PHYLOGENOMIC APPROACHES TO THE ANALYSIS OF FUNCTIONAL DIVERGENCE AND SUBCELLULAR LOCALIZATION

by

Daniel Gaston

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
February 2012

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “PHYLOGENOMIC APPROACHES TO THE ANALYSIS OF FUNCTIONAL DIVERGENCE AND SUBCELLULAR LOCALIZATION” by Daniel Gaston in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Dated: February 9, 2012

External Examiner: ______________________________

Research Supervisor: ______________________________

Examining Committee: ______________________________

____

____

Departmental Representative: ______________________________

DALHOUSIE UNIVERSITY

DATE: February 9, 2012

AUTHOR: Daniel Gaston

TITLE: PHYLOGENOMIC APPROACHES TO THE ANALYSIS OF FUNCTIONAL DIVERGENCE AND SUBCELLULAR LOCALIZATION

DEPARTMENT OR SCHOOL: Department of Biochemistry and Molecular Biology

DEGREE: PhD CONVOCATION: May YEAR: 2012

____

Signature of Author

For my family, who got me here

~

To Everett Hovey, for all the books I borrowed

And

In Memoriam to Floyd Gaston, to whom I made a promise

Table of Contents

List of Tables

List of Figures

Abstract

With rapid advances in sequencing technologies and precipitous decreases in cost, public sequence databases have increased in size apace. However, experimental characterization of novel genes and their products remains prohibitively expensive and time consuming. For these reasons, bioinformatics approaches have become increasingly necessary to generate hypotheses of biological function. Phylogenomic approaches use phylogenetic methods to place genes, chromosomes, or whole genomes within the context of their evolutionary history and can be used to predict the function of encoded proteins. In this thesis, two new phylogenomic methods and software implementations are presented that address the problems of subcellular localization prediction and functional divergence prediction within protein families respectively.

Most of the widely used programs for subcellular localization prediction have been trained on model organisms and ignore phylogenetic information. As a result, their predictions are not always reliable when applied to phylogenetically divergent eukaryotes, such as unicellular protists. To address this problem, PhyloPred-HMM, a novel phylogenomic method was developed to predict sequences that are targeted to mitochondria or mitochondrion-related organelles (hydrogenosomes and mitosomes). This method was compared to existing prediction methods using an existing test dataset of mitochondrion-targeted sequences from well-studied groups, sequences from a variety of protists, and the whole proteomes of two protists: *Tetrahymena thermophila* and *Trichomonas vaginalis*. PhyloPred-HMM performed comparably to existing classifiers on mitochondrial sequences from well-studied groups such as animals, plants, and Fungi and better than existing classifiers on diverse protistan lineages.

FunDi, a novel approach to the prediction of functional divergence was developed and tested on 11 biological datasets and two large simulated datasets. On the 11 biological datasets, FunDi appeared to perform comparably to existing programs, although performance measures were compromised by a lack of experimental information. On the simulated datasets, FunDi was clearly superior to existing methods. FunDi, and two other prediction programs, was then used to characterize the functional divergence in two groups of plastid-targeted glyceraldehyde-3-phosphate dehydrogenases (GAPDH) adapted to roles in the Calvin cycle. FunDi successfully identified functionally divergent residues supported by experimental data, and identified cases of potential convergent evolution between the two groups of GAPDH sequences.

List of Abbreviations and Symbols Used

Å	Angstroms
Π	Pi
Γ	Gamma
λ	Lambda
1,3-BPG	1,3-biphosphoglycerate
[2Fe-2S]	Two Iron Two Sulphur
6PGD	6-phosphogluconate dehydrogenase
ACC	Accuracy
AliMask-CS	Alignment Masking with Confidence Scores
ANN	Artificial Neural Network
ASCT	Acetate:succinate CoA transferase
ATP	Adenosine Triphosphate
AUC	Area Under the Curve
BD	Birth-Death
BLAST	Basic Local Alignment Search Tool
CAT	Categorical rates
CBOrg	Comparative BLAST for Organelles
CD-HIT	Cluster Database at High Identity with Tolerance
CoA	Coenzyme A
CP12	Chloroplast Protein 12
Cpn60	Chaperonin 60
DET	Difference Evolutionary-Trace

DNA	Dideoxyribonucleic acid
EGT	Endosymbiotic Gene Transfer
EST	Expressed Sequence Tag
ET	Evolutionary-Trace
E-value	Expect value
FD	Functional divergence/functionally divergent
FGAR	Phosphoribosylformylglycinamidine
Fe/S	Iron-sulphur
[Fe-S]	Iron-sulphur
[FeFe]	Iron-iron
FN	False Negative
FP	False Positive
FSA	Fast Statistical Alignment
G3P	Glyceraldehyde-3-phosphate
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GO	Gene Ontology
HClust	Hierarchical Clustering
HSP70	Heat Shock Protein 70
HMM	Hidden Markov Model
JSD	Jensen-Shannon Divergence
JSD$_{AVG}$	Average Jensen Shannon Divergence
JTT	Jones-Taylor-Thornton
LCFA	Long-chain Fatty Acid
LGT	Lateral Gene Transfer

LRT	Likelihood ratio test
MAFFT	Multiple Alignment Fast Fourier Transform
MCC	Matthews Correlation Coefficient
MCP	Mitochondrial Carrier Protein
MCL	Markov Clustering
MCL-Inf2	Markov Clustering Inflation parameter of two
ML	Maximum likelihood
MRO	Mitochondrion Related Organelle
mtHSP70	Mitochondrial Heat Shock Protein 70
NADP+	Nicotinamide adenine dinucleotide phosphate (Oxidized)
NADPH	Nicotinamide adenine dinucleotide phosphate (Reduced)
NAD+	Nicotinamide adenine dinucleotide (Oxidized)
NADH	Nicotinamide adenine dinucleotide (Reduced)
ND	Nearest Distance
ND4	Nicotinamide adenine dinucleotide dehydrogenase subunit 4
NN	Neural Network
NTL	Normalized Tree Length
P(FD)	Posterior probability of functional divergence
PFO	Pyruvate:ferredoxin oxidoreductase
pI	Isoelectric point
PR	Precision-Recall
PRK	Phosphoribulokinase
RAS	Rates-across-sites
ROC	Receiver Operator Characteristic

rvET	Real-Value Evolutionary Trace
S(FD)	Functional divergence score
SAD	Shortest Average Distance
SCPS	Spectral Clustering of Protein Sequences
Std. Dev.	Standard deviation
SVM	Support Vector Machine
T-SAD	Trimmed Shortest Average Distance
TCA	Tricarboxylic Acid Cycle
TMHMM	Transmembrane HMM
TN	True Negative
TP	True Positive
WAG	Whelan and Goldman

Chapter 1 Introduction

This chapter (section 1.3) contains material originally published in:

"Gaston D, Tsaousis AD, Roger AJ. 2009. Predicting Proteomes of Mitochondria and Related Organelles from Genomic and Expressed Sequence Tag Data. Methods in Enzymology. Mitochondrial Function, Part B: Mitochondrial Protein Kinases, Protein Phosphatases and Mitochondrial Diseases. 457:21-47"

The prediction of protein function remains an ongoing challenge in bioinformatics, and a wide variety of computational approaches has been applied to many different types of functional characterization. With massive decreases in price and advances in technology, genome sequencing projects are rapidly increasing the amount of sequence data available in public databases. As these databases grow, so too will the demand for rapid, robust, and user-friendly genome scale data processing tools and pipelines for the prediction of protein function. Although it was shown more than a decade ago that the top Basic Local Alignment Search Tool (BLAST) hit is not necessarily the closest phylogenetic relative of a sequence (Eisen *et al.* 1997; Eisen 1998b; Koski & Golding 2001) the BLAST approach is still routinely used to quickly predict functions and annotations in genome and transcriptome projects because of its wide acceptance and familiarity. The use of phylogenetic methods to infer or predict functions, originally termed 'Phylogenomics' (Eisen *et al.* 1997; Eisen 1998c) provides a more sophisticated picture of the evolution and function of protein families (Eisen *et al.* 1995; Eisen 1998a; De Grassi *et al.* 2008; Cao *et al.* 2011; Courtiade *et al.* 2011) and is often used in comparative genomics in conjunction with other experimental techniques (Schrimpf *et al.* 2009).

1.1 Phylogenomics

Phylogenomic analyses typically proceed by: 1) identifying homologous sequences of interest, 2) constructing and refining multiple sequence alignments, 3) reconstructing phylogenetic trees that relate the aligned sequences, 4) annotating the tips of the tree according to some molecular function(s), and 5) tracing (by some objective method) the evolution of that molecular function along the tree (Sjölander 2004). Once the evolutionary history of the function is known, the functional annotation of the unknown sequence can be inferred.

Some phylogenomic methods such as Orthostrapper (Storm & Sonnhammer, 2002) and RIO (Zmasek & Eddy 2002) rely on the idea that *orthologs* (Fitch, 1970), genes related by speciation events, should have more similar functions to one another than to their *paralogs* (Ohno, 1970; Koonin, 2005), genes related to one another by gene duplication events (Fitch, 1970; Zuckerkandl & Pauling, 1965). Orthology-based functional assignment requires the identification of nodes in the phylogenetic tree that correspond to speciation events, and must differentiate these from nodes corresponding to gene duplication events from which paralogs descend, which remains a very difficult problem in phylogenetics and bioinformatics. A number of programs for orthology selection and databases of orthologs have been developed including TribeMCL (Enright *et al.* 2002), OrthoMCL (Li *et al.* 2003; Chen *et al.* 2006), InParanoid (Ostlund *et al.* 2010), OMA-DB (Roth *et al.* 2008; Altenhoff *et al.* 2011), OrthologID (Chiu *et al.* 2006), OrthoSelect (Schreiber *et al.* 2009), PHOG (Datta *et al.* 2009), LOFT (van der Heijden *et al.* 2007), MetaPhoRs (Pryszcz *et al.* 2011), OrthoInspector (Linard *et al.* 2011), OrthoDB (Waterhouse *et al.* 2011), and PhylomeDB (Huerta-Cepas et al. 2011). While the majority of orthology prediction methods and databases perform reasonably well (Chen *et al.* 2007; Boeckmann *et al.* 2011), differences in the underlying datasets from which they construct ortholog groups, and even differences regarding how to best organize and present these groups can lead to different results (Boeckmann *et al.* 2011) depending on the classification strategy used. A strict reliance on orthology prediction can therefore introduce additional error in the process of functional annotation by a phylogenomic method.

There are several phylogenomic annotation methods and software packages that do not rely on known or defined orthology or paralogy definitions. For example SIFTER (Engelhardt *et al.* 2005, 2009, 2011) is an annotation method that adopts an empirical Bayes approach, creating a robust statistical framework that does not merely propagate function to entire clades of orthologs, but models additional information and accounts for scarcity of annotation within trees to make statistical phylogenetic inferences. The Gene Ontology (GO) Consortium has also built a phylogenomic pipeline, called PAINT (Gaudet *et al.* 2011a), that allows curators to propagate functional annotations within the GO network based on phylogenomic methods of inference. However, it is important that pipelines and software be able to take advantage of the most recent advances in alignment generation and phylogenetic reconstruction, including the best or most appropriate evolutionary models. Flexible and robust software packages should be easily extensible and modifiable by end users in order to remain up to date and relevant. In this thesis, we focus on two aspects of the *in silico* characterization of protein function using phylogenomic methods: the analysis of functional divergence and prediction of subcellular localization and develop novel phylogenomic software tools to carry out these analyses.

1.2 Functional Divergence

Functional divergence is the process by which proteins drift in molecular functions after speciation or gene, chromosome, or whole genome duplication events, and is perhaps the largest process contributing to the generation of molecular diversity (Ohno 1970). The hypothesis that orthologs should be more functionally similar than paralogs has recently been termed the ‘Ortholog Conjecture’ (Nehrt *et al.* 2011), but was first examined in reference to the α-, β-, and γ-hemoglobin paralogous families in jawed vertebrates (Zuckerkandl & Pauling 1965). In the context of functional divergence, orthology and paralogy are most often thought of in terms of a simple division between speciation events and a single ancestral duplication preceding the speciation events in question. However, ortholog/paralog relationships are not always consistent with these simple one-to-one mappings. Sequences frequently have multiple orthologs in another species, and there may be duplication events “nested” within overall gene trees. These complex relationships can be easily misunderstood, especially in the context of functional characterization (Fitch 2000; Jensen 2001). A set of terms for dealing with sub-types of paralogs (co-ortholog, in-paralog, out-paralog) have also been proposed (Sonnhammer & Koonin 2002), and are slowly becoming more widely adopted. These complex relationships further complicate phylogenomic inferences of protein function.

While the ortholog conjecture has almost universally been accepted, its experimental support has generally been limited to specific biological examples that have been reasonably well characterized, as with the hemoglobins discussed above. Strictly, orthologs should not always be more similar to one another in function than to paralogs, as all proteins can drift in function over evolutionary time (Jensen 2001). Indeed some recent studies have shown either greater functional divergence between orthologs than between paralogs, or at least that there is the potential for just as much divergence between orthologs as between paralogs (Studer & Robinson-Rechavi 2009, 2010; Gharib & Robinson-Rechavi 2011; Nehrt *et al.* 2011). However, it should be noted that complicated relationships with multiple duplication events and potentials for differential gene loss can greatly complicate the interpretations of divergence within protein families. Furthermore, the generality of these observations are unknown, as studies have mostly focused on protein-coding genes within animals (Studer & Robinson-Rechavi, 2010) or even more specifically on comparisons between humans and mice (Gharib & Robinson-Rechavi 2011; Nehrt *et al.* 2011). Indeed great care should be taken to place analyses of divergence, and comparisons of orthologs and paralogs, within the appropriate context of their age and biological context, such as the difference between one-to-one ortholog relationships and more complex relationships (Berglund *et al.* 2008; Hubbard *et al.* 2009; Studer & Robinson-Rechavi 2009). It is important to note, as in Jensen (2001), that after gene duplication either descendant sequence has the possibility to diverge and acquire new functions over evolutionary time. In terms of the ortholog conjecture, the canonical examples typically concern ancient duplication events where, for all taxa considered in the analysis, the relevant duplication precedes the subsequent speciation events. In this work on functional divergence we do not require this classical view of functional divergence to be true; however we are explicitly concerned with situations where monophyletic groups of sequences, whether they be descended from speciation or duplication events, share some biological similarity that can be contrasted with other sequences in the protein family. It is in this context that we can leverage the power of a phylogenomic framework to characterize functional divergence at the amino acid level. While the evolutionary pressures and processes discussed below are generally in the context of these canonical or standard ortholog/paralog relationships and evolutionary divergence, similar examples can be found for strict ortholog relationships as well. Indeed divergence between orthologs may be an important factor in the context of subcellular localization, discussed later.

1.2.1 Duplication, Divergence, and Natural Selection

The driving force of functional divergence is a shift in selective constraints on a gene in one or both descendant lineages from an ancestor. Whether this is a relaxation of selection, an increase in either diversifying or purifying selection, or a shift in the selective constraints at a site in a protein, ultimately selection or genetic drift drive any variant allele to fixation or loss in a population. Proteins are under a variety of selective constraints including their molecular function, thermodynamic constraints on protein folding (Wilke *et al.* 2005; Bloom *et al.* 2007), expression level maintenance, and the prevention of aggregation (DePristo *et al.* 2005). The latter two may be particularly important early after gene duplication, while copy-number polymorphism is not yet fixed in a population. Mutations that cause aberrations in protein folding or stability frequently can lead to aggregation or rapid degradation (Gregersen *et al.* 2005; McClellan *et al.* 2005; Cuervo *et al.* 2010). Both aggregation and premature degradation of proteins are important in a number of human diseases including cystic fibrosis, Parkinson's and Huntington's disease (DePristo *et al.* 2005) while copy-number polymorphism is implicated in an increasing number of human diseases (McCarroll & Altshuler 2007; Zhang *et al.* 2009; Ricard *et al.* 2010).

The fate of most gene duplications is pseudogenization (Lynch & Conery 2000); the introduction of nonsense substitutions create premature stop codons that truncate protein coding genes and other substitutions in regulatory regions that eventually turn off expression. Substitutions, insertions, or deletions accumulate neutrally in a pseudogene until eventually it is purged from the genome entirely, or its sequence becomes unrecognizable. Alternatively, selection may act to preserve redundant gene loci in the absence of functional divergence. Nowak *et al* (1997) put forward four models for the maintenance of such redundancy depending on the relative fitness of phenotypes with multiple gene copies and the mutation and recombination rates between loci. The gene dosage hypothesis states that redundant gene loci result in altered expression levels, which may have either positive or negative effects on fitness (Veitia 2002; Liang *et al.* 2008; Qian & Zhang 2008).

In addition to pseudogenization and the maintenance of redundant loci by selection, copies may diverge in function. Two general categories of functional divergence have been described: neofunctionalization and subfunctionalization. Neofunctionalization, as originally described by Ohno (1970), involves the gain of an entirely new, but often related, function or functions by one of the duplicated genes. On the other hand, during subfunctionalization selection acts to maintain ancestral function in the aggregate, where the ancestral protein encoded two or more distinct functions. Mutations that reduce ancestral function in one copy are tolerated so long as the total function is not lost entirely (Force *et al.* 1999; Stoltzfus 1999; Lynch & Force 2000). Various experimental studies have been conducted, indicating that either neo- or subfunctionalization is the prevailing mode of divergence after gene duplication (Zhang *et al.* 1998; Force *et al.* 1999; Betrán & Long 2003; Aguileta *et al.* 2004; Hughes & Liberles 2007). However, these two types of functional divergence are not mutually exclusive. It is also possible that one duplicate copy first undergoes a process of subfunctionalization followed by a longer period of neofunctionalization such that a related, but distinct, function is retained when compared to one of the ancestral protein's multiple functional roles (He & Zhang, 2005; Hughes & Liberles, 2007; Rastogi & Liberles, 2005).

In any case, functional divergence is likely to proceed by an initial burst of rapid evolution as selective constraints are either relaxed at previously constrained sites allowing for neutral substitutions to accumulate (under subfunctionalization scenarios) and/or changes at other positions are fixed due to positive selection for new function (i.e during neofunctionalization). This period is followed by a longer period of divergence under re-imposed purifying selection. A large-scale analysis of recent paralogs and orthologs of similar ages and level of divergence across all three domains of life showed that the strength of purifying selection was lower (but not neutral) in paralogs than orthologs, with consequently higher overall rates of evolution and an increased strength of positive selection, perhaps for gene dosage effects (Kondrashov *et al.* 2002). Under this scenario, instead of a relaxation of selection and fixation neutrally prior to functional divergence it is hypothesized that there is positive selection on most gene duplicates from the very beginning and that functional divergence is driven by diversifying selection. Recently, an analysis of lineage-specific, and thus “young”, gene duplicates in the human, macaque, mouse, and rat genomes showed evidence of positive selection in approximately 10% of paralagous pairs (Han *et al.* 2009). However, an earlier study in all vertebrates showed that while positive selection was common in vertebrate genes, there was no difference between singletons and paralagous pairs from whole genome duplication events, either the fish-specific duplication event or the two whole-genome duplications near the origin of the vertebrate lineage (Studer *et al.* 2008). The biological reality is likely that duplicate genes are fixed in a population for a variety of reasons, with both neutral evolution and positive selection having varying impact depending on the protein function, the organisms concerned and their population sizes.

1.2.2 The Impact of Gene Duplications and Divergence

Gene duplication events followed by rapid diversification has clearly played a role in the origin and diversity of many multicellular lineages (Van de Peer *et al.* 2001; Rodríguez-Trelles *et al.* 2003). Gene, chromosome, or whole genome duplication events have contributed to the pathogenicity or host specificity of several important eukaryotic pathogens including members of the genus *Phytophthora* (Martens & Van de Peer, 2010), *Trypanosoma brucei* (Jackson 2007), and *Trichomonas vaginalis* (Carlton *et al.* 2007a; Rada *et al.* 2011). These duplication events likely provided a rich source of potential variability allowing parasites to adapt to new hosts in the evolutionary arms race between a parasite’s defenses and their hosts’ immune systems, although this hypothesis has not been rigorously tested.

One way in which duplicated genes may provide increased variability is through the re-wiring of protein-protein interaction networks. Over the last decade, intrinsic structural disorder has been increasingly recognized for its importance in protein-protein interactions (Dunker *et al.* 2008; Fong *et al.* 2009; Gsponer & Babu 2009; Fong & Panchenko 2010; Turoverov *et al.* 2010), particularly in central hubs of protein-protein interaction networks (Haynes *et al.* 2006; Oldfield *et al.* 2008; Patil *et al.* 2010a, 2010b). Observed gene duplicates may be biased towards those genes encoding proteins with greater amounts of intrinsic disorder in tertiary structure or may result in a greater amount of intrinsic disorder after the duplication event (Montanari *et al.* 2011; Siltberg-Liberles 2011) as evidenced by greater degrees of intrinsic disorder in paralagous gene pairs compared to singletons. Specificity-determining sites, those sites which have undergone functional divergence and are thought to encode specificity differences in protein-protein or protein-small molecule binding between paralogs, have been observed to occur more frequently in neighbouring regions of disorder (Aharoni *et al.* 2005; Chakrabarti *et al.* 2007; Chakrabarti & Panchenko 2009). A large proportion of putative positively selected sites in 12 species of *Drosophila* have been shown to occur in intrinsically disordered regions even when considering their relative frequency in the genome (Ridout *et al.* 2010), a subset of these sites may correlate with those also undergoing functional divergence.

1.2.3 Characterizing Functional Divergence

Functional divergence results in distinctive substitution patterns in multiple alignments of protein families, which have been broadly classified as Type I and Type II functionally divergent sites (Gu 1999, 2001, 2006). Type I functional divergence (Gu 1999, 2001) or rate-shifting sites (Abhiman & Sonnhammer 2005) reflect a specific form of heterotachy (Philippe & Lopez 2001; Lopez *et al.* 2002; Philippe *et al.* 2003) characterized by a switch in evolutionary rate within a portion of the phylogenetic tree. These sites are often strongly conserved in one subtree and much more variable in the other subtree; this type of rate shift is more frequent than other types of heterotachy in paralogs undergoing functional divergence (Philippe *et al.* 2003). This pattern is generally interpreted as either a relaxation of evolutionary constraints, resulting in a shift from purifying selection to neutral evolution and a resulting higher rate of sequence substitution. Alternatively it is also possible that a formerly unconstrained site becomes fixed in a paralog neutrally and fortuitously suppresses the negative fitness effects of a mutation elsewhere in the genome and ultimately becomes maintained by purifying selection (Stoltzfus 1999). In contrast, type II functional divergence (Gu, 2006) is characterized by a high degree of conservation at a site in both subfamilies, but for amino acids with different physical and chemical properties. This pattern of substitution in its most extreme form corresponds to the so-called 'constant-but-different' (Gribaldo *et al.* 2003) or conservation-shifting (Abhiman & Sonnhammer 2005) site. An initial burst of diversifying evolution is proposed to occur immediately after duplication in this case, followed by the fixation of the radically changed amino acid residue (or residue 'type') prior to the subsequent diversification of lineages. While both type I and type II patterns are likely to arise under functional divergence, they may also occur by chance at neutrally evolving or sites under weak purifying selection and under certain phylogenetic tree shapes. Distinguishing between such neutral substitution patterns and those that are functionally meaningful is both important and difficult (Anisimova & Liberles, 2007) providing a strong rationale for a phylogenetic approach to the identification of sites undergoing functional divergence.

Predictors of functional divergence are designed to identify the amino acid positions that are contributing the most to functional divergence between paralogs and/or orthologs. These positions, also known as specificity-determining sites, are often located in, or near, active sites or protein-protein/protein-small molecule binding patches (Aharoni *et al.* 2005; Capra & Singh 2008) and can often be located on surface loops (Aharoni *et al.* 2005). Therefore, identifying functionally divergent amino acid positions can aid in the characterization of function in an unknown subfamily, predict active or binding site residues, or be used for rational drug design. Measures of sites undergoing functional divergence are, in many ways, similar to measures quantifying signals of selection. While functional divergence methods have generally focused on quantifying the patterns of substitution at the amino acid level as opposed to the codon or nucleotide level (Anisimova & Liberles, 2007), there have been recent analyses of functional divergence correlating signals of functional divergence with signals of selection using codon models showing a correlation between selection and functional divergence patterns, particularly among type II sites (Studer & Robinson-Rechavi 2010). A wide variety of approaches have been applied to the identification of residues undergoing functional divergence in protein families. Techniques can be broken into three broad and overlapping categories: phylogenetic, information theoretic, and structural with the latter category generally being used in a supplementary fashion.

1.2.4 Phylogeny-Based Functional Divergence Predictors

Perhaps the two most widely used prediction methods are those implemented in the phylogeny-based DIVERGE/DIVERGE2 (Gu, 1999, 2001, 2006) and Evolutionary Trace methods (Lichtarge *et al.* 1996; Mihalek *et al.* 2004; Yao *et al.* 2006; Ward *et al.* 2009). The Evolutionary Trace (ET) method was originally designed for identification of functionally important residues, but has also been used for identifying functional divergence. Later alterations of the method improved predictions with the addition of rank-based scoring functions (Mihalek *et al.* 2004; Yao *et al.* 2006). The Difference Evolutionary Trace method (Raviscioni *et al.* 2006; Rodriguez *et al.* 2010) formalized a procedure for the identification of sites with different degrees of functional importance between subfamilies. The Evolutionary Trace method is performed on the phylogenetic tree and sequence alignment representing the entire protein family under consideration and then for each sub-family independently. Sites that are considered functionally important with some cut-off (often the top 20% of sites) in the whole family are removed from those sites that are considered functionally important within individual subfamilies. This leaves the subset of sites considered functionally important in sub-groups but not in the family as a whole, which correspond to sites undergoing functional divergence. This method is capable of recognizing both Type I and II patterns of functional divergence.

The DIVERGE software package implements three separate methods for predicting functional divergence. The first method for Type I functional divergence, known as the Gu 99 method (Gu, 1999) implements a maximum-likelihood model for the correlation of evolutionary rates between phylogenetic groups. Sites undergoing Type I functional divergence will have less correlated evolutionary rates than those not undergoing functional divergence. Using Bayes' rule, the posterior probability can be calculated and using an arbitrary cut-off value (often 0.5), sites are predicted to belong to either the functionally divergent or non-divergent class. This subtree likelihood method was later extended to a whole tree likelihood model intended to capture both Type I and Type II functional divergence (Gu, 2001). For Type II specific functional divergence a later model was implemented (Gu, 2006) that considered the deviation from expected substitution patterns under a standard substitution matrix such as the Dayhoff (Dayhoff *et al.* 1978) or JTT (Jones *et al.* 1992) matrices. Substitutions are categorized as 'radical' or 'conserved' changes based on whether they interchange between four groups of amino acids established on the basis of shared physical properties (positive charge, negative charge, hydrophilic, hydrophobic). This model also incorporates a gamma distribution for evolutionary rates (Yang 1994). Inputs to DIVERGE are restricted to strictly bifurcating phylogenetic trees and several simple models of phylogenetic reconstruction are available to build neighbour-joining trees from input data if no tree is supplied.

Knudsen and Miyamoto (2001) developed a likelihood-ratio test method for predicting functionally divergent protein residues, as well as slowly-evolving conserved sites based on site rate differences. This likelihood-ratio test was later extended for Type II functional divergence in a similar fashion, modeling Type II divergence as the occurrence of a rate shift along the internal branch separating the two sub-families under consideration (Knudsen *et al.* 2003).

Similarly covARES (Blouin *et al.* 2003; Inagaki *et al.* 2003) implements a statistical test for a rate-shift difference between sequences in two subfamilies of a phylogenetic tree in order to identify Type I functionally divergent sites. In order to predict Type II functionally divergent sites, covARES also looks for what are called “differently-evolving” and “absolutely-differently-evolving” sites. The former may correlate with a rate-shift and thus be properly defined as Type I residues where a residue is conserved in one subtree but not the other. Absolutely-differently-evolving sites correspond to canonical Type II functional divergence patterns. The covARES method also constructs vectors of chemical properties at a site and tests for a significant difference in these vectors between subgroups.

SPEL (Pei *et al.* 2006) also implements a log-likelihood ratio based phylogenetic method for classification with simulations used to generate p-values. Unlike many predictors, SPEL is explicitly designed not to require designations of which sequences belong to which specificity group. Making it potentially powerful in cases where the sequence and functional data is less clear. The site log-likelihood is calculated given a phylogenetic tree. The amino acid column is then shuffled 100 times and the average log-likelihood of the shuffled columns is calculated. The resulting log-likelihood ratio is checked for significance by comparing to the distribution of log-likelihood ratios that result from 1000 simulated amino acid positions on the same phylogenetic tree (See Pei et al. 2006 for details). While this method is potentially quite powerful; sites that do not fit the normal distribution expected based on a substitution model such as WAG (Whelan & Goldman 2001) can arise due to various model violations not necessarily related to functional divergence, such as heterotachy. Their measure reflects how much better the data fits the phylogenetic tree compared to randomly shuffled data.

The final category of phylogeny-based methods are those that analyse proteins at the nucleotide-level employing codon models and tests for positive selection that have been extensively used in a likelihood ratio testing framework for the identification and characterization of functional divergence in protein families (Forsberg & Christiansen 2003; Bielawski & Yang 2004; Loughran *et al.* 2008). These methods, while very powerful, depend on the accurate estimation of synonymous to non-synonymous substitution rates and thus may have limited accuracy for deeper divergences or trees involving long branches where synonymous changes are saturated (Anisimova *et al.* 2002).

1.2.5 Information Theory-Based Functional Divergence Predictors

In contrast to the phylogenetic methods described above, information theoretic approaches typically ignore the phylogenetic relationship between sequences and consider only the functional clustering of sequences. This may be particularly useful in cases where the function does not map exactly with phylogenetic relationship. These information theoretic based classifiers generally employ measures such as the relative entropy (Hannenhalli & Russell 2000; Kalinina *et al.* 2004a), Mutual Information (Mirny 2002; Kalinina *et al.* 2004b), Sequence Harmony (Pirovano *et al.* 2006) (which itself an extension of the Shannon Entropy (Shannon 1948)), The Two Entropy (Shannon Entropy) measure at both the super and sub-family levels (Ye *et al.* 2006), or other measures that differentiate between within-group and between-group similarity scores (Capra & Singh 2008). Discriminatory machine learning techniques such as that implemented in Multi-RELIEF (Ye *et al.* 2008) where the objective is to discriminate between classes based on feature vectors (features being amino acids at sites in the case of Multi-RELIEF) can also be classified as information theoretic approaches. Many information theoretic approaches attempt to capture some evolutionary information, either by weighting scores according to sequence similarity as in GroupSim (Capra & Singh 2008) or by using sequence substitution matrices to account for the relative frequencies of particular amino acid substitutions (Kalinina *et al.* 2004b; Chakrabarti *et al.* 2007; Capra & Singh 2008)

While not explicitly information theoretic, there have also been developments in statistical model based classifiers for functional divergence. A Markov-Chain Monte Carlo Gibbs Sampler such as mcBPPS (Neuwald 2011) is capable of partitioning thousands of sequences, with or without “gold-standard” examples, to a specified number of functionally divergent groups and simultaneously optimizing the sequence position patterns (and thus specificity determining sites) that define them. However, as pointed out by Neuwald (2011) the mcBPPS classifier is designed to split the data into a set of statistically meaningful groups and their optimal defining patterns. While these patterns are statistically meaningful, and intended as a starting point for exploratory analysis, they are not specifically targeting sites undergoing functional divergence.

1.2.6 Combined Approaches to Functional Divergence Prediction

Some approaches attempt to combine explicitly evolutionary measures and information scores. For example SPEER (Chakrabarti *et al.* 2007) incorporates the relative entropy (Kullback-Leibler Divergence) of amino acid frequencies at a site, a Euclidean distance measure of the dissimilarity of the physicochemical properties of amino acids in groups, and the evolutionary rate; the difference in evolutionary rate being characterized in an explicit maximum-likelihood phylogenetic approach. All of these measures have been employed separately as predictors of functional divergence. In general, ensemble approaches use machine-learning techniques, or simple consensus/voting schemes, to combine the results of different prediction methods for improved performance (Opitz & Maclin 1999; Rokach 2010). This technique that has also shown promise for characterizing functional divergence (Chakrabarti & Panchenko 2009) where sites were predicted to be functionally divergent only if three top-performing methods (GroupSim (Capra & Singh 2008), Multi-RELIEF (Ye *et al.* 2008), and SPEER) all predicted a site to be divergent, or any two did.

1.2.7 Assessing the Accuracy of Functional Divergence Prediction

Comparisons between the performance of existing methods is often difficult due to the small number of biological datasets that have been used, or that are considered well characterized on an experimental level (Chakrabarti *et al.* 2007; Chakrabarti & Panchenko 2009). The performance of some programs have been reasonably well investigated on these or similar datasets, while the others have shown promise based on their ability to predict functionally divergent sites in one or a handful of handpicked datasets where the predictions can be rationalized in terms of their biological relevance. We know of no large-scale investigations conducted to date that have either attempted to evaluate performance of predictors based on extensive simulations of functional divergence or that investigate particular factors that may affect classification performance of predictors. Given the large number of predictors currently available, newly proposed methods for identifying functional divergence should be fast, sensitive, and be able to exploit the latest advances in computational phylogenetics. Additional work needs to be done to better characterize the strengths and weaknesses of various classifiers, and to assess their relative performance under differing evolutionary conditions. Current comparisons that rely on a small number of biological datasets, which contain only a small fraction of verified positions contributing to functional divergence, are insufficient.

1.3 Subcellular Localization

The subcellular location of a protein can provide useful information as to function, particularly in cases of duplication and retargeting to subcellular compartments.

Eukaryotic cells are heterogeneous and crowded environments and contain a variety of organelles and subcellular compartments with specialized functions including mitochondria, chloroplasts, the Golgi complex, the endoplasmic reticulum and other endomembrane compartments such as vacuoles. Of particular interest is the mitochondrion, an organelle of endosymbiotic origin acquired by the last common ancestor of all extant eukaryotes (Gray 1999). The progenitor of the modern mitochondrion was an alpha-proteobacterium of uncertain taxonomic affiliation (Andersson *et al.* 1998, 2003; Lang *et al.* 1999; Gray *et al.* 2001; Esser *et al.* 2004; Fitzpatrick *et al.* 2006; Brindefalk *et al.* 2011; Georgiades *et al.* 2011; Thrash *et al.* 2011) complete with a eubacterial genome. Over time most of this genome was either lost or transferred to the host nucleus as the endosymbiont lost autonomy and became an organelle (Timmis *et al.* 2004). Genes transferred to the host nucleus encode proteins that are translated in the cytoplasm but are targeted to the organelles if they are important to mitochondrial function. The ‘retargeting’ of a gene product occurs by the acquisition of targeting sequences recognized by proteins of the mitochondrial import apparatus (Pfanner *et al.* 2004; Dolezal *et al.* 2006). The second major goal of this thesis is the development of new *in silico* tools for predicting the proteomes and ultimately the functions of mitochondria and mitochondrion-related organelles (MROs) in unicellular anaerobic eukaryotes.

