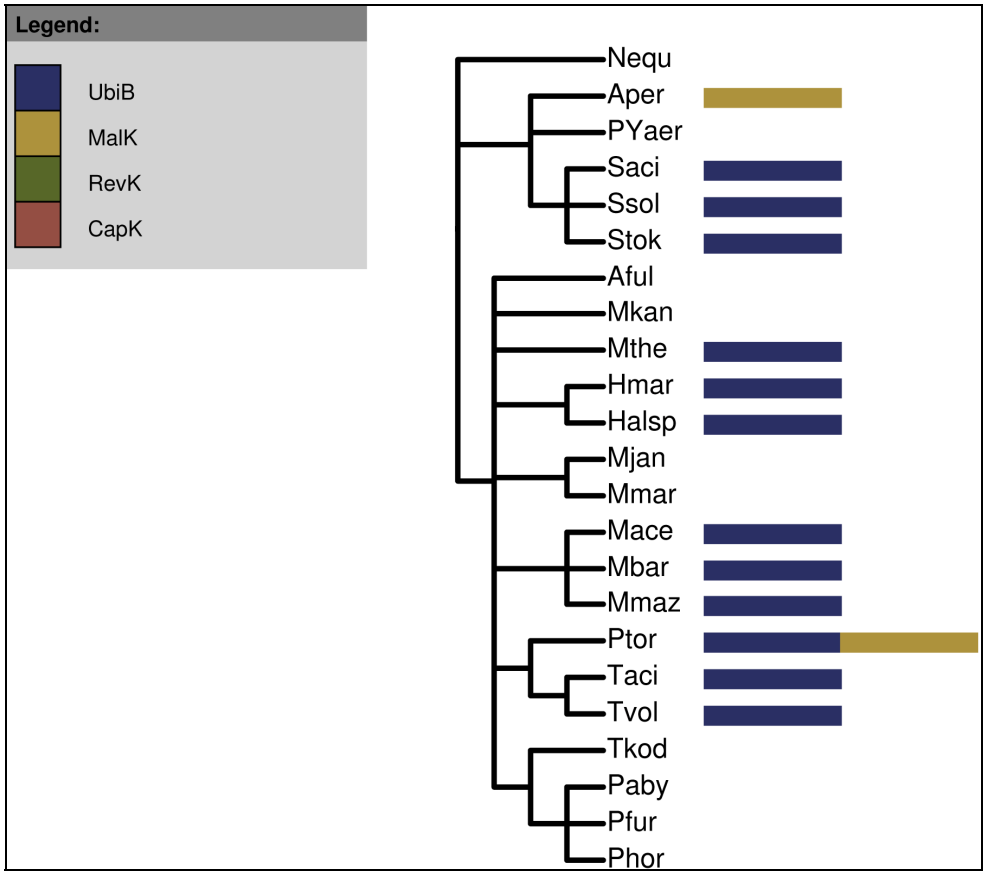

Figure 4.10 Archaeal microbial kinases, unrelated families.

Unrelated microbial kinase families present in Archaea. The blue bars are the UbiB family and the gold bars are the MalK family.

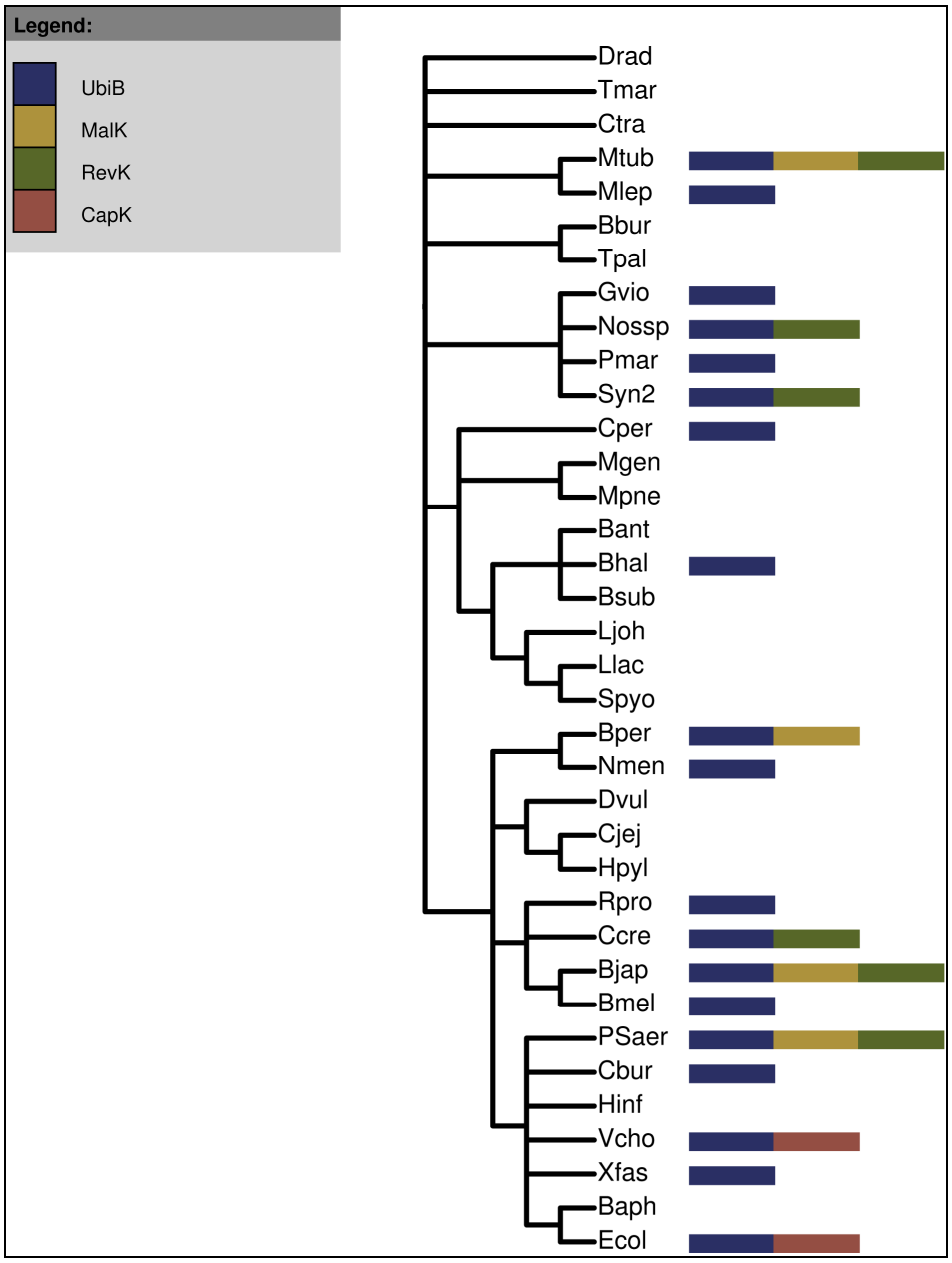

Figure 4.11 Bacterial microbial kinases, unrelated families.

Unrelated microbial kinase families present in Bacteria. The blue bars are the UbiB family, the gold bars are the MalK family, the green bars are the RevK family, and the red bars are the CapK family.

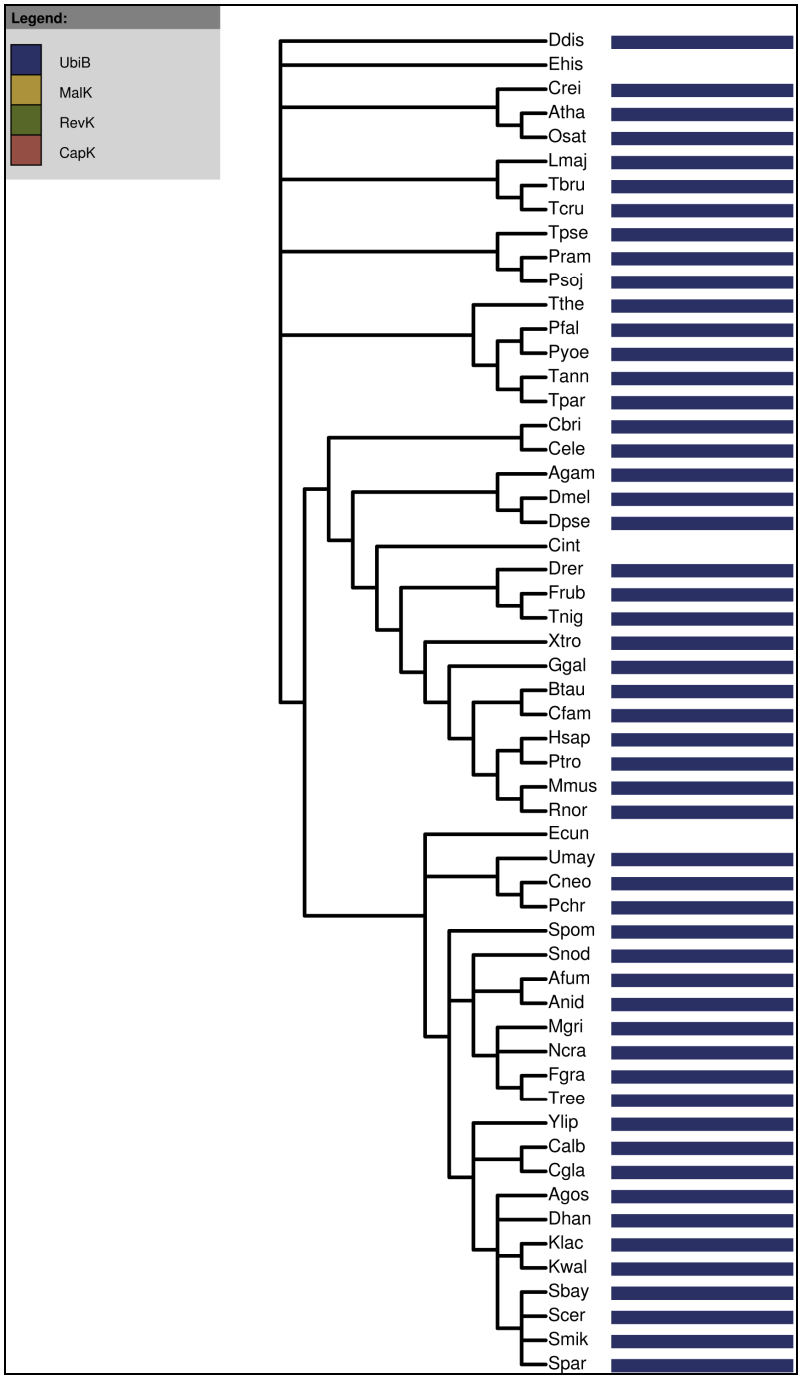

Figure 4.12 Eukaryotic microbial kinases, unrelated families.

Unrelated microbial kinase families present in Eukarya. The blue bars are the UbiB family.

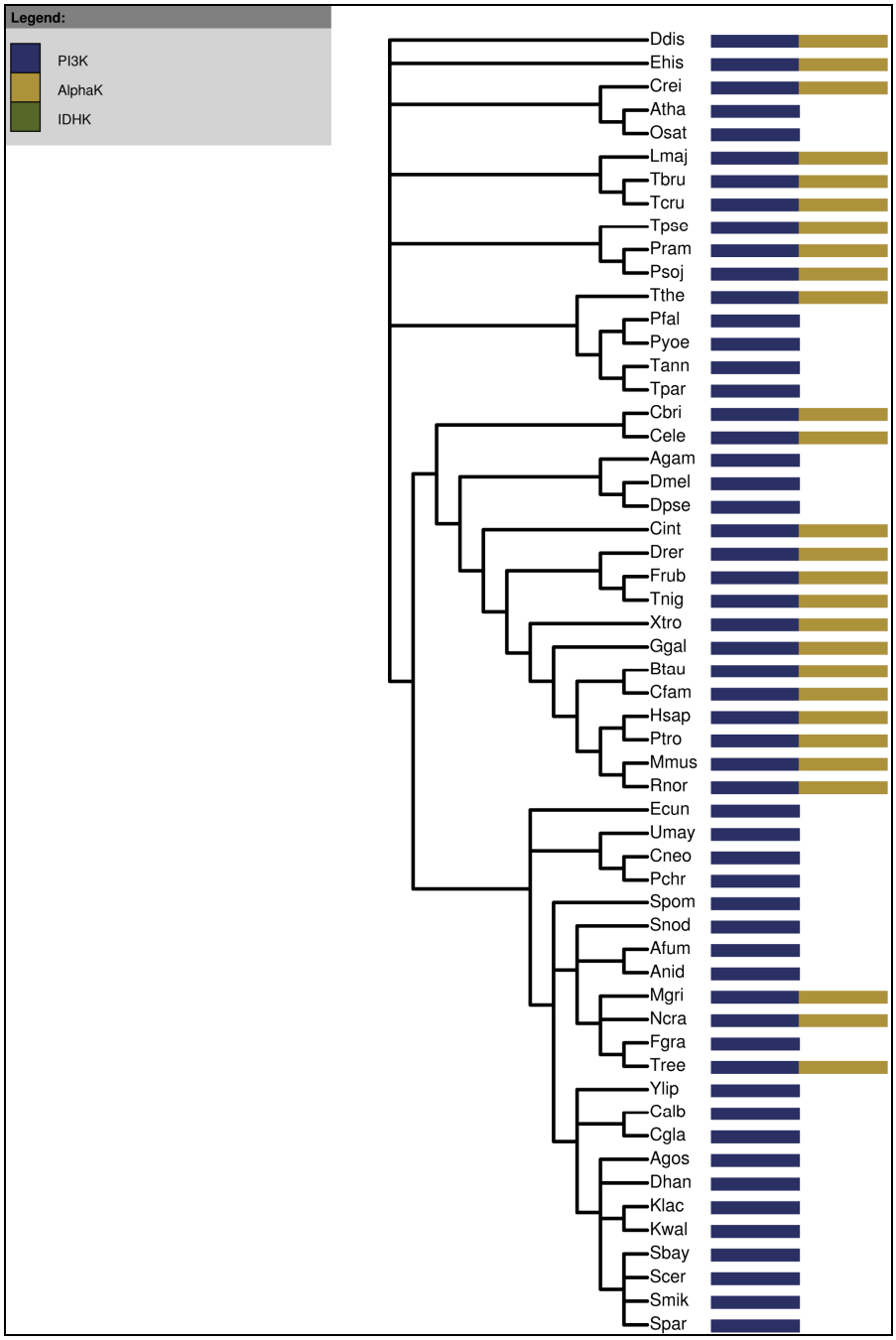

Figure 4.13 Eukaryotic PI3K and Alpha kinase families.

PI3K and Alpha kinase families present in Eukarya. The blue bars are the PI3K family and the gold bars are the Alpha kinase family.

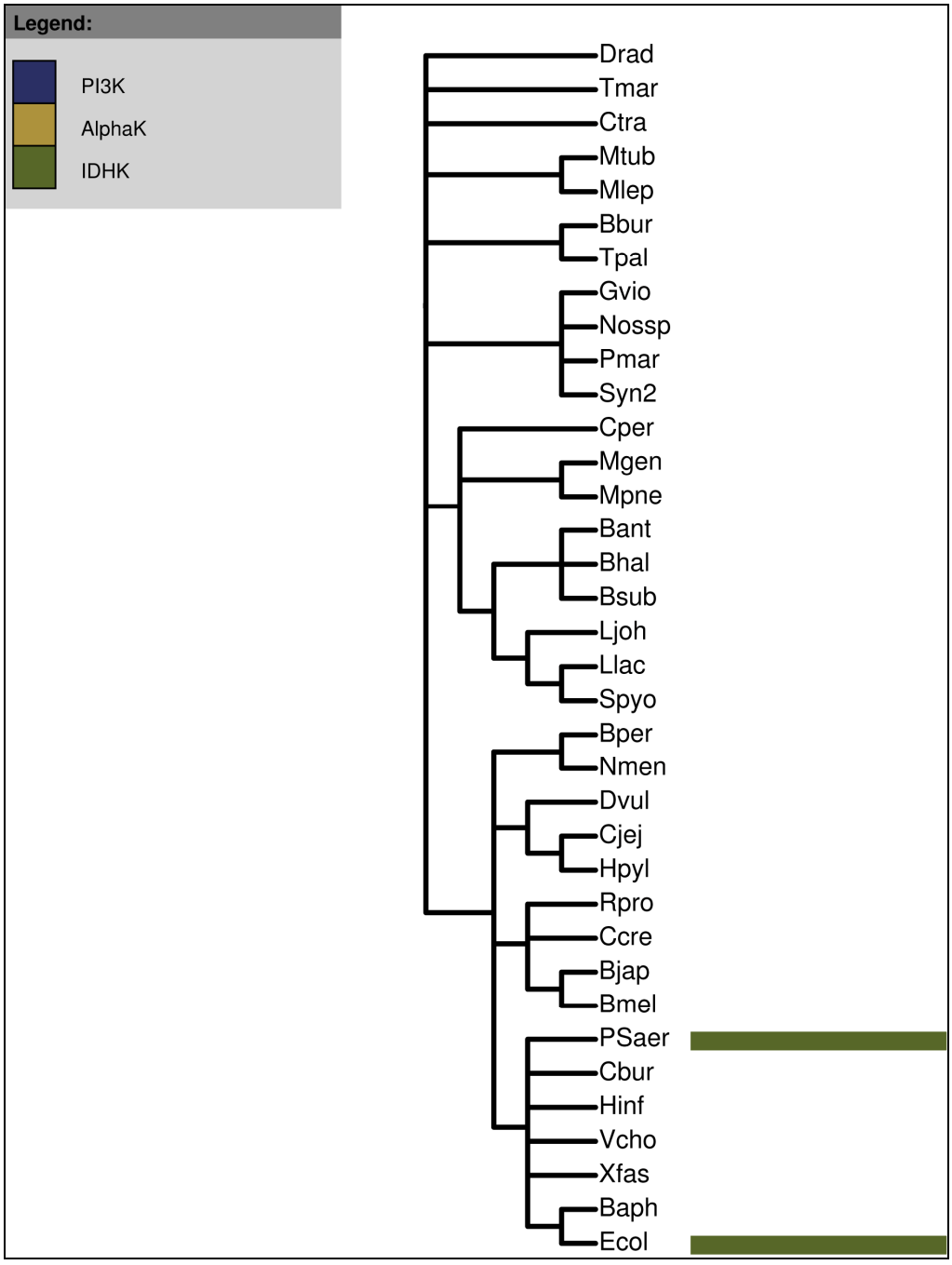

Figure 4.14 Bacterial IDHK family.

IDHK family present in Bacteria. The green bars are the IDHK family.

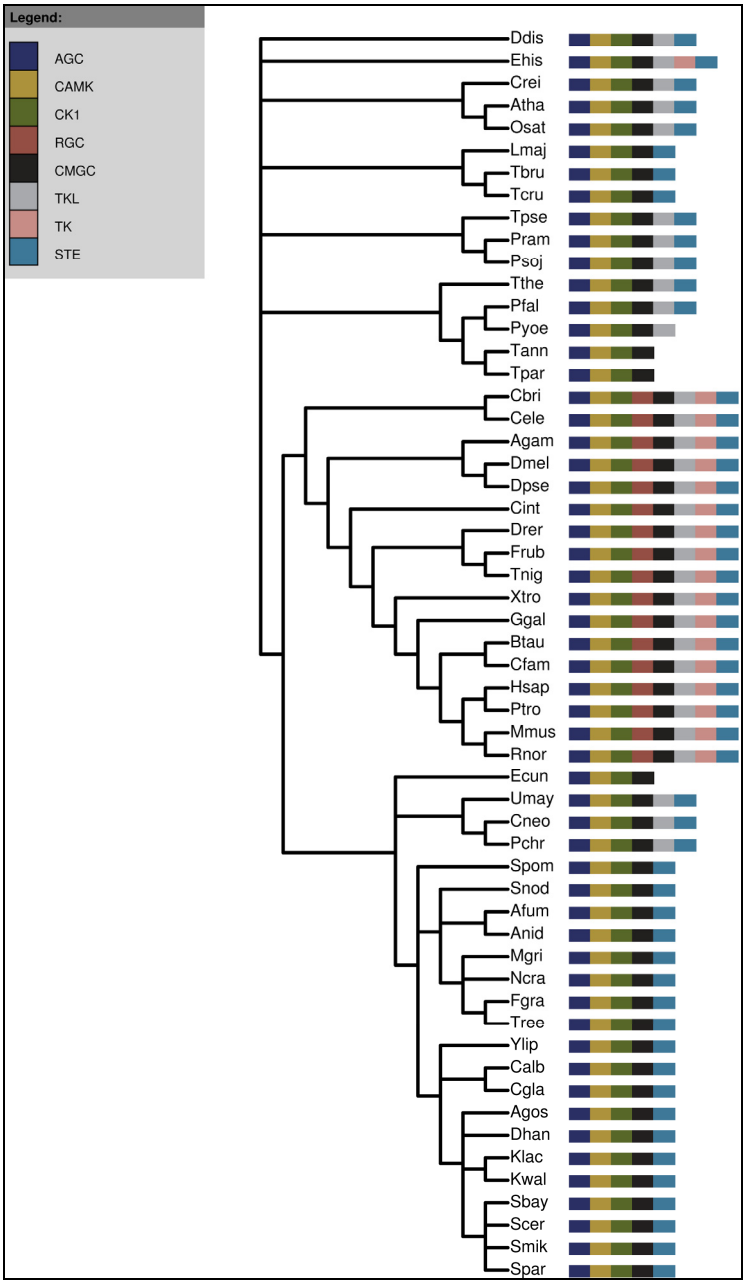

Figure 4.15 Eukaryotic ePK groups.

ePK families present in Eukarya. Dark blue bars are the AGC group, gold bars are the CAMK group, green bars are the CK1 group, red bars are the RGC group, black bars are the CMGC group, gray bars are the TKL group, pink bars are the TK group, and light blue bars are the STE group.

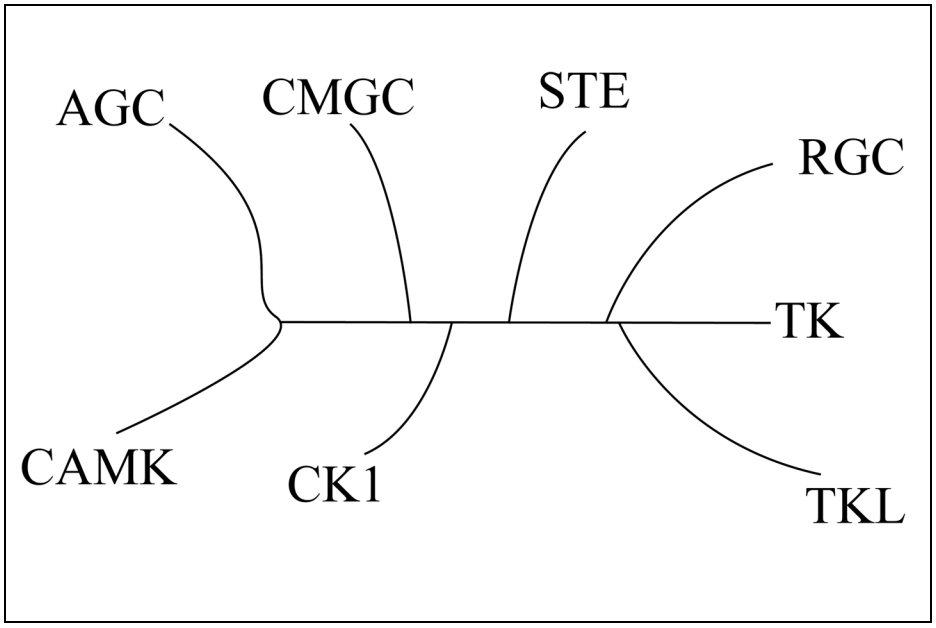

Figure 4.16 Representation of Manning human ePK phylogenetic tree.

A representation of the unrooted sequence-based eukaryotic protein kinase tree, as found by Manning *et al.* Note that phylogenetic branch length distances are not to scale.

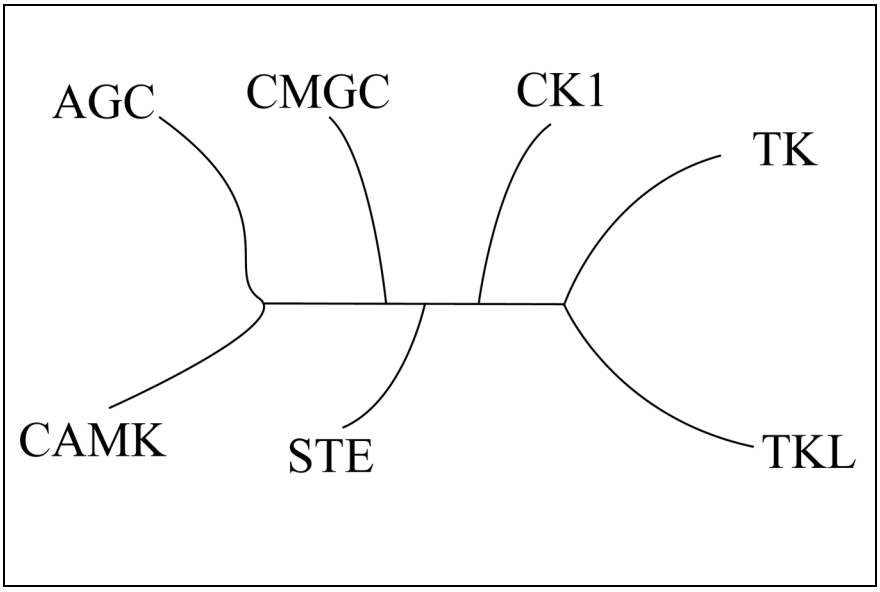

Figure 4.17 Representation of Scheeff ePK phylogentic tree.

A representation of the unrooted structure-based eukaryotic protein kinase phylogenetic tree, as found by Scheeff *et al*. Note that the RGC kinase group was not included in this study. Phylogenetic branch length distances are not to scale.

Table 4.1 Microbial kinase family names and abbreviations.

MICROBIAL KINASE FAMILY NAMES/ABBREVIATIONS
Group 1:
BLRK (Bacterial Leucine-Rich Kinase)
Bub1 (Budding Uninhibited by Benzimidazole)
PknB
HRK (Haspin-related Kinase)
GLK (Glycosylase-linked Kinase)
ePK (Eukaryotic Protein Kinase)
Group 2:
Bud32
RIO
KdoK (3-deoxy-D-manno-octulosonic acid Kinase)
Group 3:
CAK (Choline and Aminoglycoside Kinase)
HSK2 (Homoserine Kinase)
FruK (Fructosamine Kinase)
MTRK (Methylthioribose Kinase)
Individual clusters, but some similarity to above groupings:
UbiB (aka ABC1 in eukaryotes)
MalK (Maltose Kinase)
RevK (Reverse Kinase)
CapK (Capsule Kinase)
Individual clusters, most distant from other families:
PI3K (Phosphoinositide Kinase)
AlphaK
IDHK (Isocitrate Dehydrogenase Kinase)

Table 4.2 Eukaryotic protein kinase-like superfamily groups.

EUKARYOTIC PROTEIN KINASE GROUP NAMES/ABBREVIATIONS:
AGC (cAMP-dependent protein kinase A; protein kinase G; protein kinase C)
CAMK (Ca^{2+}/calmodulin-dependent Kinase)
CK1 (Casein kinase I)
CMGC (Cyclin-dependent; Mitogen-activated protein kinase; Glycogen synthase; CDK-like)
RGC (Receptor Guanylate Cyclase)
STE (Sterile yeast kinase)
TKL (Tyrosine Kinase-Like)
TK (Tyrosine Kinase)

5 Comparisons of Kinase and Phosphatase Phylogenetic Profiles

5.1 Introduction

This work has comprehensively searched over 100 genomes from the three superkingdoms for the presence of numerous kinase and phosphatase families. While understanding the evolution of these families in the context of their respective superfamilies is valuable, we can also use our data to consider kinase and phosphatase evolution in relation to each other. That is, we can compare the evolution of kinase and phosphatase families that act either on each other, or share a common substrate.

Performing such an analysis on protein kinases and phosphatases presents several challenges. Among these complications is an incomplete, or lack of, experimental functional characterization, questions as to the relevance of substrate binding studies *in vitro,* and the "promiscuity" of some kinases and phosphatases *in vivo*.

Traditionally, kinases have attracted more experimental attention than phosphatases, leading to gaps in knowledge of substrate specificity and function for some phosphatase families. Even among the kinases, some are better characterized

than others. Additionally, many experimental studies have been carried out *in vitro.* This can lead to mistaken assumptions about kinase or phosphatase function *in vivo,* as these catalytic domains can demonstrate broader substrate specificity *in vitro* than they do *in vivo* [232-234]. In the cell, kinase and phosphatase activity can be regulated by factors such as subcellular location and interactions with scaffolding, anchoring and adaptor proteins [5,235].

It has also been found that since many kinases and phosphatases function as part of signal transduction cascades, they sometimes have multiple substrates and are involved in pathways that may include a number of other kinases and phosphatases. These other proteins may impart an evolutionary pressure on the kinase or phosphatase family in question. Consequently, an analysis of "one-to-one" kinase and phosphatase pairings does not necessarily present the entire evolutionary story of a protein family. My comparison of "partnered" kinase and phosphatase families thus relies on what may be incomplete information and likely represents only part of the global evolutionary picture. However, it is still possible to glean useful and interesting evolutionary information based on current knowledge for some families, as I demonstrate here.

5.2 Methods

Pellegrini *et al.* [236] first introduced the method of analyzing the presence or absence of proteins between species to determine functional linkage. They termed this comparing "phylogenetic profiles". The technique is based on the assumption that two proteins that interact with each other experience common evolutionary pressures and need to adapt to changes in each other. As such, they are expected to evolve in a correlated manner and show a similar pattern of presence or absence in the same organisms [236-239]. This general idea has since been extended and used in predicting protein-protein interactions for large sets of data [238,240-242]. Here, I utilize this concept of coordinated phylogenetic profiles of coevolving proteins to study the evolutionary relationship between kinases and phosphatases with common substrates.

In hopes of focusing on kinase and phosphatase families that were most likely to have some experimental characterization and would provide the most interesting evolutionary events, this research focused mainly on families that exist in some, but not all, metazoan genomes as identified by Manning *et al.* [152]. In some cases, this study involved building HMMs for more specific subfamilies to supplement the previously generated data. In such instances, models were constructed as described in previous chapters from protein sequences of known family members and used to search complete proteomes. Putative family members were verified by checking sequence alignments with known family members,

reciprocal NCBI BLAST [9] searches against the non-redundant database [58] and classification by the Pfam [20] or PANTHER database [22], where available. This procedure was used for the following families presented later in this chapter: Relaxin family, Met kinases, DEP-1 phosphatase, Fer/Fes kinases, PTP-MEG2 phosphatase, LIM kinases and TES kinases.

Kinase and phosphatase "pairings" were determined through literature searches for experimentally verified interactions. Phylogenetic profiles of such families were compared, and the resulting presence and absence of the families was graphically plotted on a eukaryotic tree based on NCBI's taxonomy database. I first present here a "control" case of a non-kinase/phosphatase ligand and receptor family that is known to have coevolved with each other. I then describe my findings regarding specific kinase and phosphatase families.

5.3 Relaxin family

I initially chose a non-kinase and non-phosphatase test case of known receptor/ligand coevolution from the literature. I share that test case here as an example of what might be found in interacting kinase and phosphatase families.

The relaxin family peptides and associated receptors have an interesting evolutionary history. Relaxin was first identified in the 1920's as an important factor in the widening of the birth canal of guinea pigs. Additional relaxin-like family

members have since been identified and implicated in a number of cellular processes, including wound-healing, cardiovascular responses and mammary gland development [243]. The human genome contains seven known members of the relaxin-family: three relaxins (relaxin-1, relaxin-2 and relaxin-3), and four insulin-like (INSL3, INSL4, INSL5 and INSL6). Structural analysis of relaxin has suggested that relaxin and insulin may have evolved from a common ancestral gene [244].

Despite this long-standing knowledge of the existence of relaxin, receptors for the relaxin family were not identified until 2002 [245]. The human genome is now known to contain at least four such relaxin family receptors, all of which are G protein-coupled receptors: LGR7, LGR8, GPCR135 and GPCR142 [246].

LGR7 is known to bind relaxin-1, relaxin-2 and relaxin-3. LGR8 also has the ability to bind relaxin-1 and relaxin-2, but seems to prefer binding INSL3 [247]. GPCR135 and GPCR142 were first demonstrated to bind relaxin-3 [248,249]. INSL5 was later also identified as a strong agonist of GPCR142, but not GPCR135. Additionally, GPCR142 and INSL5 expression was found to have similar tissue expression patterns. Based on this evidence, the authors suggested that the endogenous ligand for GPCR142 is likely INSL5 [250]. The receptors for INSL4 and INSL6 are not yet known [246].

Wilkinson and colleagues have conducted an in-depth study of the evolution of the relaxin family and their receptors. Based on the aforementioned experimental

characterization of the ligand/receptor interactions, as well as sequence and phylogenetic analysis, Wilkinson *et al.* [246] have put forth the following evolutionary scenario of the relaxin family.

It is currently believed that relaxin-3 most closely represents the ancestral relaxin sequence. This ancestor appears to have emerged sometime prior to the divergence of fish [251]. Based on biochemical analysis and expression profiles, relaxin-3 is the most likely endogenous ligand of GPCR135. Wilkinson *et al.* concluded that relaxin-3 probably acquired the ability to bind GPCR142 and LGR7 at a later date. Relaxin-3 is also thought to have coevolved with the GPCR135 receptor, based in part on their correlated phylogenetic emergence and the multiple duplications of both genes in fish species [246].

The other relaxin family members seem to have diverged from relaxin-3 later in vertebrate evolution, possibly after relaxin-3 evolved its LGR7 binding function. Wilkinson *et al.* hypothesize that at some point these other proteins eventually lost the ability to bind GPCR142 and GPCR135, becoming specific substrates for LGR7 and LGR8 [246]. They also conclude that GPCR142 coevolved with INSL5, while the other two ligand/receptor pairings (relaxin-1, relaxin-2/LGR7 and INSL3/LGR8) did not coevolve. Rather, as mentioned above, they believe that the ability to bind LGR7 and LGR8 was somehow gained later in relaxin evolution [252].

Models were built for the relaxin family proteins, used to search completed eukaryotic proteome drafts, and the results were compared with that of Wilkinson *et*

al, who looked for relaxin family members using TBLASTN [244,246]. The presence and absence of relaxin family proteins was then mapped on eukaryotic phylogenetic trees and their phylogenetic profiles were compared.

As can be seen in Figure 5.1, the aforementioned hypothesized coevolution pairing of GPCR135 and relaxin-3 show the same pattern of evolution. Both appear to have emerged at some point prior to the divergence of fish.

Functional coupling can also further support this idea of coevolution. For example, INSL5 pseudogenes have been found in the rat (*Rattus norvegicus*) and dog (*Canis familiaris*) genomes [244]. As can be seen in Figure 5.2, the results show these two species have also apparently lost the corresponding receptor GPCR142 gene. Wilkinson *et al.* also noted this loss [252]. This coordinated emergence and loss lends credence to the theory that GPCR142 and INSL5 have coevolved.

Conversely, the receptor-ligand pairing of LGR8 and INSL3 is very different, as seen in Figure 5.3. Comparison of their phylogenetic profiles shows no indication of coevolution. Indeed, Wilkinson *et al.* have concluded that there is currently no specific evidence that these two proteins have coevolved [246].

I was interested in applying this idea of phylogenetic pattern analysis to kinase and phosphatase families. The next section presents case studies of related kinase and phosphatase families, compares and analyzes their phylogenetic profiles

in the eukaryotic tree, and hypothesizes what such data may indicate about their evolutionary history.

5.4 Met kinases and DEP-1 phosphatase

The Met tyrosine kinase family contains three members: Met, Ron and Sea [253]. Humans have two members of the Met tyrosine kinase family (Met and Ron), while the pufferfish *Takifugu rubripes* has been found to express all three [4,254]. These kinases are receptors for growth factors and have been implicated in several cancers, including breast, colon, lung and pancreas carcinomas [253-256]. When activated, they impact important cellular functions such as growth, motility and differentiation [257,258]. Substrate specificity of Met family members is still somewhat unclear, as most studies have been conducted *in vitro*. However, it has been suggested that these substrates include the docking protein Gab1 [259,260] and β-catenin, a protein involved in signal transduction and regulation of cell adhesion [253].

DEP-1 is a class III receptor protein tyrosine phosphatase that is expressed in a variety of cell types and has been implicated in cell differentiation and cell growth inhibition [259,261]. The human DEP-1 contains eight fibronectin type III repeats, a transmembrane domain and one cytoplasmic PTP domain [262]. While DEP-1 is thought to have several substrates, it seems to share the Gab1 and β-catenin

substrates with the Met family receptors. Additionally, DEP-1 has been shown to dephosphorylate the Met tyrosine kinase receptor itself [259]. Manning *et al.* previously noted that the Met kinase family is present in human and worm, but not fly or yeast [152]. Given the multiple aforementioned overlapping substrates between Met kinases and the DEP-1 phosphatase (Figure 5.4), I investigated whether DEP-1 showed a similar phylogenetic pattern as Met kinases.