1.3.1 Biological Diversity of Mitochondrion-Related Organelles

'Classical' aerobically-functioning mitochondria, such as those present in mammals, plants, and fungi, are the typical text-book examples of mitochondria that generate ATP by oxidative phosphorylation and carry out other biochemical functions including replication of mitochondrial DNA, transcription and translation, iron-sulfur (Fe/S) cluster biogenesis, metabolism of lipids, and the interconversion of amino acids. However, these classical mitochondria comprise only a portion of the diversity found in organelles derived from this endosymbiotic event. For instance, there are also anaerobic mitochondria that generate ATP by respiration, but use terminal electron acceptors other than oxygen (e.g. fumarate, nitrate). This type of anaerobic ATP production has been observed in a variety of eukaryotic lineages but is arguably best understood in parasitic animals such as the nematode *Ascaris* (Howe, 2008; Tielens *et al*. 2002; Takaya *et al*. 1999; Kobayashi *et al*. 1996; Finlay *et al*. 1983; Tielens and van Hellemond, 1998; van Hellemond *et al*. 1998). Amongst the unicellular eukaryotes (protists) living in low oxygen conditions, even stranger mitochondrion-related organelles (MROs) have been discovered, including hydrogenosomes (Cerkasovová *et al*. 1973; Lindmark and Müller, 1973) and mitosomes (Tovar *et al*. 1999).

Classical hydrogenosomes are genome-lacking, double-membrane bound organelles characterized by the production of molecular hydrogen as a by-product of ATP generation (Müller, 1993). They catabolize pyruvate and malate via substrate-level phosphorylation, producing ATP anaerobically using a series of enzymes, many of which are not generally found in the mitochondria of aerobic eukaryotes (e.g. pyruvate:ferredoxin oxidoreductase (PFO), [FeFe] hydrogenase and acetate:succinate CoA transferase (ASCT)).These organelles were first described in parasitic parabasalid protists (e.g. *Tritrichomonas* and *Trichomonas*) and it was initially unclear whether they represented novel non-endosymbiont-derived organelles, endosymbiotic organelles of unique origin, or they were related to mitochondria. The matter was resolved in the 1990s when nuclear-encoded mitochondrial marker proteins, such as mitochondrial-type chaperonin 60 (cpn60) and Hsp70 (mtHsp70) were discovered within the hydrogenosomal proteome and phylogenetic analysis indicated that they were most closely related to homologs existing in other mitochondria (Bui *et al*. 1996; Germot *et al*. 1996; Horner *et al*. 1996; Roger *et al*. 1996; Palmer, 1997). More recently, the hydrogenosomal localization of a number of systems including a mitochondrial carrier family (MCF) transporter and mitochondrial iron-sulfur (Fe/S) proteins as well as the presence of two subunits of complex I, indicate that hydrogenosomes are in fact highly modified mitochondria (Dolezal *et al*, 2005; Hrdy *et al*, 2004; Sutak *et al*, 2004; Tachezy *et al*, 2001; van der Giezen *et al*, 2002). Hydrogenosomes have evolved from mitochondria multiple times in a number of distantly related eukaryotic lineages including parabasalids, ciliates, and anaerobic fungi (Barberà *et al*. 2007). Recent analysis indicates that many of the unique hydrogenosomal metabolic enzymes have been transferred, via lateral gene transfer, between eukaryotes as the sequences often form monophyletic groups that are incongruent with the presumed taxonomy (Hug *et al.* 2010; Hampl *et al.* 2011; Tsaousis *et al.* 2012) Although their energy generating pathways represent a convergent adaptation to anaerobiosis, it is likely that hydrogenosomes in these different lineages (like mitochondria of different eukaryotic lineages) have distinct properties.

Mitosomes are a poorly characterized and heterogeneous category of MROs that are double-membrane bound, lack cristae and a genome, and, as far as is currently known, play no known role in energy generation. They were first discovered (and named) in the amoeboid human parasite *Entameoba histolytica* (Mai *et al*. 1999; Tovar *et al*. 1999) and were shown to contain chaperonin-60 homologs clearly related to mitochondrial homologs in their proteomes (Mai *et al*. 1999; Dolezal *et al*. 2005) suggesting a shared origin with mitochondria. Mitosomes have been shown to exist in the microsporidian *Trachipleistophora hominis* (Williams *et al*. 2002), *Giardia intestinalis* (Tovar *et al*. 2003), and a variety of other microbial eukaryotes. Like hydrogenosomes, mitosomes have distinct origins in diverse eukaryotic protists that seem to be the result of the parallel loss of many typical aerobic mitochondrial pathways. Although the exact functional roles of mitosomes are still unclear, they are thought to consume (not produce) ATP (Tsaousis *et al*. 2008) and, as a shared common function, carry out iron-sulfur cluster assembly (Roger and Silberman 2002; Goldberg *et al*. 2008), although for some taxa like *Entamoeba histolytica*, this latter function is still debated (Tsaousis *et al.* 2012).

As hallmarks of their endosymbiotic origin, classical mitochondria contain genomes of their own encoding genes that are phylogenetically most closely related to homologs in the α-proteobacteria, although gene content varies between the organelles of different lineages (Gray *et al*. 1998; Marande and Burger, 2007). Over evolutionary time, different genes have been transferred from the mitochondrial genome to the host nucleus, and their products were then re-targeted back to the organelle in a process known as Endosymbiotic Gene Transfer (EGT) (Timmis *et al*, 2004). In the case of hydrogenosomes and mitosomes where their genomes have likely been lost entirely, all essential organellar genes must be located in the host genome. However, the mitochondrion-related organelles (MROs) of *Nyctotherus* (Boxma *et al*. 2005) and *Blastocystis* (Stechmann *et al*. 2008) fall somewhere between classical mitochondria and hydrogenosomes. In both of these organisms, MROs retain their genomes, and are predicted to house part of the tricarboxylic acid (TCA) cycle, parts of the electron transport chain as well as a variety of other classical mitochondrial pathways. Curiously, both also possess a [FeFe] hydrogenase enzyme and other proteins typical of hydrogenosomes. These data, and the ongoing characterization in a variety of previously-unstudied anaerobic protists is beginning to reveal that MROs fall along a continuum of biochemical function (van der Giezen and Tovar, 2005) ranging from classical aerobically-respiring mitochondria to the mitosome. MROs, such as those found in *Blastocystis* sp. (Stechmann *et al*. 2008), *Mastigamoeba balamuthi* (Gill *et al*. 2007) and *Trimastix pyriformis* (Hampl *et al*. 2008) blur the distinctions between mitochondria, hydrogenosomes and mitosomes and illustrate this continuum of diversity.

1.3.2 Protein Import

Mitochondrial proteins may be targeted to the mitochondrial matrix, inner membrane, outer membrane, or the intermembrane space. There are two main import pathways, one mediated by N-terminal targeting sequences (presequences) and the other, known as the carrier protein import pathway, mediated by internal targeting sequences. Recent studies have shown that a significant percentage of mitochondrial proteins do not carry an N-terminal targeting sequence and carry only an internal or C-terminal sequence, if they have a detectable localization sequence at all (Marcotte *et al*. 2000; Sickmann *et al.* 2003; Bolender *et al*. 2008). The N-terminal targeting sequence and internal targeting sequence represent two distinct import pathways (Bolender *et al*. 2008).

In general, N-terminal mitochondrial targeting sequences are enriched in positive, hydrophobic, and hydroxylated amino acids with acidic residues avoided. These N-terminal targeting sequences form ampipathic alpha-helices which are recognized by the import machinery and cleaved off by processing peptidases after import (Pfanner *et al*. 2004; Dolezal *et al*. 2006; Bolender *et al*. 2008). Import requirements have been rigorously examined mostly in a small subset of eukaryotes such as yeast (Geissler *et al*. 2002; Wiedemann *et al*. 2003; Prokisch *et al*. 2004) although there have also been more recent investigations for the *Trichomonas vaginalis* (Bradley *et al.* 1997; Dolezal *et al.* 2005; Mentel *et al.* 2008; Smíd *et al.* 2008) hydrogenosome and the mitosome of *Giardia intestinalis* (Dolezal *et al.* 2005; Smíd *et al.* 2008).

The mitosomal targeting sequences in the human parasite *Giardia intestinalis* are shorter than typical mitochondrial targeting sequences and are lacking in positively charged amino acids (Smíd *et al*. 2008) with the hydrogenosomal targeting sequences of *Trichomonas vaginalis* seeming to be a mix with some short “mitosomal-like” targeting sequences as well as long sequences with motifs more like those of canonical mitochondria (Smíd *et al.* 2008). Existing “state-of-the-art” prediction algorithms tested on these divergent lineages often fail to predict targeting or localization in many putative MRO sequences (Smíd *et al*. 2008). A bioinformatic search for hydrogenosomal proteins in *Trichomonas vaginalis* predicted approximately 127 proteins (Carlton *et al*. 2007) while a more recent proteomic approach, combined with bioinformatic analyses, identified 228 (Rada *et al.* 2011).

The protein import mechanisms of the reduced MROs of *Trichomonas vaginalis*, *Giardia intestinalis*, and other microbial eukaryotes are still only partially understood. Both N-terminal targeting sequences and internal targeting motifs are clearly used (Mentel *et al.* 2008) but may be shorter and have slightly altered physicochemical properties such as lacking distal positively charged residues typically seen in canonical mitochondrial targeting sequences (Smíd *et al.* 2008).

1.3.3 Predicting Subcellular Localization

There are three broad categories of subcellular localization prediction programs: 1) N-terminal signal/targeting peptide detection, 2) full protein sequence feature based, and 3) approaches that use a mixture of the previous strategies in some fashion. N-terminal signal/targeting peptide classification is the most widely used type of subcellular prediction strategy, with the majority of available programs relying on the presence of such a signal exclusively, or using the presence/absence of a signal peptide as one of several kinds of evidence to base predictions on. Machine learning techniques such as Artificial Neural Networks (ANNs), Hidden Markov Models (HMMs), and Support Vector Machines (SVM) have all been used for the task of prediction, along with so-called 'expert systems'. The details of machine learning algorithms are outside of the scope of this work. Briefly, all of these approaches are ‘trained’ on data where targeting has been confirmed experimentally and an attempt is made to optimize the discriminatory power of classification based on that training data. In many cases, training data is split into training and testing data sets, so that performance of the trained method can be evaluated using separate test data sets. For all of these approaches, features of the input data are extracted and then clustered in some way into one or more groups, such as subcellular location, based on specified criteria. In some cases, several of these classifiers are used in multiple layers. Individual classifiers work on one component of a larger problem and a final 'master' classifier combines the outputs to make a global classification. For a more detailed review of these algorithms, see (Rabiner 1989; Gurney 1997; Cristianini and Shawe-Taylor 2000; Kecman 2001).

Feature based prediction programs encompasses a broad range of strategies that may be employed either singly or collectively. These programs may use information that can be obtained directly from the protein sequence entry in a database, such as Gene Ontology (GO) terms, sequence annotation, and structural data. Where this kind of information is not available, these methods may attempt to assign data such as GO terms via homology and make predictions about secondary structure. Most of these classifiers also make use of data that can be extracted from the sequence alone: amino acid composition, di- or tri-peptide compositions, presence/absence of particular domains, secondary structure, hydrophobicity, or the predicted isoelectric point. As well as direct sequence features, this class of predictors also encompasses sequence comparison methods that use homology or phylogenetic data to make classifications. Homology-based prediction programs are limited by the presence of sequence similar homologs in the databases that are searched. Although the existence of known homologs certainly improves the prediction accuracy, examples from the full mass-spectrometry based characterization of the *Tetrahymena thermophila* mitochondrion (Smith *et al*. 2007) have demonstrated that many hypothetical MRO-targeted proteins of unknown function could be organism specific or be phylogenetically restricted to a narrow group of eukaryotes. Sampling of organellar diversity across a wider array of taxa may reduce the number of lineage specific proteins found with unknown homologs, expanding the biochemical repertoire of mitochondria and increasing our understanding of the properties of those proteins.

’Mixed’ approaches typically use several different types of protein sequence feature-based classifiers and are frequently coupled with N-terminal signal peptide and cleavage site prediction strategies. The results are combined, usually by use of a voting system or master classifier, into a final predicted subcellular location. This mixed classifier strategy is used in the programs SherLoc/SherLoc2 (Shatkay *et al.* 2007; Briesemeister *et al.* 2009), MitoLoc/MitoLoc2 (Höglund *et al.* 2006; Blum *et al.* 2009), MITOPRED (Guda *et al*. 2004a; Guda *et al*. 2004b), PA-SUB (Lu *et al*. 2004), the PSORT family of programs (Nakai and Horton 1999; Bannai *et al*. 2002; Horton *et al*. 2006; Horton *et al*. 2007), and others. These mixed approaches are useful for overcoming the shortcomings of individual localization programs, significantly improving performance as shown recently with yimLOC (Shen and Burger 2007). Integrating the results from separate programs can be a challenge, especially if missing data from some programs is to be allowed. Flexibility is often preferred to minimize the impact of the limitations of individual classifiers. If, for instance, the input data is a large number of expressed sequence tags then sequences that are missing N-terminal sequences that could contain targeting peptides should not be removed from consideration if some of the classifiers that are part of the 'mixed' predictor do not require this information to make a prediction.

Gene Ontology (GO) terms, when available, have also proven useful for classification (Chou & Shen 2007, 2010; Huang *et al.* 2008; Blum *et al.* 2009; Mei *et al.* 2011). For unknown query sequences, assignment of GO terms suffers from some of the same problems as assignment of subcellular localization itself, and is often accomplished with simple BLAST-based searching approaches (Chou & Shen, 2010; Chou & Shen, 2007; Huang *et al*. 2008), although newer machine-learning (Mei *et al.* 2011) and phylogenetic (Gaudet *et al*. 2011) approaches also exist. Gene network (Tung & Lee 2009) and protein-interaction (Shin *et al*. 2009) data have also been used to predict subcellular localization but robust experimental data of this kind is only available in already well characterized model organisms and will not necessarily be well conserved across broader swathes of eukaryotic diversity.

In addition to the broad types of classification methods discussed above, classifiers can also fall into two different general categories of specialist predictors, or predictors that handle multiple locations. Specialist classifiers are trained and designed to predict localization to only one compartment, such as the mitochondrion, chloroplast, or secretory system whereas multi-location “general” classifiers prediction localization to the various compartments in a single step. In the latter case, the classifier may or may not attempt to deal with ‘multiplex’ proteins, those that are targeted to more than one subcellular localization. This dual-targeting, or “eclipsed distribution” (Regev-Rudzki & Pines 2007) may be uneven with minor and major localizations. Dual-targeting in photosynthetic eukaryotes often involves proteins of endosymbiotic origin, which have gained N-terminal targeting sequences with dual-specificity to the two kinds of primary endosymbiont-derived organelles: chloroplasts and mitochondria (Carrie *et al*. 2009). Other proteins have been characterized that undergo a process of reverse translocation (Ben-Menachem, et al. 2011) whereby, after import into the mitochondrion and subsequent cleavage of the N-terminal targeting peptide, a fraction of the imported proteins are translocated back across the mitochondrial membrane system to the cytosol. These complex targeting dynamics may complicate predictions of subcellular localization, particularly in the case of predictors that handle multiple locations and do not allow for multiple locations to be genuinely assigned (Chou & Shen 2010; Chou *et al*. 2011; Chou & Shen 2007).

The datasets used to train many prediction programs (e.g.the datasets for the training of MultiLoc (Höglund *et al*. 2006) or BaCelLo (Pierleoni *et al.* 2006)) are often restricted to sequences from animals, plants, and Fungi and ignore the bulk of eukaryotic diversity. Additionally, sampling bias within these groups, as well as a historical and ongoing experimentation bias in terms of localization and in-depth molecular characterization leaves the data skewed towards the multicellular members of these two groups. The remaining diversity of almost exclusively microbial eukaryotes includes anaerobic and microaerophilic members with reduced MROs like hydrogenosomes and mitosomes. Key enzymes in their unique metabolic pathways are not part of typical subcellular localization training datasets but are of interest for identification in large-scale transcriptome and genome projects of microbial eukaryotes. There have been some exceptions to this trend, for instance, the larger training datasets used for Euk-mPLoc (Chou & Shen, 2007) and Euk-mPLoc 2.0 (Chou & Shen, 2010) cover eukaryotic diversity and include the hydrogenosome as one of 22 distinct subcellular localizations. However, including hydrogenosomes as a distinct localization from mitochondria, and relying on Uniprot/Swiss-Prot to distinguish the two consistently, may have a negative impact on performance (See Chapter 2 for more details).

In order to build high-throughput pipelines and software tools for phylogenetic based analysis of protein function, it is important to maximize flexibility and modularity to more easily allow inclusion of new methods of alignment and phylogenetic reconstruction as they are developed. This ‘method-agnostic’ approach extends the usable shelf-life of scientific software so that they are not quickly outdated by rapid advances in core methods. In addition, because all phylogenetic methods can be potentially misled by systematic error, methodological artifacts, or model violation, it is important to explore (and quantify where possible) the types of phylogenetic conditions where function prediction can also go wrong.

1.4 Overview of Thesis Chapters

This thesis explores the application of phylogenomic methods to the prediction of subcellular protein localization and of functional divergence in protein families. These topics are split over three data chapters in addition to a final discussion.

In Chapter 2, two novel methods for the prediction of subcellular localization are presented. The first, Comparative BLAST for Organelles (CBOrg) is a non-phylogenomic method that uses a BLAST-based scoring system to rapidly screen sequences at genomic and transcriptomic level scales. The second, PhyloPred-HMM is a full phylogenomic method that uses a background database of sequences clustered *de novo* at the protein family level. It makes use of Hidden Markov Model (HMM) based search strategies and phylogenetic distance measured to annotate query sequences with their most probable subcellular location with special attention paid to diverse microbial eukaryotes not normally included in the training datasets of other classifiers.

Chapter 3 describes FunDi, a maximum-likelihood based phylogenetic method for predicting the protein residues in a multiple sequence alignment undergoing functional divergence between two or more groups in a protein family. The performance of FunDi is compared to that of several other prediction programs on several biological datasets, and two large simulated datasets, constructed using two different methods. These simulated datasets also allow aspects of phylogenetic trees (e.g. size, shape, and tree length) that affect classification performance to be analyzed.

Finally, in Chapter 4, FunDi and two other functional divergence prediction programs are used to characterize functional divergence and convergent evolution in the plastid-targeted glyceralhdehyde-3-phosphate dehydrogenase (GAPDH) enzyme family of two eukaryotic lineages: the Archeaplastida and the Chromalveolata. Predictions of functional divergence, especially of potentially convergent evolution to function in an NADPH-dependent manner in the Calvin cycle, are discussed within the context of the structure, function, and regulation of GAPDH.

1.5 Author Contributions

For sections previously published DG prepared manuscript, ADT and AJR provided editorial feedback. All authors agreed on final manuscript submission.

Chapter 2 Prediction of Subcellular Localization

This chapter (sections 2.1.1, 2.2.1, Figure 2.1, Table 2.1) contains material originally published in:

“Gaston D, Tsaousis AD, Roger AJ. 2009. Predicting Proteomes of Mitochondria and Related Organelles from Genomic and Expressed Sequence Tag Data. Methods in Enzymology. Mitochondrial Function, Part B: Mitochondrial Protein Kinases, Protein Phosphatases and Mitochondrial Diseases. 457:21-47”

2.1 Introduction

Prediction of subcellular localization remains a significant challenge in bioinformatics because of the complexity of subcellular targeting pathways in a given organism as well as the diversity of these pathways across the tree of eukaryotes. Although complex machine-learning algorithms trained on a variety of sequence features seem to be the most common approach taken to predict subcellular localization, N-terminal targeting sequence predictors remain the most widely used tools for this purpose. This is despite the fact that it is now well known that there are multiple organellar import pathways, some of which do not involve N-terminal targeting sequences (Pfanner & Geissler 2001) but instead employ ‘cryptic’ internal targeting sequences or C-terminal targeting sequences to direct proteins to the mitochondrial (or MRO) compartment. A second general problem for *in silico* MRO proteome prediction arises from the fact that training sets are often comprised of relatively narrow subsets of data from Uniprot (Jain *et al.* 2009; Consortium 2011) and are generally restricted to sequences from well-studied model organisms or groups, such as animals, Fungi, and plants, although there have been some exceptions (Gschloessl *et al.* 2008; Danne & Waller 2011; Delage *et al.* 2011). More recently, in-depth experimental determinations of MRO proteomes, combined with bioinformatic analyses, have given us a more robust picture of the diversity of protein content of the MROs of several microbial eukaryotes including *Tetrahymena thermophila* (Smith *et al.* 2007), *Trichomonas vaginalis* (Rada *et al.* 2011), *Entamoeba histolytica* (Mi-ichi *et al.* 2009), *Chlamydomonas reinhardtii* (Atteia *et al.* 2009), and *Giardia intestinalis* (Jedelský *et al.* 2011). This wealth of data from functionally and taxonomically diverse organelles makes phylogenomic approaches to subcellular localization prediction much more attractive.

As discussed in 1.2.1 and 1.2.2, there are a number of MROs of diverse metabolic function and proteomic content. These range from anaerobic organelles that use an alternative terminal electron acceptor instead of oxygen to hydrogenosomes and mitosomes. Hydrogenosomes and mitosomes in particular tend to be reduced in proteomic content compared to canonical mitochondria. However, hydrogenosomes also contain a number of unique metabolic pathways and alternative enzymes involved in anaerobic ATP synthesis that were acquired by their host organism by lateral gene transfer from anaerobic bacteria; these enzymes were not likely present in the original mitochondrial endosymbiont (Hug *et al.* 2010; Hampl *et al.* 2011; Stairs *et al.* 2011; Tsaousis *et al.* 2012). Mitosomes are highly reduced MROs, and in parasites such as *Giardia intestinalis,* proteins imported into these organelles via the N-terminal targeting peptide import pathway have shorter targeting peptide sequences with slightly altered physicochemical properties (Smíd *et al.* 2008). *Trichomonas vaginalis*, which possesses hydrogenosomes, features a mix of sequences some of which have more 'mitosomal-like' short N-terminal targeting peptides and others with longer canonical mitochondrial-type N-terminal targeting sequences (Smíd *et al.* 2008). These diverse MROs of anaerobic microbial eukaryotes contain sequences that are not typically included in the training sets of existing subcellular localization classifiers and thus those that do possess N-terminal targeting sequences may not properly be detected by these methods.

Existing prediction methods, often employing machine-learning algorithms and/or N-terminal targeting sequence detection, make use of several different characteristic features of proteins to predict localization, including homology. Before introducing the two novel approaches to MRO proteome prediction that we have developed, a number of the most widely used tools for *in silico* MRO proteome prediction are reviewed.

2.1.1 Existing Prediction Methods

2.1.1.1 MitoProt

MitoProt is one of the earliest-developed and most widely used programs designed to predict the presence of an N-terminal mitochondrial targeting sequence on a protein (Claros and Vincens 1996). MitoProt considers the amino acid composition of the N-terminal portion of the amino acid sequence evaluating the hydrophobic character, net charge, and number of acidic residues as well as trying to detect a cleavage site. In total, 47 individual parameters based on these general characteristics are used to calculate the final score of any given sequence using linear discriminant analysis.

2.1.1.2 TargetP

Like MitoProt, TargetP also predicts N-terminal targeting sequences; however, TargetP is also capable of predicting chloroplast and secretory system targeting signals which are also typically N-terminal sequences (Nielsen *et al.* 1997 and Emanuelsson *et al.* 2000). The most recent version of TargetP (Emanuelsson *et al.* 2000) features a dual-layer neural network, with the first layer comprised of a dedicated network for the prediction of each subcellular localization, while a second layer integrates the output of the first layer to make a final prediction.

2.1.1.3 PSORT

PSORT was first developed in 1990 and has since been updated to produce a family of variants such as WoLF-PSORT (Horton *et al.* 2006; Horton *et al.* 2007), PSORT II (Nakai and Horton, 1999), and iPSORT (Bannai *et al.* 2002). WoLF-PSORT is based on PSORT II but optimized for eukaryotic sequences, and includes features for N-terminal sequence prediction from iPSORT (below). PSORT II uses a wide variety of sequence features including the presence of any sort of N-terminal targeting sequence as well specific signals for different subcellular locations. It also includes predictors for RNA/DNA and actin binding motifs, transmembrane helices, secondary structure elements, and dozens of other specific sequence composition features. A k-nearest neighbor machine-learning classifier is used to predict the final subcellular location. The iPSORT variant uses a decision list architecture to classify sequences possessing a signal peptide, mitochondrial targeting sequence, or chloroplast targeting sequence. Sequences that possess none of these are classified as 'others'. The nodes of the decision tree perform rule-based decision making with the first node being for the presence or absence of a signal peptide based on an amino acid indexing method. Sequences that lack a signal peptide are further checked for mitochondrial or chloroplast targeting sequences, again based on amino acid indexing as well as the overall amino acid pattern at the N-terminal end. If the source organism possibly contains a chloroplast then sequences with a targeting sequence are further subjected to a similar set of rules to discriminate between chloroplast and mitochondrial targeting. Amino acid indexing is a method of converting an amino acid character into a set of numeric values based on physico-chemical properties of the amino acid. Substrings can then be averaged for these properties and to be classified must meet or exceed some given threshold.

2.1.1.4 Predotar

Predotar (Prediction of Organelle Targeting Sequences) employs a neural network to detect and classify N-terminal targeting sequences (Small *et al.* 2004). The neural network operates on three basic parameters: 1) the charge of the amino acid side chain for each residue in the putative N-terminal signal/targeting peptide region, 2) the hydrophobicity score for each amino acid in this same region, and 3) the amino acid composition in the two halves of the sequence.

2.1.1.5 MITOPRED

MITOPRED is a mixed classifier based on the presence (and occurrence pattern) of ‘subcellular location-specific’ Pfam domains, amino acid composition, and isoelectric point (pI) (Guda *et al.* 2004a; Guda *et al.* 2004b). Query sequences are compared to each subcellular location and scores are calculated individually for the amino acid composition/pI and the Pfam domain occurrence.

2.1.1.6 PProwler

PProwler (Bodén and Hawkins 2005; Hawkins and Bodén, 2006) is a prediction program based on TargetP but, unlike TargetP, prediction of an N-terminal targeting sequence is not restricted to a simple linear analysis of the N-terminal region of the input protein. PProwler examines the N-terminal region using a non-linear recurrent neural network. This recurrent network is structured such that the classification of any given residue as belonging to a signal peptide depends on the state of the amino acid residues immediately up- and downstream. These residues, in turn, are dependent on the state of the residues up- and downstream of them, and so on in a recursive pattern. This recursion is designed to overcome limitations of looking at the linear sequence alone, such as longer-range interactions important in the three-dimensional structure such as the amphipathic helix of the signal peptide. A unique neural network exists for each subcellular location (mitochondria, chloroplast, secretory system, and other) each window in the N-terminal 100 amino acids is classified, by each network, using a sliding window approach. The output of each location-specific neural network becomes the input to a final Support Vector Machine (SVM).

2.1.1.7 PA-SUB

PA-SUB (Lu *et al.* 2004) is a collection of five different classifiers for sub-cellular localization. Each classifier is trained and designed to classify proteins from a particular group of organisms: animals, plants, fungi, Gram-negative bacteria and Gram-positive bacteria. The authors of this program report a prediction accuracy of 81% for fungi and 92-94% for the other categories. PA-SUB performs classification based on presence/absence of annotated features in the SwissProt entries of the top-scoring BLAST hits to the query sequence. PA-SUB is unable to classify sequences if no homologous sequences can be found in SwissProt or if no suitable feature annotations can be found in the SwissProt entries of homologs.

2.1.1.8 yimLOC

This classifier uses a different approach to those described above. Instead of designing new predictors based on sequence features, yimLOC uses a ‘mixture-of-experts’ approach. In this strategy, the results of several different classifiers are combined and analyzed by one global 'expert' classifier to determine the overall prediction, in yimLOC’s case by using a decision-tree. The programs integrated into predictions made by the online version of yimLOCm are: SubLoc, pTARGET, SherLoc, CELLO, PA-SUB, TargetP, Predotar, PProwler, SOSUI, MitoProt, Phobins, and TMHMM (Transmembrane HMM). The programs that detect N-terminal mitochondrial targeting sequences are combined into one decision tree, with the output of that tree combined with the outputs of the other programs (based on sequence features). To improve prediction of mitochondrial membrane proteins, which are typically more poorly predicted than matrix proteins, several tools for the prediction of transmembrane domains were integrated into the decision trees (e.g TMHMM).

2.1.1.9 CELLO

CELLO (subCELlular LOcalization) is a prediction program originally designed for use on Gram-negative bacterial sequences but has since been extended for use on sequences from both Gram-positive bacteria and eukaryotes (Yu *et al.* 2006). This program uses a SVM on several n-peptide composition vectors. Where the n = 1 vector is the amino acid composition of the entire sequence, the n = 2 is the composition of di-peptides in the sequence, *etc*. CELLO uses two SVM layers; the first layer is built of classifiers that operate on individual sequence composition vectors while the second layer acts as a master classifier.

2.1.1.10 SherLoc, MultiLoc, and MultiLoc2

SherLoc (Höglund *et al.* 2006a; Shatkay *et al.* 2007) implements a novel mixture based approach that combines both sequence features, such as amino acid composition, and text features from the literature in order to predict the subcellular location; classification is performed using an SVM. Instead of using location references directly from the database entry, the authors of SherLoc designed it to use textual references extracted from titles and abstracts associated with the database entry. SherLoc is an example of a multi-tiered classifier with four sequence-based classifiers and one text-based classifier; the predictions of these individual classifiers can then be integrated into a final result. The sequence-based classifiers are based on those in MultiLoc (Höglund *et al.* 2006b) and are each based on a different sequence feature: amino acid composition, N-terminal targeting sequence, internal anchor sequences, and sorting sequence motifs. The text-based classifier operates on data extracted from the titles and abstracts on Pubmed from the Swiss-Prot entry for the query sequence. For query sequences without a Swiss-Prot entry, an attempt is made to find close homologs in the database and use the textual information associated with those entries. If no PubMed entries can be found, the text classifier is not used for that query, with the prediction then based only upon the four sequence feature classifiers.

MultiLoc2 (Blum *et al.* 2009) added two additional sub-classifiers to the original MultiLoc system, called PhyloLoc and GO-Loc. PhyloLoc is a profile of presence/absence of a protein in a set of 78 completely sequenced genomes. The phylogenetic profile is thus composed of the ratio of the best-hit and self-hit bit scores based on Basic Local Alignment Search Tool (BLAST) matches. The Go-Loc classifier uses Gene Ontology (GO) terms. Each of these sub-classifiers, and the ones already present in MultiLoc, make a prediction of subcellular localization which is then fed to the master SVM classifier for overall prediction.

2.1.1.11 PredSL

PredSL (Petsalaki *et al.* 2006) is another N-terminal targeting sequence based classifier, specifically optimized for eukaryotic sequences. PredSL was trained on eukaryotic sequences from release 3.5 of Uniprot and, unlike most of the other tools described above, is not taxonomically restricted. It uses a mix of layered neural networks, Markov chains, scoring matrices, and HMMs for prediction in a manner similar to many other ensemble based classifiers.

All of the classifiers described above, with the exception of yimLOC, tend to use similar sequence features, including (or exclusively), the presence of an N-terminal targeting sequence. The major differences between these methods lie in subtle differences in the training sets (often just different releases of Uniprot/Swiss-Prot) and slightly different machine-learning algorithms or procedures applied. Approaches that include some type of homology or “phylogenetic” approach as part of their classification strategy (e.g. MultiLoc2) do so using BLAST, or Psi-BLAST and presence/absence of a sequence in taxa but lack a robust phylogenomic approach. While BLAST results can act as a proxy for phylogenetic inference, it fails to adequately model the evolutionary process and place sequences within the context of their shared evolutionary history. Indeed, the best-scoring BLAST matches are not always the most closely related sequences (Koski & Golding 2001).

2.1.2 New Classifiers Introduced in this Work

In this chapter, we describe two new programs for the prediction of subcellular localization: CBOrg (Comparative BLAST for Organelles) and PhyloPred-HMM (Phylogenetic Prediction of subcellular localization with Hidden Markov Models). These new programs were developed to address several shortcomings in existing prediction methods. The majority of prediction programs available operate only on complete protein sequences, especially those that perform N-terminal localization sequence based predictions. Additionally, many prediction programs are not suitable for performing high-throughput predictions at the level of transcriptome or genome sequencing projects. Finally those programs that do use homology information to make predictions do not use phylogenetics and rely on simple BLAST-based scoring systems and use training data from a limited taxonomic subset of eukaryotic diversity. CBOrg relies on organellar proteome and “subtractive” (whole nuclear genome-encoded cellular proteomes minus organellar) databases of several organisms and uses BLAST-based analyses allowing the similarity profiles of sequences to be rapidly screened to determine whether or not their closest homologs are organellar. This approach is especially useful for screening genes in partial transcriptomic and genomic data, where complete gene sequences are rarer and thus the characteristics of the N-termini of encoded proteins cannot always be determined. The second method that I introduce, PhyloPred-HMM, provides a more robust methodology for the prediction of subcellular localization using the 'phylogenomic method' (Eisen, 1998; Eisen, *et al*. 1997) and a large backing/training database of HMMs based on MRO and non-MRO sequences that are sampled broadly from eukaryotic diversity.

2.2 Materials and Methods

2.2.1 CBOrg

CBOrg uses simple comparative BLAST (Altschul *et al.* 1997) searches of query sequences (nucleotide or amino acid) against a user-defined set of organellar and subtractive proteome databases from several organisms to predict potential subcellular localization. CBOrg takes as input a set of sequences of interest in FASTA format and outputs a list of putatively targeted sequences at various confidence thresholds. Our initial version of CBOrg (v1.0) (Gill et al. 2007; Stechmann et al. 2008) contained cellular and mitochondrial proteomes from Human, Mouse, *Arabidopsis*, Yeast, *Trichomonas vaginalis* (Carlton *et al*. 2007), and *Tetrahymena thermophila* (Smith *et al.* 2007). We later later updated with a more recent experimentally derived hydrogenosomal proteome of *Trichomonas vaginalis* (Rada *et al.* 2011), the mitochondrial proteome of *Chlamydomonas reinhardtii* (Atteia *et al.* 2009), the mitosome of *Entamoeba histolytica* (Mi-ichi *et al.* 2009), and the mitosome of *Giardia intestinalis* (Jedelský *et al.* 2011) along with the accompanying predicted cellular proteomes based on full genome sequences for each of these organisms. Comparisons on the larger datasets, and whole proteomes, were made using this latest version of CBOrg.

In both versions query sequences are searched against the mitochondrial and 'subtractive' cellular proteome databases with BLAST, and the best aligning sequence (henceforth referred to as the 'best hit') is returned for each organism represented in the total set of databases. If the best hit against the organellar database has a better score than the best hit against the subtractive cellular proteome database, the input sequence is classified as ‘organelle localized’ for that organism. In its most liberal ‘screening’ mode, CBOrg only requires a best hit to the organellar proteome of a single organism to be classified as putatively organelle localized. Hits can be evaluated based solely on the raw BLAST alignment score or the expect-value (e-value). To handle a wide range of input sequence types, including clustered expressed sequence tags (ESTs) from transcriptome projects (nucleotide), CBOrg can do BLAST-based comparisons using amino acid sequence queries against protein sequence databases (blastp), six-frame translated nucleotide sequences against protein sequence databases (blastx), or six-frame translated nucleotide sequences against six-frame translated nucleotide sequence databases (tblastx). In this chapter, and with the default source proteomes and genomes, only the blastp and blastx options are used. **Figure 2.1** depicts a flow-chart summarizing the CBOrg method.