The results confirm Manning *et al.'s* observation that no Met family members are present in *Drosophila melanogaster* (Figure 5.5) [152]. Additionally, no Met kinases were found in the mosquito *Anopheles gambiae* or *Drosophila pseudoobscura*, suggesting that this absence is likely not due to a genome sequencing error and may in fact be true of insects in general. Met kinases were found in all other metazoan genomes, although the *Ciona intestinalis* hit was somewhat questionable. This protein showed a strong hit to the PANTHER Met family model and returned hits to other Met family proteins in a BLAST search of NCBI's non-redundant database. However, the domain arrangement differs from mammalian Met proteins, as the *C. intestinalis* protein is missing the N-terminal Sema domain commonly conserved in other Met kinases. It is unknown whether the *C. intestinalis* gene prediction is mistaken, or if this domain is truly absent.

DEP-1 phosphatases were found only in vertebrate genomes including *Canis familiaris, Rattus norvegicus, Mus musculus, Pan troglodytes, Homo sapiens* and *Bos taurus* (Figure 5.5). Additionally, receptor protein tyrosine phosphatases in the

same class (type III/R3) were identified in worm, *Ciona intestinalis*, fly and mosquito genomes. While humans contain five class III RPTPs, fly appears to contain only two class III ancestral RPTPs. Fly also contains at least four other RPTPs, but they are more difficult to correlate with the human subfamilies [263]. A similar, but singular copy, type III RPTP ancestor is also present in worm [264]. I confirmed these findings, and also noted single copies in *Anopheles gambiae* and *Ciona intestinalis*. Thus, these results support previous suggestions that the type III RPTPs were present in an ancestral state prior to vertebrate divergence and have since diversified into the present-day five distinct genes seen in humans [263].

Due to the late radiation of the type III RPTPs, the loss of Met kinases in insects cannot be correlated with any loss of DEP-1, though it can be stated that RPTPs in the same receptor class were retained in insects. It is unclear why insects have lost the Met family, but considering these results in light of previous experimental studies produces the following interesting evolutionary scenario.

One function that may have evolved after the RPTP type III duplication is the ability of Met kinase to directly bind and phosphorylate Gab1. While Gab1 adaptor proteins are involved in signaling cascades initiated by other receptor tyrosine kinases, it is thought they usually bind the Grb2 adapter protein and not the receptor tyrosine kinase directly [265]. There is evidence, though, that Gab1 is a direct substrate of Met kinase [266]. Gab1 homologs exist in worm (*soc-1*) and fly (*dos*), but they are not thought to contain the same direct receptor tyrosine kinase binding

ability [265,267,268]. Dos is later dephosphorylated by the non-receptor tyrosine phosphatase Corkscrew [269,270]. Additionally, evidence indicates that Gab1 can be dephosphorylated by the Corkscrew homolog SHP-2 [268]. Thus, current experimental data suggests that prior to RPTP type III duplication, Gab1 homologs were recruited and bound only indirectly to RTKs.

Additionally, the phosphatase DEP-1 has been shown to preferentially dephosphorylate a human Met phosphotyrosine residue that is critical to Gab1 docking. Evidence from this study also suggested that DEP-1 dephosphorylation of the Grb2 binding site and the Met kinase catalytic loop was much more gradually affected by increasing concentrations of DEP-1 [259]. Thus, it is possible that DEP-1, Gab1 and Met have coevolved to some extent as a result of Met directly binding Gab1. Given that the worm Gab1 homolog is not thought to bind to the Met family kinases [265,268], it would be interesting to study whether there is any direct, significant dephosphorylation functional effect of the *C. elegans* class III ancestral receptor protein tyrosine phosphatase on the worm Met family kinases. Such a study may provide further support for the idea that these three proteins have coevolved. It would also be worthwhile to expand this study to other possible substrates of these families to further enhance our understanding of their evolutionary history.

5.5 PTP-MEG2 phosphatase and Fer/Fes kinases

PTP-MEG2 is a member of the non-receptor protein tyrosine kinase family [271]. It is one of the few tyrosine kinases to reside on an internal membrane (specifically, of secretory vesicles). PTP-MEG2 contains a Sec14p homology domain. This domain, similar to the yeast protein Sec14p, is responsible for targeting PTP-MEG2 to the secretory vesicle membrane, binding phosphoinositides and contributing to the regulation of PTP-MEG2 [272,273].

PTP-MEG2 is involved in the regulation of vesicle fusion in hematopoetic cells and has the ability to dephosphorylate N-ethylmaleimide-sensitive factor (NSF) [272,274]. NSF is a cytosolic protein involved in the disassembly of the soluble NSF attachment receptor (SNARE) complexes present between fusing membranes [272]. Previous studies have shown that NSF may be regulated by serine/threonine phosphorylation [275]. At least two tyrosine kinases in mammals are also thought to have the ability to phosphorylate NSF—Fes (aka Fps) and Fer (Figure 5.6). When phosphorylated at Tyr-83, NSF is functionally inactive. Thus, PTP-MEG2 is a positive regulator of vesicle fusion [272]. PTP-MEG2 has also recently been implicated in insulin signaling, though its precise substrate target in this process is not yet clear [271].

As seen in Figure 5.7, the results indicate PTP-MEG2 phosphatases are present in vertebrates and insects. No PTP-MEG2s were located in either the original *Ciona intestinalis* proteome draft or an updated genome version. PTP-

MEG2s were also not found in either of our worm genomes. Assuming the branch point of insects before *C. intestinalis* is correct, these results suggest PTP-MEG2 may have emerged after the divergence of worm, and perhaps has been lost in the *C. intestinalis* genome.

The Fer/Fes kinase family, conversely, is present in all metazoans, including the worms and *C. intestinalis* genomes. It is worth noting that this kinase family has an interesting evolutionary history. While vertebrates contain two family members (Fer and Fes), insects and *C. intestinalis* apparently only have one. This single copy present in *D. melanogaster* has been previously noted [276,277]. The results indicate this may hold true for other insects as well. Additionally, only one Fer/Fes gene was found in the *Ciona intestinalis* genome. Thus, the ancestral Fer/Fes gene present in insects and *C. intestinalis* may have duplicated following the divergence of *C. intestinalis* and diverged into separate Fer and Fes genes. The worm genomes, however, underwent a significant family expansion with multiple copies of Fer/Fes-related genes present [153]. Given the aforementioned gene copy numbers and phylogenetic pattern of evolution from the results, this gene expansion apparently took place after worms diverged.

While evidence of coevolution is difficult to discern from these phylogenetic profiles, they may still teach us something about the evolution of each individual family. Comparing the phylogenetic profiles of PTP-MEG2 phosphatases and Fer/Fes kinases, it can be seen that the Fer/Fes kinase family appears to have

emerged prior to PTP-MEG2. This raises the question of whether NSF is regulated by tyrosine phosphorylation in worms and *C. intestinalis*. If so, an interesting avenue of research may be to isolate the phosphatase responsible for dephosphorylating NSF in these genomes. If such a phosphatase were located, comparison studies with PTP-MEG2 may provide insight into the origins and evolution of PTP-MEG2. Conversely, if it is not regulated by tyrosine phosphorylation, perhaps the emergence of PTP-MEG2 was in response to an acquired ability of Fer/Fes to phosphorylate NSF.

5.6 LIM and TES kinases and Slingshot phosphatases

5.6.1 Regulation of ADF/Cofilin family proteins

Actin reorganization is critical to cell shape and motility. Cells move forward when actin polymerizes at the leading edge and disassembles at the rear of the actin network [83]. LIM kinases are involved in the regulation of the actin network by phosphorylating serine-3 in proteins of the actin-depolymerizing factor (ADF)/cofilin family [80]. This inactivates ADF/cofilin proteins and inhibits their filament-severing activity. Similar to LIM kinases, TES kinases are also involved in cytoskeleton regulation and are able to phosphorylate ADF/cofilin proteins [278,279]. ADF/cofilin proteins are dephosphorylated, and thus activated, by slingshot (SSH) phosphatases and chronophin (discussed later) (Figure 5.8) [79,84].

5.6.2 LIM kinases

The human genome contains two LIMK family members, LIMK1 and LIMK2 [280]. LIMK1 was first depicted in 1994 [281,282]. LIMK2 was published the following year [281]. Vertebrate LIM kinases contain multiple protein domains. These include two N-terminal LIM domains, a PDZ domain, and a C-terminal kinase domain [283]. The kinase domains of LIMK1 and LIMK2 share a 70% sequence identity, while the overall sequence identity is 50% [281]. There has been some confusion over the kinase activity of LIM kinases. Sequence analysis alone is inconclusive. LIM kinases have a DLNSHN motif, which does not match either the common serine/threonine motif (DLKXXN) or tyrosine kinase motifs (DLAARN or DLRAAN) [281]. Experimental evidence originally showed LIM kinases had serine/threonine kinase activity, but a subsequent study demonstrated LIMK1 can also phosphorylate tyrosine residues *in vitro* [284].

5.6.3 TES kinases

TESK1 (testis-specific protein kinase 1) was first reported in 1995 in rat, mouse, and human [285]. A second TES kinase, TESK2, was later found in rat and human [279,286]. The two TES kinase domains are 71% similar in sequence identity to each other [279]. Both TESK1 and TESK2 have the ability to

phosphorylate cofilin/ADF family proteins, limiting their ability to sever actin filaments and inhibiting the process of actin rearrangement [278,279].

The LIM kinase and TES kinase domains share roughly 40-50% sequence identity. However, the overall domain structure of the two families differs. TESKs contain no LIM motifs and also have a proline-rich C-terminal extension [285].

5.6.4 LIM kinase and TES kinase evolution

As seen in Figure 5.9, our LIMK models found LIM kinases in higher eukaryotes. LIMKs were found in insects (*Drosophila* and mosquito), mammals (human, chimp, mouse, rat and cow), chicken, fish (zebrafish and pufferfish), frog and sea squirt. The TES kinases show an identical phylogenetic presence and absence pattern (Figure 5.9). Putative TESKs were identified in, among other species, *Drosophila melanogaster, Canis familiaris, Homo sapiens, Mus musculus, Gallus gallus, Xenopus tropicalis, Takifugu rubripes* and *Ciona intestinalis*.

I further delved into these two families by attempting to quantify the number of LIM and TES kinase genes present in each genome. Protein hits to the LIMK and TESK models were validated through sequence analysis and PANTHER and NCBI BLAST characterization, and then traced back to their respective gene predictions.

Interestingly, a similar gene copy number pattern was also found in these two families (Figure 5.9). Only one TESK and LIMK gene were found in the insects

(*Drosophila melanogaster, Drosophila pseudoobscura* and *Anopheles gambiae*) and in *Ciona intestinalis*. The results indicate that both the TES and the LIM kinase families underwent an apparent gene family expansion after the divergence of *C. intestinalis*. The human, dog, chimp, mouse, rat and chicken genomes all had two TESK genes and two LIMK genes. Two TESKs, but only one LIMK gene was found in the *Bos taurus* genome draft. *Xenopus tropicalis* had one TESK and two LIMKs.

The results also indicate that fish genomes may contain an even greater number of LIM and TES kinases. For example, *Danio rerio* and *Tetraodon nigroviridis* both had three putative TESK genes. Two TESKs were found in *Takifugu rubripes*. The LIM kinase search showed similar results. All three of the aforementioned fish genomes contained three putative LIMK genes.

5.6.5 Dephosphorylation of ADF/Cofilin proteins

As previously discussed in this work (section 3.5.7), the slingshot phosphatases are able to dephosphorylate ADF/cofilin proteins, inducing actin depolymerization activities [79,83]. It has also been shown that SSHs may additionally downregulate LIM kinase activity through dephosphorylation of LIMKs themselves [80]. Slingshot phosphatases were present in only a small number of metazoan genomes, including *Homo sapiens, Bos taurus, Gallus gallus, Xenopus*

tropicalis, Drosophila melanogaster and *Danio rerio* (Figure 5.9). No clear slingshot phosphatases were found in plants, worm or fungi genomes.

More recently, a second phosphatase family was found to have the ability to dephosphorylate ADF/cofilin family proteins. Chronophin was reported as a HAD-type serine protein phosphatase by Gohla *et al.* in 2005. They identified putative orthologs in human, rat, mouse, zebrafish, fly, worm, yeast, *E. coli* and *Arabidopsis thaliana.* An alignment of sequences from these species revealed three conserved HAD motifs. Motif I: ψψψDXDX(V/T); motif II: ψψψ(S/T); motif III: K(X_n)ψψψGDXXXX(D/E) (where ψ is a hydrophobic residue) [84]. The serine/threonine phosphatases PP1 and PP2A have also been shown to dephosphorylate cofilin [287], although they are not thought to be indispensible since Bamburg *et al.* demonstrated that cofilin can still be rapidly dephosphorylated when these proteins are inhibited [288].

As expected, almost all eukaryotic genomes had a hit to the chronophin models and were confirmed to match the commonly conserved motifs described above. However, there were a few exceptions. The *Gallus gallus* hit was strong, but the sequence appeared to be an incomplete protein prediction and was too short to contain motif I. *Magnaporthe grisea* had a short gap at the end of motif I, missing the conserved (V/T) residue. *Encephalitozoon cuniculi*, *Entamoeba histolytica*, *Theileria annulata* and *Theileria parva* did not have any strong chronophin matches. Several possible reasons exist for this occurrence. The current proteome drafts may

not yet contain the chronophin sequence, the sequence may be too divergent for the models to find, or the genomes may simply not contain chronophin. This last conjecture seems unlikely, given chronophin's widespread conservation in eukaryotic genomes. However, all of the above four genomes are obligate parasites. Thus, it is possible that the genomes have lost the chronophin gene at some point and utilize an alternative mechanism, possibly even a pathway dependent on the host cell. *E. cuniculi*, in particular, has undergone gene loss and now has a very compact genome [146]. It is the smallest known eukaryotic genome to this point [147].

5.6.6 Comparing the evolution of LIMK, TESK, SSH and Chronophin

As can be seen in Figure 5.9, the LIM kinase, TES kinase and slingshot phosphatase families all share the same evolutionary pattern. That is, all three families are present in the same species and appear to have emerged in metazoan genomes at the same time. The results indicate that the LIMKs, TESKs and SSHs arose at some point after worm diverged. Chronophin seems to be a much older protein family. Chronophin phosphatases are spread throughout the Eukarya superkingdom, as previously mentioned.

Given these results of identical phylogenetic patterns between the LIMKs, TESKs and SSHs, I sought to compare the gene copy numbers of LIMK and TESK to that of SSH. As previously described, I used the results of the HMM searches to

trace back the proteins found in each family to their respective chromosomal locations and counted the total number of genes present in every organism. An interesting pattern was found in the metazoan genomes.

Ciona intestinalis and all insects studied (*Drosophila melanogaster, Drosophila pseudoobscura* and *Anopheles gambiae*) contained only one gene for each family. Most other genomes contained multiple genes of each family (Figure 5.9). *Canis familiaris, Homo sapiens, Pan troglodytes, Mus musculus* and *Rattus norvegicus* all have two LIMK genes, two TESK genes and three SSH genes. Only one gene copy of LIM kinase and TES kinase was found in *Bos taurus* and *Xenopus tropicalis,* respectively. There is, however, a second candidate LIMK in the cow genome. It returns strong BLAST hits to LIMKs in other species, but the protein ends just prior to where the kinase catalytic domain is expected to be. It was also found that the scaffold on which this protein prediction is located contains a nearby kinase domain which best matches other LIM kinase domains. Thus, it is suspected this may be a protein prediction error and the cow genome likely contains two LIMK genes. The three fish genomes all contained multiple genes of each family. Three LIMKs, three TESKs and two SSHs were found in *Danio rerio*; three LIMKs, two TESKs and three SSHs in *Takifugu rubripes*; and three LIMKs, three TESKs and four SSHs in *Tetraodon nigroviridis*.

This indicates that one or more gene duplication events took place in each protein family following the divergence of *Ciona intestinalis*. A 2007 study of LIM

domain evolution also concluded a LIM kinase duplication took place after *C. intestinalis* divergence [289]. This is an interesting duplication time point, as it has previously been suggested that one or two rounds of whole genome duplications may have taken place in eukaryotes sometime during early vertebrate evolution, after the divergence of sea squirts. This theory has been criticized in light of the completed human genome sequencing and subsequent discovery of far fewer genes than previously expected, but other analysis of different gene clusters has supported this idea [290]. If this "2R" hypothesis is true, many duplicated genes have since been lost [148]. But imagining for a moment that the 2R scenario is correct, the fact that the TESK, LIMK and SSH families have all retained multiple gene copies is intriguing and perhaps further supportive of the idea that these families have coevolved.

The concurrent emergence of the LIM kinase, TES kinase and SSH phosphatase families, as well as their "coordinated" gene duplication and subsequent retention of multiple gene copies suggests these families may have coevolved. No structure of the TES kinases has yet been solved. However, a LIMK structure was published in 2005 and a SSH structure was recently released [31,291]. In the future, we hope to use this data to compare the sequences and structures of these different families and move beyond simple phylogenetic analysis to search for signs of coevolution on a molecular scale.

5.7 Conclusion

I have presented here analysis of how subsets of kinase and phosphatase families have evolved in relation to each other. Clearly, there is much that can be learned when we move beyond individual families and interpret evolution in the context of related systems and pathways. The advent of complete genome sequencing will inevitably lead to the defining of complete kinomes and phosphatomes in more and more species. We hope to utilize such information in the future to further our study of kinase and phosphatase evolution, both apart and in association with each other. Additionally, it would be interesting to undertake such a study in the context of entire signal transduction pathways as opposed to just part of the larger system.