2.2.2 PhyloPred-HMM

PhyloPred-HMM is a phylogeny-based method developed for the prediction of MRO proteins that can be extended to any organelle with the creation of appropriate databases. PhyloPred-HMM employs a two-stage prediction process. Each sequence is first treated as a query and searched with hmmsearch from the HMMER3 package (Eddy 1998, 2011) against a database of HMM profiles generated from alignments of homologous sequences clustered approximately at the superfamily level (see Database Construction). The top scoring profile match for each sequence (if below a user-defined e-value threshold (Default: 10)) is selected as the assigned superfamily/sequence cluster and the sequence is added to a seed alignment using the cluster's HMM profile with HMMER3 (Eddy 1998; 2011). The resulting multiple sequence alignments are then automatically trimmed using a custom in-house alignment masking algorithm (described in detail below). A maximum-likelihood phylogenetic tree is then estimated using FastTree2 (Price *et al*. 2009; 2010) and subcellular localization information is predicted for the sequence of interest from annotation information associated with known sequences in the cluster. If all members of the cluster have the same subcellular localization annotation (‘non-mixed’ clusters), then the novel sequence is assigned this annotation. Otherwise, if the cluster contains both organellar and non-organellar sequences the annotation of the novel sequence is estimated by using one of three possible phylogenetic distance metrics (described in detail below). Figure 2.2 graphically summarizes the PhyloPred-HMM method. Sequences that cannot be assigned to a cluster are compared to the unclustered sequences from the dataset using phmmer, a search algorithm from the HMMER3 software package. Annotations for these sequences are then based on the annotation of their best matches in the phmmer search. Finally, remaining sequences, which have no hit using hmmsearch or phmmer default significance cut-offs, cannot be classified and are annotated as non-MRO sequences.

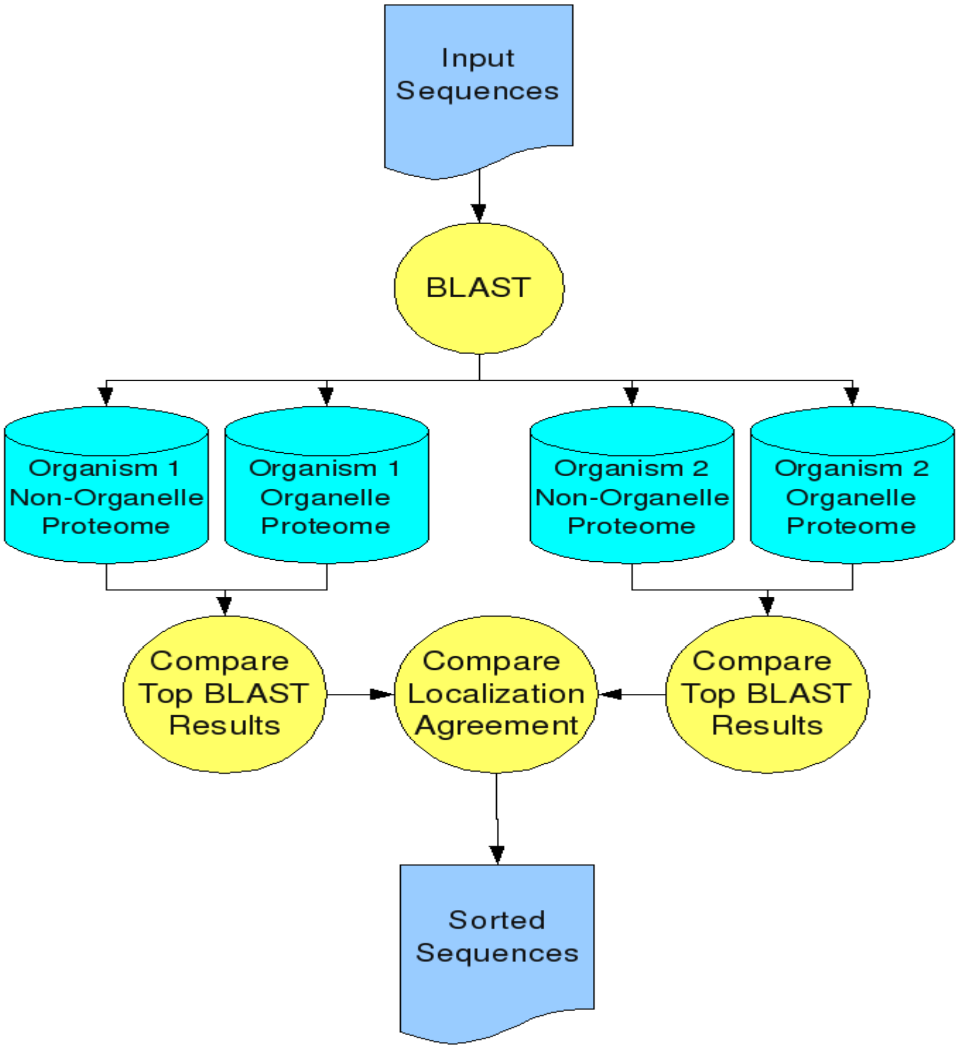

Input Sequences
BLAST
Organism 1 Non-Organelle Proteome
Organism 1 Organelle Proteome
Organism 2 Non-Organelle Proteome
Organism 2 Organelle Proteome
Compare Top BLAST Results
Compare Localization Agreement
Compare Top BLAST Results
Sorted Sequences

FIGURE 2.1 A SCHEMATIC REPRESENTATION OF THE COMPARATIVE BLAST FOR ORGANELLES (CBORG) METHOD; non-organelle proteome refers to the set of all nucleus-encoded proteins with proteins from the organellar (mitochondrial) proteome removed. Top scoring results from BLAST searches are compared between databases for localization within that organism. Localization data is then compared between organisms in order to sort input sequences. Where best-hit localizations differ amongst databases, input sequences are sorted based on the level of agreement between the results from the different databases. The level of agreement required to determine the annotation is an adjustable variable of the method.

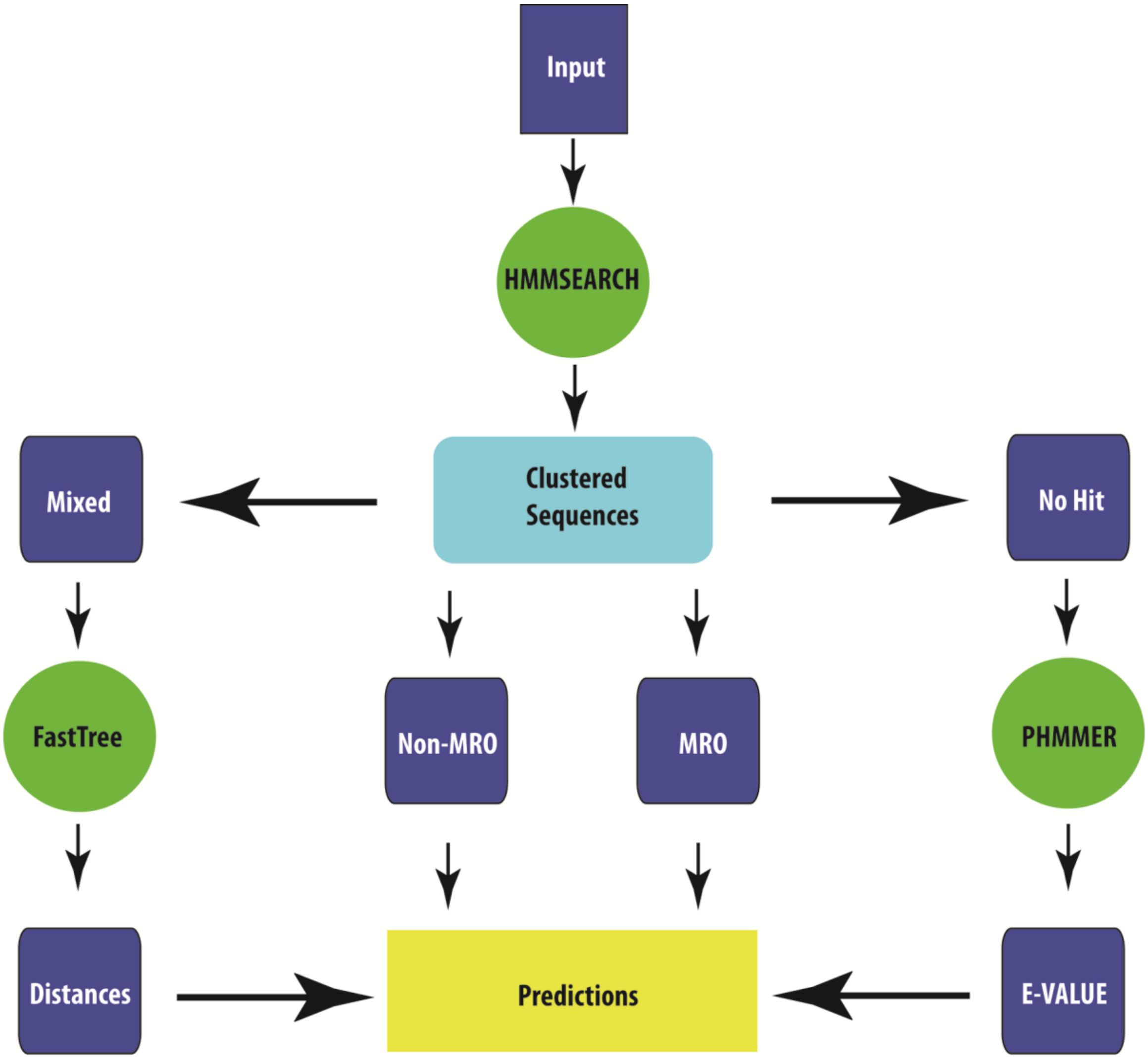


FIGURE 2.2 A SCHEMATIC REPRESENTATION OF THE PHYLOPRED-HMM METHOD. Query sequences are compared against pre-generated protein family clusters using hmmsearch. Sequences with no significant hit to a cluster are compared to unclustered sequences using phmmer. Sequences with hits to mixed-member clusters (i.e. clusters with both organellar and non-organellar members) are automatically aligned and included in phylogenetic trees. E-values, phylogenetic distances, and cluster membership are used to predict localization.

2.2.3 Alignment Masking

Although software tools such as GBlocks (Castresana, 2000; Talavera & Castresana, 2007) exist for trimming ambiguous regions of multiple alignments, they are based on arbitrary metrics and cutoff criteria and do not edit alignments in a similar way to the human expert. Using robust statistical confidence scores based on a modeling approach and methods used for constructing alignments (such as the seed HMMs used with HMMER3) should provide a superior alternative . To this end, we devised a multiple alignment masking tool, AliMask-CS (Alignment Masking with Confidence Scores), that takes confidence scores for the columns of a multiple sequence alignment as output by HMMER3, FSA (Bradley *et al.* 2009), GUIDANCE (Penn *et al.* 2010), or other tools and evaluates them along with the percentage of gap characters at a site to determine whether a site should be kept or removed from the final masked alignment. A sliding window is used to calculate the weighted average of column confidence scores and its size can be adjusted by the end user along with all relevant thresholds for the inclusion of columns in the final alignment. Informal testing of this method suggested that it could be adjusted to more closely match masks generated by human inspection (data not shown). For the following analyses, HMMER3 scores were used with a sliding window of size seven. Sites kept in the final alignment had a percentage of gap characters less than 70%, a raw column confidence value greater than or equal to six, and a sliding window average score of at least eight. This high-throughput, automated alignment masking tool is similar to TrimAL (Capella-Gutiérrez *et al*. 2009) in its use of a sliding window calculation (if desired by the user), and the fact that TrimAL's similarity and consistency scores should be positively correlated with the statistical support values output by FSA (Fast Statistical Alignment) and HMMER3. However, because HMMs generated from pre-aligned clusters were used for the final alignments in this study, and the statistical support value output by HMM measures the fit of the alignment to the HMM itself, the AliMask-CS method based on these values was employed instead of TrimAL.

2.2.4 Assigning Localization Annotation Using Phylogenetic Distances

While more complex methods, such as ancestral state reconstruction, can be used to assign annotations associated with leaves (sequences) on phylogenetic trees to unknown sequences, here we propose several simpler methods based on phylogenetic distances. These methods have the advantage that they are very fast and they can be easily and quickly calculated when the number of query sequences, and thus the number of phylogenetic trees to be reconstructed, may be large (i.e. hundreds to thousands of alignments). Phylogenetic distance, in this chapter, is defined as the tip-to-tip sum of branch lengths separating two sequences on a phylogenetic tree, a metric that is also known as the patristic distance between terminal nodes. Three different phylogenetic distance-based measures were used in this work. Given a phylogenetic tree with N taxa (T_i’s), one of which corresponds to an unannotated sequence of interest (T_1) we can calculate the average tip-to-tip phylogenetic distance between T_1 and all MRO and non-MRO T_i's. This measure is the shortest-average-distance (SAD) method. Highly divergent sequences, which result in abnormally long branches on phylogenies compared to the average sequence, may skew the SAD values inappropriately, so a modified form of the SAD method is calculated by first removing outliers (top and bottom 5% of sequences by length from both MRO and non-MRO groups), resulting in a outlier-trimmed shortest-average-distance (T-SAD). A third measure for assignment of localization is to determine the nearest tip-to-tip neighbor of the sequence of interest and assign to the latter the annotation associated with this neighbor (i.e. assign the annotation of the T_i that has the smallest distance to T_1). This latter method will be referred to throughout the chapter as the nearest distance (ND) assignment.

2.2.5 Database Construction

To construct an appropriate source database for PhyloPred-HMM, all Eukaryotic MRO (Subcellular Localization: Mitochondrion, Hydrogenosome, Mitosome) and non-MRO sequences were retrieved from Uniprot release 2011_06 (Jain *et al.* 2009; Consortium

2011) and clustered at the 100% redundancy level. Only entries with a status of reviewed, that is sequences whose annotations were reviewed by a biocurator instead of simply automated predictions, were selected, and any confidence level (experimental, possible, probable, by similarity) for MRO localization was allowed. In addition the annotated MRO and non-MRO sequences from *Tetrahymena thermophila* (Smith *et al.* 2007) and the *Trichomonas vaginalis* genome project (Carlton *et al.* 2007a) were included for an overall total of 13,652 MRO and 147,669 non-MRO sequences (161,321 sequences total). Fragments were not initially excluded from the search parameters in order to increase taxonomic coverage of the sequences that were retrieved, especially for poorly sampled microbial eukaryotic groups.

2.2.5.1 Sequence Clustering and Alignment

Sequences were clustered using several *de novo* clustering algorithms implemented in the Spectral Clustering of Protein Sequences (SCPS) program (Nepusz *et al.* 2010) including: hierarchical clustering (agglomerative, e-value cutoff of 10^{-6}), the Markov Clustering algorithm (MCL) with an inflation parameter of 2 (MCL-Inf2), and MCL with default settings (MCL-Default). MCL, as implemented in SCPS is essentially identical to TribeMCL (Enright *et al.* 2002). The input to each of these clustering algorithms is a table of all-versus-all BLAST e-values (Altschul *et al.* 1997). For MCL, e-values are transformed by SCPS into Euclidean distances prior to clustering (see Nepusz, et al. 2010 for details). SCPS offers a number of possible transformations; the default method was used in this work. Output clusters with two or fewer sequences were separated from the main group as being unsuitable for phylogenetic analysis and placed with sequences that could not be clustered. There were no clusters containing only two sequences that contained both one MRO and one non-MRO sequences. Clustered sequences were aligned using the multi-threaded version of MAFFT 6 (Katoh 2002; Katoh *et al.* 2005; Katoh & Toh 2008, 2010). Because there are no support values for the initial alignments of source clusters, MAFFT (Multiple Alignment Fast Fourier Transform) alignments were trimmed using a gap percentage cutoff (60%) for every aligned column. Trimmed alignments were input into HMMER3 to generate a library of HMM profiles.

2.2.5.2 Expanded and Cleaned Dataset

After initial testing, the Uniprot dataset created as described above was cleaned by removing all sequences longer than 6000 amino acids (80 sequences), a filtering procedure that removed only two mitochondrial sequences (Nesprin-2 in Human and Mouse). Polypeptide fragments that were less than 50% of the average sequence length of the cluster (with the MCL-Inf2 clustering algorithm) were also removed. To improve the taxonomic coverage of microbial eukaryotes, complete predicted proteome sets of sequences for *Giardia intestinalis* (Jedelský *et al.* 2011), *Entamoeba histolytica* (Mi-ichi *et al.* 2009), and *Chlamydomonas reinhardtii* (Atteia *et al.* 2009) were added all of which included sequences with experimentally validated MRO localization annotations. *Giardia intestinalis* sequences were retrieved from genome build 2.5, Assemblage A, at GiardiaDB (Aurrecoechea *et al.* 2009, 2010). In addition, a proteomic survey of the hydrogenosome of *Trichomonas vaginalis* was recently published leading us to update the number of hydrogenosome localized proteins in its predicted proteome from 138 to 228 (Rada *et al.* 2011). These sequences, and the rest of the proteome, were retrieved from build 1.3 at TrichDB (Aurrecoechea *et al.* 2009, 2010). All sequences from the appropriate genera were removed from the initial dataset prior to addition of the new sequences. This enhanced dataset contained 276,262 sequences, 14,871 of which are mitochondrial or MRO-derived.

2.2.6 Testing Predictions

CBOrg was tested using a small dataset of 15 representative sequences from a *Blastocystis* transcriptome project (Stechmann *et al.* 2008); eight of which were MRO localized and seven that were non-MRO (Stechmann et al. 2008; A. Tsaousis and A. J.

Roger, unpublished data). This small dataset was chosen to reflect several MRO localized sequences present both in canonical mitochondria and in reduced hydrogenosomes. A small dataset was also necessary to facilitate comparison against several other classifiers that have web-only interfaces and hence cannot be used to analyze large data sets.

PhyloPred-HMM performance was evaluated with the starting Uniprot dataset using several differing approaches. First, in order to assess the clustering algorithms, annotated sequences from the Uniprot dataset were divided into three non-overlapping sets of randomly selected sequences to form three jack-knifed test sets. Sequences from the appropriate test set were removed from their respective multiple sequence alignments and HMM profiles were re-generated using HMMER3. Query sequences were assigned to clusters using HMMER3, with the same parameters used in the full PhyloPred-HMM method. Additionally, to test our three proposed phylogenetic assignment methods (i.e. SAD, T-SAD and ND), for all mixed-member clusters, we treated each sequence in turn as an unknown query and calculated each of the three distance measures.

To compare PhyloPred-HMM performance to existing programs such as MultiLoc2 (Blum *et al.* 2009), the Höglund reduced homology test dataset was used. This dataset contains 5959 reviewed sequences from release 42 of Uniprot, 510 of which are mitochondrial. All sequences in this dataset have been reviewed and represent proteins from only three eukaryotic groups: the animals, Fungi, and plants. To test this dataset, identical matches (based on the Uniprot ID) were removed from the PhyloPred-HMM sequence set as described above. A second test was done by removing all sequences with more than 80% sequence identity.

Because the majority of subcellular localization prediction methods have been almost exclusively trained and tested on a very narrow range of eukaryotic diversity, we tested PhyloPred-HMM, CBOrg, and MultiLoc2 on the complete proteomes of two microbial eukaryotes: *Trichomonas vaginalis* and *Tetrahymena thermophila*. These organisms possess two very distinct MRO types: *T. thermophila* has canonical mitochondria, albeit with some unique adaptations to its lifestyle (Smith *et al.* 2007), whereas *Trichomonas vaginalis* possesses anaerobically-functioning hydrogenosomes, have reduced proteomes resulting from the loss of many aerobic mitochondrial functions, but that also contain a number of distinct metabolic pathways which function in low oxygen conditions (Carlton *et al.* 2007b; Smíd *et al.* 2008; Rada *et al.* 2011; Schneider *et al.* 2011).

A smaller protist-only test set was constructed for the comparison of the performance of PhyloPred-HMM to CBOrg, MultiLoc2, PredSL, and iPSORT in predicting MRO proteomes in these non-model system eukaryotes. This dataset was constructed by randomly sampling 500 MRO and 5000 non-MRO sequences from the complete proteomes of microbial eukaryotes included in this study including: *Tetrahymena thermophila*, *Trichomonas vaginalis*, *Chlamydomonas reinhardtii*, *Entamoeba histolytica*, and *Giardia intestinalis*. Sequences were randomly chosen to prevent any investigator bias, to ensure that MRO sequences with and without N-terminal targeting peptides would be selected, and to ensure that non-MRO sequences would cover a range of other possible localizations. With respect to the latter point, the inclusion of *Chlamydomonas reinhardtii* sequences could potentially result in some chloroplast targeted proteins being included in the dataset, which represents a potential pitfall for exclusively N-terminus sequence-based classifiers.

2.2.7 Assessing Performance

Several standard statistical measures of performance were used including precision, recall, the F_1 score, the Matthews Correlation Coefficient (MCC) (Baldi *et al.* 2000), and the accuracy (ACC) to compare CBOrg and PhyloPred-HMM to other prediction programs with respect to their true positive (TP), false positive (FP), true negative (TN) and false negative (FN) classification scores. These measures are defined in equations 1-5 below. Note that the recall is also known as the sensitivity or statistical power of a test.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad \text{Eqn. 1}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad \text{Eqn. 2}$$

$$F_1 = 2\ \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision+\ Recall}} \qquad \text{Eqn. 3}$$

$$MCC = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})}{\surd(\mathrm{TP+FP})(\mathrm{TP+FN})(TN+FP)(TN+FN)} \qquad \text{Eqn. 4}$$

$$ACC = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+\ FN}} \qquad \text{Eqn. 5}$$

The precision is the probability that a sequence predicted to be MRO-localized is truly MRO-localized, while the recall is the probability that a truly MRO-localized sequence will be correctly predicted. It is possible to enhance recall at the expense of precision, which results in an increase in the number of false positive predictions. The F_1 score and MCC are two different measures that balance precision and recall scores, acting as overall assessments of performance along both axes. Additionally the MCC is considered to be a more balanced measure of performance, compared to simple percentages or the accuracy especially when the membership of the two classes are very different as is the case here in the MRO protein set which is small relative to the non-MRO set which is much larger (Baldi *et al.* 2000). Balanced in this sense applies to the trade-off between the sensitivity and specificity of a measure. However, the MCC does tend to favour minimizing the number of false positives as it will return a high correlation when the number of false positives is low but the number of true positives is also low (Baldi *et al.* 2000). The MCC returns a value of 1.0 in the case of a perfect predictor, 0 for random, and -1.0 in the case of a perfectly inverse predictor.

2.3 Results

2.3.1 CBOrg

CBOrg's performance was assessed on a set of 15 sequences from a *Blastocystis* transcriptome project (Stechmann *et al.* 2008) and compared against the performance of twelve other classifiers: TargetP, SherLoc, Predotar, MitoProt, CELLO, PProwler, PA-SUB, WoLF-PSORT, PSORT II, iPSORT, Mitopred, and yimLOC. (**Table 2.1**) In this analysis CBOrg was tested in its least stringent "screening" mode which only requires a better hit to the MRO database than non-MRO database for one (or more) organism(s) (note however that CBOrg output also ranks results by the number of organisms where the MRO hit was better than the best non-MRO hit, which can be used as a kind of confidence measure). CBOrg performed comparably to many of the classifiers tested against on this small, limited set of protein sequences but did yield a large number of false positives (3). WoLF-pSORT, yimLOC, and iPSORT also had high performance.

2.3.2 Sequence Clustering

SCPS (Nepusz *et al.* 2010) implements several different *de novo* clustering methods including spectral clustering, hierarchical, the Markov Cluster algorithm (MCL), and connected components analysis. All of these clustering methods can potentially produce

TABLE 2.1 PERFORMANCE OF CBORG AND SEVERAL OTHER SELECTED CLASSIFIERS ON A DATASET OF SEQUENCES FROM A BLASTOCYSTIS TRANSCRIPTOME SEQUENCING PROJECT. Default options were used for all prediction programs through their respective online interfaces. M indicates an MRO localization whole O is for other.

	True MRO Localized							
	[FeFe]-Hydrogenase	PFO	ND4	Serine Dehydrogenase	mIF2-Like	[2Fe-2S] Ferredoxin	mtHSP70	MCP
TargetP	M	M	O	O	O	M	M	N
SherLoc	M	M	O	O	M	M	M	M
Predotar	M	M	O	O	O	M	M	O
MitoProt	M	M	M	M	O	M	M	O
CELLO	O	O	O	M	O	O	M	O
PProwler	M	O	O	O	O	M	M	O
PA-SUB	M	M	M	O	O	M	O	M
WoLF-PSORRT	M	M	M	M	O	M	M	M
PSORTII	M	M	O	M	O	M	M	O
iPSORT	M	M	O	M	O	M	M	O
Mitopred	M	M	O	M	O	O	M	O
yimLOC	M	M	M	M	O	M	M	M
CBOrg	O	M	M	O	M	M	M	M

	True Non-MRO Localized						
	RNAse-L-Inhibitor	6PGD	LCFA-CoA-Ligase	Acyl-CoA-Binding Protein	FGAR Synthase	Hydroxypyruvate Reductase	Sodium/Chloride Transport
TargetP	O	O	O	O	O	O	O
SherLoc	O	O	O	O	O	O	M
Predotar	O	O	O	O	O	O	O
MitoProt	O	O	O	O	O	O	O
CELLO	O	O	M	O	O	O	O
PProwler	O	O	O	O	O	O	O
PA-SUB	M	O	O	O	O	O	O
WoLF-PSORRT	O	O	M	O	O	O	M
PSORTII	O	O	O	O	O	O	O
iPSORT	M	O	O	O	O	O	O
Mitopred	O	O	O	O	O	O	O
yimLOC	O	O	O	O	O	O	O
CBOrg	M	O	M	M	O	O	O

very different clustering results, which can affect predictive accuracy. Additionally each of these clustering methods has different parameters that can be adjusted. For these analyses, two simple widely-used clustering methods were employed: hierarchical and MCL. For the MCL method two different inflation (Inf) parameters were selected: the default setting of 1.2 and a larger inflation parameter of two. This inflation parameter controls the granularity of clustering and can range from a low of one to a high of five (Enright *et al.* 2002). The larger the inflation parameter, the smaller and tighter the clusters will tend to be. **Table 2.2** summarizes the results, in terms of numbers and sizes of clusters, as well as the number of sequences that were not included in clusters with at least three members, for each of the three clustering methods used on the original Uniprot dataset.

TABLE 2.2 A COMPARISON OF THE CLUSTERING METHODS USED TO GROUP EUKARYOTIC SEQUENCES TOGETHER AT THE FAMILY AND SUPERFAMILY LEVEL. Total number of clusters, number of sequences not placed in clusters, and average cluster size are indicated.

Clustering Method	**Number of Clusters**	**Number Unclustered**	**Average Number of Sequences per Cluster**
Hierarchical	10642	29525	12
MCL-Default	5673	9646	26
MCL (Inf = 2)	8890	15708	16

2.3.3 Testing PhyloPred-HMM Predictions

2.3.3.1 Cluster Assignment Accuracy

We used several complementary techniques to evaluate: i) the performance of PhyloPred-HMM as a general framework, ii) the individual clustering methods used, and iii) our proposed phylogenetic distance-based annotation assignment methods. First, three jack-knifed test sets were searched against the HMM profiles of the clustered sequences using HMMER3. Several possible results could be obtained. For sequences that belong to a cluster, HMMER3 can assign the query sequence correctly to a cluster, to an incorrect cluster, or incorrectly to no cluster. Conversely, for query sequences that do not belong to a cluster, HMMER3 can assign the query incorrectly to a cluster or correctly by not returning a significant hit. **Table 2.3** shows the accuracy (calculated using equation 5) for each test set and each clustering method. Cluster assignment accuracy is virtually identical for both of the Markov Clustering Algorithm (MCL) methods and lowest for hierarchical clustering (HClust).

TABLE 2.3 ACCURACY ON THREE JACK-KNIFE TEST SETS FOR ALL THREE CLUSTERING METHODS USED TO CONSTRUCT PHYLOPRED-HMM DATABASES. Best performing results are depicted in bold.

	MCL-Default	**MCL-Inf2**	**HClust**
Test Set 1	**0.94**	**0.94**	0.86
Test Set 2	**0.94**	**0.94**	0.86
Test Set 3	**0.95**	0.94	0.85

To determine whether any biases existed in cluster assignment accuracy with HMM profiles between MRO, non-MRO, or mixed clusters, the accuracy was re-evaluated, but was limited only to sequences that belong to clusters (i.e. excluding unclustered sequences). Sequences that were not clustered by the clustering method, or that were placed in clusters too small for phylogenetic analysis were not included in order to assess only HMMER3's ability to correctly place sequences in to the appropriate cluster. Cluster assignment accuracy is again virtually identical between the two MCL-based clustering methods and lowest for HClust. Performance was virtually identical across test sets (Table 2.4). MCL with default parameters performs slightly better on the MRO and non-MRO only datasets, but MCL with an inflation parameter of two produces slightly better accuracy for mixed clusters. This may be a result of the difference in the number of mixed clusters between the two methods (665 with the default parameters, 728 with an inflation parameter of two) or because the resulting clusters are small and tighter which results in better HMMs, or both.

Table 2.4 Cluster assignment accuracy of sequences using HMM profiles and HMMSearch from HMMER3 package across three jack-knife datasets and three different clustering method and parameter settings. Best performing clustering method in bold for each dataset.

	MRO		
	MCL-Default	**MCL-Inf2**	**Hclust**
Test Set 1	**0.98**	0.95	0.85
Test Set 2	**0.97**	0.96	0.86
Test Set 3	**0.98**	0.97	0.86
Non-MRO			
Test Set 1	**0.95**	0.94	0.88
Test Set 2	**0.95**	0.94	0.88
Test Set 3	**0.95**	0.94	0.88
Mixed			
Test Set 1	0.93	**0.95**	0.73
Test Set 2	0.93	**0.96**	0.73
Test Set 3	0.93	**0.95**	0.73

2.3.3.2 Phylogenetic Distance Measures

To test the hypothesis that, in general, sequences will have the same subcellular localization as their closest homologs the three proposed distance measures: shortest average distance (SAD), trimmed shortest average distance (T-SAD), and nearest distance (ND) were calculated for every sequence in a mixed-member cluster. Localization was assigned by one of these measures for each sequence in turn, treating it as a novel query. The predicted localization could then be classified as true positive (TP), false positive (FP), true negative (TN), or false negative (FN). Positives belong to the MRO class and negatives non-MRO. Performance was assessed by calculating the Precision, Recall, F_1 score, and Matthews Correlation Coefficient (MCC) (**Table 2.5**).

TABLE 2.5 PERFORMANCE OF EACH OF THE THREE PROPOSED PHYLOGENETIC DISTANCE METHODS FOR ANNOTATIONS AS EVALUATED BY PRECISION, RECALL, F1 SCORE, AND MCC. Each measure was evaluated on the mixed clusters that result from each of the three tested clustering methods. The best performing measure in each category, for each clustered set, is indicated in bold.

Clustering	**Distance**	**Precision**	**Recall**	**F_1 Score**	**MCC**
HClust	SAD	0.6157	0.8557	0.7161	0.5734
	T-SAD	0.6107	0.8521	0.7115	0.5660
	ND	**0.8769**	**0.9030**	**0.8898**	**0.8389**
MCL-Inf2	SAD	0.5808	0.9146	0.7104	0.6032
	T-SAD	0.5847	0.9129	0.7128	0.6062
	ND	**0.9112**	**0.9246**	**0.9182**	**0.8880**
MCL-Default	SAD	0.4577	0.9159	0.6104	0.5361
	T-SAD	0.4596	0.9128	0.6114	0.5367
	ND	**0.9274**	**0.9390**	**0.9332**	**0.9176**

The F_1 score and MCC are both widely considered superior to Precision, Recall, or Accuracy alone when assessing the performance of a classifier, particularly when the size of the classification categories are of different sizes as is the case for MRO and Non-MRO sequences (Baldi *et al.* 2000; Carugo 2007). Performance across the three clustering algorithms is highly similar as measured by either the MCC or F_1 score. For a given clustering method performance differs sharply between the SAD/T-SAD measures and the ND method of annotation assignment with the latter generally performing the best. This difference is primarily driven by the precision scores; there are markedly fewer false positive predictions using the ND method for classification. No significant difference was observed between the trimmed (T-SAD) and standard (SAD) forms of the shortest average distance.

2.3.3.3 Comparison on Höglund Dataset

The foregoing results suggest that the PhyloPred-HMM method works reasonably well and validates the use of the proposed phylogenetic distance methods, with the nearest distance (ND) measure showing surprisingly better overall performance. To compare PhyloPred-HMM to several existing prediction methods, such as MultiLoc (Höglund et al., 2006b), we evaluated their performance on the Höglund test dataset of mitochondrial and non-mitochondrial sequences (**Table 2.6**). No sequence in this dataset is more than 80% identical to any other. After removing identical sequences (by Uniprot ID) from the PhyloPred-HMM database, this test set was also used to evaluate the relative performance of PhyloPred-HMM and CBOrg. In the case of CBOrg the human and yeast proteomes were not included as source databases to avoid any potential exact matches. The reported performance values for MultiLoc and PSORT in table 2.6 come from the MultiLoc publication (Höglund et al., 2006b).

TABLE 2.6 PERFORMANCE OF VARIOUS CLASSIFIERS AS MEASURED BY THE PRECISON, RECALL, F1 SCORE, AND MCC ON THE HÖGLUND DATASET. Both the SAD and ND measures are reported for PhyloPred-HMM

	HClust		**MCL-Default**		**MCL-Inf2**		**CBOrg**	**MultiLoc**	**PSORT**
	SAD	ND	SAD	ND	SAD	ND			
Precision	0.73	**0.87**	0.45	**0.79**	0.59	**0.83**	0.43	-	-
Recall	0.92	0.94	0.92	0.94	0.92	0.93	**1.00**	0.70	0.88
F_1	0.81	**0.90**	0.60	0.86	0.72	**0.88**	0.60	-	-
MCC	0.80	**0.89**	0.60	**0.85**	0.71	**0.87**	0.61	0.83	0.58

PhyloPred-HMM was the best performing classifier using this test set, with an MCC score that was slightly higher than that of MultiLoc and much higher, regardless of the clustering method used, than CBOrg or PSORT. In all cases the ND measure outperformed SAD (T-SAD was not reported as it was virtually identical to SAD). CBOrg had perfect recall, all 510 mitochondrial proteins from the dataset were recovered; however, it had low precision (683 false positives), which is reflected in the MCC. The MultiLoc and PSORT values are taken from the MultiLoc publication (Höglund et al., 2006b) where only recall and the MCC were reported and it should be noted that no sequence in the test set was more than 30% identical to any sequence in the training set. In the context of a machine learning classifier this is desirable, to ensure the test data is sufficiently distinct enough from the training data set. Note; however, in the case of PhyloPred-HMM and CBOrg we are explicitly using homology information to guide our predictions.