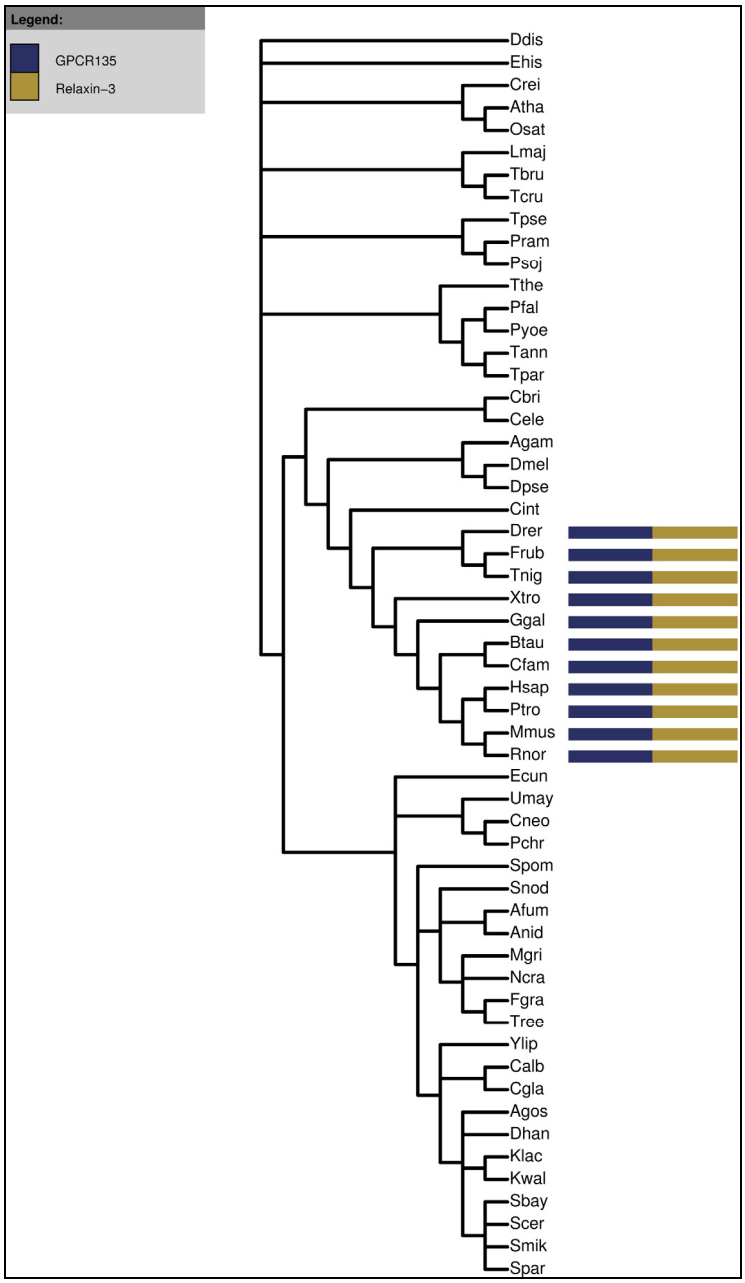

Figure 5.1 Eukaryotic GPCR135 and Relaxin-3 families.

Presence and absence of GPCR135 and Relaxin-3 families in eukaryotic genomes. Blue bars represent the GPCR135 family and gold bars represent the Relaxin-3 family.

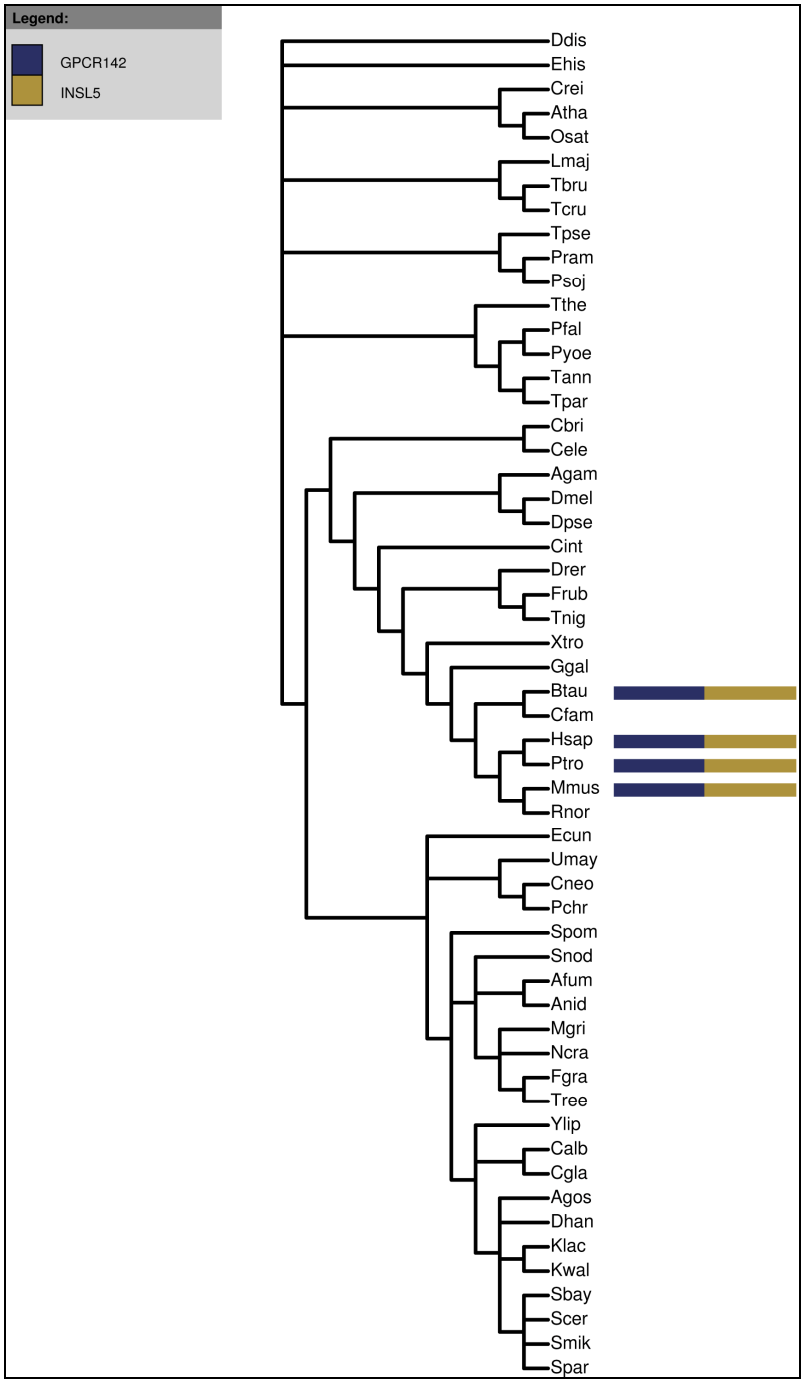

Figure 5.2 Eukaryotic GPCR142 and INSL5 families.

Presence and absence of GPCR142 and INSL5 families in eukaryotic genomes. Blue bars represent the GPCR142 family and gold bars represent the INSL5 family.

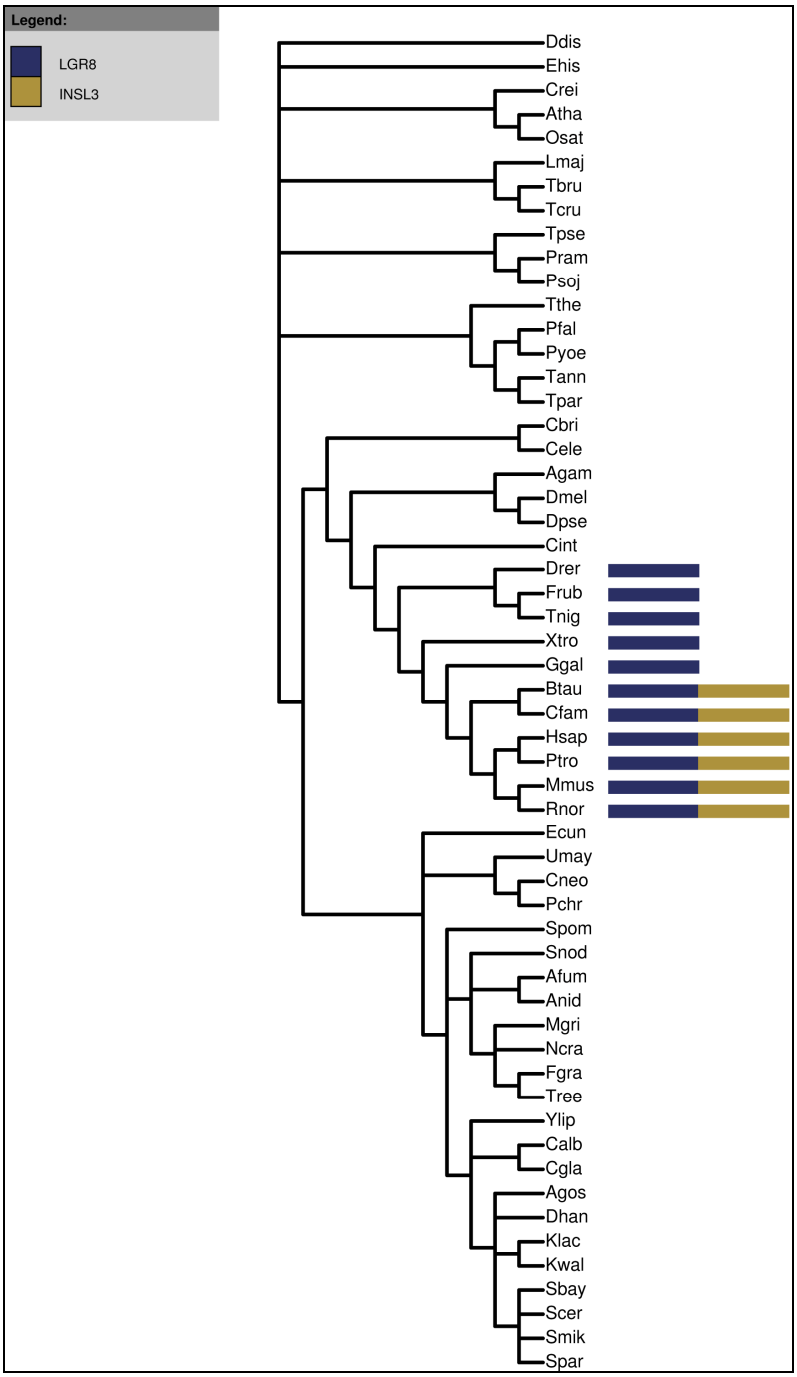

Figure 5.3 Eukaryotic LGR8 and INSL3 families.

Presence and absence of LGR8 and INSL3 families in eukaryotic genomes. Blue bars represent the LGR8 family and gold bars represent the INSL3 family.

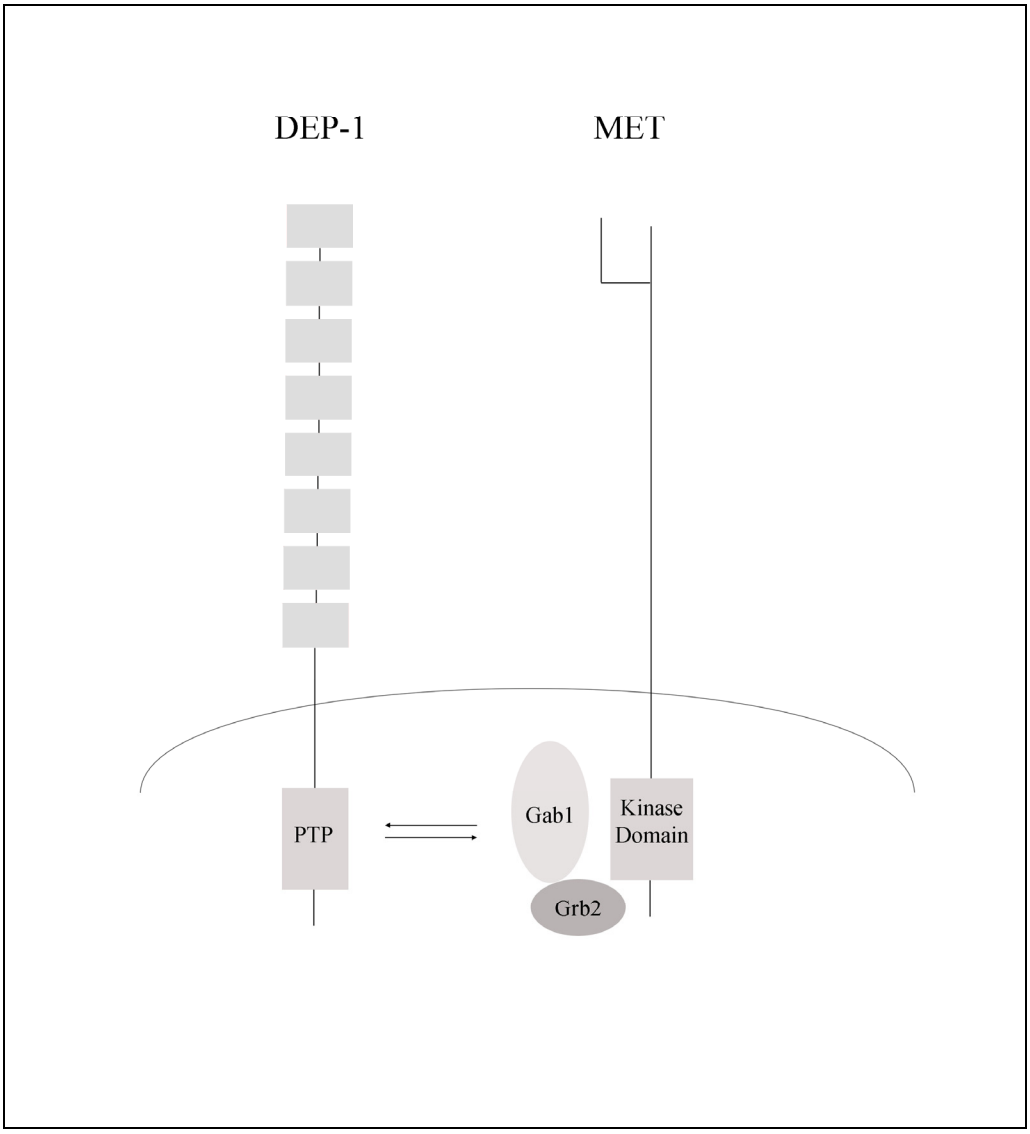

Figure 5.4 DEP-1 phosphatase and Met kinase substrates.

DEP-1 phosphatases and Met kinases interact with Gab1, Grb2, and each other.

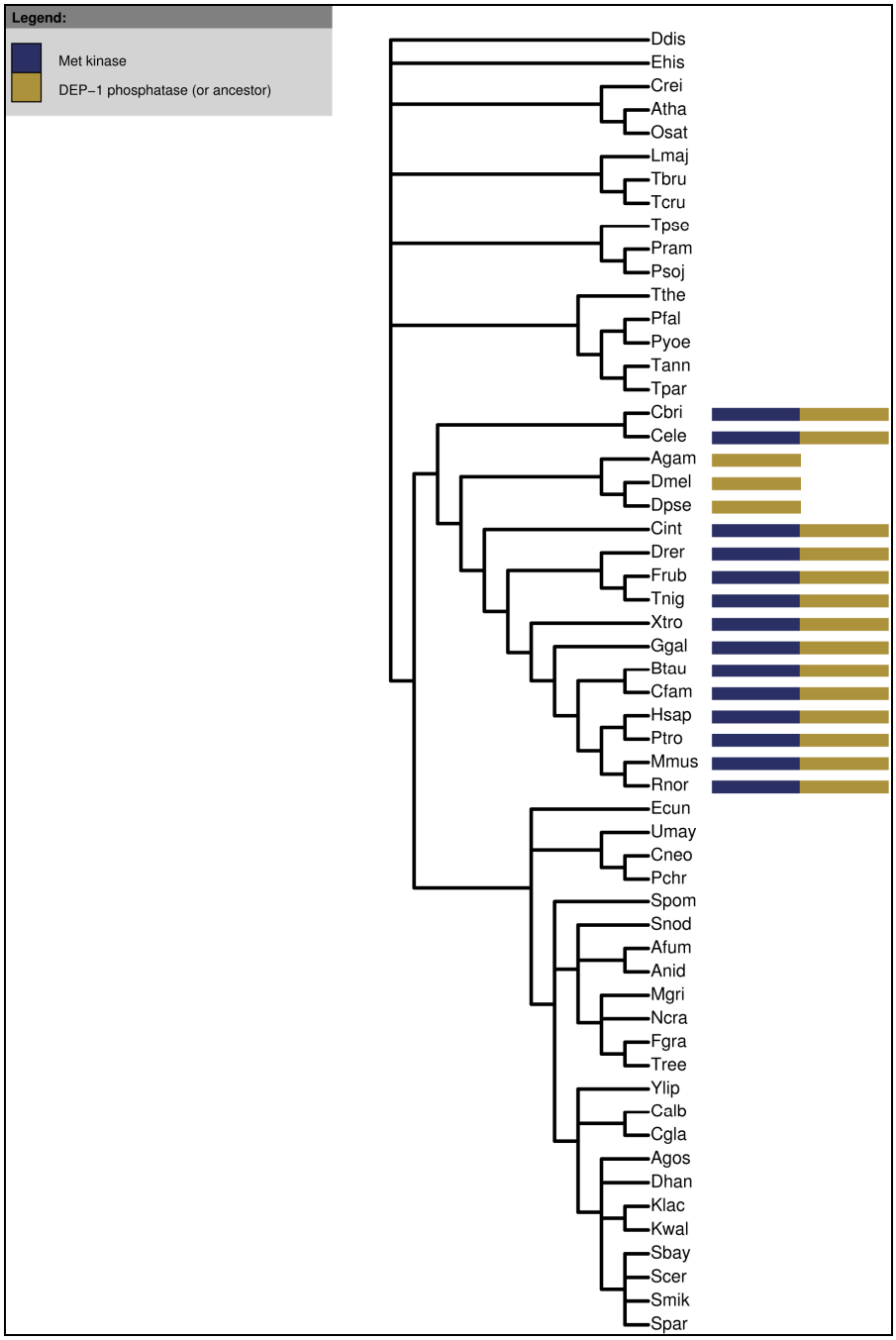

Figure 5.5 Met kinase and DEP-1 phosphatase families.

Presence and absence of Met kinase and DEP-1 phosphatase families in eukaryotic genomes. Blue bars represent the Met kinase family and gold bars represent the DEP-1 phosphatase family.