2.3.4 Improved Dataset

To improve the quality of the PhyloPred-HMM clusters, as well as the taxonomic and functional coverage, the original Uniprot dataset was improved upon as described in section 2.2.5.2. All-versus-all BLAST comparisons were then re-computed and the sequences were re-clustered using MCL with an inflation parameter of two, as this method seemed to result in an intermediate number and size of clusters (compared to default MCL parameters and hierarchical clustering) while still allowing for a high-level of performance. All subsequent comparisons of PhyloPred-HMM were done using this new set of sequences and clusters. Performance of PhyloPred-HMM with the new sequence database and clusters was largely unchanged on the Höglund dataset, although there was a drop in precision using the shortest-average-distance measurement (SAD) but a gain using the nearest distance (ND) measure. Precision, recall, F_1 score, and MCC using SAD were 0.47, 0.92, 0.62, and 0.61 while for the nearest ND measure they were 0.86, 0.93, 0.89, and 0.88 respectively. The performance gain in the ND measure may be explainable by the removal of long sequences, which may link together distinct clusters depending on what domains are present, and the removal of short fragments. These factors could have a profound impact on both the alignment estimation and phylogenetic reconstruction.

2.3.4.1 Comparison on the Höglund-80 Dataset

To demonstrate how performance is affected by the taxonomic coverage and the presence of close homologs in the database, the same Höglund test dataset was again used, but this time any sequences more than 80% identical to the queries were removed from the PhyloPred-HMM database of clustered and aligned sequences (Höglund-80). Precision, recall, F_1 score, and MCC using SAD were 0.41, 0.80, 0.54, and 0.52 while for the ND measure they were 0.75, 0.75, 0.75, and 0.73 respectively. These values are contrasted with those reported in table 2.6.

2.3.4.2 Microbial Eukaryote Whole Proteome Test

Table 2.7 shows the clustering statistics for the new dataset and the changes in cluster composition (MRO, non-MRO, and Mixed clusters) when several different test datasets are constructed. In addition to the two tests described above (Höglund and Höglund-80) we also analysed performance on the *Tetrahymena thermophila* and *Trichomonas vaginalis* proteomes. These two proteomes are from two distantly related microbial eukaryotes in two very different “super-groups”. *Tetrahymena* is a ciliate belonging to the Alveolata that contains a ‘canonical’ aerobic mitochondrion while *Trichomonas* contain hydrogenosomes and belongs to the Excavata. These two organisms are much more distantly related than any two organisms in the test and training sets upon which most subcellular prediction programs are trained. Furthermore their MROs represent a much broader range of organellar diversity and are potentially the most difficult organellar proteomes to predict.

2.3.4.2.1 Comparison on Two Complete Microbial Eukaryotic Proteomes

Several other subcellular localization prediction methods were selected to compare to PhyloPred-HMM on these two microbial eukaryotic proteomes based on their previously published performance and their suitability for analysis on complete proteomes of this size (*Tetrahymena thermophila* has 27,410 protein sequences in its proteome whereas *Trichomonas vaginalis* has 59,672). In these analyses, we also included the type I error rate (false positive rate, equation 6) for all classifiers. The false positive rate was not calculated on the previous datasets as exact false and negative counts were not available for MultiLoc2.

$$False\ Positive\ Rate = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \qquad \text{Eqn. 6}$$

TABLE 2.7 NUMBER OF CLUSTERS AND CLUSTER COMPOSITION FOR THE WHOLE DATASET AND CHANGES WHEN SPECIFIC SEQUENCES FROM TEST SETS ARE REMOVED.

Dataset	No. Clusters	No. Mito	No. non-Mito	No. Mixed	Average Size (Std Dev)
Whole Dataset	20769	409 (0.020)	18892 (0.909)	1468 (0.071)	12.427 (50.607)
No *Trichomonas*	17793	415 (0.023)	16035 (0.902)	1343 (0.075)	11.224 (30.982)
No *Tetrahymena*	18862	431 (0.023)	17136 (0.908)	1295 (0.069)	12.486 (51.484)
No Höglund	20583	404 (0.020)	18724 (0.909)	1455 (0.071)	12.271 (50.549)
No Höglund 80	20299	385 (0.019)	18499 (0.911)	1415 (0.070)	11.928 (50.205)

Performance on both proteomes is not very good for any classifier (**Table 2.8**), but both CBOrg and PhyloPred-HMM outperform MultiLoc2, a relatively high performance classifier according to previous comparisons in the literature (Blum *et al.* 2009). Again, the ND measure outperforms SAD, driven by an increase in precision and only a slight decrease in recall. As expected, performance on *Tetrahymena thermophila* is higher than that on *Trichomonas vaginalis*, regardless of the method used. Because *Trichomonas* possesses a hydrogenosome instead of mitochondria, it contains many proteins that are components of its unique metabolic pathways. Organisms with these anaerobic organelles are poorly represented in public databases such as Uniprot and in

TABLE 2.8 COMPARISON OF PERFORMANCE MEASURED BY PRECISION, RECALL, THE F1 SCORE, AND THE MCC ON TWO WHOLE PROTEOME DATASETS FROM MICROBIAL EUKARYOTES. For PhyloPred-HMM both the Shortest-Average-Distance (SAD) and Nearest Distance (ND) predictive measures are used.

	Tetrahymena thermophila					***Trichomonas vaginalis***				
	Precision	Recall	FPR	F_1 Score	MCC	Precision	Recall	FPR	F_1 Score	MCC
MultiLoc2	0.07	0.56	0.16	0.12	0.15	0.008	0.25	0.12	0.015	0.025
CBOrg	0.08	0.51	0.13	0.14	0.17	0.014	0.34	0.12	0.027	0.047
PhyloPred-HMM (SAD)	0.16	**0.6**	0.067	0.25	0.28	0.03	**0.35**	0.037	0.06	0.09
PhyloPred-HMM (ND)	**0.43**	0.53	**0.015**	**0.47**	**0.46**	0.12	0.20	**0.0057**	**0.15**	**0.15**

addition, when homologous sequences are present in Uniprot they may not be annotated with subcellular localization data, especially in the reviewed category. Inclusion of the *Trichomonas* proteome in the PhyloPred-HMM database should improve predictions for data from other anaerobic eukaryotes. PhyloPred-HMM using the SAD measure produces better results, in both precision and recall, compared to MultiLoc2. In the comparison performed here the GO-Loc sub-classifier from MultiLoc2 was not used due to time constraints. As briefly described in section 2.1.1.10, MultiLoc2 contains several sub-classifiers whose outputs are fed into an overall master classifier. Two additions in MultiLoc2 are Go-Loc and Phylo-Loc. Go-Loc uses GO terms associated with a sequence for classification of subcellular location, while Phylo-Loc uses BLAST-based phylogenetic profiles; the authors of MultiLoc2 report the bulk of improvement in MultiLoc2 over MultiLoc comes from the Phylo-Loc sub-classifier (Blum *et al.* 2009). When gene ontology (GO) terms are not available, as is the case for most of the sequences tested here from *Tetrahymena* and *Trichomonas*, GO term assignment is performed as a first stage based on homology, adding an additional layer of annotation transfer in the process of subcellular location classification.

2.3.4.3 Performance on a Random Microbial Eukaryotic Test Set

Other prediction tests that were feasible for the comparison of a larger range of classifiers were conducted on a test set of randomly selected sequences from microbial eukaryotes with completed genome sequences and experimentally determined MRO proteomes (**Table 2.9**). These organisms represent broad taxonomic diversity and MRO function. While this dataset will be somewhat biased towards sequences from the larger proteomes of *Tetrahymena*, *Trichomonas*, and *Chlamydomonas* it should not be unfairly balanced in terms of internal homology nor towards sequences with or without N-terminal targeting sequences. This random test set allows us to compare the performance of PhyloPred-HMM to several different classifiers for which prediction of entire proteomes is not feasible because the classifier software is only as a web-based tool or because of run-time constraints. For CBOrg, only the yeast and human proteomes were used as the comparative databases.

Similar to the whole proteome comparisons, the performance of PhyloPred-HMM in these trials, using both the SAD- and ND-based annotations and all measures of performance, is superior to that of either MultiLoc2 or iPSORT. While iPSORT and PredSL only predict the presence of N-terminal targeting signals, and therefore cannot successfully predict any sequences with an internal targeting sequence, MultiLoc2 is a mixed classifier and does not rely exclusively on the presence of N-terminal targeting signals for prediction.

TABLE 2.9 COMPARISON OF SEVERAL SUBCELLULAR LOCALIZATION CLASSIFIERS ON A TEST DATASET OF 500 RANDOMLY SELECTED MRO AND 5000 RANDOMLY SELECTED NON-MRO SEQUENCES. Sequences were extracted from several whole proteomes of microbial eukaryotes and performance measured by the Precision, Recall, F_1 Score, and MCC. The best score in each category is depicted in bold.

	Precision	Recall	FPR	F1 Score	MCC
CBOrg	0.30	0.32	0.075	0.31	0.24
PhyloPred-HMM (SAD)	0.47	**0.43**	0.047	**0.45**	0.40
PhyloPred-HMM (ND)	**0.82**	0.30	**0.0063**	0.44	**0.47**
MultiLoc2	0.20	0.36	0.14	0.26	0.17
iPSORT	0.20	0.21	0.087	0.20	0.12
PredSL	0.28	0.22	0.057	0.25	0.18

2.3.5 Properties of Clusters and Phylogenetic Trees and Predictive Accuracy

To improve performance of a phylogenomic prediction method it is helpful to identify any potential properties of clusters, or their resulting phylogenetic trees, that affect the ability to correctly predict localization. *Tetrahymena thermophila* has 3779 sequences out of 27410 that had their best match to a mixed cluster using HMMER3, while *Trichomonas vaginalis* had 4475 of 59672 (**Table 2.10**). For all hits we recorded the number of sequences in the matching cluster, the proportion of MRO sequences, the proportion of cases where the ND or the SAD measure correctly predicted the localization, the normalized tree length (the sum of all branch lengths divided by the number of taxa), the difference between MRO and non-MRO distance for both phylogenetic measures, and the true localization of the query sequence. Using the R statistical environment and rattle (Williams 2009), an R-plugin for exploratory data analysis, we conducted an exploratory data analysis to identify characteristics of clusters and phylogenetic trees where our simple phylogenetic distance measures failed to correctly predict localization, and potentially identify clusters where the SAD measure provided the correct localization but the ND measure did not. Consistent with previous experiments, the SAD measure returns more positive MRO predictions, but at the expense of an increase in false positives.

TABLE 2.10 SUMMARY OF PREDICTIONS OF SAD AND ND PHYLOGENETIC MEASURES ON TWO MICROBIAL EUKARYOTIC PROTEOMES. The number of query sequences from the proteome with hits to clusters containing both MRO and non-MRO sequences is shown, as well as the number of those hits that are truly MRO query sequences. The total number of correct predictions (MRO and non-MRO queries) for both the SAD and ND measures and the number of correct predictions when only MRO queries are considered are also shown

	Mixed Hits	**MRO Query**	**SAD Correct (Total)**	**ND Correct (Total)**	**SAD Correct (MRO)**	**ND Correct (MRO)**
Tetrahymena thermophila	3779	259	1936 (51%)	3302 (87%)	203 (78%)	165 (64%)
Trichomonas vaginalis	4475	79	1557 (50%)	2870 (92%)	68 (86%)	36 (46%)

The number of sequences in a cluster and the proportion of those sequences that belong to an MRO are slightly negatively correlated with one another, in both *Tetrahymena* and *Trichomonas* (**Figure 2.3**). One possible combination of factors that may explain the difference between the SAD and ND measures is the proportion of MRO sequences and the normalized tree length (**Figure 2.4**). Both the ND and SAD measures will be affected by the overall length of the phylogenetic tree, but they may react differently to the presence of long-branching sequences. This is particularly true as the proportion of MRO sequences in clusters changes. When the proportion of MRO and non-MRO sequences are nearly even, the presence of longer branching sequences can be averaged out in the SAD measure, but not the ND measure. However, when these factors were plotted against one another for both *Tetrahymena thermophila* and *Trichomonas vaginalis*, no clear trend or region could be identified where the SAD measure consistently outperformed the ND measure (**Figure 2.4**), although there was a slight tendency for the longest average normalized branch length clusters with the smallest proportion of MRO sequences to be best predicted by SAD.

We also examined in the context of the proportion of MRO sequences and the normalized tree length, cases where the SAD and ND measures predicted the same localization, either correctly or incorrectly (**Figure 2.5**). As with the previous analysis, no obvious trend could be found in the data. Both correct and incorrect predictions were scattered along both axes.

One factor that may influence the predictive accuracy of these measures is the absolute value of the difference between the MRO and non-MRO distance for the proposed phylogenetic distance measure. If this value were very small, for example, we might expect both measures to frequently make mistakes because of random errors in the distance values, whereas larger differences should perhaps lead to less noisy predictions. We examined whether this was the case and plotted the results in **Figure 2.6**. The cases where the SAD measure correctly predicted localization and ND did not correspond to cases where the absolute difference between the closest MRO and closest Non-MRO was quite small. The region of the plot for the case where ND was correct and SAD incorrect, overlaps with the converse, however it seems to include more cases where larger nearest and average distances were correctly predicted.

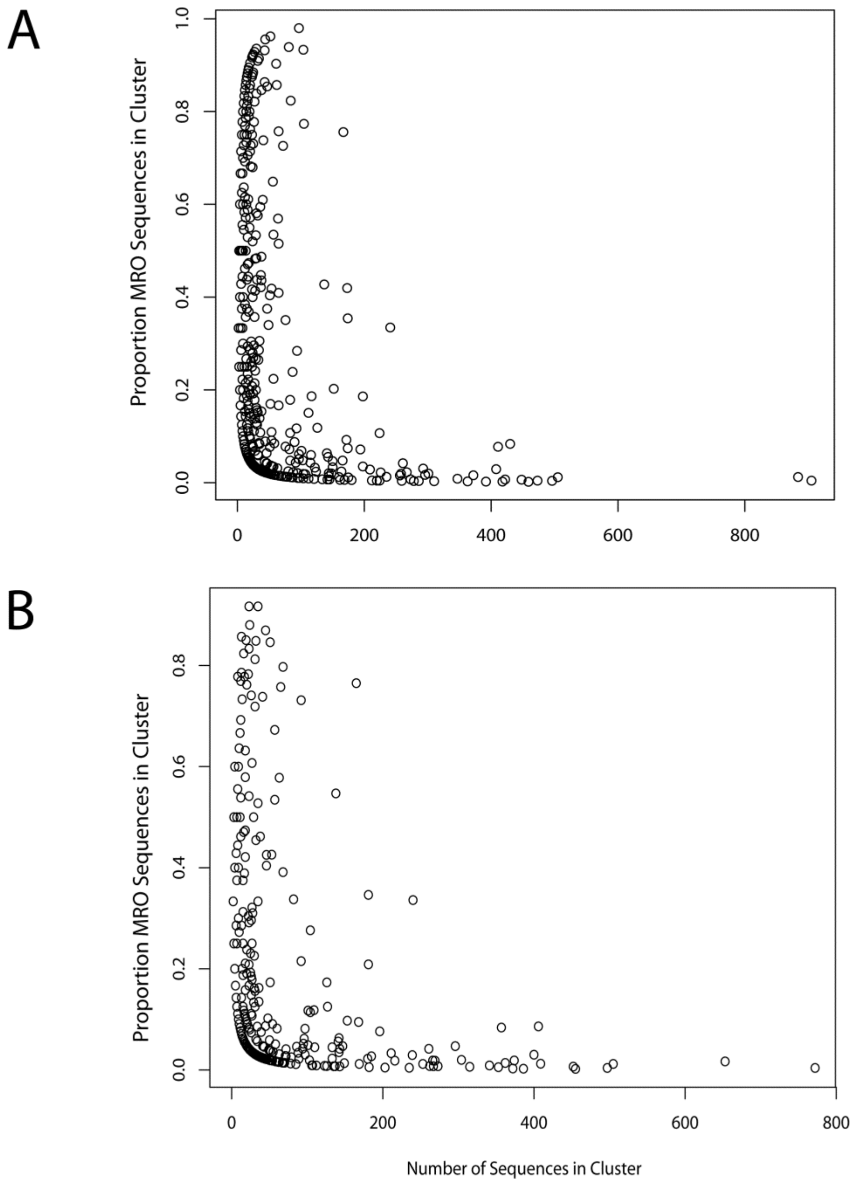


FIGURE 2.3 THE PROPORTION OF MRO SEQUENCES IN A CLUSTER AS A FUNCTION OF CLUSTER SIZE FOR BOTH TETRAHYMENA THERMOPHILA (A) AND TRICHOMONAS VAGINALIS (B). In both cases there is a slight negative correlation with fewer MRO sequences, by proportion, in larger clusters. The majority of the data however, is clustered between 0 and 200 sequences per cluster with a broad distribution for the proportion of sequences.

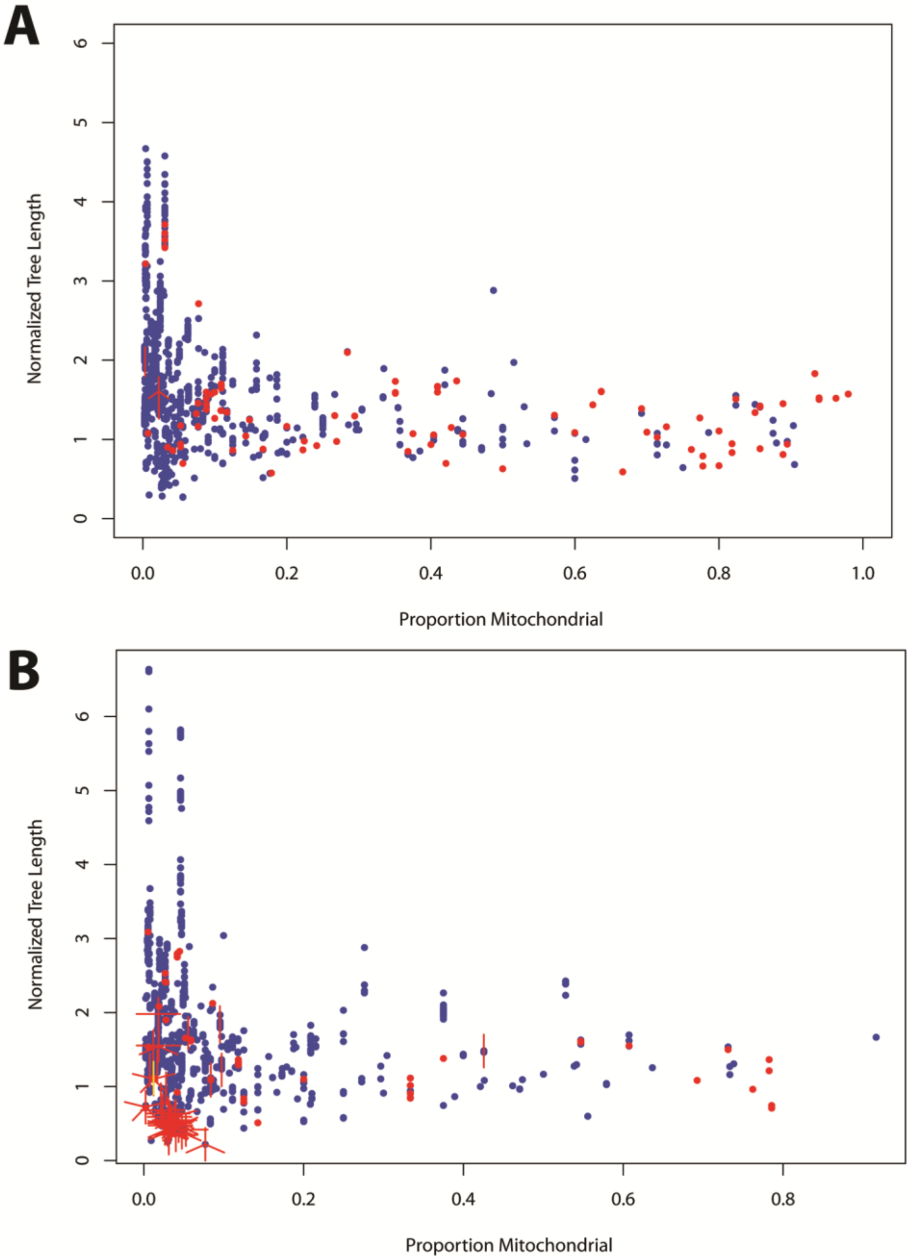


FIGURE 2.4. THE NORMALIZED TREE LENGTH VERSUS THE PROPORTION OF MRO SEQUENCES IN A CLUSTER, AS SUNFLOWER PLOTS, FOR BOTH TETRAHYMENA THERMOPHILA (A) AND TRICHOMONAS VAGINALIS (B). Blue points are cases where the ND Measure was correct while red points are cases where the SAD measure was correct. Cases where both measures were correct or incorrect are not shown. Red spoke lines from blue dots and yellow from red dots quantify the density of overlapping points in that region with the more spokes corresponding to higher density.

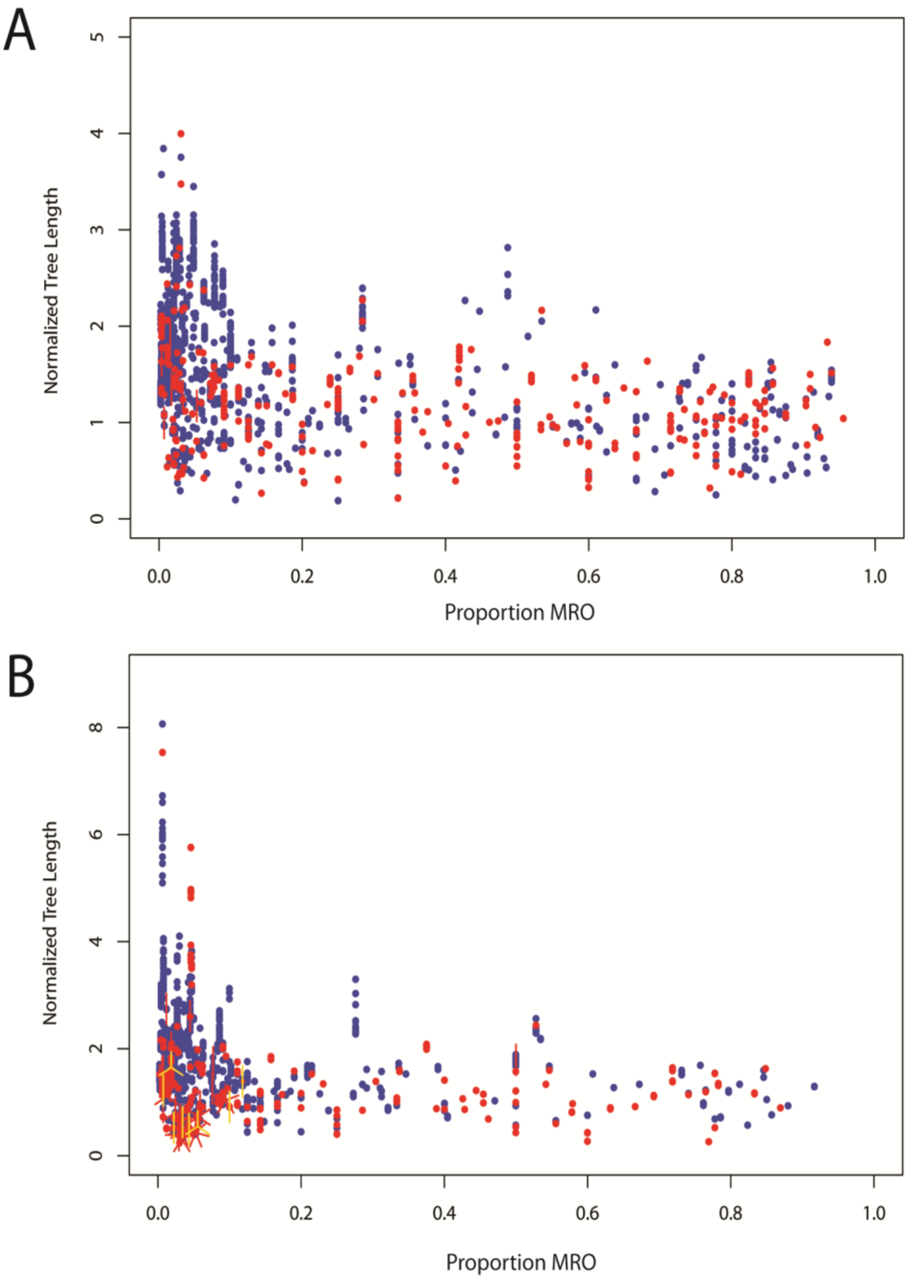


FIGURE 2.5 THE NORMALIZED TREE LENGTH VERSUS THE PROPORTION OF MRO SEQUENCES IN A CLUSTER, AS SUNFLOWER PLOTS, FOR BOTH TETRAHYMENA THERMOPHILA (A) AND TRICHOMONAS VAGINALIS (B). Blue points are cases where both the SAD and ND measures were correct while red points are cases where both were incorrect. Red spoke lines from blue dots and yellow from red quantify overlap of points with length of the spoke proportional to density

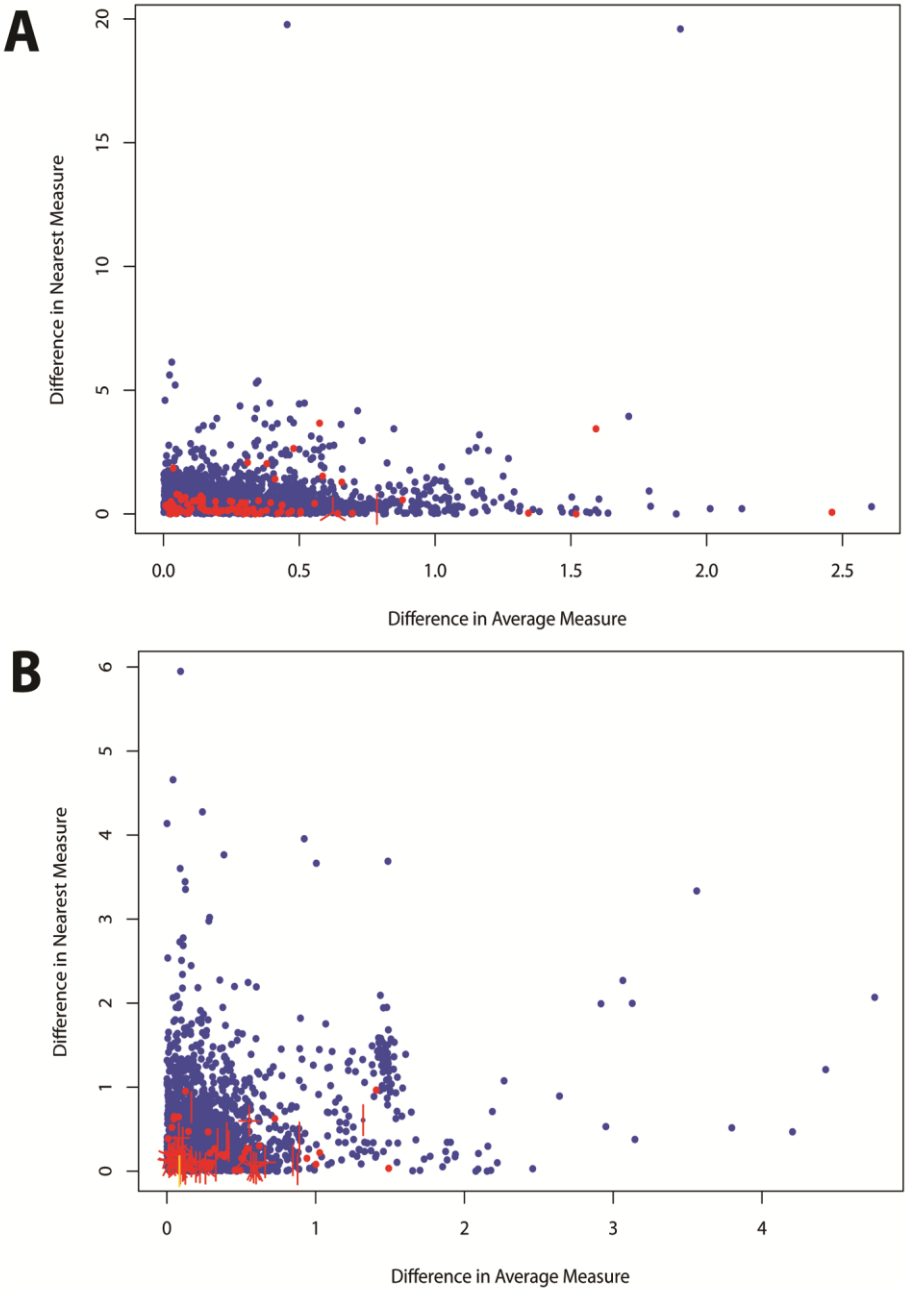


FIGURE 2.6 DIFFERENCE BETWEEN MRO AND NON-MRO DISTANCE FOR THE ND MEASURE VERSUS SAD MEASURE FOR BOTH TETRAHYMENA (A) AND TRICHOMONAS (B) AS SUNFLOWER PLOT AS IN FIGURE 4. Cases where both the SAD and ND measure agreed (correct or incorrect) not shown. ND correct are in blue while SAD correct are in red. Red spoke lines from blue dots and yellow from red quantify degree of overlap, with length of the spokes proportional to the density in the region.

It should be noted that this region of the nearest distance versus average distance plot also overlaps with the region where both measures predict incorrectly (**Figure 2.7**). Note however, that both measures tend to both incorrectly predict predominantly in cases where the nearest absolute distances between MRO and non-MRO sequences are very small and, conversely both agree and are correct when this measure is large (**Figure 2.7**).

It is likely that as the difference between distances decreases, the difference approaches the expected range of branch length estimation error on phylogenetic trees (although the latter is strongly influenced by the number of taxa and number of aligned positions used to reconstruct that phylogeny). The SAD and ND measures are weakly correlated with one another according to a linear regression, although the degree of correlation is less strong in *Trichomonas* (R^2=0.24) compared to *Tetrahymena* (R^2=0.38), which may be a function of *Trichomonas* tending to form long-branches as a result of being highly divergent from most organisms in the dataset on the sequence level (**Figure 2.8**). Despite any clear identification of a region of parameter space where the SAD method will consistently perform better than the ND method, there are some measures that indicate whether both measures are more likely to perform equally poorly or well. In these cases, it is worth examining some particular clusters and predictions of interest in more detail.

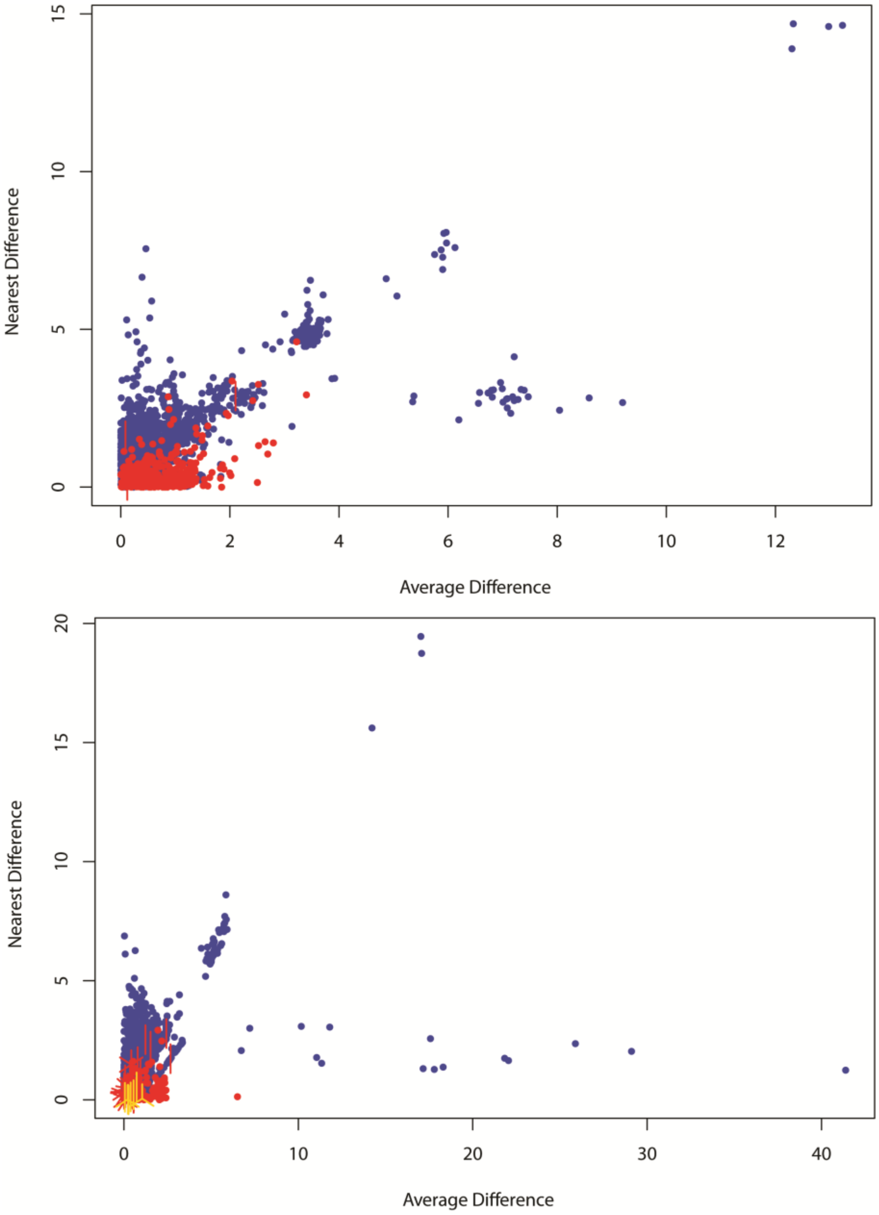


FIGURE 2.7 THE DIFFERENCE BETWEEN MRO AND NON-MRO SEQUENCES IN TETRAHYMENA (A) AND TRICHOMONAS (B) IN CASES WHERE THE SAD AND ND MEASURES MAKE THE SAME PREDICTIONS, EITHER CORRECTLY (BLUE) OR INCORRECTLY (RED). Points in the lower left of both plots overlap significantly although incorrect predictions tend to be more concentrated in the lower left regions of the plot, where the differences are small.