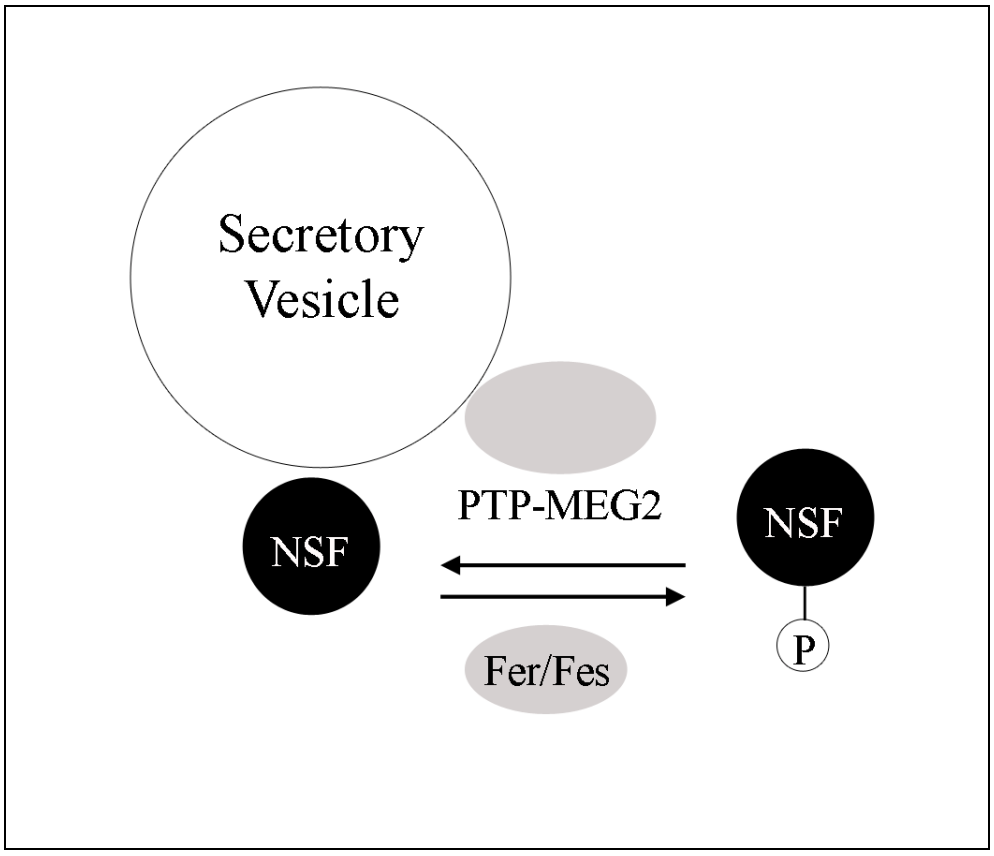

Figure 5.6 PTP-MEG2 phosphatase and Fer/Fes kinase interactions.

PTP-MEG2 phosphatases dephosphorylate NSF, while Fer/Fes kinases phosphorylate NSF.

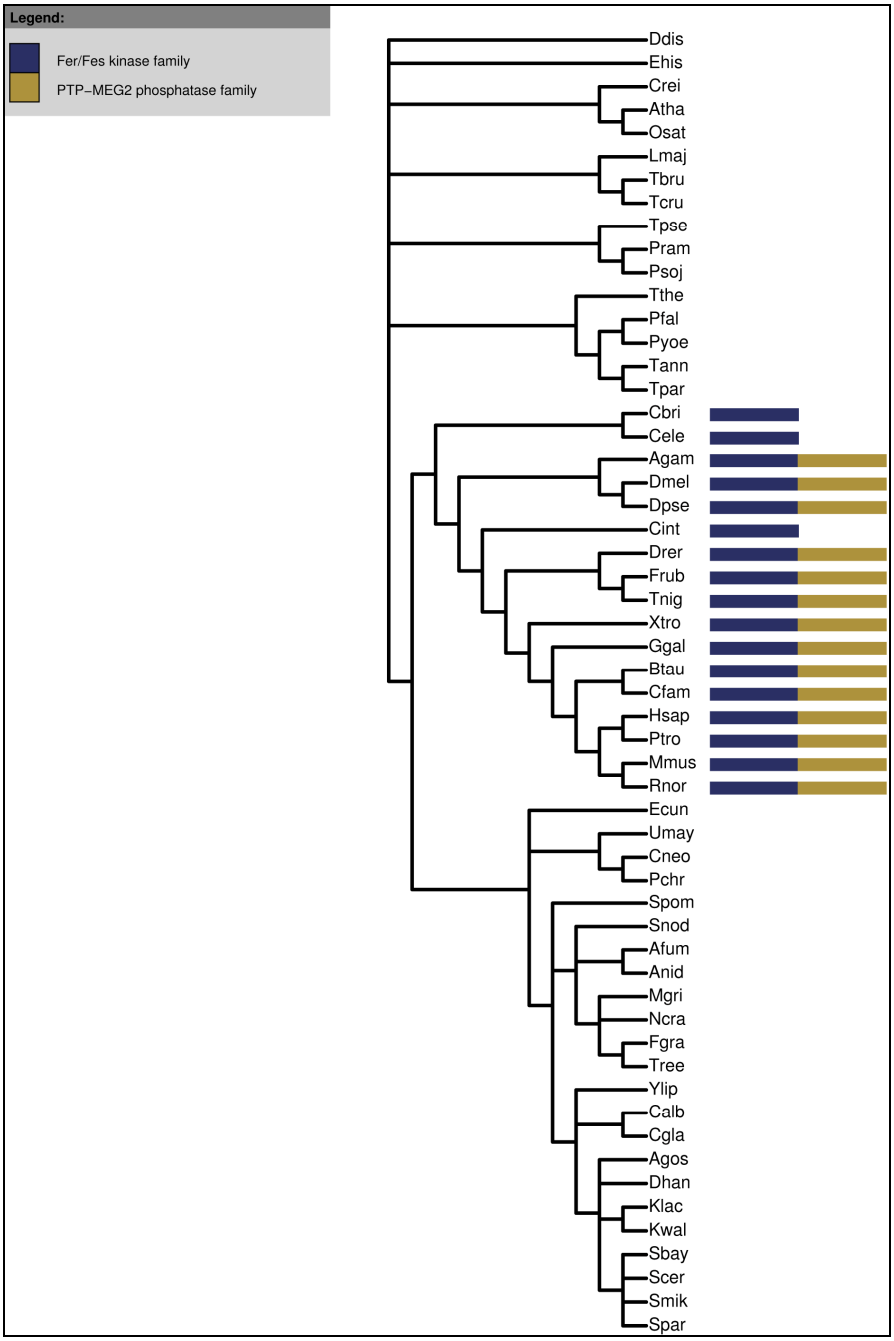

Figure 5.7 Fer/Fes kinase and PTP-MEG2 phosphatase families.

Presence and absence of Fer/Fes kinase and PTP-MEG2 phosphatase families in eukaryotic genomes. Blue bars represent the Fer/Fes kinase family and gold bars represent the PTP-MEG2 phosphatase family.

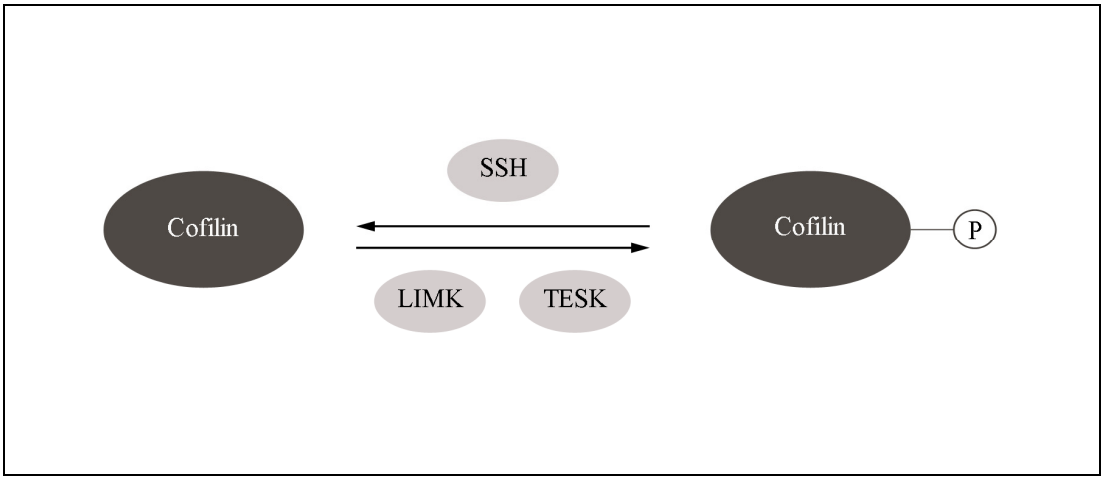

Figure 5.8 LIM and TES kinase and SSH phosphatase reactions.

LIM and TES kinases phosphorylate cofilin proteins, and the SSH phosphatases dephosphorylate cofilin proteins.

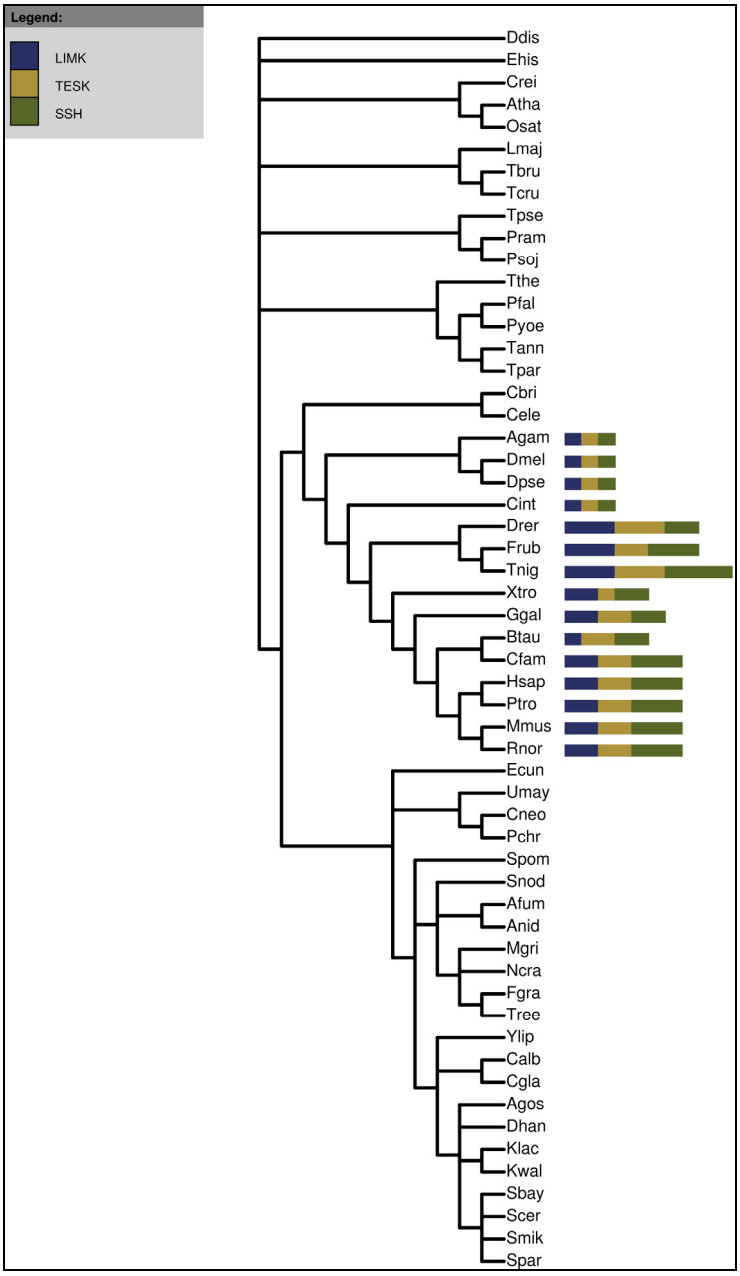

Figure 5.9 LIM and TES kinase and SSH phosphatase families.

Presence and absence of LIM and TES kinase and SSH phosphatase families in eukaryotic genomes. Blue bars represent the LIM kinase family, gold bars represent the TES kinase family, and green bars represent the SSH family. Length of bar corresponds to gene copy number, as described in text.

6 Thoughts for Future Studies

The research herein presents several intriguing directions for future study. Discussed below are several such areas, including the continued generation of new genome sequence data, metagenomics environmental studies, the definition of "complete" kinomes and phosphatomes, comparison of protein domain combinations in the kinase and phosphatase families, and continued evolutionary analysis of these proteins.

The number of completely sequenced genomes will likely continue to grow at an ever-increasing rate. Each newly sequenced genome presents a new opportunity to contribute to our knowledge of protein kinases and phosphatases. Particularly interesting results may be obtained from studying Bacteria or Archaea in currently understudied lineages.

Likewise, data being generated by ongoing metagenomics projects could prove to be a valuable resource. To date, such studies have included sequencing of open ocean water [161,292], soil [293,294], coral atolls [295], the human gut [296] and hot springs at Yellowstone National Park [297]. This sequencing data can be used to study such questions as how kinases and phosphatases differ not only between species, but between different environments. Do kinomes and phosphatomes show any major adaptations based on temperature, pH or nutrient

conditions? Do any such modifications result in mechanistic or structural differences from "normal" kinases and phosphatases?

Another particularly interesting line of research is to move beyond the presence or absence study presented in this research, and attempt to quantify the complete number of kinases and phosphatases in a variety of species. Such "complete" kinomes have been previously published for some organisms such as humans [4], fly [152], *Dictyostelium* [156] and worm [153]. Although phosphatases have traditionally received less attention than kinases, the last few years have seen an increase in the number of publications focusing on large-scale annotation efforts of phosphatomes [2,72,112,134].

Continued efforts in defining the complete collections of kinases and phosphatases in complementary organisms will allow researchers to directly compare and contrast kinases and phosphatases and how they coexist. It also presents the opportunity to cluster and identify any novel subfamilies that may be specific to those groupings that have not yet been as intensely studied as metazoans. Several instances of apparent kinase family expansion have been noted in eukaryotic genomes. The aforementioned line of study would allow us to discover whether similar such expansions are also found in phosphatase families.

The generation of such a large collection of protein kinase and phosphatase sequences presents the opportunity to study other protein domains that are present in

these proteins. A few such studies of kinase domain combinations have been undertaken in humans [298], plants [299] and prokaryotes [300].

The research presented in this dissertation could be extended to study such domain combinations present in both kinases and phosphatases in a wider variety of organisms. Presumably, the potential identification of novel kinase and phosphatase subgroups in this data may also include proteins with previously unseen domains present in conjunction with the kinase or phosphatase domain. Such findings could lead to the identification of kinases and phosphatases with novel physiological roles. It would also be interesting to investigate the domain combinations present in kinase and phosphatase families that act on the same substrate.

This dissertation also compared the evolutionary patterns of several kinases and phosphatases that are functionally connected. A major challenge to such an undertaking is the still incomplete substrate and functional knowledge of kinases and phosphatases. However, efforts will certainly continue to more fully define such features of these protein families.

Recently, an attempt at large-scale computational assignment of kinases and substrates was published by Linding *et al.* [301]. Their method predicted the kinase-substrate interactions for 62% of all currently known phosphorylation sites. Two such predictions were then experimentally confirmed.

While it remains to be seen precisely how accurate such *in silico* methods can be made, continued efforts in this area will provide experimental researchers with interesting targets of study. As further advances in kinase and phosphatase substrate study occur, this idea of comparing evolutionary patterns could be expanded into a large-scale study.

The aforementioned avenues of research are not minor undertakings. They will require intensive computational approaches, the continued genome sequencing of diverse organisms, and would be best executed in conjunction with coordinated wet lab experimental tests. However, the continued dedication by researchers worldwide to kinase and phosphatase study can certainly overcome these obstacles and lead to new, exciting breakthroughs in both the field of medicine and our understanding of evolution.

Appendix A Eukaryotic genomes included in this study

Abbreviation	Name	Phylum	Class
Pyoe	Plasmodium yoelii ssp. yoelii 1	Alveolata	Apicomplexa
Pfal	Plasmodium falciparum 1	Alveolata	Apicomplexa
Tann	Theileria annulata	Alveolata	Apicomplexa
Tpar	Theileria parva	Alveolata	Apicomplexa
Tthe	Tetrahymena thermophila	Alveolata	Ciliophora
Anid	Aspergillus nidulans 1 r3.1	Fungi	Ascomycota
Afum	Aspergillus fumigatus	Fungi	Ascomycota
Fgra	Fusarium graminearum 1	Fungi	Ascomycota
Tree	Trichoderma reesei	Fungi	Ascomycota
Mgri	Magnaporthe grisea 7 r2.3	Fungi	Ascomycota
Ncra	Neurospora crassa 3	Fungi	Ascomycota
Snod	Stagonospora nodorum	Fungi	Ascomycota
Ylip	Yarrowia lipolytica	Fungi	Ascomycota
Calb	Candida albicans	Fungi	Ascomycota
Cgla	Candida glabrata	Fungi	Ascomycota
Dhan	Debaromyces hansenii	Fungi	Ascomycota
Agos	Ashbya gossypii 1.0	Fungi	Ascomycota
Klac	Kluyveromyces lactis	Fungi	Ascomycota
Kwal	Kluyveromyces waltii	Fungi	Ascomycota
Scer	Saccharomyces cerevisiae	Fungi	Ascomycota
Sbay	Saccharomyces bayanus MIT	Fungi	Ascomycota
Smik	Saccharomyces mikatae MIT	Fungi	Ascomycota
Spar	Saccharomyces paradoxus MIT	Fungi	Ascomycota
Spom	Schizosaccharomyces pombe	Fungi	Ascomycota
Umay	Ustilago maydis 1 r2	Fungi	Basidiomycota
Cneo	Cryptococcus neoformans	Fungi	Basidiomycota
Pchr	Phanerochaete chrysosporium	Fungi	Basidiomycota
Ecun	Encephalitozoon cuniculi	Fungi	Microsporidia

Appendix A Eukaryotic genomes included in this study, continued

Abbreviation	Name	Phylum	Class
Agam	Anopheles gambiae 22.2b	Metazoa	Arthropoda
Dmel	Drosophila melanogaster 3.2	Metazoa	Arthropoda
Dpse	Drosophila pseudoobscura	Metazoa	Arthropoda
Drer	Danio rerio 22.3b	Metazoa	Chordata
Frub	Fugu rubripes 22.2c	Metazoa	Chordata
Tnig	Tetraodon nigroviridis	Metazoa	Chordata
Xtro	Xenopus tropicalis 2.0	Metazoa	Chordata
Ggal	Gallus gallus 22.1	Metazoa	Chordata
Hsap	Homo sapiens 22.34d	Metazoa	Chordata
Ptro	Pan troglodytes 22.1	Metazoa	Chordata
Mmus	Mus musculus 22.32b	Metazoa	Chordata
Rnor	Rattus norvegicus 22.3b	Metazoa	Chordata
Btau	Bos taurus	Metazoa	Chordata
Cfam	Canis familiaris	Metazoa	Chordata
Cint	Ciona intestinalis 1.0	Metazoa	Chordata
Cbri	Caenorhabditis briggsae Aug03	Metazoa	Nematoda
Cele	Caenorhabditis elegans WS123	Metazoa	Nematoda
Ddis	Dictyostelium discoideum 2	Mycetozoa	Dictyosteliida
Atha	Arabidopsis thaliana 5	Viridiplantae	Streptophyta
Osat	Oryza sativa ssp. japonica 2.0	Viridiplantae	Streptophyta
Crei	Chlamydomonas reinhardtii	Viridiplantae	Chlorophyta
Ehis	Entamoeba histolytica	Entamoebidae	Entamoeba
Lmaj	Leishmania major	Euglenozoa	Kinetoplastida
Tcru	Trypanosoma cruzi	Euglenozoa	Kinetoplastida
Tbru	Trypanosoma brucei	Euglenozoa	Kinetoplastida
Tpse	Thalassiosira pseudonana	Stramenopiles	Bacillariophyta
Pram	Phytophthora ramorum	Stramenopiles	Oomycetes
Psoj	Phytophthora sojae	Stramenopiles	Oomycetes