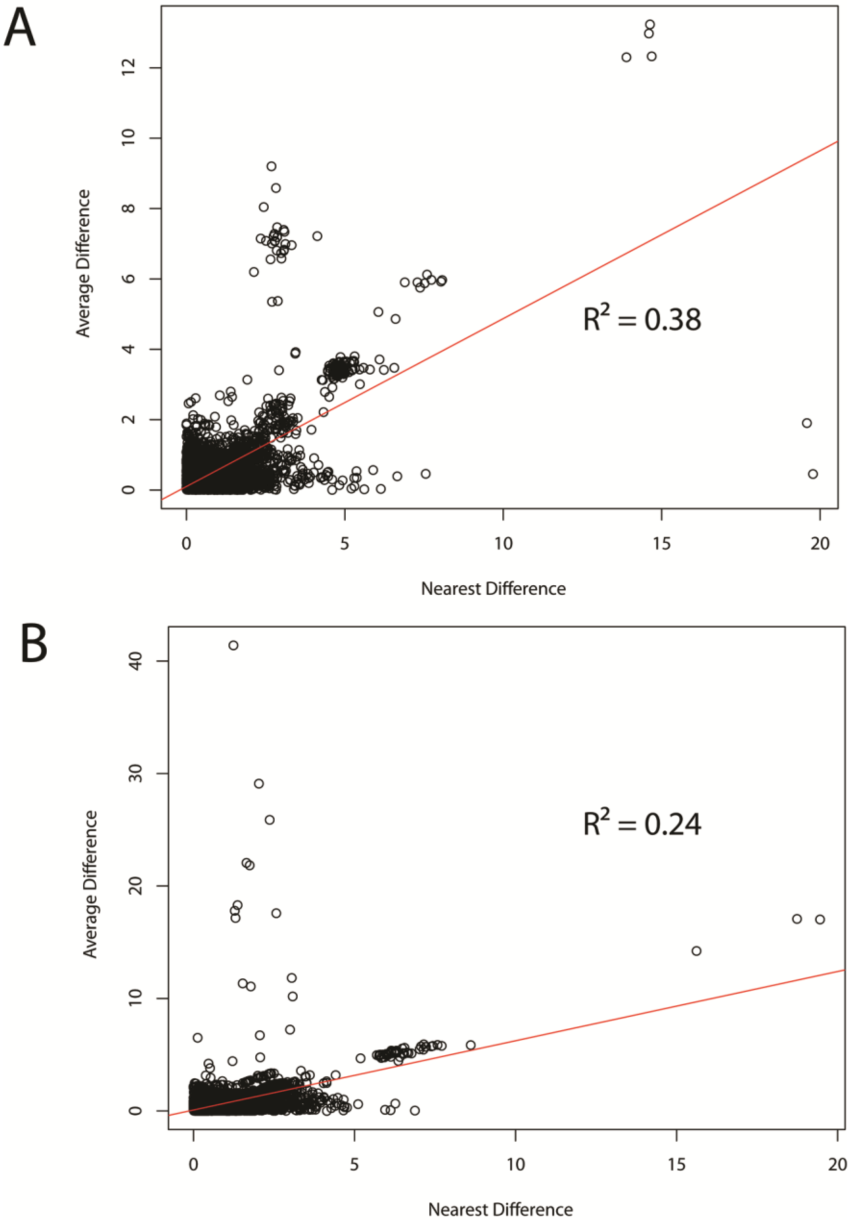


FIGURE 2.8 CORRELATION OF THE DIFFERENCE BETWEEN MRO AND NON-MRO SEQUENCES FOR THE ND AND SAD MEASURES IN TETRAHYMENA (A) AND TRICHOMONAS (B). Linear regression line is shown in red along with the linear regression correlation coefficient.

2.3.6 Specific Prediction Examples

If we restrict attention to queries where the sequence was assigned to a mixed sequence cluster, for *Tetrahymena* there were 1837 (153 MRO/1684 non-MRO) instances where both the SAD and ND measures correctly classified the query sequence and 377 (44 MRO/337 non-MRO) cases where both measures made an incorrect classification. In *Trichomonas* the both measures predicted 2201 (35 MRO/2166 non-MRO) correctly and 303 (10 MRO/293 non-MRO) incorrectly respectively. With respect to differences between the SAD and ND measures there were 99 (50 MRO/49 non-MRO) instances, in *Tetrahymena*, where the SAD measure correctly predicted localization and the ND did not (SAD-right/ND-wrong), and 1465 (12 MRO/1453 non-MRO) where the reverse was true (ND-right/SAD-wrong). In *Trichomonas* there were 57 (33 MRO/24 non-MRO) SAD-right/ND-wrong and 1913 (1 MRO/1912 non-MRO) ND-right/SAD wrong cases. We selected two cases from *Tetrahymena thermophila* to examine reasons why one or both measures of phylogenetic distance failed.

2.3.6.1 Tetrahymena thermophila *Aminotransferase Class I and II family (XP_001024634)*

This unknown member of the aminotransferase family is localized to the MRO of *Tetrahymena thermophila* (Smith *et al.* 2007) and its closest BLAST matches are annotated as hypothetical proteins, putative aminotransferases, or alanine aminotransferase. This sequence was correctly placed by HMMER3 into the same cluster from which it was taken in the original overall database of clusters created by MCL. This cluster is made up of 28 other sequences including four from *Trichomonas vaginalis*, three from *Entamoeba histolytica*, two from *Chlamydomonas reinhardtii*, and one from *Giardia intestinalis*. In *Chlamydomonas* and *Trichomonas* there are both MRO and non-MRO localized proteins while none of the sequences are localized to the mitosomes of *Entamoeba* or *Giardia*. In an unrooted phylogenetic tree, the MRO localized sequences are distributed in a patchy or punctuated fashion throughout the tree (**Figure 2.9**) and are, in some cases intermingled with non-MRO localized sequences. The difference between the MRO and non-MRO distance to *T. thermophila* using the SAD measure was 0.033 and 0.032 using the ND measure. However, the query was correctly annotated as belonging to the MRO only by the former. This phylogenetic tree illustrates an extremely difficult situation for a phylogenomic (or any similarity-based) method to resolve. There are no close relatives of the *Tetrahymena* sequence in the cluster, there are many lineage-specific duplications events whose descendants have mixed localizations, and, more generally, subcellular targeting appears to have changed frequently across the tree. It is particularly problematic that, in this case, the closest branching sequence to the putative alanine aminotransferase of *Tetrahymena thermophila* is from *Giardia intestinalis*; there will be many MRO localized proteins in *Tetrahymena* that are not found in the functionally, hence proteomically, reduced *Giardia* mitosome. In order to better analyse the particular relationships observed in this cluster we performed a bootstrap analysis with 100 bootstrap replicates using FastTreeMP. Bootstrap support for the relationship between the sequence from *Tetrahymena* and *Giardia* is low (42) with no support deeper within the backbone of the tree for the placement of these sequences as a clade.

The lineage-specific expansions in *Trichomonas vaginalis* and *Entamoeba histolytica* form monophyletic groups, with high bootstrap support in the case of the *Trichomonas* expansion (100) and two of the *Entamoeba* sequences (87), but not for the placement of the long-branching *Entamoeba* sequence with the other two (38). The two sequences from *Giardia intestinalis* do not group closely together, with one appearing to be most closely related to the sequence from *Chlamydomonas reinhardtii* and the other with the *Tetrahymena thermophila* query. Again, placement of the *Giardia* sequences is not supported, with a bootstrap support value of 46 for the grouping with *Chlamydomonas*.

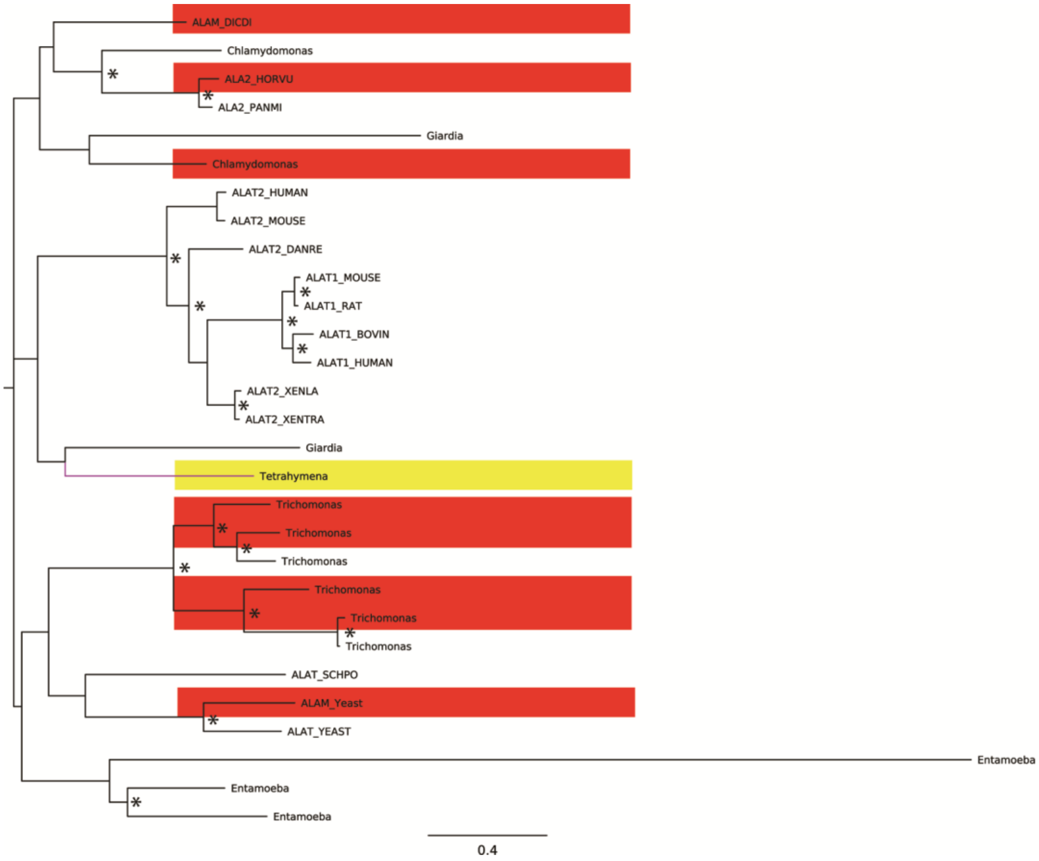


FIGURE 2.9 UNROOTED MAXIMUM LIKELIHOOD PHYLOGENETIC TREE OF THE AMINOTRANSFERASE CLUSTER PRODUCED BY FASTTREEMP AS DESCRIBED IN THE METHODS. MRO-localized sequences are highlighted in red and the query sequence from *T. thermophila* (also mitochondrion-localized) is highlighted in yellow. Nodes with bootstrap values greater than 80 are marked with asterisks.

This could be a result of mis-alignment, although the HMMER3 trimmed alignment has no obvious errors (See Appendix B2 Figure 2.1) and neither forms an abnormally long or short branch in the tree, which could have induced a phylogenetic artifact. Because other taxa also have multiple representatives, it could reflect more ancient duplications and differential loss. It is also possible that this cluster should be merged with another to form a larger superfamily, which would result in a better, more robust, phylogenetic tree. In any case, this serves as an excellent example of functional characterization that is difficult to resolve using phylogenetic methods alone, especially in a large-scale automated fashion. The SAD measure did correctly predict the localization of the protein as the average branch length of the MRO localized sequences was shorter than for non-MRO sequences, which were skewed by the long branch length of one of the *Entamoeba histolytica* sequences. If the trimmed SAD measure had been used instead, with the shortest and longest 5% of branch lengths removed prior to averaging, the predicted localization would have been of a non-MRO protein. In a sense, the SAD method made the right prediction for the wrong reasons since there is no coherent phylogenetic signal in the dataset that would indicate an organelle-localization for the *T. thermophila* homolog.

In contrast, this sequence does possesses a N-terminal targeting sequence and is correctly predicted as being MRO-localized by Predotar (Non-Plant), MitoProt 2, iPSORT, and MultiLoc2. Interestingly, of the other seven MRO-localized sequences in this cluster, no N-terminal targeting sequence could be predicted in three of the four sequences from *Trichomonas vaginalis* but could be detected in the others, including the sequences from *Saccharomyces cerevisiae*, *Dictyostelium discoideum*, and *Chlamydomonas reinhardtii*.

2.3.6.2 Putative Adenylyl/Guanylyl Cyclase Family Member (XP_001033203)

This *Tetrahymena thermophila* query sequence is not MRO-localized but belongs to a diverse cluster of uncharacterized hypothetical proteins as well as members of the adenylyl/guanylyl cyclase family. The cluster, which contains 274 protein sequences in total, includes 15 sequences from *Tetrahymena thermophila*, 114 from *Trichomonas vaginalis*, and 75 from *Chlamydomonas reinhardtii*. This family contains both membrane bound receptors and soluble forms that participate in a wide variety of molecular processes, predominantly as members of signaling pathways according to their Uniprot entries (Jain *et al.* 2009; Consortium 2011). Although this is technically a mixed member cluster, only a single sequence from *Trichomonas vaginalis* is localized to an MRO (the hydrogenosome); all other sequences are non-MRO localized. Both the SAD and T-SAD measures make an incorrect classification suggesting the *T. thermophila* sequence is MRO-localized (the T-SAD method cannot "trim" outliers in cases where there is only a single sequence) but the ND measure makes the correct classification of a non-MRO protein as its closest branching relative is a non-MRO localized homolog from *Trichomonas vaginalis* (**Figure 2.10**). The sequences from *Tetrahymena* and *Trichomonas* are extremely closely related to one another but sit at the extreme end of a very long branching clade of highly divergent sequences compared to the rest of the tree. Inspection of the automated alignment revealed that the *Trichomonas vaginalis* sequence in question is probably not a member of this family and was not aligned in any of the core blocks of the alignment, with only a single position remaining in the masked alignment for this sequence. The next closest sequence, from *Chlamydomonas reinhardtii*, and other sequences that form this long-branching group also follow this pattern. This long-branching clade of sequences is an artifact of the pipeline and automated alignment. The ND measure gets the prediction correct, but, again, for the wrong reasons. Without the presence of those extremely long branches, the SAD and T-SAD measures

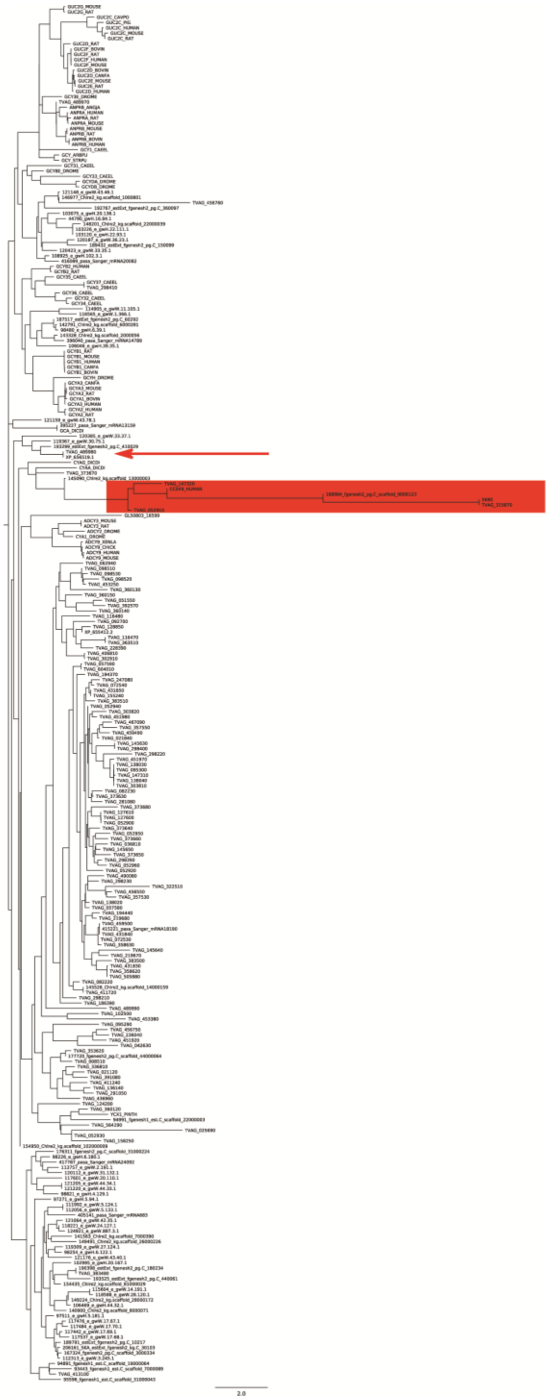


FIGURE 2.10 PHYLOGENETIC TREE OF THE ADENYLYL/GUANYLYL CYCLASE FAMILY MEMBERS. THE LONG-BRANCHING CLADE TO WHICH THE TETRAHYMENA QUERY SEQUENCE BELONGS IS HIGHLIGHTED IN RED. The *Tetrahymena* sequence (labeled 5690)) branches sister to a non-MRO sequence from *Trichomonas vaginalis* at the extreme tip. Arrow indicates the position of the sole MRO localized sequence from *Trichomonas*.

may have been able to correctly predict the localization as the ND measure did, although it appears as if the MCL algorithm (Inf = 2) incorrectly grouped together two distinct clusters that only poorly align with one another in some regions.

This sequence is weakly predicted as being located in the cytoplasm by MultiLoc2 (0.35) with nuclear (0.26) and mitochondrial (0.22) being the next highest predicted localizations. It is not predicted to be MRO localized by N-terminal targeting prediction methods such as Predotar or iPSORT, or MitoProt2. The sole MRO localized sequence in this cluster, from *Trichomonas vaginalis*, appears to either lack an N-terminal targeting sequence entirely, or the targeting sequence is very short and is not predicted to be localized to the MRO by any of the programs above.

2.4 Discussion

The performance of any large-scale phylogenomic method depends heavily on several factors including: the method by which sequences were placed together into clusters, the quality of the multiple sequence alignments, the alignment masking method and the resolution of the phylogenetic trees, the distribution of the functional annotation of interest within the tree, and the method by which functional annotations are propagated from known to unknown sequences.

2.4.1 The Impact of Clustering Method and Parameters

Clusters with too little sequence diversity will produce inferior HMMs that are less capable of detecting remote homologs. On the other hand, clusters that are too large may contain sequences that are not truly homologous, or are so divergent in sequence that an automated alignment of the cluster is not of sufficiently high quality for phylogenomic analysis. Because the method of transferring annotation from one sequence to another relies on the quality of the phylogenetic tree, and the accurate estimation of the tree, in turn, relies heavily on the quality of the alignment, the potential errors introduced in each step in the chain from gathering clusters to building trees may be compounded and compromise the overall performance of the method.

The choice of clustering method parameters can also have a large impact on sequence clustering, such as the ability to recover known manually curated gene families (Frech & Chen 2010). Indeed, with Tribe-MCL, the optimal inflation parameter for recovering a manually curated family has been noted to differ between different gene families even within a closely related group of organisms (Frech & Chen 2010). Here two different inflation parameters were tested, the default parameter of 1.2 and a larger value of 2.0 based on several previous studies (Stein *et al.* 2003; Tekaia & Latgé 2005; Wall *et al.* 2008; Frech & Chen 2010). Hierarchical clustering was also used due to its simplicity and popularity with a frequently used e-value cut-off of 10^{-6}. While the two sets of parameters used with the MCL method resulted in differences in performance between the SAD and ND measures MCC of 0.6 and 0.85 for Default, 0.71 and 0.87 with inflation parameter of two), hierarchical clustering showed a less dramatic effect (0.80 and 0.89). On the other hand, hierarchical clustering also had the worst accuracy in terms of assigning a query correctly to its source cluster using HMMER3 with an accuracy of 0.85. This may be a function of hierarchical clustering producing many small clusters, sometimes splitting orthologs and paralogs. If these small clusters contain predominantly extremely similar sequences, then queries against their HMM profiles with homologous but much more divergent sequences will not yield significant e-values; such homologs could potentially be missed entirely. The MCL method with an inflation parameter of two represented an ideal compromise between the MCL method with default parameters and hierarchical clustering. It produced more clusters of intermediate size compared to either of the other two methods as well as an intermediate number of clusters in general. Based on our analyses using the refined

Uniprot + proteomes dataset, this method produced adequate results on highly divergent microbial eukaryotes of interest, although there may be some instances of clustering errors, such as with the adenylyl/guanylyl cyclase family discussed above. Future developments of these methods should focus on efficient methods of identifying and correcting such errors.

The choice of phylogenetic distance measure also had an impact on performance, with the simple ND measure performing the best overall in these comparisons. It appears as if in the majority of cases the ND measure is the most likely to be correct, although it is still worth investigating regions of tree/cluster parameter space where the SAD, or another appropriate measure, may offer superior performance. With the current implementation PhyloPred-HMM using the ND measure offers the best performance and should be used with that setting.

2.4.2 Sequence Diversity and Uniprot Annotation

It is notable that in the case of *Trichomonas,* only seven sequences are annotated in Uniprot as being localized to the hydrogenosome (HSP60; Ferredoxin; Adenylate kinase; Succinyl-CoA ligase subunit alpha 1,2, and 3; and Succinyl-CoA ligase subunit beta) and two are annotated as mitochondrial (both are GrpE homologs) although 228 sequences were found to be hydrogenosomal based on organellar proteomics (Schneider *et al.* 2011). Only one sequence (Flap endonuclease 1) is annotated in Uniprot as belonging to the mitosome of *Giardia intestinalis*. While experimental data exists for these organisms, providing a wealth of data, Uniprot is often slow to incorporate these new findings, particularly in the more stringently controlled ‘reviewed’ subset of sequences.

The better studied eukaryotic groups commonly used for training and testing of other MRO-localization classifiers are quickly becoming saturated in terms of sequence diversity within Uniprot. The majority of experimentally annotated localizations are from taxa in these groups, but even more sequences have been annotated with potential localizations with less stringent evidence (e.g. probable, potential, by similarity), providing a wealth of data that can aid in predictions. In PhyloPred-HMM we used all sequences annotated as being localized to the MRO regardless of the annotation stringency. As demonstrated here, relatively simple phylogenomic approaches to determining localization do at least as well as arguably more sophisticated machine-learning approaches and actually outperform these machine-learning algorithms on more divergent data from less well-studied eukaryote groups. As sequence data coupled with improved annotations from microbial eukaryotes accumulates, we can expect phylogenomic-based annotation to improve for these organisms as well. Because the presence of an N-terminal targeting sequence is seen as solid evidence of MRO targeting, combining a fast and accurate N-terminal targeting prediction algorithm with PhyloPred-HMM could substantially improve performance for difficult to classify phylogenetic situations like those discussed previously. However, some improvement in training data would be desired for any N-terminal targeting sequence prediction algorithm that may potentially be used because of the presence of shorter, idiosyncratic targeting sequences found in some organisms with reduced MROs (Dolezal *et al.* 2005; Smíd *et al.* 2008).

2.4.3 The Impact of Phylogenetic Reconstruction Method and Alignment

For the phylogenetic tree inferences in this study, speed and accuracy were the primary concerns. FastTree2 provides rapid maximum-likelihood inference of phylogenetic trees, with performance largely comparable to RAxML based on benchmarking analyses (Price *et al.* 2010), at least for recovering well supported groups. An alternative approach that was not explored in this study would be to pre-compute phylogenetic trees for all clusters using RAxML (Stamatakis 2006b) and use the novel environmental sequence placement method implemented in this program to place query sequences on the predefined backbone topology. Because branch length measures are being used to determine subcellular localization, accuracy of the maximum-likelihood estimated branch lengths may have a large impact on performance, particularly in cases such as the *Tetrahymena thermophila* aminotransferase described above, where the difference between MRO and Non-MRO sequences were minute. While maximum likelihood methods are expected to be less error prone in terms of branch length estimation than methods such as maximum parsimony or distance-based methods, the complexity of the datasets, problems with automated alignment generation/masking and resulting phylogenetic tree estimation all produce some possibility of error (Schwartz & Mueller 2010).

FastTree2 uses several heuristics to produce trees, with generally good accuracy in comparison to PhyML 3 and RAxML (Price *et al.* 2010), and it has fast run-times especially when the number of sequences are large. However, there is potential for significant error to enter the pipeline at this stage. This may be compounded by errors during the multiple sequence alignment phase, which is further compounded by potential errors from the original sequence clustering. While the results for sequence from well-studied groups are comparable to those from other published classifiers, and superior in the case of query sequences from diverse microbial eukaryotes, significant improvements could be made in identifying potential sources of clustering or alignment errors. Construction of robust, pre-defined reference phylogenies for clusters would also reduce the instances of phylogenetic artifacts that affect phylogenomic inference.

2.4.4 The Type of Dataset Used to Evaluate Performance Matters

There are two commonly-used methods of assessing the performance of a subcellular localization classifier. The first is to construct either a non-redundant (and reduced homology) test dataset based on the newest release of Uniprot (or another database of interest) while training on all sequences from previous releases. The second, and perhaps more widely used, method is to use a jack-knife approach to iteratively build training and test datasets, often using a complete sequence-by-sequence jack-knife. A less often used approach, at least in published performance evaluations, is to systematically predict the subcellular localizations of the entire proteome of an organism with rigorous experimental data available to validate predictions. In the cases where the latter has been done, it has typically been based on members of the same well-studied groups of eukaryotes (e.g. typically humans, yeast, nematodes, or *Arabidopsis*) that were used to train the classifiers.

In this study, we evaluated performance both on a reduced homology dataset, a randomly generated test set of sequences from microbial eukaryotes with completely determined MRO proteomes, and on two whole proteomes from distantly related eukaryotic microbes with functionally distinct MROs. This approach more clearly demonstrates the situations where performance will be high and the cases where use of a single prediction method is likely to produce very poor results. In particular, we showed that there are significant differences in the performance measures, such as the MCC, when analyzing a random subset of data from multiple organisms and when analyzing whole proteomes. This performance discrepancy was particularly apparent with the poor performance on the whole proteome of *Trichomonas vaginalis*. The genome of *Trichomonas vaginalis* is nearly three times larger than the genome of humans and features many lineage specific expansions of gene families. This could be seen clearly in the examples of predictions of *Tetrahymena thermophila* sequences that we examined in detail in which *Trichomonas vaginalis* often made up a substantial proportion of the total number of sequences in the cluster. In cases such as this, when predicting the subcellular localization of those *Trichomonas vaginalis* sequences, whole sets of paralagous sequences will either be correctly, or more likely, incorrectly classified.

To summarize, it is important that future prediction programs developed for subcellular localization incorporate a broader taxonomic range of sequences in their training sets. We have likely reached, or nearly reached, the saturation point for the better studied taxonomic groups, except perhaps for proteins with multiple localizations where the number of cases identified is still increasing. Substantial improvements are still being made in prediction methods capable of classifying these kinds of proteins.

2.4.6 Comparison to Other Classifiers

While we could only directly compare PhyloPred-HMM performance against a small subset of existing classifiers, we expect the results to be comparable to other classifiers as well. In terms of performance on mitochondrial proteins, as opposed to performance values when trying to predict multiple possible localizations, other classifiers self-report similar values to MultiLoc2 on mitochondrial proteins and several were tested either on the Höglund dataset or very similar datasets, such as the BaCello Independent dataset (Blum *et al.* 2009).

Although the majority of classifiers have been taxonomically restricted in terms of their test and training data, a notable case where this is not true is Euk-mPLoc (Chou & Shen 2007) and the newer Euk-mPLoc 2.0 (Chou & Shen 2010). These classifiers, which allow for multiple localizations of proteins, include a broad list of possible subcellular localizations and a dataset that included all eukaryotic sequences with experimental localization data. However, in these cases the hydrogenosome was treated as a unique subcellular localization, and not grouped together with the mitochondrion. This means that the training set for hydrogenosomal proteins was very small (10). In jack-knife tests only 2 of the 10 proteins were correctly predicted to be localized in the hydrogenosome for Euk-mPLoc 2.0 and none with the first version of Euk-mPLoc (Chou & Shen 2007, 2010). The small sample size is due to annotation issues in Uniprot as described in section 2.4.2; only a small subset of hydrogenosomal proteins are annotated as such while many others are annotated as mitochondrial, resulting in a training set far too small for proper use in a machine-learning algorithm.

2.4.7 PhyloPred-HMM as a General Phylogenomic Framework

PhyloPred-HMM is a fast and flexible annotation pipeline for predicting the subcellular localization of novel amino acid input sequences. The pipeline combines the power of Hidden Markov Model (HMM) based profile searching against an automatically *de novo* clustered dataset of annotated proteins from Uniprot with rapid multiple-sequence alignment, posterior probability and gap-based alignment trimming, and rapid phylogenetic inference. PhyloPred-HMM leverages the computational speedups of parallel processing wherever possible, making the automatic generation of phylogenetic trees and automated annotation possible for a large number of input sequences. PhyloPred will also perform rough/approximate 6-frame translations of input nucleotide sequences allowing for annotation based on preliminary expressed-sequence tag (EST) and genomics data. The PhyloPred-HMM framework relies on several external programs for alignment and phylogenetic analysis including MAFFT, HMMER3, and FastTree2 for this analysis and it can be easily adapted to interface with other programs. The choice of clustering methodology in this work was based on a desire for larger more diverse clusters wherever possible, and not to restrict clusters to the more stringent requirements of orthology. Any clustering approach could be used, and only a handful of *de novo* techniques were tested here. The MCL algorithm with an inflation parameter of two seemed to provide the best trade-off in terms of the number and size of resulting clusters and measures of prediction performance in terms of the F_1 score and Matthews Correlation Coefficient. The annotation used was specifically to identify putative MRO localized proteins, but any phylogenetically meaningful annotation suitable for the phylogenomic approach could be used. The simplest assignment technique, using the annotation of the nearest neighbour as measured by tip-to-tip phylogenetic distance also proved to be the most precise in terms of performance measures. The comparatively poor performance of the shortest-average and trimmed shortest-average distance metrics may have been impacted by heavily skewed clusters where the majority of sequences, with the exception of one or two, had predominantly one of the two localizations. More robust metrics involving ancestral state reconstruction by, for example, maximum parsimony or likelihood-based methods may improve performance.

Overall PhyloPred-HMM provides a robust phylogenomic platform for the prediction of subcellular localization to MROs. Performance as measured by several factors, notably the Matthews Correlation Coefficient, showed comparable performance to established machine-learning classifiers on sequences from animals, plants, and fungi and superior performance on less well studied microbial eukaryotes with highly divergent MROs in terms of their function, and highly divergent sequences in general.

2.5 Author Contributions

For sections previously published DG prepared manuscript and conducted experiments. Test sequences for CBOrg comparisons selected by DG and ADT. ADT and AJR provided editorial comments. All authors agreed on final submission of manuscript.

Chapter 3 Predicting Functionally Divergent Protein Residues

This chapter was originally published as:

“Gaston D, Susko E, Roger AJ. 2011. A Phylogenetic Mixture Model for the Identification of Functionally Divergent Protein Residues. Bioinformatics. 27:2655-2663”

3.1 Introduction

Functional divergence in proteins over evolutionary time includes the processes of sub- and neo-functionalization after gene duplication, as well as specialization or loss of functions of proteins in distinct organismal lineages (Li 1983; Henikoff *et al*. 1997). Two main patterns of functional divergence at the amino acid residue level have been described in the literature and were classified by Gu (1999, 2001) as Type I and Type II. In the case of a protein family composed of two subgroups, Type I functional divergence is characterized by greater conservation at a site in one subfamily versus the other subfamily, indicating a difference in evolutionary rate between them due to fewer selective constraints in the more rapidly evolving group. For Type II divergence sequence conservation at a site is observed in both sub-families but with a marked preference for different amino acids, generally with very different physicochemical properties in each group. Accurate prediction of functionally divergent residues, also known as ‘specificity determining sites’ in the case where divergence changes the substrate that is bound (Gerlt and Babbitt 2000), leads to an enhanced understanding of the mechanisms underlying functional diversification.

Three main approaches have been used for the prediction of functionally divergent protein residues that, in broad terms, can be classified as primarily phylogenetic, information theoretic, or biophysical. Phylogenetic approaches such as Evolutionary Trace (Lichtarge *et al.* 1996),DIVERGE (Gu 1999, 2001; Gu and Vander Velden 2002), and various Likelihood Ratio Test/Rate Shift based tests (Knudsen and Miyamoto 2001; Knudsen *et al.* 2003; Susko *et al.* 2002) explicitly take into account a phylogenetic tree that describes the evolutionary relationships among the sequences in the protein family under consideration. In general, phylogenetic methods for functional divergence prediction correlate observed patterns of amino acid substitution at a site in a multiple sequence alignment across subgroups within a phylogenetic tree. Local conservation (i.e. within a subgroup on a phylogenetic tree) relative to other sequences reflect probable functional specificity of that subgroup if the degree of conservation is large relative to the overall divergences of the sequences within that subgroup. This general case can be extended to more rigorous statistical models of functional divergence such as the type I and type II specific prediction methods employed by DIVERGE (Gu 1999, 2001; Gu and Vander Velden, 2002).

In contrast, information theoretic approaches do not generally explicitly consider the relationship between sequences, only the known, or predicted divisions into functional subgroups and perhaps some weighting based on overall sequence distances as in GroupSim (Capra and Singh 2008). These approaches contrast information theoretic measures of variation of site profiles within a subgroup to those observed at a site across the whole multiple sequence alignment. These profiles may most commonly be represented by some information theoretic measure of variability among residues at a site such as the Jensen-Shannon Divergence (Lin 1991), Relative Entropy/Kullback-Leibler Divergence (Kullback and Leibler 1951), Sequence Harmony (Piravano *et al*. 2006) and/or simple Shannon Entropy.

Biophysical/structural methods may also include some measures of sequence diversity/information content as above, with a greater focus on the physico-chemical properties of structurally conserved residue positions. Active Sites Modeling and Clustering (de Milo-Minardi *et al.* 2010) compares profiles of structurally aligned and modeled active sites to identify specificity-determining residues. Surface map techniques have also been developed (Pawlowski and Godzik, 2001, Sael *et al.* 2008) that compare properties such as charge and hydrophobicity of surface proteins of two proteins. Other methods have been used to predict the substrate-specificity of unknown family members in cases where annotation transfer from paralogs may not be adequate (Caffrey *et al.* 2008) and the related task of identifying functional sites (Capra *et al.* 2009; Sankararaman *et al.* 2010). No structural methods were evaluated as part of this study due to the lack of adequate methods for simulating evolutionary divergence in the context of protein structure.

Here we introduce a new phylogeny-based method, called FunDi, for detecting functionally divergent sites across a phylogenetic split in a protein family tree. By explicitly modeling type I and type II functional divergence using a mixture model, FunDi provides a maximum-likelihood phylogenetic framework to predict functionally divergent sites using specific models of amino acid substitution. As an open framework for functional divergence classification, FunDi is easily extended to accommodate the latest methods/programs for maximum-likelihood based phylogenetic reconstruction and new, more accurate models of amino acid substitution. We also evaluate whether a weighted average of FunDi's score and the Jensen-Shannon Divergence scores of surrounding residues (Capra and Singh 2008) improves performance.

A number of well-characterized protein datasets have been used for evaluations of the performance of some functional divergence/specificity-determining classifiers (FD classifiers) (Chakrabarti *et al.* 2007). One limitation of these biological datasets is the difficulty in assigning the labels of "true negative" or “false positive” to sites that are not involved in functional divergence. Thorough molecular characterization of every amino acid position in a protein family is practically infeasible; requiring mutagenesis and functional studies not only on a single representative sequence, but also over the biological sequence diversity represented by the protein family. While these biological datasets are unavoidably ‘noisy’ for testing the efficacy of functional divergence predictors for these reasons, their true positive sites are often well supported with robust experimental validation. We evaluated the performance of FunDi, and several other FD classifiers over 11 of these biological datasets with two phylogenetically distinct sub-families each (Chakrabarti *et al.* 2007). In order to provide a more robust estimate of performance on less noisy data, we also introduce two alternative frameworks for simulating functional divergence. In this framework, we examine the impact of taxon sampling and the scale of branch lengths on the predictive performance of functional divergence classifiers because under-sampling of phylogenetic diversity (taxon sampling) and overall sequence divergence are two well-known factors influencing the accuracy and error associated with phylogenetic reconstructions (Susko *et al.* 2005; Zwickl and Hillis 2002). In the case of functional divergence, it seems likely that under sampling of meaningful phylogenetic diversity can lead to incorrect observations of substitution patterns and sequence conservation levels (Blouin *et al.* 2005). By explicitly taking the phylogeny of protein families into account with an appropriate model of functional divergence, we expect improved predictive performance relative to programs that do not use this information.

3.2 Methods

3.2.1 FunDi

We assume that a given multiple sequence alignment is composed of sites that fall in to two classes, those contributing to functional divergence and non-divergent sites. To capture the dynamics of the functionally divergent class we construct a two component phylogenetic mixture model where non-functionally divergent sites evolve across a shared phylogenetic tree (standard evolutionary model/dependent component), while functionally divergent sites are treated as being evolutionarily ‘uncoupled’, evolving on independent subtrees (FD component).

Specifically, the dependent component models amino acid residues whose evolutionary constraints remain similar across a single phylogenetic tree. This is captured by a standard substitution model of protein evolution such as JTT (Jones *et al.* 1992), WAG (Whelan and Goldman, 2001), or LG (Le and Gascuel 2008) with rates across sites (RAS) modeled using a discrete rate approximation to the gamma distribution.

During functional divergence, this 'standard' model of evolution is violated. Under type I functional divergence a rate shift has occurred (heterotachy) such that a site can no longer be adequately modeled by the same rate categories in different lineages of the tree, similarly for type II functional divergence where a site has undergone a shift in the amino acid preferences across a phylogenetic tree. In both cases, the normal assumption of a homogeneous substitution process across lineages no longer holds. In order to capture functional divergence we introduce an ‘independent component' approximation where sites in subtrees are modeled as if they were completely independent observations. In the maximum likelihood (ML) framework, model parameters such as the alpha shape parameter, amino acid frequencies, and branch lengths are allowed to be independently optimized in each subgroup. The total likelihood of a site under this simplified approximate functional divergence model will therefore be the product of the site likelihoods for each subgroup. Note that this is equivalent to assuming that the length of the internal branch length between the subtrees, *b,* is effectively infinite and approximates the period of rapid evolution that immediately follows the changes in functional constraints at a site associated with functional divergence. For two subgroups the likelihood of a site x under this independence model is given by:

$$L_x = P(X_1|T_1)\ P(X_2|T_2) \qquad \text{Eqn. 1}$$

T_1 and T_2 are the phylogenies and associated branch lengths for each of the two subtrees while X_1 and X_2 are the data patterns in the two subgroups at that site. In a mixture model context the likelihood of a site is given as the weighted sum of the dependent and functional divergence components:

$$L_x = \rho P(X_1, X_2|T_, b) + (1\text{-}\rho)(P(X_1|T_1)\ P(X_2|T_2)) \qquad \text{Eqn. 2}$$

where ρ represents the optimized class weight parameter and T refers to the entire phylogenetic tree comprised of T_1 and T_2 linked by an internal branch of length *b*. The site likelihoods of each component are calculated by standard ML phylogenetic estimation software using a supplied tree and alignment along with subgroup assignments for taxa. The ρ parameter is optimized using a two-step line-search procedure to two decimal places of precision.

In this framework, we assume, when used in an appropriate biological context, that the independent component is modeling functional divergence, other violations of the 'standard' model will also be captured, such as heterotachy and potentially certain types of amino acid composition biases. FunDi, as currently described has for instance also been used to identify positions in multiple sequence alignments that do not support a particular phylogenetic reconstruction, more specifically sites that are incongruent with a particular monophyletic grouping of sequences (data unpublished).

While the observed functional divergence site patterns in the two subtrees are not expected to be completely independent (i.e they retain a shared evolutionary history/trajectory), approximating them as independent offers several advantages. First, it allows for maximum flexibility of ML model choice. Any phylogenetic software tool that outputs site likelihoods can be used as the back-end engine for likelihood calculations. Currently, FunDi can accept site log-likelihood values from TREE-PUZZLE (Schmidt *et al*. 2002), RAxML version 7.2.6 (Stamatakis 2006b), QmmRAxML (Wang *et al*. 2008), or FastTree (Price *et al*. 2009, 2010). This 'Plug-And-Play' utility allows FunDi to rapidly accommodate new, complex models of sequence evolution as they are developed. In addition, by implementing FunDi as a mixture model containing both independent and dependent components, the shared evolutionary history of functionally divergent sites is not completely ignored. All sites will be modeled with likelihood contributions from both components. It is the relative contribution of the independent component, measured by the site-wise posterior probability of the functionally divergent class, that serves as an estimator of the functional divergence character for a given site.

Here the performance of FunDi using either the "base" RAxML v.7.2.6 (called FunDi-RAxML) or QmmRAxML (called FunDi-QmmRAxML) is evaluated. In brief, QmmRAxML is a mixture-model of a user-defined number of rate matrix classes (Q_i's) each with an associated weight (w_i) that is optimized by ML. Here for each class *i* we define entries of an instantaneous rate matrix $Q_{jk}(i) = R_{jk}\Pi_k(i)$ for all pairwise combinations of amino acids k and j. R_{jk} is the standard amino acid exchangeability of amino acid j for amino acid k from an exchangeability matrix (WAG in this case) and the Π_i's represent nine commonly occurring amino acid frequency profiles estimated by Sjölander and colleagues (Sjölander et al. 1996). The WAG database frequencies form a 10th, catch-all, class. $\Pi_k(i)$ is therefore the frequency of the amino acid k in the frequency profile class *i*. Under the functional divergence model, this model has the advantage that each subgroup can optimize towards different class (profile) preferences and rates, allowing for functional shifts at particular sites across the split.

FunDi outputs the posterior probability of each site belonging to the functional divergent class. FunDi can optionally be run with the ConsWin windowing method (FunDi-ConsWin) as described in Capra and Singh (2008), which weights site scores based on the Jensen-Shannon Divergence of surrounding amino acid residues. Based on previous results (Capra & Singh 2008) the ConsWin windowing method improved performance of both GroupSim and other tested classifiers of functional divergence due to a presumed bias for functionally divergent residues to be located preferentially within more conserved regions, likely ones that serve a role in either enzymatic behavior (active site) or those involved with other protein-protein or protein-small molecule interactions. The Jensen-Shannon divergence score for all sites is calculated using a python script as detailed by Capra and Singh (2007). The average Jensen-Shannon divergence score of a window of surrounding columns in the alignment is then weighted and added to the posterior probability of functional divergence:

$$S(FD_{x}) = \lambda P(FD_{x}) + (1 - \lambda)JSD_{avg} \qquad \text{Eqn. 3}$$

Where $S(FD_x)$ is the functional divergence score at site x, λ is the weight for the posterior probability of functional divergence (P(FD)) at site x, and JSD_{avg} is the average Jensen-Shannon Divergence score of the window. Here we used the recommended optimal values (Capra and Singh 2007) of 0.7 for λ and a window size of three to either side of the column under consideration. This sliding-window scheme has been shown to improve predictive performance both in the GroupSim method and with other classifiers (Capra and Singh 2008).

3.2.2 Simulations

We have implemented two simulation strategies for functional divergence in order to evaluate the relative performance of various FD classifiers. Alignments containing both functionally divergent (FD) and non-divergent (non-FD) sites were simulated over a variety of tree topologies in order to provide a comprehensive analysis of performance.

3.2.2.1 Strategy I*: Site-Specific Amino Acid Profiles*

INDELible (Fletcher and Yang 2009) was used to simulate alignments consisting of both functionally divergent and non-divergent sites using the 10-component QmmRAxML mixture model described above. In order to conform to GroupSim assumptions, functionally divergent sites were required to be located within windows of non-FD sites in the primary amino acid sequence and all sites were selected, as described below, to have specific Jensen-Shannon divergence score distributions. Distributions for each site type were estimated from biological datasets that have been used in previous studies (Chakrabarti *et al*. 2007), with five sets of distributions used. Jensen-Shannon Divergence scores were calculated for all functionally divergent sites, sites located in a three residue windows on either side of a functionally divergent site, and all other sites separately. This was done for all of the two-family alignments used in Chakrabarti *et al*. 2007. Four sets of the above estimates were used directly while a fifth set of score distributions was set to be of intermediate values compared to the other four. The divergence score distributions for Window and other non-FD sites in this set were equal to one another (**Table 3.1**). One-hundred random trees and corresponding alignments were simulated under each of these five sets.

3.2.2.1.1 Phylogenetic Trees

Random phylogenetic trees were generated using INDELible with a birth-death (BD) process. Trees were randomly chosen to have be-tween 10 and 50 taxa. Birth, death, and mutation rates were randomly selected from a uniform distribution between 0 and 1 while the sampling parameter was constrained to a value of 1 (for details about the BD process see Yang and Rannala (1997)). For each of the five sets of Jensen-Shannon divergence scores one thousand individual trees were simulated and from these 100 pairs were randomly selected and joined by a midpoint rooted internal branch of length 1 expected substitution per site (0.5 on either side of root). This represented 500 protein family trees undergoing functional divergence across a central split.

TABLE 3.1 THE FIVE JENSEN-SHANNON DIVERGENCE SCORE DISTRIBUTIONS USED IN SIMULATION OF 500 FUNCTIONALLY DIVERGENT ALIGNMENTS USING INDELIBLE SELECTED FROM CHAKRABARTI *ET AL.* (2007). The intermediate case represents values selected to fall within the range of values drawn from the real datasets

Set	FD Mean	FD Std Dev	Win Mean	Win Std Dev	Other Mean	Other Std Dev	Alignment
1	0.5	0.1	0.62	0.14	0.62	0.14	Intermediate
2	0.74	0.07	0.66	0.06	0.56	0.16	Nucleotidyl cyclase
3	0.73	0.02	0.44	0.11	0.39	0.13	cd00985
4	0.71	0.09	0.82	0.06	0.81	0.07	Smad
5	0.41	0.12	0.63	0.14	0.51	0.21	G-proteins

3.2.2.1.2 Simulated Alignments

Ten-thousand non-functionally divergent sites were simulated from a random ancestral root sequence under each of the 10 mixture model components. Sites were then sampled from these sets randomly to construct both the non-FD windows around divergent sites and the remainder of non-FD sites in an alignment. To generate functionally divergent sequence data, we simulated, from a shared ancestral sequence, over each subtree independently with a zero internal branch length between the subtrees. Subtrees of non-FD sites were separated by an internal branch length of one.

For all site types four discrete Γ site-rate categories were used based on an α shape parameter of 0.5.

Type I divergent sites (i.e. rate-shifted sites), were simulated using the standard WAG model of evolution. A root sequence was sampled from the WAG model frequencies and sequences for each subtree were simulated separately, with the same root sequence to allow for independence of rates. The simulated pool of sites was then filtered to remove all columns where an identical evolutionary rate was randomly assigned by INDELible.

For Type II divergent sites, 9 pairs of mixture-model components were selected from the mixture model such that amino acids with a high frequency in one component of the pair will have a low frequency in the other and vice versa. For each component pair we simulated 10,000 sites. Root sequences were randomly sampled from each of the two amino acid distributions of the components used. As for Type I sites, an alignment was simulated for each subtree independently with the same ancestral sequence. To simulate the effect of selection for differing physico-chemical properties, each branch in the subtree was allowed to evolve according to the proportional model (i.e. rates of interchange are proportional to the frequency of the target amino acid, similar to the CAT-Poisson model of Lartillot and Philippe, 2004) using pairs of the 10 component amino acid profiles selected as described above. The resulting 90,000 simulated sites were combined into a single pool of type II divergent sites. To accentuate the differences between subtrees, type I and type II simulated datasets were then filtered to remove any columns where the most prevalent amino acid in one subtree represented 30% or more of sites in its counterpart. Alignment columns were also sampled from the type I and type II pools to have appropriate Jensen-Shannon divergence score distributions comparable to one of five biological datasets. To accomplish this all site types (Type I FD, Type II FD, non-FD) were simulated in excess as described above and a subset was sampled so that proportions of Jensen-Shannon divergence scores in a given subset roughly matched those of the biological datasets. The final alignments used were

400 residues in length, 40 of which were functionally divergent sites (20 Type I and 20 Type II). Each functionally divergent site was given a window of three non-FD sites to either side in the final alignment to conform to the assumptions made by the GroupSim ConsWin (Capra and Singh 2008) method as described above.

3.2.2.1.3 Taxon Sampling

To test the impact of taxon sampling on the prediction of functional divergence, two phylogenetic trees were chosen to represent best and worst-case examples based on the performance of FunDi relative to other classifiers. Taxa were randomly re-sampled from these datasets in groups of 10, 15, 20, 25, 30, and 35 with the only constraint being that a minimum of 4 taxa were present in each subgroup. Ten replicate samplings were conducted for each number of taxa. A phylogenetic tree was then re-estimated from the data using RAxML version 7.2.6 (Stamatakis 2006b) and predictions of functional divergence made with each of the tested prediction methods.

3.2.2.1.4 Branch Length Scaling

The two tree topologies discussed above were again used as best and worst-cases to investigate the impact of branch lengths on the predictive performance of functional divergence detection. For each of the two trees the branches in the subtrees were re-scaled by a factor of 0.5, 1.5, 2, 3, 4, 5, or 10, or the internal branch separating the two subtrees was set to a length of 1.5, 3, 5, or 10. A simulated dataset was generated as described above on this new phylogenetic tree and evaluated using each of the chosen prediction methods.

3.2.2.2 Strategy II: Defined Motifs

The ‘evolutionary motif’ method in Indel-Seq-Gen version 2 (Strope *et al*. 2007; 2009) was also used. In brief, the sequence motifs of functionally divergent sites from select datasets used in prior performance evaluations (Chakrabarti *et al*. 2009; Chakrabarti and Panchenko 2009) were compiled for each of the two subgroups in a given family (**Table 3.2**).

Functionally divergent sites were constrained to the motifs found in the biological datasets selected as recommended by Strope *et al*. (2009). An ancestral character state for functionally divergent sites in both subtrees was randomly selected and evolved according to the differing motifs of the subtrees (*e.g.* Once a site fits the defined motif no mutation that does not fit the motif is tolerated at that position during the remainder of the simulation). As before, window sites were constructed surrounding each functionally divergent residue, but in this case were constrained to be 100% conserved in order to provide optimal conditions for the ConsWin windowing method and GroupSim. Non-functionally divergent sites for the remainder of the sequence length were simulated with INDELible using the 10-component amino acid profile mixture with WAG exchangeabilities, four Γ rate categories and an α shape parameter of 0.5 using the original tree from the protein family. Non-FD sites were simulated under each of these 10 components, 25 sites per component for 250 non-FD positions unconstrained in their conservation level. For each of the selected biological datasets 10 independent simulations were performed, to yield 70 simulated alignments over seven different phylogenetic trees. This simulation strategy allows true functionally divergent sites to be simulated based on known biologically derived parameters/motifs while removing the serious problem, in biological datasets, of undetected positives from being incorrectly labeled as negatives (false negatives) by simulating non-FD sites under substitution regimes that should correspond to neutral evolution (the standard WAG matrix for example).

3.2.3 Testing Divergence

For all programs, where appropriate, default values and raw scores were used to produce ordered lists with sites labeled zero or one according to whether they were a truly functionally divergent (1) versus a non-FD (0) site. When necessary, raw scores of programs were rescaled to be between zero and one, with high scores being indicative of functional divergence.

For Evolutionary Trace the Real-Value Evolutionary Trace score (rvET) (Mihalek *et al.* 2004; Yao *et al.* 2006) was re-scaled to be between 0 and 1, high scores being better (indicative of functional divergence), to be comparable and in the same format as the other predictors for AUC-PR and AUC-ROC calculation. The real-value evolutionary trace score combines both the evolutionary rank of a column in and alignment and the information entropy. For the Difference Evolutionary Trace (DET) (Lichtarge *et al.* 1996; Madabushi *et al.* 2004; Raviscioni *et al.* 2006) the rvET and ranks were compared between evolutionary trace runs on the whole tree and independently on each subtree. If the rank was lower in the whole tree than either subtree the site was given a score of zero. Otherwise, the rvET value from the lower ranking of either subtree was taken and rescaled as described above. Several other scoring schemes were tested to simulate DET as performed by biologists with comparable results.

For the likelihood ratio test (LRT) of Knudsen *et al*. (2001, 2003) rescaling was also performed to provide a single uniform score for every site between 0 and 1, with 1 being the "best" functional divergence score. Here we looked at the likelihood ratio statistic, U, for each hypothesis (Type I divergence, Type II divergence, Type I/II, and Slow/Fast) and compared it to the cut-off for that type. The category with the largest difference from the cutoff was chosen as the classification for that site. However, if U was not above the cut-off in this case, or Slow/Fast was the categorization, the score was set to 0. Otherwise, scores were re-scaled between 0 and 1 based on the minimum and maximum range of U for that category.

To evaluate overall performance receiver operator characteristic (ROC) and precision-recall (PR) curves, as well as the total Area under the curve (AUC) for both curves (AUC-PR and AUC-ROC) values, were calculated using AUCCalculator 0.2 (Davis and Goadrich, 2006). The AUC values each yield a single relative performance score for evaluation of the overall classification performance; the greater the AUC, the better the predictor averaged over all thresholds. An AUC value equal to one indicates perfect performance according to the criterion. We also used the 'average ranks' evaluation method that averages the rank of all true positive sites (ordered by decreasing FD score) in a tested dataset or series of datasets (in this case all 500 or 70 datasets for a simulation method). The lower the average rank the better the performance of the method. AUC values and

TABLE 3.2 THE 7 BIOLOGICAL DATASETS SELECTED FROM CHAKRABARTI *ET AL* (2007) TO SIMULATE ALIGNMENTS WITH DEFINED MOTIFS FOR SUBTREES USING INDEL-SEQ-GEN-V2. Datasets selected cover a range of alignment lengths, number of functionally divergent sites, taxa, and alignment lengths.

Set	# Subsites	# Taxa	Simulated Alignment Length
cbm9	7	19	300
cd00333	12	27	335
cd00423	4	33	279
cd00985	3	180	272
CNmyc	11	34	328
MDH-LDH	1	44	258
Nucleotidyl cyclase	2	49	365

calculated average ranks were then used to generate boxplots using the R statistical package. All programs were evaluated over the larger 500 alignment and tree set (simulation set 1) and the smaller 70 alignment set with motifs (simulation set 2) as well as a set of 11 biological datasets (Chakrabarti *et al*. 2007).

3.3 Results

We investigated and compared the performance of the FunDi methods and several other methods for functional divergence site prediction over a range of tree-topologies and sizes, identifying particular tree topologies that may prove problematic for prediction of functional divergence. The impact of taxon sampling (recovery of true molecular diversity), length of the branch separating the subtrees, and overall tree length was also investigated in order to build a robust picture of the behavior of the various functional divergence prediction algorithms and their performance over phylogenetically diverse data.

Several programs were selected based on their performance in previous studies (Capra and Singh 2008; Chakrabarti and Panchenko 2009; Brandt *et al*. 2010) as well as their ability to be used in a large-scale testing pipeline. We tested the performance of our own method, FunDi (using both QmmRAxML and RAxML for site log-likelihood calculation), FunDi+ConsWin, SPEER (Chakrabarti *et al*. 2007), GroupSim (Capra and Singh 2008), Sequence Harmony, and Multi-RELIEF. Sequence Harmony and Multi-RELIEF were both used as implemented in Multi-Harmony (Brandt *et al*. 2010). These programs represent the top performing methods as determined by prior studies and include both phylogenetic and information theoretic approaches. A likelihood ratio test method for functional divergence detection was also evaluated (Knudsen and Miyamoto 2001; Knudsen *et al.* 2003) on both the simulated and real biological datasets. However, as it had poor performance in initial tests we did not do a complete set of analyses.

Similarly, the Difference Evolutionary-Trace and Real-Value Evolutionary-Trace methods could only be carried out on the 11 biological datasets tested (see below).

Eleven two subfamily biological datasets were selected from Chakrabarti *et al*. (2007) and performance evaluated. The 11 datasets were selected with the requirement that each subfamily had to be phylogenetically distinct and contain a minimum of four taxa per subfamily. The datasets selected feature a broad range in terms of number of taxa and number of functionally divergent sites. The performance of most classifiers as measured by AUC-ROC was very similar, with more variation seen in the AUC-PR metrics (**Figure 3.1**). Using the medians of the AUC-PR distribution to judge the overall performance, GroupSim appeared to have the overall best performance with FunDi-ConsWin and Multi-RELIEF as the next best performers in the AUC-PR plots. Median performance as measured by AUC-ROC is slightly higher on these 11 biological datasets than observed under either of the two simulation conditions examined below, but not significantly, except in the case of SPEER in the case of simulated dataset 2. We also compared the performance of the Real-Value Evolutionary Trace (Mihalek *et al*. 2004) and Difference-ET methods (Madabushi *et al*. 2004; Raviscioni *et al*. 2006) here but due to technical constraints were unable to perform those evaluations on our larger simulated datasets.

In simulation Set 1 across all 500 datasets the performances of FunDi using either QmmRAxML or RAxML were highly similar, outperforming all other methods tested as measured by the area under the Precision-Recall curve (AUC-PR), the area under the ROC curve (AUC-ROC), and the average rank of true positive functionally divergent sites (**Figure 3.2**). The program GroupSim (Capra and Singh 2008) applies a simple windowing method (ConsWin) for adjusting scores of functional divergence based on neighboring residues in the primary sequence (described previously). We applied this same method to the posterior probabilities of functional divergence P(FD) generated by FunDi with QmmRAxML to see if it yielded an improvement and to ensure that our simulation

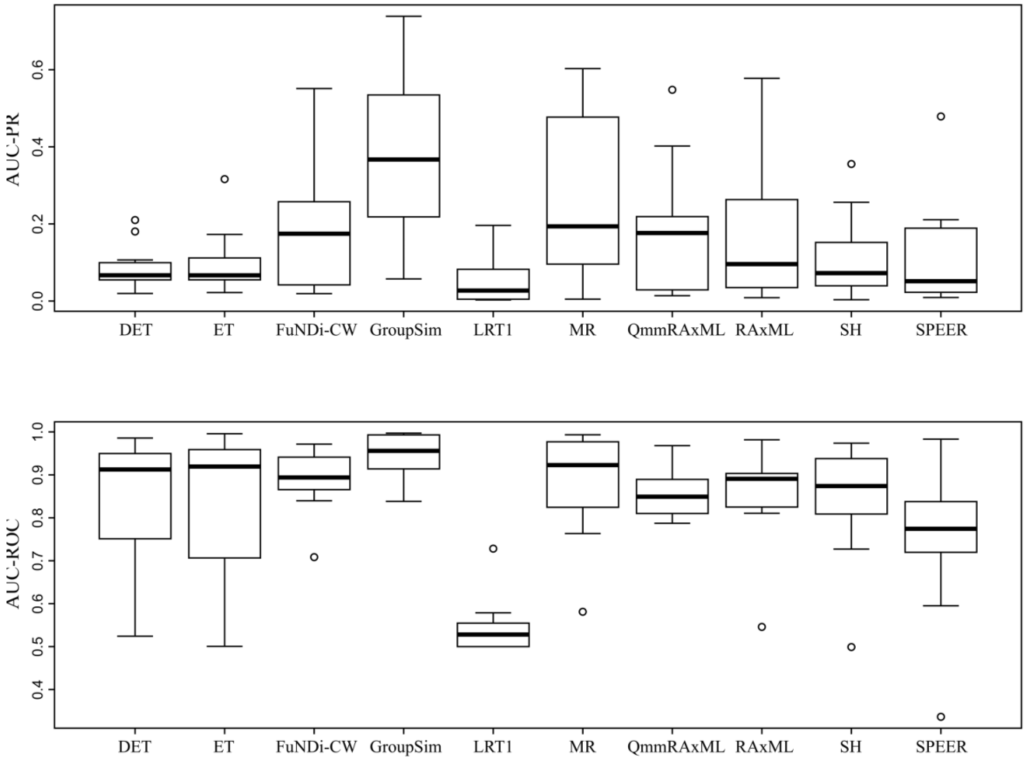


FIGURE 3.1 PERFORMANCE ON 11 BIOLOGICAL DATASETS TAKEN FROM CHAKRABARTI *ET AL* (2007) (See Table 3.3). ET (Evolutionary Trace) and DET (Difference-ET), and the Likelihood ratio test (LRT1) included. Results as measured by the area under the Precision-Recall (**A**) and ROC (**B**) curves. FunDi+CW is FunDi+QmmRAxML with the ConsWin windowing method applied as in GroupSim.

TABLE 3.3 THE 11 BIOLOGICAL DATASETS FROM CHAKRABARTI ET AL (2007) USED TO TEST PREDICTION PROGRAMS AND THE NORMALIZED TREE LENGTH AND SUBTREE LENGTHS.

Dataset	Sub-tree1	Sub-tree2	Difference
cbm9	0.3761	0.7081	0.3320
cd00264	2.0427	2.4408	0.3981
cd00333	0.9780	0.8154	0.1626
cd00365	0.7390	0.4668	0.2721
cd00423	0.8300	1.2568	0.4262
cd00985	0.5589	0.5831	0.0243
CNmyc	0.1815	0.0839	0.0977
MDH_LDH	0.2524	0.4172	0.1648
nucleotidyl_cyclase	0.5515	0.2626	0.2889
Rab56	0.0568	0.1348	0.0780
RasRal	0.2004	0.0422	0.15822
Tree A	0.3503	0.0508	0.2995
Tree B	0.2564	0.2596	0.0031

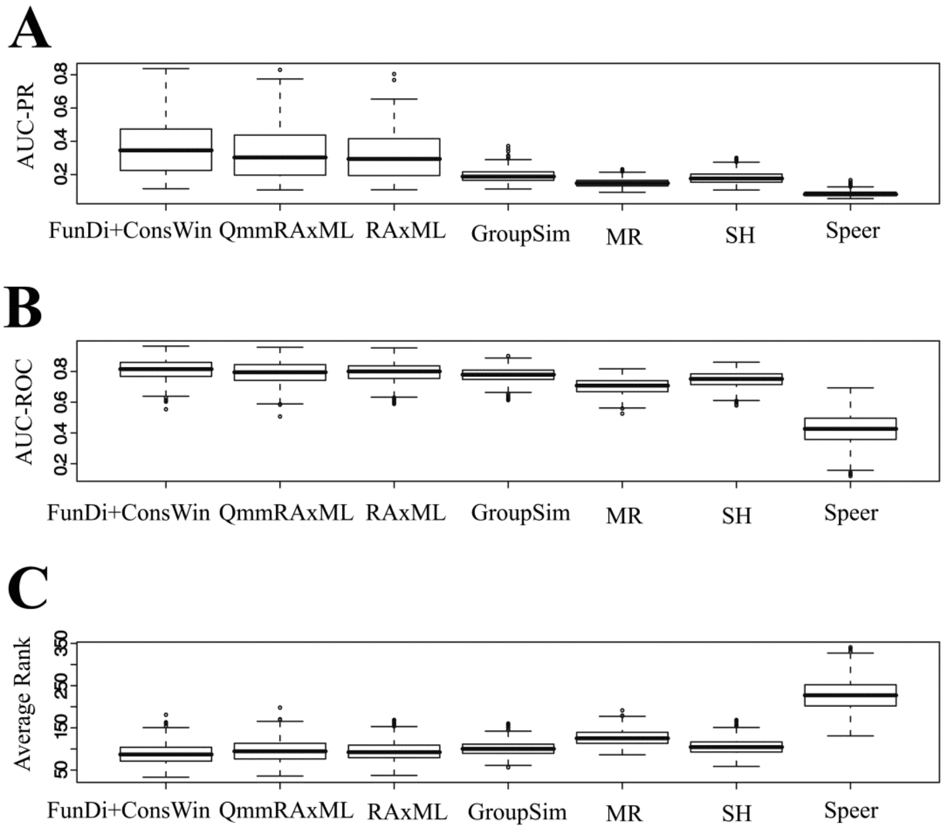


FIGURE 3.2 BOXPLOTS SHOWING PERFORMANCE OF SEVERAL FUNCTIONAL DIVERGENCE CLASSIFIERS ON 500 SIMULATED DATASETS AS MEASURED BY THE AREA UNDER THE PRECISION-RECALL (A), AND RECEIVER OPERATING CHARACTERISTIC (B) CURVES. Higher values reflect increased performance with a maximum value of 1.0. Additionally performance was characterized by the average rank of true positive functionally divergent sites (**C**) with sites ordered by the respective FD score of the program tested. All scores transformed (if required) to be between 0 and 1 with high scores reflecting a better functional divergence score. For Average Rank lower median values show increased performance. The methods evaluated in all cases are FunDi with QmmRAxML, FunDi with QmmRAxML and the ConsWin windowing method, FunDi with RAxML, GroupSim, Multi-Harmony (MR), Sequence-Harmony (SH), and SPEER. The 500 datasets were simulated with varying conditions over randomly generated tree topologies.

settings were appropriate for Group-Sim's prediction strategy. FunDi+ConsWin displayed an increase in predictive performance compared to the non-windowed P(FD) scores alone (**Figure 3.2**). GroupSim was the next best classifier after the three FunDi-based methods across all datasets. Surprisingly given previous studies (Chakrabarti and Panchenko 2009), SPEER appeared to have the lowest performance for all three scoring metrics. The distribution of AUC-PR values for FunDi was significantly different from the other predictors in all pairwise comparisons by the Wilcoxon Signed-Rank Test with a Bonferroni correction for multiple comparisons (p-value < 8.8e-16). All other tests were also significantly different from one another with Bonferroni corrected p-values << 0.05.

3.3.1 The Impact of Phylogenetic Tree Shape and the Number of Taxa

We also investigated performance of the classifiers as a function of several common phylogenetic tree shape statistics. The clearest trend indicated that performance was greatly influenced by normalized tree length (i.e. overall sum of branch lengths divided by the number of taxa) as shown in Figure 3.3. All programs that we examined exhibited some increase in AUC-PR as the normalized tree length increased; however, this trend was much stronger in the three variations of FunDi tested. As the normalized tree length increased the performance gap between FunDi and the other prediction programs increases, with the FunDi-based methods doing much better in general at longer tree lengths. We also identified two tree topologies with identical normalized tree lengths (NTL = 0.18) where the performance of FunDi differs significantly (**Figure 3.4**). For the tree in figure 3.4A, FunDi performs poorly (AUC-PR=0.33) whereas in the tree in Figure 3.4B, it performs much better (AUC-PR = 0.51). Curiously, the AUC-PR results under the poorly performing tree are nearly indistinguishable from those of Group-Sim (second best performing method overall), while there is a large difference in AUC-PRs under the ‘high-performance’ tree. These two topologies differ mainly in their tree shapes, with a

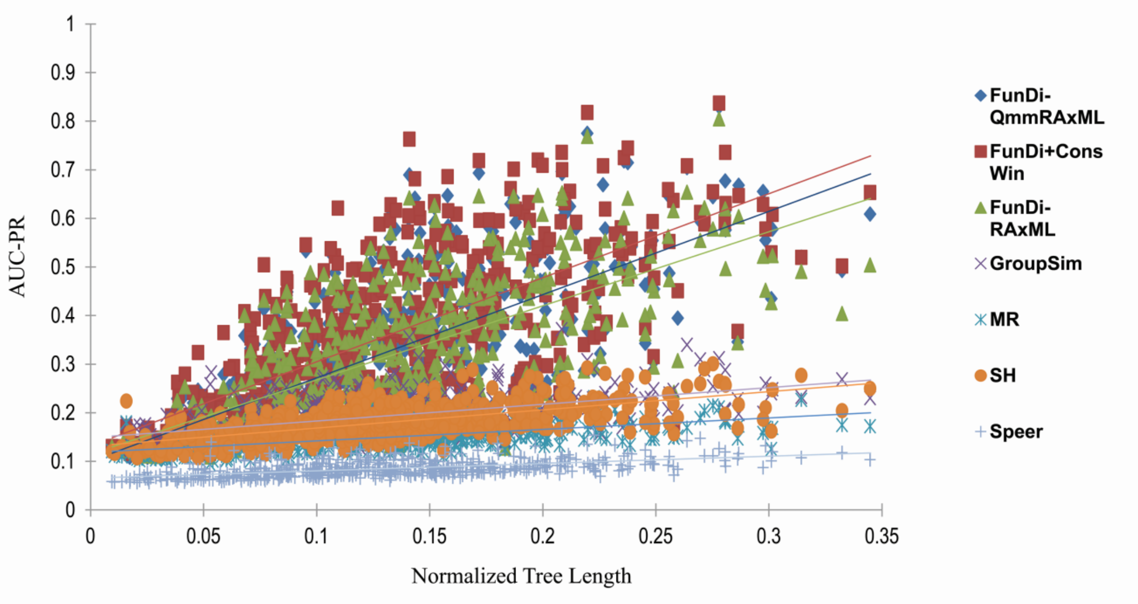


FIGURE 3.3 PERFORMANCE (AREA UNDER THE PRECISION RECALL CURVE) VERSUS NORMALIZED TREE LENGTH (TOTAL SUMMED LENGTH OF ALL BRANCHES IN THE PHYLOGENETIC TREE DIVIDED BY THE NUMBER OF TAXA) FOR SEVERAL FUNCTIONAL DIVERGENCE CLASSIFIERS EVALUATED OVER 500 RANDOMLY SIMULATED DATASETS. Linear trend lines for the data points are also shown for each classifier. Larger normalized tree lengths result in increased predictive performance, particularly for the three versions of FunDi tested here.

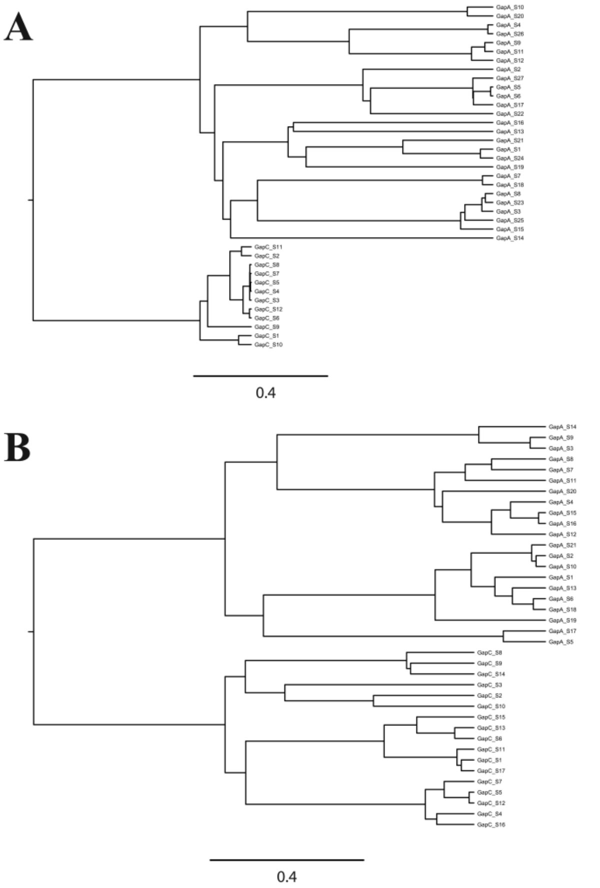


FIGURE 3.4 TREES USED FOR SIMULATIONS THAT LED TO (A) POOR PERFORMANCE OF FUNDI (RELATIVE TO OTHER CLASSIFIERS) AND (B) GOOD PERFORMANCE. These trees were selected for further analyses of the impact of taxon sampling and branch length re-scaling as best and worst-case examples of phylogenetic tree shapes, balance, and differences between subtrees. Both tree topologies have an identical normalized tree length of 0.18. Performance of FunDi on tree topology A (as measured by the Area Under the Precision Recall Curve) was 0.33, while it was 0.51 for tree topology B.

large discrepancy between the branch lengths in the subtrees. These two tree topologies are best- and worst-case examples for FunDi and were selected for further analyses on the effect of branch length and taxon sampling on functional divergence prediction. While large discrepancies between branch lengths of subtrees in datasets with functional divergence is not uncommon, in this case a large branch length discrepancy is compounded with relatively short branches throughout the tree when compared to similar trees from 11 biological datasets examined (**Table 3.3**). This may explain why performance increases so dramatically when branches in the subtrees are made longer (See Below).