Appendix B Bacterial genomes included in this study

Abbreviation	Name	Phylum	Class
Mlep	Mycobacterium leprae	Actinobacteria	Actinobacteridae
Mtub	Mycobacterium tuberculosis H37Rv	Actinobacteria	Actinobacteridae
Ctra	Chlamydia trachomatis	Chlamydiae	Chlamydiales
Gvio	Gloeobacter violaceus	Cyanobacteria	Chroococcales
Syn2	Synechocystis sp. PCC 6803	Cyanobacteria	Chroococcales
Nossp	Nostoc sp. PCC 7120	Cyanobacteria	Nostocales
Pmar	Prochlorococcus marinus ssp. marinus CCMP1375	Cyanobacteria	Prochlorophytes
Drad	Deinococcus radiodurans R1	Deinococcus-Thermus	Deinococci
Bant	Bacillus anthracis Ames	Firmicutes	Bacillales
Bhal	Bacillus halodurans	Firmicutes	Bacillales
Bsub	Bacillus subtilis ssp. subtilis 168	Firmicutes	Bacillales
Cper	Clostridium perfringens 13	Firmicutes	Clostridia
Ljoh	Lactobacillus johnsonii NCC 533	Firmicutes	Lactobacillales
Llac	Lactococcus lactis ssp. lactis	Firmicutes	Lactobacillales
Spyo	Streptococcus pyogenes M1 GAS	Firmicutes	Lactobacillales
Mgen	Mycoplasma genitalium	Firmicutes	Mollicutes
Mpne	Mycoplasma pneumoniae	Firmicutes	Mollicutes
Ccre	Caulobacter crescentus CB15	Proteobacteria	Alphaproteobacteria

Appendix B Bacterial genomes included in this study, continued

Abbreviation	Name	Phylum	Class
Bjap	Bradyrhizobium japonicum USDA 110	Proteobacteria	Alphaproteobacteria
Bmel	Brucella melitensis 16M	Proteobacteria	Alphaproteobacteria
Rpro	Rickettsia prowazekii	Proteobacteria	Alphaproteobacteria
Bper	Bordetella pertussis Tohama I	Proteobacteria	Betaproteobacteria
Nmen	Neisseria meningitidis MC58	Proteobacteria	Betaproteobacteria
Dvul	Desulfovibrio vulgaris ssp. vulgaris Hildenborough	Proteobacteria	Deltaproteobacteria
Cjej	Campylobacter jejuni ssp. jejuni NCTC 11168	Proteobacteria	Epsilonproteobacteria
Hpyl	Helicobacter pylori 26695	Proteobacteria	Epsilonproteobacteria
Baph	Buchnera aphidicola Bp	Proteobacteria	Gammaproteobacteria
Ecol	Escherichia coli K12	Proteobacteria	Gammaproteobacteria
Cbur	Coxiella burnetii RSA 493	Proteobacteria	Gammaproteobacteria
Hinf	Haemophilus influenzae Rd KW20	Proteobacteria	Gammaproteobacteria
PSaer	Pseudomonas aeruginosa PAO1	Proteobacteria	Gammaproteobacteria
Vcho	Vibrio cholerae O1 biovar eltor N16961	Proteobacteria	Gammaproteobacteria
Xfas	Xylella fastidiosa 9a5c	Proteobacteria	Gammaproteobacteria
Bbur	Borrelia burgdorferi B31	Spirochaetes	Spirochaetales
Tpal	Treponema pallidum ssp. pallidum Nichols	Spirochaetes	Spirochaetales
Tmar	Thermotoga maritima	Thermotogae	Thermotogales

Appendix C Archaeal genomes included in this study

Abbreviation	Name	Phylum	Class
Aper	Aeropyrum pernix	Crenarchaeota	Thermoprotei
Saci	Sulfolobus acidocaldarius DSM 5348	Crenarchaeota	Thermoprotei
Ssol	Sulfolobus solfataricus	Crenarchaeota	Thermoprotei
Stok	Sulfolobus tokodaii	Crenarchaeota	Thermoprotei
PYaer	Pyrobaculum aerophilum IM2	Crenarchaeota	Thermoprotei
Aful	Archaeoglobus fulgidus DSM 4304	Euryarchaeota	Archaeoglobi
Halsp	Halobacterium sp. NRC-1	Euryarchaeota	Halobacteria
Hmar	Haloarcula marismortui ATCC 43049	Euryarchaeota	Halobacteria
Mthe	Methanothermobacter thermautotrophicus Delta H	Euryarchaeota	Methanobacteria
Mjan	Methanocaldococcus jannaschii	Euryarchaeota	Methanococci
Mmar	Methanococcus maripaludis	Euryarchaeota	Methanococci
Mace	Methanosarcina acetivorans C2A	Euryarchaeota	Methanomicrobia
Mbar	Methanosarcina barkeri fusaro	Euryarchaeota	Methanomicrobia
Mmaz	Methanosarcina mazei Goe1	Euryarchaeota	Methanomicrobia
Mkan	Methanopyrus kandleri AV19	Euryarchaeota	Methanopyri
Paby	Pyrococcus abyssi	Euryarchaeota	Thermococci
Pfur	Pyrococcus furiosus DSM 3638	Euryarchaeota	Thermococci
Phor	Pyrococcus horikoshii	Euryarchaeota	Thermococci
Tkod	Thermococcus kodakarensis KOD1	Euryarchaeota	Thermococci
Ptor	Picrophilus torridus DSM 9790	Euryarchaeota	Thermoplasmata
Taci	Thermoplasma acidophilum	Euryarchaeota	Thermoplasmata
Tvol	Thermoplasma volcanium	Euryarchaeota	Thermoplasmata
Nequ	Nanoarchaeum equitans Kin4-M	Nanoarchaeota	Nanoarchaeum

References

1. Bradham CA, Foltz KR, Beane WS, Arnone MI, Rizzo F, et al. (2006) The sea urchin kinome: a first look. Dev Biol 300: 180-193.

2. Alonso A, Sasin J, Bottini N, Friedberg I, Friedberg I, et al. (2004) Protein tyrosine phosphatases in the human genome. Cell 117: 699-711.

3. Cohen PTW (2004) I Overview of protein serine/threonine phosphatases. Protein Phosphatases.

4. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298: 1912-1934.

5. Bauman AL, Scott JD (2002) Kinase- and phosphatase-anchoring proteins: harnessing the dynamic duo. Nat Cell Biol 4: E203-206.

6. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36: D475-479.

7. Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res 29: 126-127.

8. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides NC (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34: D332-334.

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

10. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.

11. Frishman D, Argos P (1992) Recognition of distantly related protein sequences using conserved motifs and neural networks. J Mol Biol 228: 951-962.

12. Ben-Hur A, Brutlag D (2003) Remote homology detection: a motif based approach. Bioinformatics 19 Suppl 1: i26-33.

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

14. Altschul SF, Koonin EV (1998) Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23: 444-447.

15. Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 22: 1315-1316.

16. Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14: 846-856.

17. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, et al. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12: 327-345.

18. Krogh A (1998) An introduction to hidden Markov models for biological sequences. Computational Methods in Molecular Biology: 45–63.

19. Park J, Karplus K, Barrett C, Hughey R, Haussler D, et al. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284: 1201-1210.

20. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-288.

21. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903-919.

22. Mi H, Guo N, Kejariwal A, Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res 35: D247-252.

23. Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311: 347-351.

24. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85-94.

25. Yang AS, Honig B (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 301: 679-689.

26. Brenner SE, Chothia C, Hubbard TJ (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A 95: 6073-6078.

27. Hon WC, McKay GA, Thompson PR, Sweet RM, Yang DS, et al. (1997) Structure of an enzyme required for aminoglycoside antibiotic resistance reveals homology to eukaryotic protein kinases. Cell 89: 887-895.

28. Holm L, Sander C (1995) Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme. Embo J 14: 1287-1293.

29. Bartlett GJ, Todd AE, Thornton JM (2003) Inferring protein function from structure. Methods Biochem Anal 44: 387-407.

30. Li WW, Quinn GB, Alexandrov NN, Bourne PE, Shindyalov IN (2003) A comparative proteomics resource: proteins of Arabidopsis thaliana. Genome Biol 4: R51.

31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.

32. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.

33. Alexandrov N, Shindyalov I (2003) PDP: protein domain parser. Bioinformatics 19: 429-430.

34. Lopez R, Silventoinen V, Robinson S, Kibria A, Gish W (2003) WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res 31: 3795-3798.

35. Alexandrov NN, Nussinov R, Zimmer RM (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. Pac Symp Biocomput: 53-72.

36. Cheek S, Ginalski K, Zhang H, Grishin NV (2005) A comprehensive update of the sequence and structure classification of kinases. BMC Struct Biol 5: 6.

37. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucleic Acids Res 34: D556-561.

38. Hanks SK, Hunter T (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. Faseb J 9: 576-596.

39. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35: D5-12.

40. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.

41. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34: D187-191.

42. Ye X, Ji C, Zhou C, Zeng L, Gu S, et al. (2004) Cloning and characterization of a human cDNA ACAD10 mapped to chromosome 12q24.1. Mol Biol Rep 31: 191-195.

43. Brenner S (1987) Phosphotransferase sequence homology. Nature 329: 21.

44. Yuan C, Kent C (2004) Identification of critical residues of choline kinase A2 from Caenorhabditis elegans. J Biol Chem 279: 17801-17809.

45. Scheeff ED, Bourne PE (2005) Structural Evolution of the Protein Kinase-Like Superfamily. PLoS Comput Biol 1: e49.

46. Aoyama C, Liao H, Ishidate K (2004) Structure and function of choline kinase isoforms in mammalian cells. Prog Lipid Res 43: 266-281.

47. Glunde K, Jie C, Bhujwalla ZM (2004) Molecular causes of the aberrant choline phospholipid metabolism in breast cancer. Cancer Res 64: 4270-4276.

48. Ackerstaff E, Glunde K, Bhujwalla ZM (2003) Choline phospholipid metabolism: a target in cancer cells? J Cell Biochem 90: 525-533.

49. Daigle DM, McKay GA, Thompson PR, Wright GD (1999) Aminoglycoside antibiotic phosphotransferases are also serine protein kinases. Chem Biol 6: 11-18.

50. Kannan N, Neuwald AF (2005) Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? J Mol Biol 351: 956-972.

51. Kennelly PJ (2001) Protein phosphatases--a phylogenetic perspective. Chem Rev 101: 2291-2312.

52. Wishart MJ, Guan KL (2005) Protein phosphatases. Encyclopedia of Life Sciences John Wiley & Sons, Ltd doi 10: 5.

53. Andersen JN, Mortensen OH, Peters GH, Drake PG, Iversen LF, et al. (2001) Structural and evolutionary relationships among protein tyrosine phosphatase domains. Mol Cell Biol 21: 7117-7136.

54. Alonso A, Rojas A, Godzik A, Mustelin T (2003) The dual-specific PTP family. Top Curr Genet 5: 333–358.

55. Pils B, Schultz J (2004) Evolution of the multifunctional protein tyrosine phosphatase family. Mol Biol Evol 21: 625-631.

56. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129-2141.

57. Labarga A, Valentin F, Anderson M, Lopez R (2007) Web services at the European bioinformatics institute. Nucleic Acids Res 35: W6-11.

58. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36: D13-21.

59. Kennelly PJ (2002) Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. FEMS Microbiol Lett 206: 1-8.

60. Ramponi G, Stefani M (1997) Structure and function of the low Mr phosphotyrosine protein phosphatases. Biochim Biophys Acta 1341: 137-156.

61. Fauman EB, Cogswell JP, Lovejoy B, Rocque WJ, Holmes W, et al. (1998) Crystal structure of the catalytic domain of the human cell cycle control phosphatase, Cdc25A. Cell 93: 617-625.

62. Jemc J, Rebay I (2007) The eyes absent family of phosphotyrosine phosphatases: properties and roles in developmental regulation of transcription. Annu Rev Biochem 76: 513-538.

63. den Hertog J (2004) 13 Receptor protein tyrosine phosphatases. In: Arino J, Alexander D, editors. Protein Phosphatases: Springer-Verlag. pp. 253-274.

64. Tonks NK (2006) Protein tyrosine phosphatases: from genes, to function, to disease. Nat Rev Mol Cell Biol 7: 833-846.

65. Tertoolen LG, Blanchetot C, Jiang G, Overvoorde J, Gadella TW, Jr., et al. (2001) Dimerization of receptor protein-tyrosine phosphatase alpha in living cells. BMC Cell Biol 2: 8.

66. Groen A, Overvoorde J, van der Wijk T, den Hertog J (2008) Redox regulation of dimerization of the receptor protein-tyrosine phosphatases RPTPalpha, LAR, RPTPmu and CD45. Febs J.

67. den Hertog J, Ostman A, Bohmer FD (2008) Protein tyrosine phosphatases: regulatory mechanisms. Febs J 275: 831-847.

68. Lee S, Faux C, Nixon J, Alete D, Chilton J, et al. (2007) Dimerization of protein tyrosine phosphatase sigma governs both ligand binding and isoform specificity. Mol Cell Biol 27: 1795-1808.

69. Nam HJ, Poy F, Krueger NX, Saito H, Frederick CA (1999) Crystal structure of the tandem phosphatase domains of RPTP LAR. Cell 97: 449-457.

70. Nam HJ, Poy F, Saito H, Frederick CA (2005) Structural basis for the function and regulation of the receptor protein tyrosine phosphatase CD45. J Exp Med 201: 441-452.

71. Muller CI, Blumbach B, Krasko A, Schroder HC (2001) Receptor protein-tyrosine phosphatases: origin of domains (catalytic domain, Ig-related domain, fibronectin type III module) based on the sequence of the sponge Geodia cydonium. Gene 262: 221-230.

72. Byrum CA, Walton KD, Robertson AJ, Carbonneau S, Thomason RT, et al. (2006) Protein tyrosine and serine-threonine phosphatases in the sea urchin, Strongylocentrotus purpuratus: identification and potential functions. Dev Biol 300: 194-218.

73. Nordle AK, Rios P, Gaulton A, Pulido R, Attwood TK, et al. (2007) Functional assignment of MAPK phosphatase domains. Proteins 69: 19-31.

74. Keyse SM (2008) The regulation of stress-activated MAP kinase signalling by protein phosphatases. TOPICS IN CURRENT GENETICS 20: 33.

75. Stegmeier F, Amon A (2004) Closing mitosis: the functions of the Cdc14 phosphatase and its regulation. Annu Rev Genet 38: 203-232.

76. Gray CH, Good VM, Tonks NK, Barford D (2003) The structure of the cell cycle protein Cdc14 reveals a proline-directed protein phosphatase. Embo J 22: 3524-3535.

77. Ah Fong AM, Judelson HS (2003) Cell cycle regulator Cdc14 is expressed during sporulation but not hyphal growth in the fungus-like oomycete Phytophthora infestans. Mol Microbiol 50: 487-494.

78. Kerk D, Templeton G, Moorhead GB (2008) Evolutionary radiation pattern of novel protein phosphatases revealed by analysis of protein data from the completely sequenced genomes of humans, green algae, and higher plants. Plant Physiol 146: 351-367.

79. Niwa R, Nagata-Ohashi K, Takeichi M, Mizuno K, Uemura T (2002) Control of actin reorganization by Slingshot, a family of phosphatases that dephosphorylate ADF/cofilin. Cell 108: 233-246.

80. Soosairajah J, Maiti S, Wiggan O, Sarmiere P, Moussi N, et al. (2005) Interplay between components of a novel LIM kinase-slingshot phosphatase complex regulates cofilin. Embo J 24: 473-486.

81. Ohta Y, Kousaka K, Nagata-Ohashi K, Ohashi K, Muramoto A, et al. (2003) Differential activities, subcellular distribution and tissue expression patterns of three members of Slingshot family phosphatases that dephosphorylate cofilin. Genes Cells 8: 811-824.

82. Maciver SK, Hussey PJ (2002) The ADF/cofilin family: actin-remodeling proteins. Genome Biol 3: reviews3007.

83. Huang TY, DerMardirossian C, Bokoch GM (2006) Cofilin phosphatases and regulation of actin dynamics. Curr Opin Cell Biol 18: 26-31.

84. Gohla A, Birkenfeld J, Bokoch GM (2005) Chronophin, a novel HAD-type serine protein phosphatase, regulates cofilin-dependent actin dynamics. Nat Cell Biol 7: 21-29.

85. Stephens BJ, Han H, Gokhale V, Von Hoff DD (2005) PRL phosphatases as potential molecular targets in cancer. Mol Cancer Ther 4: 1653-1661.

86. Kim KA, Song JS, Jee J, Sheen MR, Lee C, et al. (2004) Structure of human PRL-3, the phosphatase associated with cancer metastasis. FEBS Lett 565: 181-187.

87. Zeng Q, Hong W, Tan YH (1998) Mouse PRL-2 and PRL-3, two potentially prenylated protein tyrosine phosphatases homologous to PRL-1. Biochem Biophys Res Commun 244: 421-427.

88. Bessette DC, Qiu D, Pallen CJ (2008) PRL PTPs: mediators and markers of cancer progression. Cancer Metastasis Rev 27: 231-252.

89. Pendyala PR, Ayong L, Eatrides J, Schreiber M, Pham C, et al. (2008) Characterization of a PRL protein tyrosine phosphatase from Plasmodium falciparum. Mol Biochem Parasitol 158: 1-10.

90. Fiordalisi JJ, Keller PJ, Cox AD (2006) PRL tyrosine phosphatases regulate rho family GTPases to promote invasion and motility. Cancer Res 66: 3153-3161.

91. Wang Y, Li ZF, He J, Li YL, Zhu GB, et al. (2007) Expression of the human phosphatases of regenerating liver (PRLs) in colonic adenocarcinoma and its correlation with lymph node metastasis. Int J Colorectal Dis 22: 1179-1184.

92. Taylor GS, Dixon JE (2003) PTEN and myotubularins: families of phosphoinositide phosphatases. Methods Enzymol 366: 43-56.

93. Maehama T (2007) PTEN: its deregulation and tumorigenesis. Biol Pharm Bull 30: 1624-1627.

94. Bonifant CL, Kim JS, Waldman T (2007) NHERFs, NEP, MAGUKs, and more: interactions that regulate PTEN. J Cell Biochem 102: 878-885.