3.3.2 The Impact of Taxon Sampling

To test the effects of taxon sampling on functional divergence prediction, random taxon subsets were created from the simulated datasets from trees A and B in Figure 3.4. Results are shown in figure 3.5. For both trees, addition of more data in the form of additional taxonomic coverage improved the performance of FunDi relative to other tested methods, although the trend is much more pronounced for tree B. On the other hand, information theoretic methods appear to be relatively insensitive to taxon sampling, showing only moderate performance increases (or some apparent decreases, e.g. 3.5B).

3.3.3 The Impact of Branch Length Scaling

We also investigated the impact of branch lengths on performance
using trees A and B as examples. We present only the results for AUC-PR as they showed the clearest trends. For both trees, either the branch lengths in the subtrees or the internal branch length separating the two subtrees were rescaled and for each case, a dataset was simulated and tested. The branch length effect is dependent on which

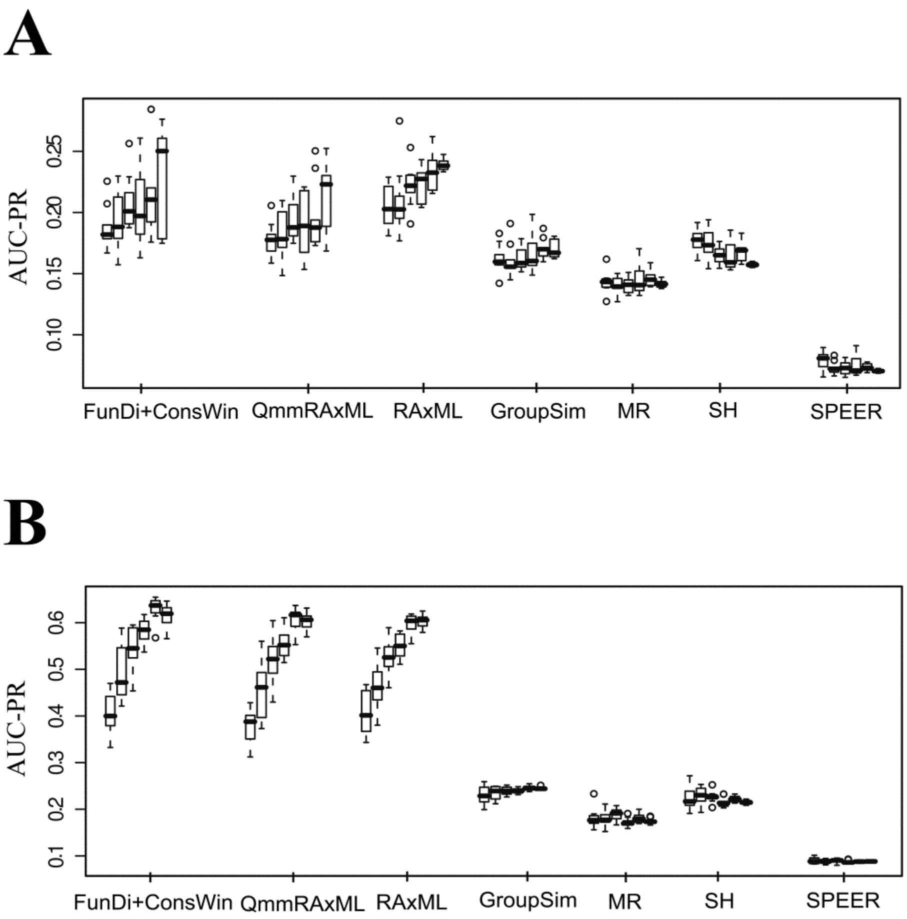


FIGURE 3.5 BOXPLOTS SHOWING THE IMPACT OF TAXON SAMPLING ON PERFORMANCE AS MEASURED BY THE AREA UNDER THE PRECISION-RECALL CURVE (PANELS A AND B) ON TREE TOPOLOGIES A (A) AND B (B) FROM FIGURE 3. For each tree, 10 sub-sampled replicates of 10, 15, 20, 25, 30, or 35 taxa were constructed and the performances of each of the listed classifiers assessed. AUC-ROC results are shown in Appendix B3, Supplementary Figure 3.3

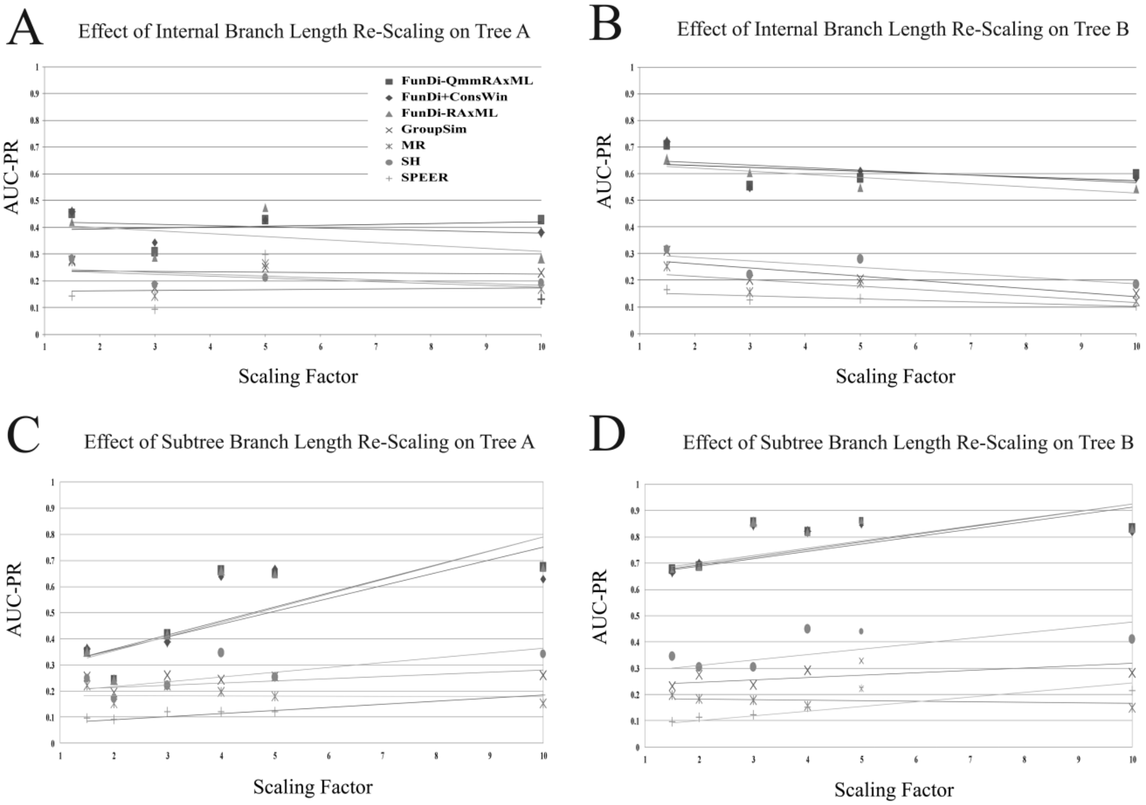


FIGURE 3.6 IMPACT OF BRANCH LENGTH RE-SCALING ON INTERNAL (A,B) OR SUB-TREE (C,D) BRANCHES AS MEASURED BY THE AREA UNDER THE PRECISION-RECALL CURVE.

Re-scaling was applied to both tree topologies A and B from figure 3. Scaling factors are shown on the x-axis with AUC-PR scores on the y-axis. For each of the two topologies re-scaling was applied to the indicated branch length(s) and a random dataset simulated and performance of the tested functional divergence classifiers assessed as described in the Methods.

branch is being re-scaled (**Figure 3.6**), as well as the given tree. Increasing the internal branch (panels A and B) separating subtrees results in a decrease in predictive performance for all classifiers for tree B. For tree A, it is difficult to discern a clear trend as the performances of most classifiers do not change dramatically, although at the longest branch length setting (10) all methods do generally poorer than at shorter lengths. This general effect is expected in terms of the performance of FunDi; as the internal branch between subgroups increases, the whole tree becomes closer and closer to the independence model, decreasing the distinction (and separatability) of the two components of the mixture model.

In the case of varying branch lengths in subtrees (**Figure 3.6C** and **D**), AUC-PR clearly increases for all classifiers and on either tree as the branch lengths are increased. The trend is most dramatic for all three versions of FunDi relative to the other programs. As each subtree is evolving under its own evolutionary model, longer branches in subtrees provide additional time for substitutions that allow discrimination between the two site types (functionally divergent versus non-FD) to appear (if such a substitution did not occur along the internal branch) and provide more information upon which a classification can be made. This may be particularly true for Type I functional divergence as longer branches lead to more scrambling of the amino acid states at that site in the subtree with relaxed selective pressures. In tree A (the 'poorly' performing tree topology) we see rapid increases in performance for the FunDi-based methods as branch lengths increase.

3.3.4 Prediction of Type I versus Type II FD Sites

We evaluated whether there were differences in the ability of methods to predict Type I versus Type II FD sites. While better performance is observed for Type II sites (**Figure**

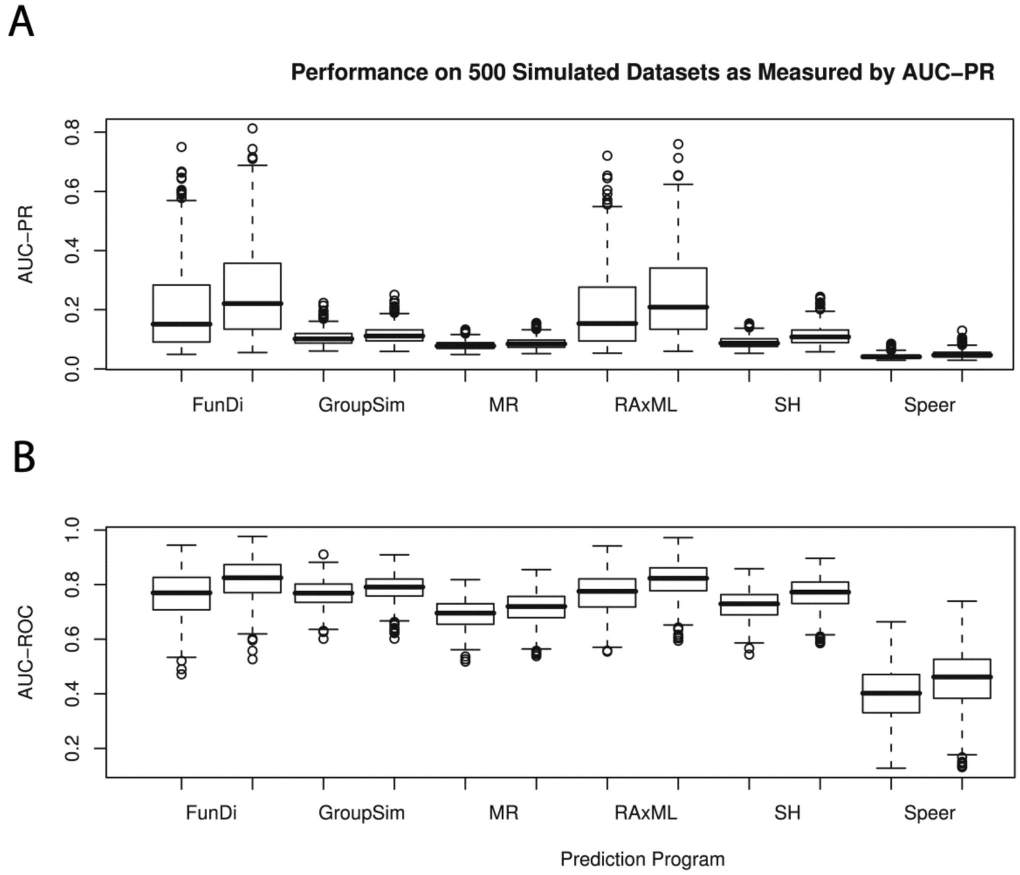


FIGURE 3.7 PERFORMANCE DIFFERENCES BETWEEN TYPE I (LEFT) AND TYPE II (RIGHT) SITES FOR EACH OF THE TESTED PREDICTION PROGRAMS ACROSS 500 RANDOMLY SIMULATED DATASETS AS MEASURED BY AUC-PR (A) AND AUC-ROC (B).

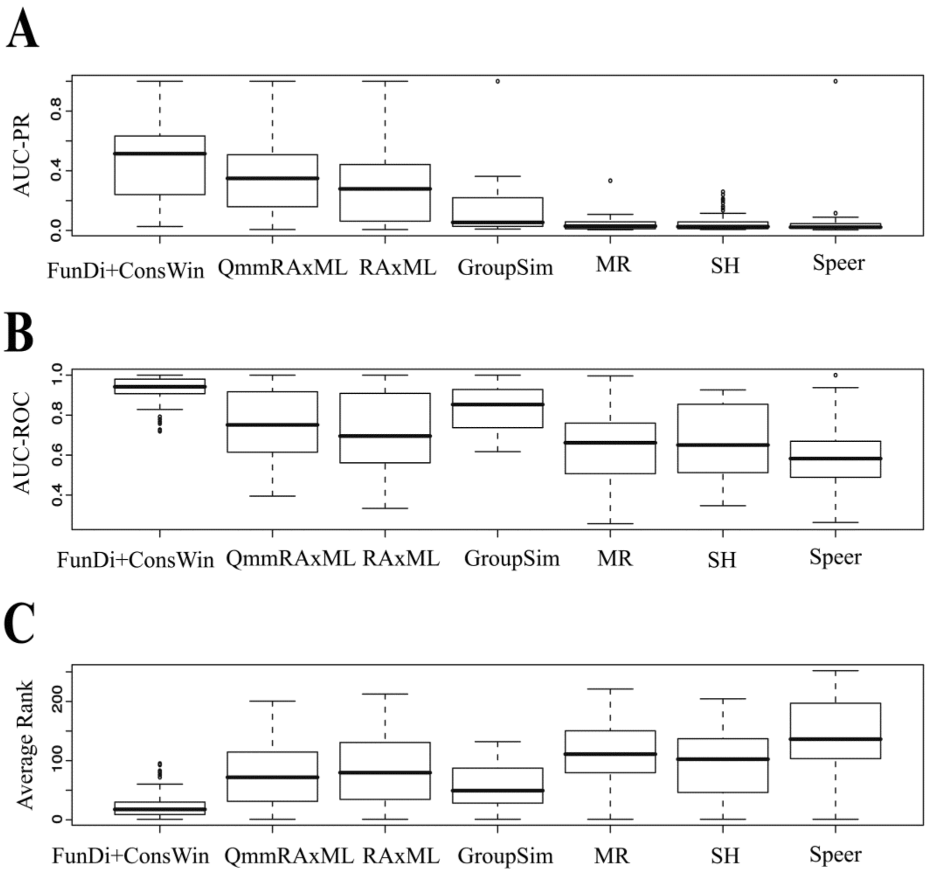


FIGURE 3.8 BOXPLOTS OF PERFORMANCE OF FUNCTIONAL DIVERGENCE CLASSIFIERS ON 70 DATASETS SIMULATED WITH DEFINED MOTIFS USING INDEL-SEQ-GEN-V2 AS MEASURED BY AREA UNDER THE PRECISION-RECALL (A) AND RECEIVER OPERATOR CHARACTERISTIC (B) CURVES AS WELL AS THE AVERAGE RANK (C). The 70 simulated datasets are simulated with 10 replicates over each of seven tree topologies (and with defined motifs) from real biological datasets from Chakrabarti *et al*. (2007 and 2009).

3.7), the difference is not great as median values for AUC-PR or AUC-ROC metrics for Type I or Type II sites fall within the other site type's interquartile range.

3.3.5 Performance With Defined Evolutionary Motifs

The functional divergence prediction programs were tested using the same performance metrics on the 70 alignments from simulated dataset 2, which used the 'defined evolutionary motifs' simulation strategy. Overall, the same general trends in relative performance are observed. The median AUC-PR was highest for FunDi (RAxML, QmmRAxML, and QmmRAxML + ConsWin); with GroupSim the next best performing prediction method (Figure 3.8A). The large range of performance scores observed with the AUC-PR data can best be explained by the varying performance on individual motifs and phylogenetic tree shapes used (data not shown). When evaluated using AUC-ROC, FunDi + ConsWin was the best performing prediction method, followed by GroupSim, then by FunDi (QmmRAxML or RAxML) without the windowing method (**Figure 3.8B**). Performance of FunDi (AUC-PR values) is significantly different from other predictors as measured by the Wilcoxon Signed-Rank Test. (p-value < 2.65e-12). If non-FD windows that are more highly conserved compared to the majority of non-FD sites do tend to be located around FD sites, a windowing method such as ConsWin is of clear benefit, regardless of the testing methodology, as described previously (Capra and Singh 2008).

3.4 Discussion

Although not dramatic, there is an apparent discrepancy between the performance results for the 11 real datasets versus those of the two simulation studies. GroupSim does best overall for the real datasets with FunDi-based methods amongst the next best performers whereas in the two kinds of simulations, FunDi-based methods (especially FunDi+ConsWin) typically outperform GroupSim and other methods. The source of the discrepancy is not very clear, but we would suggest that all of the performance metrics are inherently less trustable for the 11 real data sets because, for these, only the true positive class of sites is known with certainty. As the relative rank of performance of the various methods is similar over two completely distinct functional divergence simulation settings and multiple performance indicators, we believe that the simulation results are more representative of the true performance properties of the various methods.

While FunDi performs better overall, its predictive power, as measured by precision-recall curves, remains relatively low (although performance as measured by ROC curves appears quite strong). This low predictive power is due to a variety of factors. Even under our simulation conditions, some functional shifts may result in relatively subtle amino acid substitutions, especially in complex situations that are an apparent mix of type I and type II functional divergence types, or what have been termed ‘Marginally Conserved’ sites (Chakrabarti *et al*. 2007). Marginally conserved sites, in this context, refer to sites that do not confirm exactly to canonical Type I or Type II site-patterns. It is these marginally conserved sites that prove to be the most difficult in terms of prediction. In addition, the majority of existing approaches essentially search for particular patterns of amino acid usage, patterns which can arise in an evolutionary context due simply to stochastic neutral changes over the underlying phylogeny of the protein family without any functional shift occurring, making adjustments for the underlying phylogeny of great importance. FunDi can also leverage improved models of amino acid evolution, such as the 10 component amino acid profile mixture models implemented in QmmRAxML (Wang *et al*. 2008) using WAG (used in this study), JTT, LG, or user-supplied exchangeabilities. Judicious model selection allows the incorporation of some prior knowledge of the protein's evolutionary history, structure, function, and amino acid frequencies.

We have also introduced two new simulation strategies for functional divergence that are useful for benchmarking new prediction programs, and improving existing ones, especially in phylogenetic "trouble spots" (e.g. worst-case tree used in our analyses of the impact of branch lengths and taxon sampling). Unfortunately, we were unable to compare our results with some of other phylogenetically-based methods such as DIVERGE (Gu 1999; Gu 2001; Gu and Vander Velden 2002) because the software implementation and run-times of the latter precluded analyses of large simulated datasets. Furthermore, the Gu 2001-based predictions could not be obtained for several of the biological datasets examined. Since FunDi is: i) scalable to the analysis of multiple (potentially thousands) of large protein datasets. ii) has a single coherent framework for the prediction of type I and type II functionally divergent sites and iii) can be used with any phylogenetic model of protein evolution implemented in a maximum likelihood framework, it has distinct advantages as compared to other phylogenetic-based functional divergence predictors currently in use.

Our analysis on a large, phylogenetically diverse set of simulated functionally divergent datasets shows that taking into account the phylogeny of a protein family is an important part of the prediction of functionally divergent sites. While non-phylogenetically aware prediction schemes such as GroupSim can be characterized as insensitive to issues of taxon sampling and phylogenetic tree topologies, they also do not increase in predictive accuracy under appropriate phylogenetic conditions and show generally poorer performance under a wide range of conditions. FunDi, as a phylogenetically aware prediction program, shows marked improvement in the quality of its predictions under increased taxon sampling (recovery of true biological diversity) as well as increased evolutionary time as measured by the normalized tree length and illustrated in our branch length re-scaling experiments.

The main problem with using real biological data to evaluate the performance of functional divergence methods is the infeasibility of experimentally testing false and true negative prediction; there are simply too many sites and too many character state combinations to test comprehensively. The two simulation strategies described here therefore provide much cleaner data, with less noise than true biological data and so can be used to evaluate the performance of functional divergence methods over a wide range of possible evolutionary conditions such as tree topologies and taxon sampling. The ability to specify particular sequence motifs for functionally divergent residues based on observed biological data, as we have done here in the second set of simulations, may be useful in developing methods that have better performance on difficult-to-classify functionally divergent residues.

3.5 Acknowledgements

DG would like to thank William Fletcher for implementing several suggested changes to INDELible, to BW Brandt for providing a script for the Multi-Harmony web-server that allowed testing of a large number of datasets, and to Olivier Lichtarge and Angela Dawn Wilkins for running real value ET on the 11 Biological datasets

Funding: DG is supported by a Nova Scotia Health Research Foundation graduate student research award. This work was supported by an NSERC Discovery Grant awarded to AJR

3.6 Author Contributions

DG programmed all scripts, designed FunDi, designed and performed experiments, and wrote manuscript. AJR conceived of project and the two-component mixture model for functional divergence. ES provided advice and support for statistical tests. AJR and ES provided editorial comments. All authors agreed on final form of manuscript.

Chapter 4 Analysis of Functionally Divergent and Convergent Evolution in the Plastid-Targeted Glyceraldehyde-3-Phosphate Dehydrogenases of the Archeaplastida and Chromalveolata

4.1 Introduction

4.1.1 Glyceraldehyde-3-phosphate Dehydrogenase

Cytosolic glyceraldehyde-3-phosphate dehydrogenase (GAPDH) reversibly catalyzes the sixth step of glycolysis, the conversion of glyceraldehyde 3-phosphate to D-glycerate 1,3-bisphosphate reducing the coenzyme NAD+ to NADH in the process (Cerff & Chambers 1978; Ferri *et al.* 1978). Structurally, cytosolic GAPDH canonically functions as a homotetramer with each monomer composed of two domains, the N-terminal coenzyme binding domain, which includes the first 147 amino acids as well as the final 20 (314 to 334), and the C-terminal catalytic domain, including the catalytic cysteine 149 (residue numbering is based on the *Bacillus stearothermophilus* homolog as in (Biesecker *et al.* 1977)). The catalytic domain contains the P_s and P_i sites, which bind the $C_{(3)}$ phosphate of the substrate and the inorganic phosphate ion respectively during the phosphorylation step carried out by the enzyme (Moras *et al.* 1975). Another important structural feature, the S-loop (177 to 203), folds over in close proximity to the bound cofactor. A structural overview is shown in **Figure 4.1**.

Archeaplastida (land plants, green algae, red algae, and glaucophytes) are eukaryotes that have acquired their chloroplasts via primary endosymbiosis of a cyanobacterium (Palmer 2003; Reyes-Prieto *et al.* 2007; Gould *et al.* 2008) and have a plastid-targeted GAPDH, homologous to the Gap2 of cyanobacteria (Brinkmann *et al.* 1987). This plastid GAPDH is active in the Calvin cycle and the fixation of CO_2 during photosynthesis where it preferentially carries out the reverse of the glycolytic reaction (Henze *et al.* 1995).

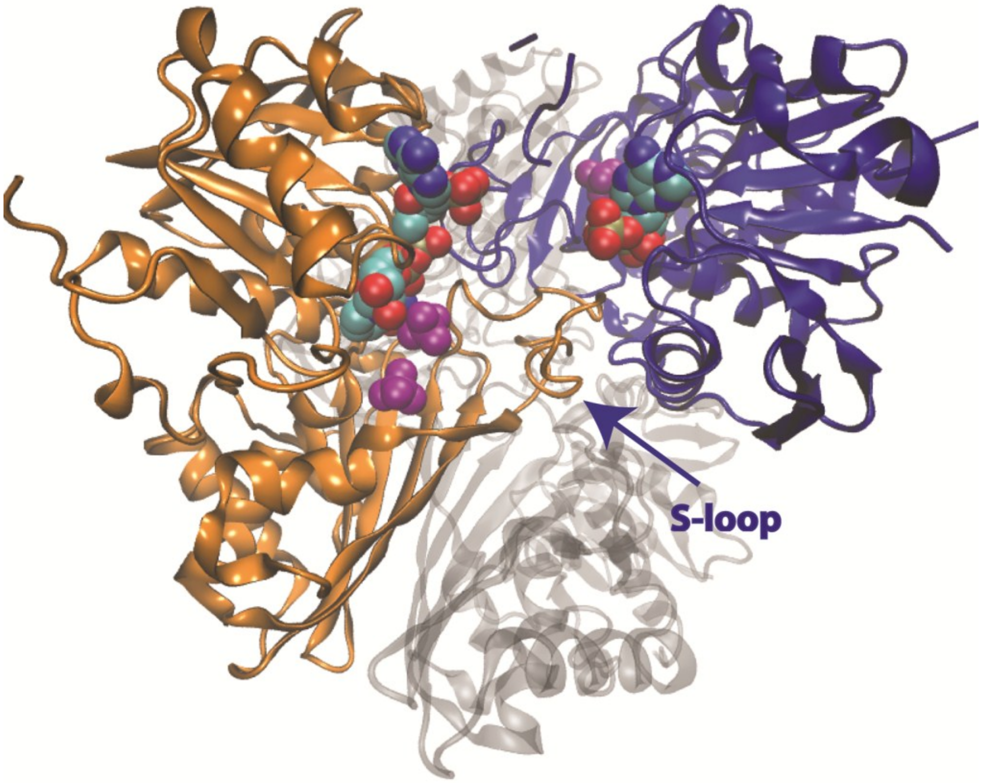


FIGURE 4.1 QUATERNARY STRUCTURE OF THE GAPDH A2B2 HETEROTETRAMER (PDB: 2PKQ) FROM SPINACIA OLERACEA (SPINACH). Crystallographic subunits O (GapB) and R (GapA) are show in blue and orange respectively, with the S-loop of the R subunit indicated. Bound NADPH in the co-enzyme binding domain drawn as a van der Waals space-filling model coloured by element type according to the Visual Molecular Dynamics (VMD) program. Sulfate ions in the P_s and P_i in purple, also as space-filling models. The O and R subunit are related by the R axis of symmetry. The P and Q subunits are shown in the background.

Unlike its glycolytic homologs, the plastid-targeted GAPDH, GapA, has a dual coenzyme specificity with a marked preference for binding (and oxidizing NADPH) in the Calvin cycle reaction, reducing 1,3 biphosphoglycerate (1,3BPGA) and producing glyceraldehyde-3-phosphate (G3P). In another group of photosynthetic eukaryotes, the Chromalveolata, a duplicate copy of cytosolic GAPDH (GapC1) was re-targeted to the chloroplast (Liaud *et al.* 1997; Fagan *et al.* 1998; Fast *et al.* 2001). This enzyme also features dual coenzyme specificity with a preference for NADPH when functioning in the Calvin cycle. Plastid targeted forms of GAPDH can also participate in the oxidative pentose phosphate pathway (OPP) (Buchanan 1980, 1984; Plaxton 1996); because of these dual-roles, NADPH-dependent GAPDH must be differentially regulated during light and dark cycles to prevent futile enzymatic cycling. This regulation is thioredoxin-mediated through CP12, a small intrinsically unstructured protein that forms a protein complex with both GAPDH and ribulophosphokinase (RPK) (Wedel & Soll 1998; Graciet *et al.* 2004; Marri *et al.* 2005, 2008; Lebreton *et al.* 2006; Trost *et al.* 2006; Erales *et al.* 2009).

4.1.1.1 GapA and GapB

GapA is the plastid-targeted isoform of GAPDH shared by all members of the Archeaplastida and is the most closely related to the Gap2 sequence of cyanobacteria. GapB is a land plant specific duplicate of GapA (Petersen *et al.* 2006a) with a C-terminal extension consisting of approximately 30 amino acids that is homologous to the CP12 regulatory enzyme (Pohlmeyer *et al.* 1996). In land plants, active GAPDH in the plastid is primarily found the form of A_2B_2 tetramers (Ferri *et al.* 1978; Cerff 1979; Cerff & Chambers 1979) although A_4 tetramers are also known (Cerff 1979; Sparla *et al.* 2004). Crystal structures for the A_2B_2 tetramer and mutant A_4 tetramers (Fermani *et al.* 2001; Sparla *et al.* 2004) have been solved, including fragments of the C-terminal extension of GapB that is homologous to CP12. Comparisons with crystal structures of GapC have yielded insights into the structural changes required for the discrimination between

NADH and NADPH cofactor binding and the CP12 interactions necessary to form the CP12/GAPDH/PRK supramolecular complex (Gardebien *et al.* 2006; Lebreton *et al.* 2006; Marri *et al.* 2008; Groben *et al.* 2010).

4.1.1.2 GapC1

The Chromalveolata is a proposed monophyletic 'super-group' of microbial eukaryotes comprised of the stramenopiles, haptophytes, cryptophytes, and alveolates (Cavalier-Smith 1999, 2002, 2003). Several lineages within each of these monophyletic groups contain a red-algal derived plastid of secondary endosymbiotic origin (Daugbjerg & Andersen 1997; Douglas & Penny 1999; Oliveira & Bhattacharya 2000; Zhang *et al.* 2000; Archibald *et al.* 2001). The presumed rarity of successful secondary endosymbiotic integration, along with several molecular phylogenies, including that of GapC1, have been used as support for this 'chromalveolate hypothesis' (Fast et al. 2001; Harper and Keeling 2003; Patron et al. 2004; Petersen et al. 2006b; Yoon et al. 2002). Recent work by Takeshita and colleagues (2009) has shown that extensive LGT within this group produces a phylogeny inconsistent with the presumed organismal phylogeny and may make its usefulness for determining chromalveolate monophyly less clear. In addition, various phylogenetic studies recover support for alternative groupings that modify or exclude Chromalveolata monophyly, such as the SAR (stramenopile, alveolate, and rhizaria) clade (Burki *et al.* 2007, 2008, 2009) and the inconsistent placement of the cryptophytes and haptophytes (Hackett *et al.* 2007; Patron *et al.* 2007). A recent multi-gene phylogenetic analysis based on comparisons of support for the chromalveolate hypothesis between plastid, mitochondrial, and nuclear sequences favoured an explanation involving multiple serial eukaryote-to-eukaryote enodymbiotic events over common linear descent of the chromalveolates (Baurain *et al.* 2010).

Regardless of whether the chromalveolate taxa truly form a monophyletic group, the transition of GapC1 from a cytosolic NADH-binding GAPDH to a plastid-targeted, NADPH-dependent GAPDH appears to have occurred only once, and GapC1 sequences are monophyletic. While the evolutionary history of the red-algal derived plastid and their host organisms is likely very complex, we expect there to have been relatively uniform evolutionary pressures on GAPDH function within the plastid in some respects. However, because the host organisms have widely varying habitats and metabolic lifestyles, there appear to be many differences in terms of GapC1 regulation (discussed below). GapC1 represents a secondary event of plastid targeting (if we consider GapA and GapB as the primary event) and also a secondary case of adaptation to NADPH-dependent function. Because the organisms within which it is found are comparatively less well studied, correspondingly less molecular data exists regarding GapC1 function and regulation, and fewer sequences from this subfamily are available in public databases. Like GapA and GapB, GapC1 shows dual specificity for both NADH and NADPH but with a preference for NADPH (Liaud *et al.* 1997). A large scale comparison of GAPDH and PRK regulation in algae including chromalveolates (Maberly *et al.* 2010) has shown variation within the latter group, and between chromalveolates and other algae with canonical GapA plastid-targeted sequences. Unfortunately, many of the species for which these studies of regulation have been completed are not the same species for which we have molecular sequence data for GapC1. In this chapter, we analyze the functional divergence and shifts in selective pressures of GapA/B and GapC1 relative to GapC within this enzyme's complex phylogenetic distribution.

4.1.2 Regulation of NADPH-Dependent GAPDH in Chloroplasts

Except in the chloroplasts of the Streptophyta (land plants), NADPH-dependent GAPDH typically exists as a homotetramer of GapA subunits. Redox regulation of NADPH-dependent GAPDH activity by CP12 is ancestrally inherited from cyanobacteria (Pohlmeyer *et al.* 1996; Wedel & Soll 1998) and is widespread throughout photosynthetic eukaryotes, including some chromalveolates (Boggetto *et al.* 2007; Erales *et al.* 2008; Maberly *et al.* 2010). Under light conditions, thioredoxins are constantly reduced by ferredoxin-thioredoxin reductase, which in turn, reduces target proteins such as CP12 (Buchanan 1980, 1984; Martin *et al.* 2004; Buchanan & Balmer 2005). The redox state of CP12 has a significant impact on its structure; oxidation of CP12 results in the formation of two internal disulphide bridges, reducing the overall disorder of CP12 and increasing the helical content (Graciet *et al.* 2003), although CP12 still remains highly disordered. Oxidized CP12 then acts as a scaffold, forming a supramolecular complex of GAPDH and phosphoribulokinase (PRK) whereby two molecules of CP12, two GAPDH homotetramers, and two of PRK form an inactivated complex, with each CP12 molecule binding a single PRK and a GAPDH homotetramer (Wedel *et al.* 1997; Wedel & Soll 1998; Graciet *et al.* 2003). Reduced thioredoxin, in turn, reduces CP12, which releases PRK from the complex although CP12 is still bound to GAPDH, which is only fully activated in the presence of the substrate 1,3-biphosphoglycerate and when NADPH to NADP+ ratios shift in favour of NADPH (Wolosiuk & Buchanan 1978; Wedel & Soll 1998). In the streptophytes, the C-terminal extension (CTE) of GapB fulfills the regulatory role of CP12, mediating the aggregation of A_2B_2 tetramers in to higher order, inactive A_8B_8 complexes (Zapponi *et al.* 1993; Baalmann *et al.* 1996). The formation of higher order complexes of GAPDH, and its enzymatic activity are both tightly controlled by light conditions and substrate availability, at least in the Archeaplastida.

In contrast, GAPDH regulation in members of the chromalveolates is less clear. In the diatoms *Odontella sinensis* (Michels *et al.* 2005), *Thalassiosira pseudonana*, and *Phaeodactylum tricornutum* (Gruber *et al.* 2009) the entire oxidative pentose phosphate pathway appears to be cytosol localized instead of occurring in the plastid. On the other hand, in the freshwater diatom *Asterionella formosa* no regulation of PRK was observed in cellular extracts (Boggetto *et al.* 2007) but GAPDH regulation appears to function much like that in *Chlamydomonas reinhardtii* via CP12 (Boggetto *et al.* 2007; Erales *et al.* 2008). Additionally, putative CP12 homologs have been identified by bioinformatic means in the chromalveolates *Thalassiosira pseudonana* and *Emiliania huxleyi* (Groben *et al.* 2010). Bioinformatic annotation and experimental studies fusing putative N-terminal targeting sequences from thioredoxins with green fluorescent protein in *Phaeodactylum* have also identified several plastid-targeted thioredoxins (Weber *et al.* 2009), including two canonically cytosolic proteins that appear to be localized to the periplastidal space (i.e. in the space between the innermost two membranes and the outer third membrane). Given their complex evolutionary history and acquisition of plastids by secondary endosymbiosis, it is reasonable to expect large differences in plastid associated metabolism and regulation, including the regulation of GAPDH function.