95. Maehama T, Dixon JE (1999) PTEN: a tumour suppressor that functions as a phospholipid phosphatase. Trends Cell Biol 9: 125-128.

96. Laporte J, Blondeau F, Buj-Bello A, Mandel JL (2001) The myotubularin family: from genetic disease to phosphoinositide metabolism. Trends Genet 17: 221-228.

97. Leslie NR, Yang X, Downes CP, Weijer CJ (2007) PtdIns(3,4,5)P(3)-dependent and -independent roles for PTEN in the control of cell migration. Curr Biol 17: 115-125.

98. Walker SM, Downes CP, Leslie NR (2001) TPIP: a novel phosphoinositide 3-phosphatase. Biochem J 360: 277-283.

99. Hafizi S, Ibraimi F, Dahlback B (2005) C1-TEN is a negative regulator of the Akt/PKB signal transduction pathway and inhibits cell survival, proliferation, and migration. Faseb J 19: 971-973.

100. Mouneimne G, Brugge JS (2007) Tensins: a new switch in cell migration. Dev Cell 13: 317-319.

101. Lo SH (2004) Tensin. Int J Biochem Cell Biol 36: 31-34.

102. Li L, Ernsting BR, Wishart MJ, Lohse DL, Dixon JE (1997) A family of putative tumor suppressors is structurally and functionally conserved in humans and yeast. J Biol Chem 272: 29403-29406.

103. Choudhury P, Srivastava S, Li Z, Ko K, Albaqumi M, et al. (2006) Specificity of the myotubularin family of phosphatidylinositol-3-phosphatase is determined by the PH/GRAM domain. J Biol Chem 281: 31762-31769.

104. Laporte J, Hu LJ, Kretz C, Mandel JL, Kioschis P, et al. (1996) A gene mutated in X-linked myotubular myopathy defines a new putative tyrosine phosphatase family conserved in yeast. Nat Genet 13: 175-182.

105. Robinson FL, Dixon JE (2006) Myotubularin phosphatases: policing 3-phosphoinositides. Trends Cell Biol 16: 403-412.

106. Begley MJ, Dixon JE (2005) The structure and regulation of myotubularin phosphatases. Curr Opin Struct Biol 15: 614-620.

107. Laporte J, Bedez F, Bolino A, Mandel JL (2003) Myotubularins, a large disease-associated family of cooperating catalytically active and inactive phosphoinositides phosphatases. Hum Mol Genet 12 Spec No 2: R285-292.

108. Wishart MJ, Dixon JE (2002) PTEN and myotubularin phosphatases: from 3-phosphoinositide dephosphorylation to disease. Trends Cell Biol 12: 579-585.

109. Musumeci L, Bongiorni C, Tautz L, Edwards RA, Osterman A, et al. (2005) Low-molecular-weight protein tyrosine phosphatases of Bacillus subtilis. J Bacteriol 187: 4945-4956.

110. Rudbeck L, Johnsen A, Dissing J (2003) Evolutionary aspects of the gene for the classical enzyme polymorphism, ACP1. International Congress Series 1239: 733-736.

111. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, et al. (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34: D257-260.

112. Brenchley R, Tariq H, McElhinney H, Szoor B, Huxley-Jones J, et al. (2007) The TriTryp phosphatome: analysis of the protein phosphatase catalytic domains. BMC Genomics 8: 434.

113. Bennett MS, Guan Z, Laurberg M, Su XD (2001) Bacillus subtilis arsenate reductase is structurally and functionally similar to low molecular weight protein tyrosine phosphatases. Proc Natl Acad Sci U S A 98: 13577-13582.

114. Boutros R, Dozier C, Ducommun B (2006) The when and wheres of CDC25 phosphatases. Curr Opin Cell Biol 18: 185-191.

115. Boutros R, Lobjois V, Ducommun B (2007) CDC25 phosphatases in cancer cells: key players? Good targets? Nat Rev Cancer 7: 495-507.

116. Rudolph J (2007) Inhibiting transient protein-protein interactions: lessons from the Cdc25 protein tyrosine phosphatases. Nat Rev Cancer 7: 202-211.

117. Kristjansdottir K, Rudolph J (2004) Cdc25 phosphatases and cancer. Chem Biol 11: 1043-1051.

118. Rudolph J (2007) Cdc25 phosphatases: structure, specificity, and mechanism. Biochemistry 46: 3595-3604.

119. Boudolf V, Inze D, De Veylder L (2006) What if higher plants lack a CDC25 phosphatase? Trends Plant Sci 11: 474-479.

120. Dhankher OP, Rosen BP, McKinney EC, Meagher RB (2006) Hyperaccumulation of arsenic in the shoots of Arabidopsis silenced for arsenate reductase (ACR2). Proc Natl Acad Sci U S A 103: 5413-5418.

121. Duan GL, Zhou Y, Tong YP, Mukhopadhyay R, Rosen BP, et al. (2007) A CDC25 homologue from rice functions as an arsenate reductase. New Phytol 174: 311-321.

122. Bordo D, Bork P (2002) The rhodanese/Cdc25 phosphatase superfamily. Sequence-structure-function relations. EMBO Rep 3: 741-746.

123. Draetta G, Eckstein J (1997) Cdc25 protein phosphatases in cell proliferation. Biochim Biophys Acta 1332: M53-63.

124. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33: D284-288.

125. Tootle TL, Silver SJ, Davies EL, Newman V, Latek RR, et al. (2003) The transcription factor Eyes absent is a protein tyrosine phosphatase. Nature 426: 299-302.

126. Li X, Oghi KA, Zhang J, Krones A, Bush KT, et al. (2003) Eya protein phosphatase activity regulates Six1-Dach-Eya transcriptional effects in mammalian organogenesis. Nature 426: 247-254.

127. Hsiao FC, Williams A, Davies EL, Rebay I (2001) Eyes absent mediates cross-talk between retinal determination genes and the receptor tyrosine kinase signaling pathway. Dev Cell 1: 51-61.

128. Takeda Y, Hatano S, Sentoku N, Matsuoka M (1999) Homologs of animal eyes absent (eya) genes are found in higher plants. Mol Gen Genet 262: 131-138.

129. Rayapureddi JP, Kattamuri C, Chan FH, Hegde RS (2005) Characterization of a plant, tyrosine-specific phosphatase of the aspartyl class. Biochemistry 44: 751-758.

130. Farkas I, Dombradi V, Miskei M, Szabados L, Koncz C (2007) Arabidopsis PPP family of serine/threonine phosphatases. Trends Plant Sci 12: 169-176.

131. Stern A, Privman E, Rasis M, Lavi S, Pupko T (2007) Evolution of the metazoan protein phosphatase 2C superfamily. J Mol Evol 64: 61-70.

132. Andreeva AV, Kutuzov MA (2004) Widespread presence of "bacterial-like" PPP phosphatases in eukaryotes. BMC Evol Biol 4: 47.

133. Gallego M, Virshup DM (2005) Protein serine/threonine phosphatases: life, death, and sleeping. Curr Opin Cell Biol 17: 197-202.

134. Kerk D, Bulgrien J, Smith DW, Barsam B, Veretnik S, et al. (2002) The complement of protein phosphatase catalytic subunits encoded in the genome of Arabidopsis. Plant Physiol 129: 908-925.

135. Dickman MB, Yarden O (1999) Serine/threonine protein kinases and phosphatases in filamentious fungi. Fungal Genet Biol 26: 99-117.

136. Schweighofer A, Hirt H, Meskiene I (2004) Plant PP2C phosphatases: emerging functions in stress signaling. Trends Plant Sci 9: 236-243.

137. Kennelly PJ (2003) Archaeal protein kinases and protein phosphatases: insights from genomics and biochemistry. Biochem J 370: 373-389.

138. Shi L, Carmichael WW, Kennelly PJ (1999) Cyanobacterial PPP family protein phosphatases possess multifunctional capabilities and are resistant to microcystin-LR. J Biol Chem 274: 10039-10046.

139. Tabernero L, Aricescu AR, Jones EY, Szedlacsek SE (2008) Protein tyrosine phosphatases: structure-function relationships. Febs J 275: 867-882.

140. Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134: 191-203.

141. Huang JF (2003) Different protein tyrosine phosphatase superfamilies resulting from different gene reading frames. Mol Biol Evol 20: 815-820.

142. Jia Z (1997) Protein phosphatases: structures and implications. Biochem Cell Biol 75: 17-26.

143. Shi L, Potts M, Kennelly PJ (1998) The serine, threonine, and/or tyrosine-specific protein kinases and protein phosphatases of prokaryotic organisms: a family portrait. FEMS Microbiol Rev 22: 229-253.

144. Ramponi G, Stefani M (1997) Structural, catalytic, and functional properties of low M(r), phosphotyrosine protein phosphatases. Evidence of a long evolutionary history. Int J Biochem Cell Biol 29: 279-292.

145. Mukhopadhyay R, Rosen BP (2001) The phosphatase C(X)5R motif is required for catalytic activity of the Saccharomyces cerevisiae Acr2p arsenate reductase. J Biol Chem 276: 34738-34742.

146. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, et al. (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature 414: 450-453.

147. Miranda-Saavedra D, Stark MJ, Packer JC, Vivares CP, Doerig C, et al. (2007) The complement of protein kinases of the microsporidium Encephalitozoon cuniculi in relation to those of Saccharomyces cerevisiae and Schizosaccharomyces pombe. BMC Genomics 8: 309.

148. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3: e314.

149. Baselga J (2006) Targeting tyrosine kinases in cancer: the second wave. Science 312: 1175-1178.

150. Vlahovic G, Crawford J (2003) Activation of tyrosine kinases in cancer. Oncologist 8: 531-538.

151. Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci U S A 101: 11707-11712.

152. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 27: 514-520.

153. Manning G (2005) Genomic overview of protein kinases. WormBook: 1-19.

154. Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T (1999) The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. Proc Natl Acad Sci U S A 96: 13603-13610.

155. Hunter T, Plowman GD (1997) The protein kinases of budding yeast: six score and more. Trends Biochem Sci 22: 18-22.

156. Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, et al. (2006) The dictyostelium kinome--analysis of the protein kinases from a simple model organism. PLoS Genet 2: e38.

157. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, et al. (2006) Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol 4: e286.

158. Anamika K, Bhattacharya A, Srinivasan N (2008) Analysis of the protein kinome of Entamoeba histolytica. Proteins 71: 995-1006.

159. Ward P, Equinet L, Packer J, Doerig C (2004) Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. BMC Genomics 5: 79.

160. Parsons M, Worthey EA, Ward PN, Mottram JC (2005) Comparative analysis of the kinomes of three pathogenic trypanosomatids: Leishmania major, Trypanosoma brucei and Trypanosoma cruzi. BMC Genomics 6: 127.

161. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.

162. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5: e16.

163. Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G (2007) Structural and functional diversity of the microbial kinome. PLoS Biol 5: e17.

164. Tang Z, Shu H, Oncel D, Chen S, Yu H (2004) Phosphorylation of Cdc20 by Bub1 provides a catalytic mechanism for APC/C inhibition by the spindle checkpoint. Mol Cell 16: 387-397.

165. Tyler-Smith C, Floridia G (2000) Many paths to the top of the mountain: diverse evolutionary solutions to centromere structure. Cell 102: 5-8.

166. Young TA, Delagoutte B, Endrizzi JA, Falick AM, Alber T (2003) Structure of Mycobacterium tuberculosis PknB supports a universal activation mechanism for Ser/Thr protein kinases. Nat Struct Biol 10: 168-174.

167. Boitel B, Ortiz-Lombardia M, Duran R, Pompeo F, Cole ST, et al. (2003) PknB kinase activity is regulated by phosphorylation in two Thr residues and dephosphorylation by PstP, the cognate phospho-Ser/Thr phosphatase, in Mycobacterium tuberculosis. Mol Microbiol 49: 1493-1508.

168. Fernandez P, Saint-Joanis B, Barilone N, Jackson M, Gicquel B, et al. (2006) The Ser/Thr protein kinase PknB is essential for sustaining mycobacterial growth. J Bacteriol 188: 7778-7784.

169. Kang CM, Abbott DW, Park ST, Dascher CC, Cantley LC, et al. (2005) The Mycobacterium tuberculosis serine/threonine kinases PknA and PknB: substrate identification and regulation of cell shape. Genes Dev 19: 1692-1704.

170. Higgins JM (2003) Structure, function and evolution of haspin and haspin-related proteins, a distinctive group of eukaryotic protein kinases. Cell Mol Life Sci 60: 446-462.

171. Dai J, Higgins JM (2005) Haspin: a mitotic histone kinase required for metaphase chromosome alignment. Cell Cycle 4: 665-668.

172. Dai J, Sultan S, Taylor SS, Higgins JM (2005) The kinase haspin is required for mitotic histone H3 Thr 3 phosphorylation and normal metaphase chromosome alignment. Genes Dev 19: 472-488.

173. Leonard CJ, Aravind L, Koonin EV (1998) Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res 8: 1038-1047.

174. Lopreiato R, Facchin S, Sartori G, Arrigoni G, Casonato S, et al. (2004) Analysis of the interaction between piD261/Bud32, an evolutionarily conserved protein kinase of Saccharomyces cerevisiae, and the Grx4 glutaredoxin. Biochem J 377: 395-405.

175. Facchin S, Lopreiato R, Ruzzene M, Marin O, Sartori G, et al. (2003) Functional homology between yeast piD261/Bud32 and human PRPK: both phosphorylate p53 and PRPK partially complements piD261/Bud32 deficiency. FEBS Lett 549: 63-66.

176. Angermayr M, Roidl A, Bandlow W (2002) Yeast Rio1p is the founding member of a novel subfamily of protein serine kinases involved in the control of cell cycle progression. Mol Microbiol 44: 309-324.

177. LaRonde-LeBlanc N, Wlodawer A (2004) Crystal structure of A. fulgidus Rio2 defines a new family of serine protein kinases. Structure 12: 1585-1594.

178. Laronde-Leblanc N, Guszczynski T, Copeland T, Wlodawer A (2005) Structure and activity of the atypical serine kinase Rio1. Febs J 272: 3698-3713.

179. Krupa A, Srinivasan N (2002) Lipopolysaccharide phosphorylating enzymes encoded in the genomes of Gram-negative bacteria are related to the eukaryotic protein kinases. Protein Sci 11: 1580-1584.

180. Zhao X, Wenzel CQ, Lam JS (2002) Nonradiolabeling assay for WaaP, an essential sugar kinase involved in biosynthesis of core lipopolysaccharide of Pseudomonas aeruginosa. Antimicrob Agents Chemother 46: 2035-2037.

181. White KA, Lin S, Cotter RJ, Raetz CR (1999) A Haemophilus influenzae gene that encodes a membrane bound 3-deoxy-D-manno-octulosonic acid (Kdo) kinase. Possible involvement of kdo phosphorylation in bacterial virulence. J Biol Chem 274: 31391-31400.

182. Tran AX, Stead CM, Trent MS (2005) Remodeling of Helicobacter pylori lipopolysaccharide. J Endotoxin Res 11: 161-166.

183. Yamashita S, Hosaka K (1997) Choline kinase from yeast. Biochim Biophys Acta 1348: 63-69.

184. Sohlenkamp C, Lopez-Lara IM, Geiger O (2003) Biosynthesis of phosphatidylcholine in bacteria. Prog Lipid Res 42: 115-162.

185. Serino L, Virji M (2000) Phosphorylcholine decoration of lipopolysaccharide differentiates commensal Neisseriae from pathogenic strains: identification of licA-type genes in commensal Neisseriae. Mol Microbiol 35: 1550-1559.

186. Kim K, Kim KH, Storey MK, Voelker DR, Carman GM (1999) Isolation and characterization of the Saccharomyces cerevisiae EKI1 gene encoding ethanolamine kinase. J Biol Chem 274: 14857-14866.

187. Singh SK, Yang K, Karthikeyan S, Huynh T, Zhang X, et al. (2004) The thrH gene product of Pseudomonas aeruginosa is a dual activity enzyme with a novel phosphoserine:homoserine phosphotransferase activity. J Biol Chem 279: 13166-13173.

188. Patte JC, Clepet C, Bally M, Borne F, Mejean V, et al. (1999) ThrH, a homoserine kinase isozyme with in vivo phosphoserine phosphatase activity in Pseudomonas aeruginosa. Microbiology 145 (Pt 4): 845-853.

189. Fortpied J, Gemayel R, Stroobant V, van Schaftingen E (2005) Plant ribulosamine/erythrulosamine 3-kinase, a putative protein-repair enzyme. Biochem J 388: 795-802.

190. Delpierre G, Van Schaftingen E (2003) Fructosamine 3-kinase, an enzyme involved in protein deglycation. Biochem Soc Trans 31: 1354-1357.

191. Gemayel R, Fortpied J, Rzem R, Vertommen D, Veiga-da-Cunha M, et al. (2007) Many fructosamine 3-kinase homologues in bacteria are ribulosamine/erythrulosamine 3-kinases potentially involved in protein deglycation. Febs J 274: 4360-4374.

192. Delplanque J, Delpierre G, Opperdoes FR, Van Schaftingen E (2004) Tissue distribution and evolution of fructosamine 3-kinase and fructosamine 3-kinase-related protein. J Biol Chem 279: 46606-46613.

193. Ku SY, Yip P, Cornell KA, Riscoe MK, Behr JB, et al. (2007) Structures of 5-methylthioribose kinase reveal substrate specificity and unusual mode of nucleotide binding. J Biol Chem 282: 22195-22206.

194. Sekowska A, Mulard L, Krogh S, Tse JK, Danchin A (2001) MtnK, methylthioribose kinase, is a starvation-induced protein in Bacillus subtilis. BMC Microbiol 1: 15.

195. Sauter M, Cornell KA, Beszteri S, Rzewuski G (2004) Functional analysis of methylthioribose kinase genes in plants. Plant Physiol 136: 4061-4071.

196. Kushad MM, Richardson DG, Ferro AJ (1982) 5-Methylthioribose kinase activity in plants. Biochem Biophys Res Commun 108: 167-173.

197. Riscoe MK, Ferro AJ, Fitchen JH (1988) Analogs of 5-methylthioribose, a novel class of antiprotozoal agents. Antimicrob Agents Chemother 32: 1904-1906.

198. Poon WW, Davis DE, Ha HT, Jonassen T, Rather PN, et al. (2000) Identification of Escherichia coli ubiB, a gene required for the first monooxygenase step in ubiquinone biosynthesis. J Bacteriol 182: 5139-5146.

199. Do TQ, Hsu AY, Jonassen T, Lee PT, Clarke CF (2001) A defect in coenzyme Q biosynthesis is responsible for the respiratory deficiency in Saccharomyces cerevisiae abc1 mutants. J Biol Chem 276: 18161-18168.

200. Drepper A, Peitzmann R, Pape H (1996) Maltokinase (ATP:maltose 1-phosphotransferase) from Actinoplanes sp.: demonstration of enzyme activity and characterization of the reaction product. FEBS Lett 388: 177-179.

201. Niehues B, Jossek R, Kramer U, Koch A, Jarling M, et al. (2003) Isolation and characterization of maltokinase (ATP:maltose 1-phosphotransferase) from Actinoplanes missouriensis. Arch Microbiol 180: 233-239.

202. Jarling M, Cauvet T, Grundmeier M, Kuhnert K, Pape H (2004) Isolation of mak1 from Actinoplanes missouriensis and evidence that Pep2 from Streptomyces coelicolor is a maltokinase. J Basic Microbiol 44: 360-373.

203. Maestro B, Sanz JM (2007) Novel approaches to fight Streptococcus pneumoniae. Recent Patents Anti-Infect Drug Disc 2: 188-196.

204. Brey RN (2005) Molecular basis for improved anthrax vaccines. Adv Drug Deliv Rev 57: 1266-1292.

205. Reid AN, Whitfield C (2005) Functional analysis of conserved gene products involved in assembly of Escherichia coli capsules and exopolysaccharides: evidence for molecular recognition between Wza and Wzc for colanic acid biosynthesis. J Bacteriol 187: 5470-5481.

206. Jouravleva EA, McDonald GA, Garon CF, Boesman-Finkelstein M, Finkelstein RA (1998) Characterization and possible functions of a new filamentous bacteriophage from Vibrio cholerae O139. Microbiology 144 (Pt 2): 315-324.

207. Barylko B, Wlodarski P, Binns DD, Gerber SH, Earnest S, et al. (2002) Analysis of the catalytic domain of phosphatidylinositol 4-kinase type II. J Biol Chem 277: 44366-44375.

208. Walker EH, Pacold ME, Perisic O, Stephens L, Hawkins PT, et al. (2000) Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine. Mol Cell 6: 909-919.

209. Guo J, Wenk MR, Pellegrini L, Onofri F, Benfenati F, et al. (2003) Phosphatidylinositol 4-kinase type IIalpha is responsible for the phosphatidylinositol 4-kinase activity associated with synaptic vesicles. Proc Natl Acad Sci U S A 100: 3995-4000.

210. Bosotti R, Isacchi A, Sonnhammer EL (2000) FAT: a novel domain in PIK-related kinases. Trends Biochem Sci 25: 225-227.

211. Templeton GW, Moorhead GB (2005) The phosphoinositide-3-OH-kinase-related kinases of Arabidopsis thaliana. EMBO Rep 6: 723-728.

212. Hiom K (2005) DNA repair: how to PIKK a partner. Curr Biol 15: R473-475.

213. Drennan D, Ryazanov AG (2004) Alpha-kinases: analysis of the family and comparison with conventional protein kinases. Prog Biophys Mol Biol 85: 1-32.

214. Yamaguchi H, Matsushita M, Nairn AC, Kuriyan J (2001) Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity. Mol Cell 7: 1047-1057.

215. Singh SK, Matsuno K, LaPorte DC, Banaszak LJ (2001) Crystal structure of Bacillus subtilis isocitrate dehydrogenase at 1.55 A. Insights into the nature of substrate specificity exhibited by Escherichia coli isocitrate dehydrogenase kinase/phosphatase. J Biol Chem 276: 26154-26163.

216. Singh SK, Miller SP, Dean A, Banaszak LJ, LaPorte DC (2002) Bacillus subtilis isocitrate dehydrogenase. A substrate analogue for Escherichia coli isocitrate dehydrogenase kinase/phosphatase. J Biol Chem 277: 7567-7573.

217. Oudot C, Cortay JC, Blanchet C, Laporte DC, Di Pietro A, et al. (2001) The "catalytic" triad of isocitrate dehydrogenase kinase/phosphatase from E. coli and its relationship with that found in eukaryotic protein kinases. Biochemistry 40: 3047-3055.

218. Parker PJ, Parkinson SJ (2001) AGC protein kinase phosphorylation and protein kinase C. Biochem Soc Trans 29: 860-863.

219. Kannan N, Haste N, Taylor SS, Neuwald AF (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. Proc Natl Acad Sci U S A 104: 1272-1277.

220. Braun AP, Schulman H (1995) The multifunctional calcium/calmodulin-dependent protein kinase: from form to function. Annu Rev Physiol 57: 417-445.

221. Xu RM, Carmel G, Sweet RM, Kuret J, Cheng X (1995) Crystal structure of casein kinase-1, a phosphate-directed protein kinase. Embo J 14: 1015-1023.

222. Eide EJ, Woolf MF, Kang H, Woolf P, Hurst W, et al. (2005) Control of mammalian circadian rhythm by CKIepsilon-regulated proteasome-mediated PER2 degradation. Mol Cell Biol 25: 2795-2807.

223. Wedel B, Garbers D (2001) The guanylyl cyclase family at Y2K. Annu Rev Physiol 63: 215-233.

224. Aparicio JG, Applebury ML (1996) The photoreceptor guanylate cyclase is an autophosphorylating protein kinase. J Biol Chem 271: 27083-27089.

225. Schulz S (1992) Guanylyl Cyclase: A Cell-Surface Receptor Throughout the Animal Kingdom. MBL. pp. 155-158.

226. Miranda-Saavedra D, Barton GJ (2007) Classification and functional annotation of eukaryotic protein kinases. Proteins 68: 893-914.

227. Zdobnov EM, Apweiler R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.

228. Rudrabhatla P, Rajasekharan R (2004) Functional characterization of peanut serine/threonine/tyrosine protein kinase: molecular docking and inhibition kinetics with tyrosine kinase inhibitors. Biochemistry 43: 12123-12132.

229. Shiu SH, Li WH (2004) Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. Mol Biol Evol 21: 828-840.

230. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451: 783-788.

231. Li W, Young SL, King N, Miller WT (2008) Signaling properties of a non-metazoan Src kinase and the evolutionary history of Src negative regulation. J Biol Chem.

232. Beullens M, Vancauwenbergh S, Morrice N, Derua R, Ceulemans H, et al. (2005) Substrate specificity and activity regulation of protein kinase MELK. J Biol Chem 280: 40003-40011.

233. Woodgett JR, Gould KL, Hunter T (1986) Substrate specificity of protein kinase C. Use of synthetic peptides corresponding to physiological sites as probes for substrate recognition requirements. Eur J Biochem 161: 177-184.

234. Sontag E (2001) Protein phosphatase 2A: the Trojan Horse of cellular signaling. Cell Signal 13: 7-16.

235. Pawson T, Scott JD (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278: 2075-2080.

236. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285-4288.

237. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) Co-evolution of proteins with their interaction partners. J Mol Biol 299: 283-293.

238. Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 352: 1002-1015.

239. Cokus S, Mizutani S, Pellegrini M (2007) An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinformatics 8 Suppl 4: S7.

240. Jothi R, Cherukuri PF, Tasneem A, Przytycka TM (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J Mol Biol 362: 861-875.

241. Zhou Y, Wang R, Li L, Xia X, Sun Z (2006) Inferring functional linkages between proteins from evolutionary scenarios. J Mol Biol 359: 1150-1159.

242. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5: R35.

243. Hsu SY, Semyonov J, Park JI, Chang CL (2005) Evolution of the signaling system in relaxin-family peptides. Ann N Y Acad Sci 1041: 520-529.

244. Wilkinson TN, Speed TP, Tregear GW, Bathgate RA (2005) Evolution of the relaxin-like peptide family. BMC Evol Biol 5: 14.

245. Hsu SY, Nakabayashi K, Nishi S, Kumagai J, Kudo M, et al. (2002) Activation of orphan receptors by the hormone relaxin. Science 295: 671-674.

246. Wilkinson TN, Bathgate RA (2007) The evolution of the relaxin peptide family and their receptors. Adv Exp Med Biol 612: 1-13.

247. Dschietzig T, Bartsch C, Baumann G, Stangl K (2006) Relaxin-a pleiotropic hormone and its emerging role for experimental and clinical therapeutics. Pharmacol Ther 112: 38-56.

248. Liu C, Chen J, Sutton S, Roland B, Kuei C, et al. (2003) Identification of relaxin-3/INSL7 as a ligand for GPCR142. J Biol Chem 278: 50765-50770.

249. Liu C, Eriste E, Sutton S, Chen J, Roland B, et al. (2003) Identification of relaxin-3/INSL7 as an endogenous ligand for the orphan G-protein-coupled receptor GPCR135. J Biol Chem 278: 50754-50764.

250. Liu C, Kuei C, Sutton S, Chen J, Bonaventure P, et al. (2005) INSL5 is a high affinity specific agonist for GPCR142 (GPR100). J Biol Chem 280: 292-300.

251. Wilkinson TN, Speed TP, Tregear GW, Bathgate RA (2005) Evolution of the relaxin-like peptide family: from neuropeptide to reproduction. Ann N Y Acad Sci 1041: 530-533.

252. Wilkinson TN, Speed TP, Tregear GW, Bathgate RA (2005) Coevolution of the relaxin-like peptides and their receptors. Ann N Y Acad Sci 1041: 534-539.

253. Danilkovitch-Miagkova A, Miagkov A, Skeel A, Nakaigawa N, Zbar B, et al. (2001) Oncogenic mutants of RON and MET receptor tyrosine kinases cause activation of the beta-catenin pathway. Mol Cell Biol 21: 5857-5868.

254. Cottage A, Clark M, Hawker K, Umrania Y, Wheller D, et al. (1999) Three receptor genes for plasminogen related growth factors in the genome of the puffer fish Fugu rubripes. FEBS Lett 443: 370-374.

255. Zhou YQ, He C, Chen YQ, Wang D, Wang MH (2003) Altered expression of the RON receptor tyrosine kinase in primary human colorectal adenocarcinomas: generation of different splicing RON variants and their oncogenic potential. Oncogene 22: 186-197.

256. Chen YQ, Zhou YQ, Fisher JH, Wang MH (2002) Targeted expression of the receptor tyrosine kinase RON in distal lung epithelial cells results in multiple tumor formation: oncogenic potential of RON in vivo. Oncogene 21: 6382-6386.

257. Chen YQ, Zhou YQ, Fu LH, Wang D, Wang MH (2002) Multiple pulmonary adenomas in the lung of transgenic mice overexpressing the RON receptor tyrosine kinase. Recepteur d'origine nantais. Carcinogenesis 23: 1811-1819.

258. Santoro MM, Collesi C, Grisendi S, Gaudino G, Comoglio PM (1996) Constitutive activation of the RON gene promotes invasive growth but not transformation. Mol Cell Biol 16: 7072-7083.

259. Palka HL, Park M, Tonks NK (2003) Hepatocyte growth factor receptor tyrosine kinase met is a substrate of the receptor protein-tyrosine phosphatase DEP-1. J Biol Chem 278: 5728-5735.

260. Furge KA, Zhang YW, Vande Woude GF (2000) Met receptor tyrosine kinase: enhanced signaling through adapter proteins. Oncogene 19: 5582-5589.

261. Keane MM, Lowrey GA, Ettenberg SA, Dayton MA, Lipkowitz S (1996) The protein tyrosine phosphatase DEP-1 is induced during differentiation and inhibits growth of breast cancer cells. Cancer Res 56: 4236-4243.

262. Kovalenko M, Denner K, Sandstrom J, Persson C, Gross S, et al. (2000) Site-selective dephosphorylation of the platelet-derived growth factor beta-receptor by the receptor-like protein-tyrosine phosphatase DEP-1. J Biol Chem 275: 16219-16226.

263. Jeon M, Nguyen H, Bahri S, Zinn K (2008) Redundancy and compensation in axon guidance: genetic analysis of the Drosophila Ptp10D/Ptp4E receptor tyrosine phosphatase subfamily. Neural Develop 3: 3.

264. Berset TA, Hoier EF, Hajnal A (2005) The C. elegans homolog of the mammalian tumor suppressor Dep-1/Scc1 inhibits EGFR signaling to regulate binary cell fate decisions. Genes Dev 19: 1328-1340.

265. Hopper NA (2006) The adaptor protein soc-1/Gab1 modifies growth factor receptor output in Caenorhabditis elegans. Genetics 173: 163-175.

266. Weidner KM, Di Cesare S, Sachs M, Brinkmann V, Behrens J, et al. (1996) Interaction between Gab1 and the c-Met receptor tyrosine kinase is responsible for epithelial morphogenesis. Nature 384: 173-176.

267. Feller SM, Wecklein H, Lewitzky M, Kibler E, Raabe T (2002) SH3 domain-mediated binding of the Drk protein to Dos is an important step in signaling of Drosophila receptor tyrosine kinases. Mech Dev 116: 129-139.

268. Gu H, Neel BG (2003) The "Gab" in signal transduction. Trends Cell Biol 13: 122-130.

269. Herbst R, Carroll PM, Allard JD, Schilling J, Raabe T, et al. (1996) Daughter of sevenless is a substrate of the phosphotyrosine phosphatase Corkscrew and functions during sevenless signaling. Cell 85: 899-909.

270. Raabe T (2000) The sevenless signaling pathway: variations of a common theme. Biochim Biophys Acta 1496: 151-163.

271. Cho CY, Koo SH, Wang Y, Callaway S, Hedrick S, et al. (2006) Identification of the tyrosine phosphatase PTP-MEG2 as an antagonist of hepatic insulin signaling. Cell Metab 3: 367-378.

272. Huynh H, Bottini N, Williams S, Cherepanov V, Musumeci L, et al. (2004) Control of vesicle fusion by a tyrosine phosphatase. Nat Cell Biol 6: 831-839.

273. Saito K, Williams S, Bulankina A, Honing S, Mustelin T (2007) Association of protein-tyrosine phosphatase MEG2 via its Sec14p homology domain with vesicle-trafficking proteins. J Biol Chem 282: 15170-15178.

274. Wang X, Huynh H, Gjorloff-Wingren A, Monosov E, Stridsberg M, et al. (2002) Enlargement of secretory vesicles by protein tyrosine phosphatase PTP-MEG2 in rat basophilic leukemia mast cells and Jurkat T cells. J Immunol 168: 4612-4619.

275. Matveeva EA, Whiteheart SW, Vanaman TC, Slevin JT (2001) Phosphorylation of the N-ethylmaleimide-sensitive factor is associated with depolarization-dependent neurotransmitter release from synaptosomes. J Biol Chem 276: 12174-12181.

276. Murray MJ, Davidson CM, Hayward NM, Brand AH (2006) The Fes/Fer non-receptor tyrosine kinase cooperates with Src42A to regulate dorsal closure in Drosophila. Development 133: 3063-3073.

277. Katzen AL, Montarras D, Jackson J, Paulson RF, Kornberg T, et al. (1991) A gene related to the proto-oncogene fps/fes is expressed at diverse times during the life cycle of Drosophila melanogaster. Mol Cell Biol 11: 226-239.

278. Toshima J, Toshima JY, Amano T, Yang N, Narumiya S, et al. (2001) Cofilin phosphorylation by protein kinase testicular protein kinase 1 and its role in integrin-mediated actin reorganization and focal adhesion formation. Mol Biol Cell 12: 1131-1145.

279. Toshima J, Toshima JY, Takeuchi K, Mori R, Mizuno K (2001) Cofilin phosphorylation and actin reorganization activities of testicular protein kinase 2 and its predominant expression in testicular Sertoli cells. J Biol Chem 276: 31449-31458.

280. Stanyon CA, Bernard O (1999) LIM-kinase1. Int J Biochem Cell Biol 31: 389-394.

281. Okano I, Hiraoka J, Otera H, Nunoue K, Ohashi K, et al. (1995) Identification and characterization of a novel family of serine/threonine kinases containing two N-terminal LIM motifs. J Biol Chem 270: 31321-31330.

282. Bernard O, Ganiatsas S, Kannourakis G, Dringen R (1994) Kiz-1, a protein with LIM zinc finger and kinase domains, is expressed mainly in neurons. Cell Growth Differ 5: 1159-1171.

283. Scott RW, Olson MF (2007) LIM kinases: function, regulation and association with human disease. J Mol Med 85: 555-568.

284. Proschel C, Blouin MJ, Gutowski NJ, Ludwig R, Noble M (1995) Limk1 is predominantly expressed in neural tissues and phosphorylates serine, threonine and tyrosine residues in vitro. Oncogene 11: 1271-1281.

285. Toshima J, Ohashi K, Okano I, Nunoue K, Kishioka M, et al. (1995) Identification and characterization of a novel protein kinase, TESK1, specifically expressed in testicular germ cells. J Biol Chem 270: 31331-31337.

286. Rosok O, Pedeutour F, Ree AH, Aasheim HC (1999) Identification and characterization of TESK2, a novel member of the LIMK/TESK family of protein kinases, predominantly expressed in testis. Genomics 61: 44-54.

287. Ambach A, Saunus J, Konstandin M, Wesselborg S, Meuer SC, et al. (2000) The serine phosphatases PP1 and PP2A associate with and activate the actin-binding protein cofilin in human T lymphocytes. Eur J Immunol 30: 3422-3431.

288. Bamburg JR (1999) Proteins of the ADF/cofilin family: essential regulators of actin dynamics. Annu Rev Cell Dev Biol 15: 185-230.

289. Te Velthuis AJ, Isogai T, Gerrits L, Bagowski CP (2007) Insights into the molecular evolution of the PDZ/LIM family and identification of a novel conserved protein motif. PLoS ONE 2: e189.

290. Kasahara M (2007) The 2R hypothesis: an update. Curr Opin Immunol 19: 547-552.

291. Jung SK, Jeong DG, Yoon TS, Kim JH, Ryu SE, et al. (2007) Crystal structure of human slingshot phosphatase 2. Proteins 68: 408-412.

292. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A 105: 3805-3810.

293. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.

294. Daniel R (2005) The metagenomics of soil. Nat Rev Microbiol 3: 470-478.

295. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, et al. (2008) Microbial ecology of four coral atolls in the northern line islands. PLoS ONE 3: e1584.

296. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

297. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, et al. (2008) Assembly of Viral Metagenomes from Yellowstone Hot Springs. Appl Environ Microbiol.

298. Krupa A, Srinivasan N (2002) The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. Genome Biol 3: RESEARCH0066.

299. Krupa A, Anamika, Srinivasan N (2006) Genome-wide comparative analyses of domain organisation of repertoires of protein kinases of Arabidopsis thaliana and Oryza sativa. Gene 380: 1-13.

300. Krupa A, Srinivasan N (2005) Diversity in domain architectures of Ser/Thr kinases and their homologues in prokaryotes. BMC Genomics 6: 129.

301. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129: 1415-1426.