Recently, a large-scale analysis of both PRK and GAPDH regulation was carried out on cellular extracts in a phylogenetically broad manner, including 12 members of the Chromalveolata (Maberly *et al.* 2010). GAPDH was activated under reducing conditions except in *Thalassiosira pseudonana* (marine diatom) and *Alexandrium minutum* (marine dinoflagellate), although the GAPDH of those organisms was activated when both reducing conditions and NADPH were present. Addition of NADPH alone to the extracts tended to result in inhibition in the case of the chromalveolates, except in the cryptophye alga *Hemiselmis rufescens*, and had no effect in the tested archeaplastidian species with the exception of *Staurastrum cingulum* which was strongly inhibited. In the presence of CP12 from *Chlamydomonas reinhardtii* GAPDH and PRK were both inhibited in *Pseudocharaciopsis ovalis*, and *Asterionella formosa* but not in the other chromalveolates tested. However, supramolecular complexes involving GAPDH and CP12 as found in the Archeaplastida have only been characterized in the freshwater diatom *Asterionella formosa* (Boggetto *et al.* 2007; Erales *et al.* 2008). It appears that, based on these experimental studies, light/dark regulation of GAPDH activity may be less tightly controlled and overall more differentiated within the chromalveolates, with large differences especially between marine and freshwater species. Additionally, not all members of the Chromalveolata are photosynthetic and some, such as the apicomplexan parasites (e.g. members of the genus *Plasmodium*) contain a relict plastid known as the apicoplast. Even among photosynthetic chromalveolates some taxa, such as the dinoflagellates may be mixotrophic predators of other unicellular organisms as well as photosynthesizing.

4.2 Methods

4.2.1 Dataset Construction

To build a phylogenetically broad dataset of GAPDH sequences all eukaryotic sequences from the Reference Sequence (RefSeq) database of the National Center for Biotechnology Information (NCBI) annotated with the keyword glyceraldehyde-3-phosphate dehydrogenase were downloaded and clustered at the 90% identity level using UCLUST (Edgar 2010), an alternative to the program CD-HIT (Cluster Database at High Identity with Tolerance) (Li *et al.* 2001, 2002; Li & Godzik 2006). UCLUST is part of package of programs including USEARCH and UBLAST (Edgar 2010) which offer various heuristic speed-ups and improved performance over BLAST searches. UCLUST clusters sequences based on percent identity cut-offs as determined by an all-versus-all USEARCH run given a set of sequences. A random sequence is then selected as a representative of the cluster. Additional non-RefSeq sequences from various microbial lineages were also assembled; these included sequences specifically annotated as GapA or GapB based on a prior phylogenetic analysis (Petersen *et al.* 2006a), GapC1 sequences from the dataset of Takeshita and colleagues (Takishita *et al.* 2009), cyanobacterial Gap2 sequences from Petersen *et al* (2006), and additional cyanobacterial GapC (cytosolic) homologs yielding a total of approximately 2000 sequences. Although these additional sequences were not clustered like the RefSeq sequences, any 100% identical sequences were screened out, preferentially retaining any sequences currently annotated as Gap2, GapA, GapB, or GapC1 from the studies cited to retain appropriate annotation data. This initial set of sequences was aligned with the Fast Statistical Alignment (FSA) program (Bradley *et al.* 2009) because of its speed and relative accuracy when aligning large groups of sequences. The preliminary multiple sequence alignment was then trimmed using AliMask-CS, an in house alignment masking script described in more detail in Chapter 2, with default parameters. After alignment and trimming, any sequences that were not well aligned (because of possible misannotations or pseudogenes), covered less than 75% of the trimmed alignment (fragments or pseudogenes), or that were 100% identical to other sequences in the dataset were removed, leaving 490 sequences. We then began an iterative removal process to remove sequences contributing little to the overall phylogenetic or sequence diversity of the dataset as defined by short terminal branch lengths.

To identify sequences with short terminal branch lengths, an initial phylogenetic tree was inferred using FastTreeMP version 2 (Price *et al.* 2009, 2010) with the maximum-Likelihood approximation to the CAT rates across sites method (Stamatakis 2006a) although the number of possible rate categories was fixed at 20 and every site assigned to the most likely rate category. The final maximum-likelihood branch length optimization used 20 gamma distributed rate categories and the JTT (Jones *et al.* 1992) amino acid substitution model. The objective was to retain a final dataset with broad taxonomic coverage as well as maximum sequence diversity. To avoid bias in the removal of short-branching taxa, sequences within the bottom 10 percentile of the terminal branch length distribution were identified from the non-GapC1 set and one was randomly deleted. This process was repeated until the final dataset size of 350 sequences was reached. The final size was selected to balance enhanced sequence diversity in the GapA/B/2 and cytosolic GAPDH groups, overall reduction of sequences with extremely short terminal branch lengths, and the speed of analyses performed with QmmRAxML (Wang *et al.* 2008).

4.2.2 Final Multiple Sequence Alignment and Phylogenetic Tree

After the final reduced set of sequences was finalized, a final multiple sequence alignment was constructed using hmmalign from the HMMER3 software package (Eddy 1998, 2011; Johnson *et al.* 2010) using an HMM profile generated from the OrthoMCL release 4 (Li *et al.* 2003; Chen *et al.* 2006, 2007) seed alignment of GAPDH (OG4_10093). The alignment was then automatically trimmed using an in-house sequence masking program called AliMask-CS (Alignment Masking with Confidence Scores, described in Chapter 2) leaving an final alignment length of 327 amino acid positions.

A final maximum-likelihood phylogenetic tree was inferred using FastTree2 selecting options to make it slightly more accurate including: slow nearest neighbour interchanges (NNIs), four rounds of sub-tree pruning and re-grafting (SPR), branch lengths optimized with 20 gamma distributed rate categories and the JTT substitution matrix as above, always optimizing all of the five relevant branch lengths during an NNI with three optimization rounds, and the slow option. The tree was displayed as arbitrarily rooted using the cyanobacterial non-Gap2, glycolytic GapC orthologs as basal taxa (all displayed trees were inferred as unrooted). In order to carry out contrasting analyses of functional divergence (below) the rooted phylogenetic tree was parsed in to two alternatives, one with the GapC1 clade removed and the other with the GapA/B/2 clade removed. These contrasting trees allow for comparison of functional divergence between the GapA/B/2 and the glycolytic GAPDH group as well as the GapC1 clade versus the glycolytic GAPDH group without either of these Calvin-cycle tuned GAPDH groups unduly influencing the analysis of the other. The aim of this study is to determine the residues within the two Calvin-cycle tuned GAPDH groups that have altered evolutionary constraints as a result of their changed cofactor-binding and regulatory properties.

4.2.3 Testing Functional Divergence

Several different programs were used to test for sites undergoing functional divergence in GapC1 and GapA/B/2 relative to other GAPDH sequences. These included: FunDi (Chapter 3) using QmmRAxML, RAxML 7.2.6 (Stamatakis 2006b), or the multi-threaded version of FastTree 2; GroupSim (Capra & Singh 2008); and the Difference Evolutionary-Trace method (Lichtarge *et al.* 1996; Madabushi *et al.* 2004; Raviscioni *et al.* 2006). Default options were used in all cases, except GroupSim where the criteria for ignoring alignment sites containing gap characters was changed in order to only ignore columns that were all gaps (none present in the tested alignments), as the AliMask-CS method was used to mask the final alignment. For FunDi-based predictions, the ConsWin windowing approach described previously (Chapter 3) was used as it increased predictive performance on both simulated and biological datasets. Default parameter choices for window size and weighting were used as described in those experiments.

To identify cases of convergent functional divergence between the GapA/B/2 group and the GapC1 group, and to avoid one group of NADPH-dependent GAPDH sequences from influencing functional divergence predictions of the other, contrast analyses were used as described above. In this technique, each group of NADPH-dependent GAPDH sequences is considered in comparison only to the cytosolic GAPDH sequences. For both FunDi and the Difference Evolutionary-Trace methods, the two contrast phylogenetic trees described above were used in the analyses.

Sites with a functional divergence score above 0.5 (FunDi, GroupSim) were considered to be functionally divergent in order to identify the maximum number of possible sites for consideration. For FunDi this represents all sites that are at least as well modeled by the independent component (in terms of their posterior probability of being functionally divergent; see Chapter 3 for more detail) as by the standard dependent component and thus provide some signal for functional divergence. For the Difference Evolutionary-Trace method sites are considered to be functionally divergent if they score within the top 20% of functionally important residues in either of the sub-groups under examination, but do not fall within the top 20% of functionally important residues when all sequences are considered in the context of the whole phylogenetic tree.

4.2.4 Constructing a Homology Model

In order to visualize the physical context of sites predicted to be functionally divergent, a homology model was constructed using the SWISS-MODEL webserver (Peitsch 1995; Arnold *et al.* 2006; Kiefer *et al.* 2009). Both guided and automated homology models were constructed using the *Ascophyllum nodosum* GapC1 sequence as input. For the guided homology models, either the 2PKQ (*Spinacia oleracea*) or 4DBV (*Geobacillus stearothermophilus*) structures from the PDB were used as templates. For the fully automated mode, which constructed a homology model that included all four subunits of the tetramer, SWISS-MODEL used the 3E5R structure (corresponding to the cytosolic GAPDH of *Oryza sativa*) as a template. To select the GapC1 sequence for comparison, all GapC1 sequences were compared to PDB sequences using BLASTP, and the sequence with the best overall blast match to a relevant structure in the PDB was selected. The *Ascophyllum nodosum* homology models based on the 2PKQ and 3E5R templates were selected for structural comparisons of NADPH/NADH binding.

4.2.5 Prediction of Structural Disorder

Several different publicly available methods were used to predict regions of structural disorder on the sequences from the PDB structure 2PKQ and the chromalveolate *Ascophyllum nodosum* GapC1 that was used for homology model construction. We used DisEMBL (Linding *et al.* 2003) with the ‘Hot Loops’ definition of disorder and the Predictor Of Naturally Disordered Regions (PONDR) VL-XT method (Romero *et al.* 1997, 2001; Li *et al.* 1999) to predict structural disorder. Default parameters were used for both programs. Access to PONDR® was provided by Molecular Kinetics (6201 La Pas Trail - Ste 160, Indianapolis, IN 46268; 317-280-8737; E-mail: main@molecularkinetics.com). VL-XT is copyright©1999 by the WSU Research Foundation, all rights reserved. PONDR® is copyright©2004 by Molecular Kinetics, all rights reserved.

4.3 Results

After retrieving GAPDH sequences from NCBI, along with the datasets of several phylogenetic analyses (Harper & Keeling 2003; Petersen *et al.* 2006a; Takishita *et al.* 2009) and removing sequences according to the criteria outlined in section 4.2, a dataset of 350 eukaryotic and cyanobacterial GAPDH sequences was assembled. This dataset size offered the best compromise between sequence diversity, taxonomic coverage, and reasonable analysis times using QmmRAxML for functional divergence prediction. After alignment and automated masking this multiple sequence alignment was used to infer a maximum-likelihood phylogenetic tree using FastTree2 with maximal accuracy settings (Figure 4.2). FastTree 2 was selected due to its speed and relative accuracy compared to RAxML (Stamatakis 2006b). FastTree consistently recovers well supported nodes in phylogenetic reconstruction (Price *et al.* 2010) and the FunDi step of functional divergence prediction re-optimizes maximum-likelihood branch lengths of a supplied topology according to the model parameters used (Chapter 3).

The phylogenetic tree was constructed to incorporate as much potential sequence diversity as possible, and as a result it contains many paralogous sequences. Presumably, the various GAPDH paralogs included that are not targeted to the chloroplast bind NADH and function in glycolysis, although it is possible that some paralogs have altered functional constraints as many side-functions have been discovered for GAPDH including virulence and adhesion (Dumke *et al.* 2011), neurodegeneration in Alzheimer's (Butterfield *et al.* 2010), carcinogenesis (Colell *et al.* 2009), and transcriptional regulation (Zheng *et al.* 2003). In this analysis we do not take this possibility into consideration under the assumption that many of these secondary functions will be unique to individual sequences and will not create a general functional divergence pattern over all the cytosolic homologs.

The phylogenetic tree did recover GapA and GapB clades branching together as a monophyletic group sister to the cyanobacterial Gap2 sequences. This larger clade (Gap2, GapA, and GapB) will be henceforth referred to as the ‘green group’. As expected the green group and GapC1 sequences fall within different locations in the overall GAPDH phylogeny. The closest branching sister taxa (yellow clade in figure 4.2) to the GapC1 sequences in this tree topology are several ciliate taxa (*Tetrahymena thermophila*, *Paramecium tetraurelia*, and *Halteria grandinella*); the *Tetrahymena* and *Paramecium* sequences were those from a previous study (Takishita *et al.* 2009) where they were used as the outgroup for determining the internal branching order within the GapC1 phylogeny. Several long-branching sequences group together basal to these sequences including several microsporidians, the chlorarachniophyte *Bigelowiella natans* (Rhizaria), and the green alga *Micromonas pusilla*. Many of these sequences appear to be paralogs and their position and grouping together may be the result of long-branch attraction, although in some cases it could reflect a legitimate phylogenetic affiliation between the rhizarians, stramenopiles and alveolates (SAR clade) and perhaps the cryptophytes + haptophytes and green algae as discussed previously. Alternative phylogenetic tree reconstruction with RAxML 7.2.6 and the LG+Γ model of sequence evolution resulted in a similar overall topology although the position of the long-branches discussed above moved basal to the nearby large assemblage of cytosolic GAPDH sequences, which contained the chromalveolate outgroup taxa (*Tetrahymena*, *Paramecium*, and *Halteria*). Because the positions of the closest outgroup sequences may impact the site- likelihood calculations and inferences of functional divergence, the

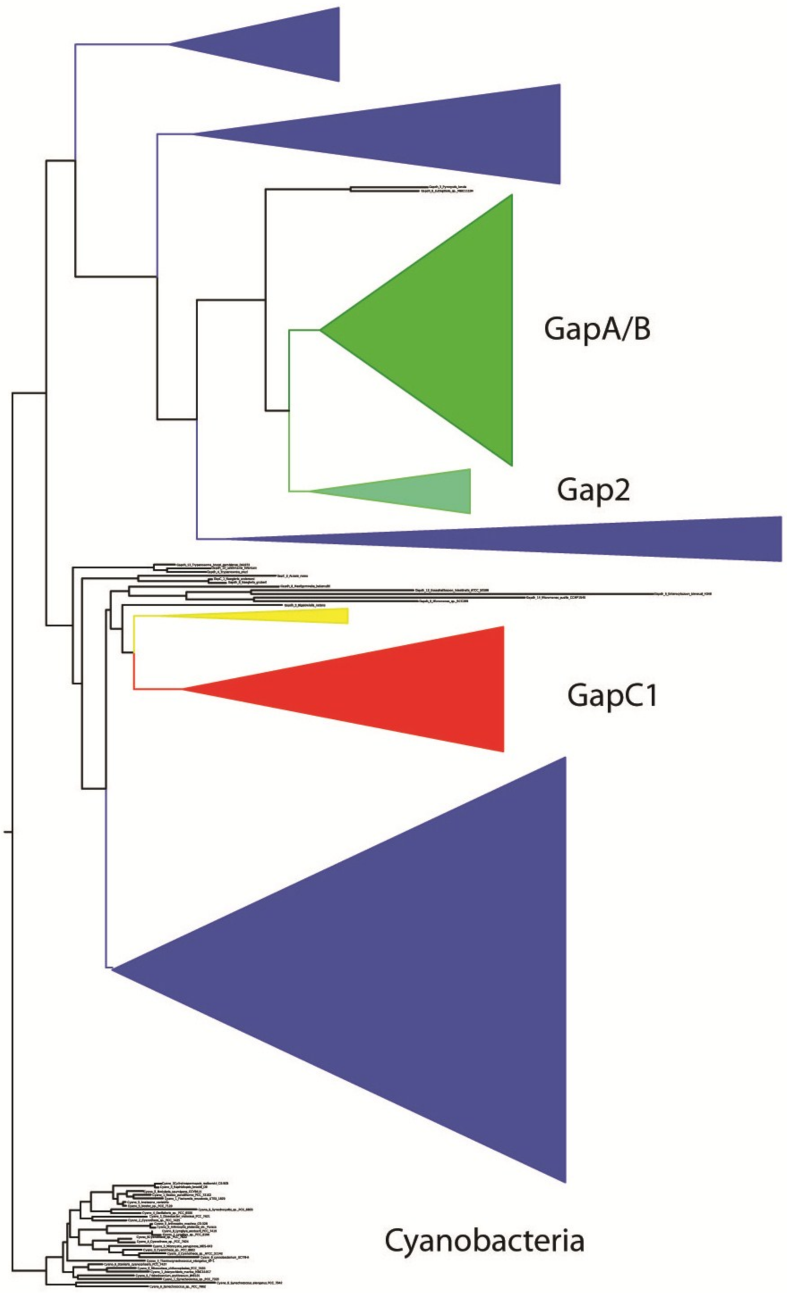


FIGURE 4.2 SIMPLIFIED MAXIMUM-LIKELIHOOD PHYLOGENETIC TREE OF GAPDH, SHOWN AS ARBITRARILY ROOTED WITH CYANOBACTERIAL SEQUENCES. Cytosolic GAPDH collapsed clades are coloured in blue. Gap A/B and Gap2 are in green, with GapC1 in red. The yellow contains several ciliate sequences that branched sister to the chromalveolate GapC1 sequences. Uncoloured long branches sister to the GapC1/ciliate group includes highly divergent microsporidians and several other long branching sequences.

first FastTree topology was selected for use in functional divergence analysis.

4.3.1 Sites Predicted to be Functionally Divergent

All predictors of functional divergence identified a number of alignment positions with significant scores for both the green group versus other and the GapC1 versus the other GAPDH comparisons. Table 4.1 lists the number of sites predicted to be functionally divergent in the green group only, GapC1 only, and those predicted as functionally divergent in both groups for each of the prediction methods used. FunDi using QmmRAxML for site-likelihood calculation predicts the largest number of sites (115) while the Difference Evolutionary Trace Method predicts the least (39), followed by GroupSim (43). The Difference Evolutionary Trace Method's scoring system is intrinsically more conservative, as only values in the top 20% of "importance" are considered potentially functionally divergent, with any sites that are important in the whole tree removed from the predicted pool. The two other FunDi sub-types, using FastTree or RAxML for site likelihood calculations, share the greatest degree of overlap, as expected given that the primary difference between the two is FastTree's use of the maximum likelihood CAT approximation to the rates-across-sites model during its heuristic maximum-likelihood search, followed by use of the JTT exchangeabilities for branch length optimizations, whereas the closely related WAG exchangeabilities are used by RAxML. QmmRAxML uses the WAG exchangeabilities and a mixture of 10 different amino acid frequency profiles described in Chapter 3 and references therein.

To place the predictions of functional divergence into context, and describe the differences in performance between FunDi, GroupSim, and the Difference Evolutionary-Trace Method, we gathered data from many functional and structural studies in the literature concerning GAPDH, paying particular attention to NADPH/NADH binding/discrimination and the regulation of function by CP12. Some sites play important roles in both, as expected, because CP12 regulation and a preference for

NADPH over NADH are intertwined with one another functionally, at least in the Archeaplastida (Pohlmeyer *et al.* 1996; Wedel *et al.* 1997; Wedel & Soll 1998; Graciet *et al.* 2003, 2004; Lebreton *et al.* 2006; Trost *et al.* 2006; Fermani *et al.* 2007; Marri *et al.* 2008).

TABLE 4.1 NUMBER OF SITES PREDICTED TO BE FUNCTIONALLY DIVERGENT IN EACH OF THE GROUPS IN QUESTION, OR IN BOTH GROUPS, FOR EACH OF THE CLASSIFIERS USED.

	Green Group Only	**GapC1 Only**	**Shared**
FunDi – FastTree	44	7	6
FunDi – RaxML	47	8	6
FunDi – QmmRAxML	69	26	20
GroupSim	37	4	2
Difference Evolutionary Trace	17	15	7

4.3.1.1 Loop positions 32-35

Aspartate 32 (Figure 4.3 and 4.4) is one of the canonical residues implicated in the discrimination between NADH and NADPH in the photosynthetic form of GAPDH (Clermont *et al.* 1993). When NADH is bound, Asp32 forms a hydrogen bond with the 2'-hydroxyl group of the adenosine in NADH (Skarzyński *et al.* 1987). When NADPH is bound, it rotates away from the cofactor thereby preventing steric clashes (Fermani *et al.* 2001; Sparla *et al.* 2004). This position is conserved in chloroplast (NADH-binding) GAPDH (e.g. GapA/B) sequences as well as Gap2 sequences and is quite highly conserved as an aspartate residue in the cytosolic GAPDH sequences as expected for NADH binding and therefore this position was not predicted as being functionally divergent by any of the tested methods in sequences from the green group. However, in the chromalveolate plastid-targeted sequences this site it is not as well conserved with amino acids such as glutamate, serine, threonine, and alanine all observed, and predicted to be functionally divergent by FunDi (all versions) but not by either GroupSim or Difference Evolutionary-Trace. In the sequences where aspartate has been substituted with glutamate, it is likely that glutamate is performing the same function. When non-acidic residues at position 32 are observed, it is likely that other residues on the flexible loop where this site is located are involved in NADH binding in those sequences.

There are also several predicted functionally divergent sites in close proximity to the conserved aspartate 32 residue, such as the valines in position 28 and 29, threonine 33, and the glycines in position 34 and 35. The two valines were predicted to be functionally divergent in both the green group and chromalveolates by all FunDi variants for 28 and by FunDi with FastTree or RAxML for 29 (or only in the greens using QmmRAxML),while Thr33 (FunDi with QmmRAxML) and Gly35 (FunDi with FastTree or QmmRAxML) were only predicted in the green group and Gly34 (FunDi with QmmRAxML) only in the chromalveolates. None of these positions were predicted as functionally divergent by GroupSim or the Difference Evolutionary-Trace methods. Previous experimental work in the NADH-binding GAPDH of the thermophilic eubacterium *Bacillus stearothermophilus* (Clermont *et al.* 1993; Didierjean *et al.* 1997; Sparla *et al.* 2004) involved mutating some of these residues (33, 34, and 35), which had previously been described as a NAD-binding ‘fingerprint’ (Wierenga *et al.* 1985; Wierenga 1986), to their counterparts in NADPH-dependent GAPDH to investigate their role in discrimination between the two co-enzymes. The effect of mutating these residues was slight, except in the case of aspartate 32. However, threonine 33, along with arginine 77 and serine 188 (below) may be involved in bonding with the 2’-phosphate of NADPH, at least under certain circumstances (Fermani *et al.* 2001; Sparla *et al.* 2004).

4.3.1.2 Arginine 77

Arginine 77 (Figures 4.3 and 4.4) is located in the “cleft” between monomers in the tetrameric complex along one of several flexible loop regions, physically near to threonine 33 and aspartate 32. In one wild-type crystal structure (Fermani *et al.* 2001) at 3.0 Å resolution this residue was highly disordered; however, in a second, more recent, 2.0 Å resolution crystal structure, Arg77 had a high degree of order and was in position to form a salt-bridge with the 2’-phosphate of NADPH (Sparla *et al.* 2004). The increased resolution of the more recent structure, combined with the fact that a conserved arginine generally is involved in binding the 2’-phosphate of NADP in other NADP-binding enzymes with Rossman folds (Carugo & Argos 1997) provides increased evidence that Arg77 is involved in NADPH binding instead of Thr33.

As well as stabilizing the 2’-phosphate of NADPH via a salt bond, Arg77 also plays an important role in the CP12 mediated regulation of NADPH-dependent activity in GAPDH (Figure 4.5). When CP12 (or the activated form of GapB’s C-terminal extension) enters in to the “cleft” region between neighbouring (R-axis symmetry related) monomers Arg77 is distracted away from potential interactions with the 2’-phosphate of NADPH and swings away, towards CP12 where it is thought to interact with the negatively charged residues of the regulatory peptide instead (Sparla *et al.* 2004, 2005; Lebreton *et al.* 2006; Trost *et al.* 2006; Fermani *et al.* 2007). Thus, Arg77 is one of the residues responsible for several key interactions unique to NADPH-dependent GAPDH.

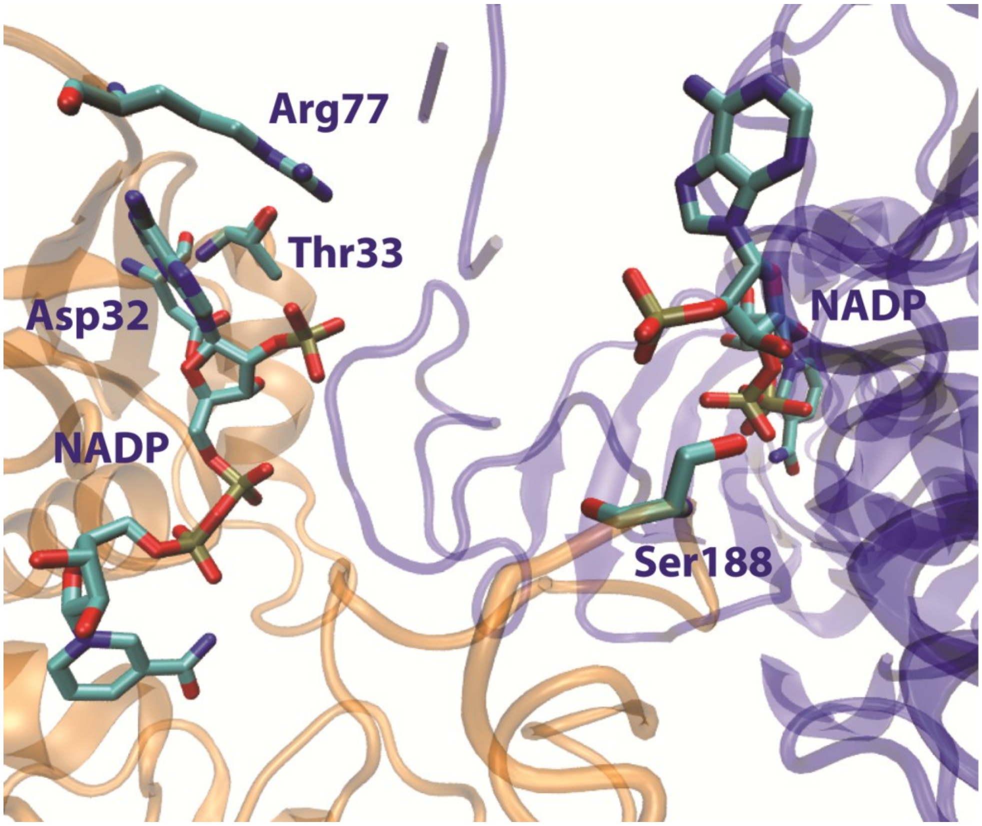


FIGURE 4.3 KEY RESIDUES FOR CO-ENZYME DISCRIMINATION IDENTIFIED BY PREVIOUS EXPERIMENTAL WORK. Cartoon representation of O and R subunits shown as in Figure 4.1. Asp32 and Thr33 located behind bound NADP molecule, with Asp32 rotated away to reduce steric clash with 2’-phosphate of NADP. Ser188 and Arg77 are shown in positions to interact with 2’-phosphate of bound NADP.

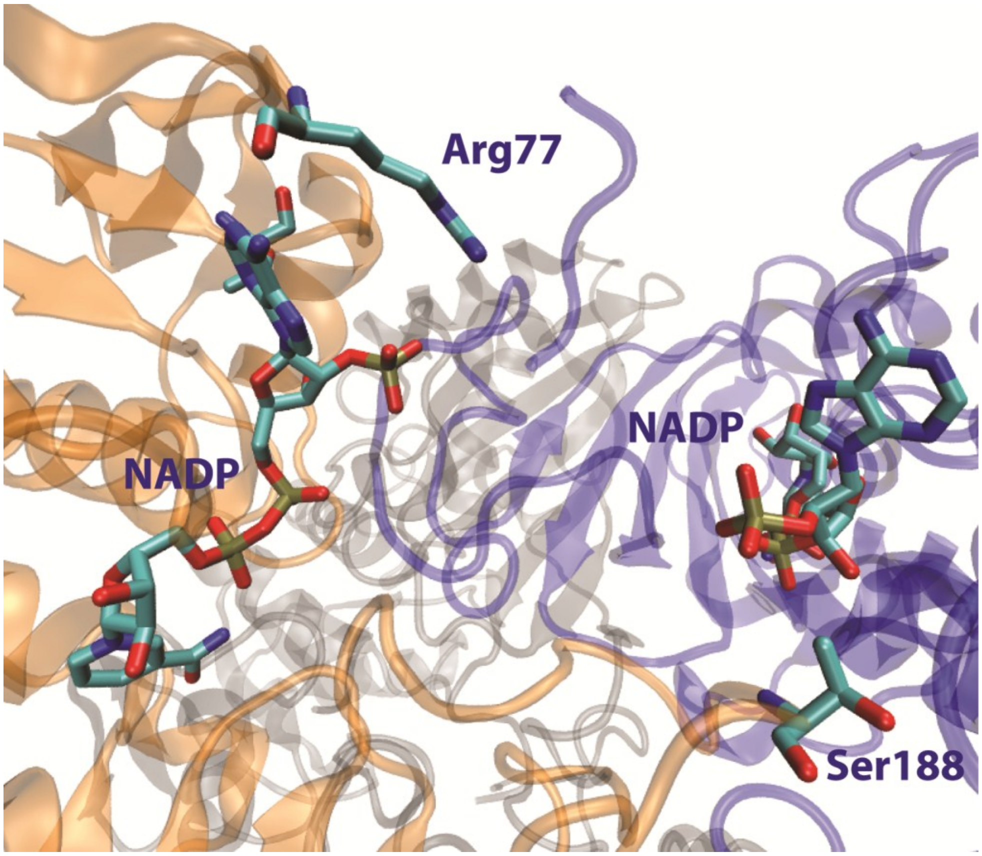


FIGURE 4.4 KEY RESIDUES INVOLVED IN NADPH/NADH BINDING AND DISCRIMINATION IN THE CHROMALVEOLATE GAPC1 PROTEIN SEQUENCE ARE SHOWN MAPPED ON THE HOMOLOGY MODEL OF ASCOPHYLLUM NODOSUM (ORANGE). In blue is the R subunit from the PDB structure 2PKQ used as template, along with bound NADP. NADP and selected residues (Ser188 and Arg77) are coloured according to element type.

This residue was predicted to be functionally divergent by both FunDi (using either FastTree or QmmRAxML) and the Difference Evolutionary Trace method, but not GroupSim, in both the green group and among the chromalveolate plastid-GAPDH sequences. In the multiple sequence alignment arginine is strictly conserved in both of these clades, but not among cytosolic GAPDH sequences, although arginine and lysine are the two most commonly occurring amino acids in that position, making this a likely case of convergent Type I functional divergence (i.e. functional divergence involving a rate shift). The high degree of conservation in the cytosolic GAPDH sequences can make this type of site-pattern difficult to identify as being functionally divergent, but the phylogenetic-based methods such as FunDi and the Evolutionary-Trace method can resolve these difficulties. As noted previously in Chapter 1, functionally divergent residues are strongly correlated with regions of intrinsic disorder, which has been observed in this loop region of residues, giving further evidence for its role in the functional differentiation of NAD+ and NADPH-dependent GAPDHs.

4.3.1.3 Serine 188

Serine 188 (Figures 4.3 and 4.4) is located on the S-Loop of GAPDH where it hydrogen bonds with the 2’-phosphate group of NADPH bound to the monomer located across the R-axis of symmetry (Clermont *et al.* 1993; Didierjean *et al.* 1997; Fermani *et al.* 2001; Sparla *et al.* 2004). When NAD+ is bound, as in the down-regulated form of plastid-targeted GAPDH, Ser188 instead forms hydrogen bonds with available water molecules and potentially with Asn39 of the opposite, R-related subgroup (Sparla *et al.* 2004).

Serine 188 was predicted to be functionally divergent, both within the green group as well as the chromalveolate GapC1 sequences, by both FunDi with QmmRAxML (the FuNDi-FastTree Method only detected functional divergence at this site in the green group) and GroupSim. This position is almost universally conserved as a serine within the green group (with the exception of a single alanine substitution in a GapA paralog of the unicellular red alga *Galdieria sulphuraria*. It is unlikely that this is a simple mis-annotation of the sequence as it groups with other sequences from the green group as expected on phylogenetic grounds. Serine at this position is almost universally conserved among the chromalveolate GapC1 sequences as well, with a handful of substitutions to alanine, and some to threonine. In mutants with Ser188 substituted by an alanine (Sparla et al. 2004) there is a reduced preference for NADPH over NAD+ and a phenotypic difference characterized by a “loosened” and enlarged conformation due to the loss of the Ser188 interaction with bound NADPH. This is probably not the case of Ser188Ala substitutions observed in some chromalveolate GapC1 sequences due to the presence of additional serine or threonine residues at positions neighbouring site 188 on the S-loop such as 187 or the insertion between 188 and 189 in chromalveolates relative to the greens. Position 187 is often a serine or threonine in the chromalveolates, compared to a conserved alanine in the green group, and no strict conservation among the other sequences in the alignment. It has been hypothesized previously that the alanine found at position 187 in non-chromalveolate plastid-targeted GAPDHs reduces steric clash with the 2’phosophate of NADPH and allows serine 188 to form the necessary hydrogen bond (Corbier *et al.* 1990; Clermont *et al.* 1993; Eyschen *et al.* 1996; Didierjean *et al.* 1997; Fermani *et al.* 2001). This position was also predicted to be functionally divergent in both the chromalveolates and the green group by FunDi using QmmRAxML and in the green group only by FunDi using FastTree2 and by GroupSim. The other neighbouring position to Ser188 on the loop, which corresponds to a gap character in the alignment in the green group (presumably because of a deletion event in their common ancestor), is also often a serine or threonine residue in the chromalveolates. While this position was not predicted to be functionally divergent by any method, in those chromalveolate sequences with a substitution of serine 188 to an alanine, this position is always either a serine or threonine residue that is likely substituting for Ser188’s normal functional role.

4.3.1.4 Other important S-loop Residues Near Serine 188

Among the canonical NADP-dependent GAPDH sequences of plants, the S-loop contains several conserved arginine residues (183, 191, 194, 195, and 197) that, along with Ser188, are thought to be important residues involved in interactions with CP12 (Figure 4.5). Due to this excess of positively-charged residues, and because the regulatory region of CP12 and the C-terminal extension of GapB contain an excess of negatively-charged residues it has been hypothesized that there is an important interaction between the two, especially given the "enlarged" phenotype often seen in Ser188Ala mutants.

Several of these residues on the S-loop were predicted to be functionally divergent. Arg183 was predicted to be functionally divergent by all of the methods, whereas Arg191 was predicted by GroupSim, and Arg195 was predicted by both Groupsim and Difference Evolutionary Trace. All three of these residues were predicted to be functionally divergent in the green group, but not in the chromalveolate GapC1 sequences. They are all strictly conserved arginine residues in the green group NADPH-dependent GAPDH sequences and not strictly conserved in cytosolic GAPDH's, although lysine and arginine are frequently observed. Within the chromalveolate GapC1 clade these positions are also not strictly conserved, although there are conserved arginine sequences very near (often at neighbouring positions) on the S-loop. One of these conserved arginine residues was predicted to be functionally divergent but only by the Difference Evolutionary-Trace method.

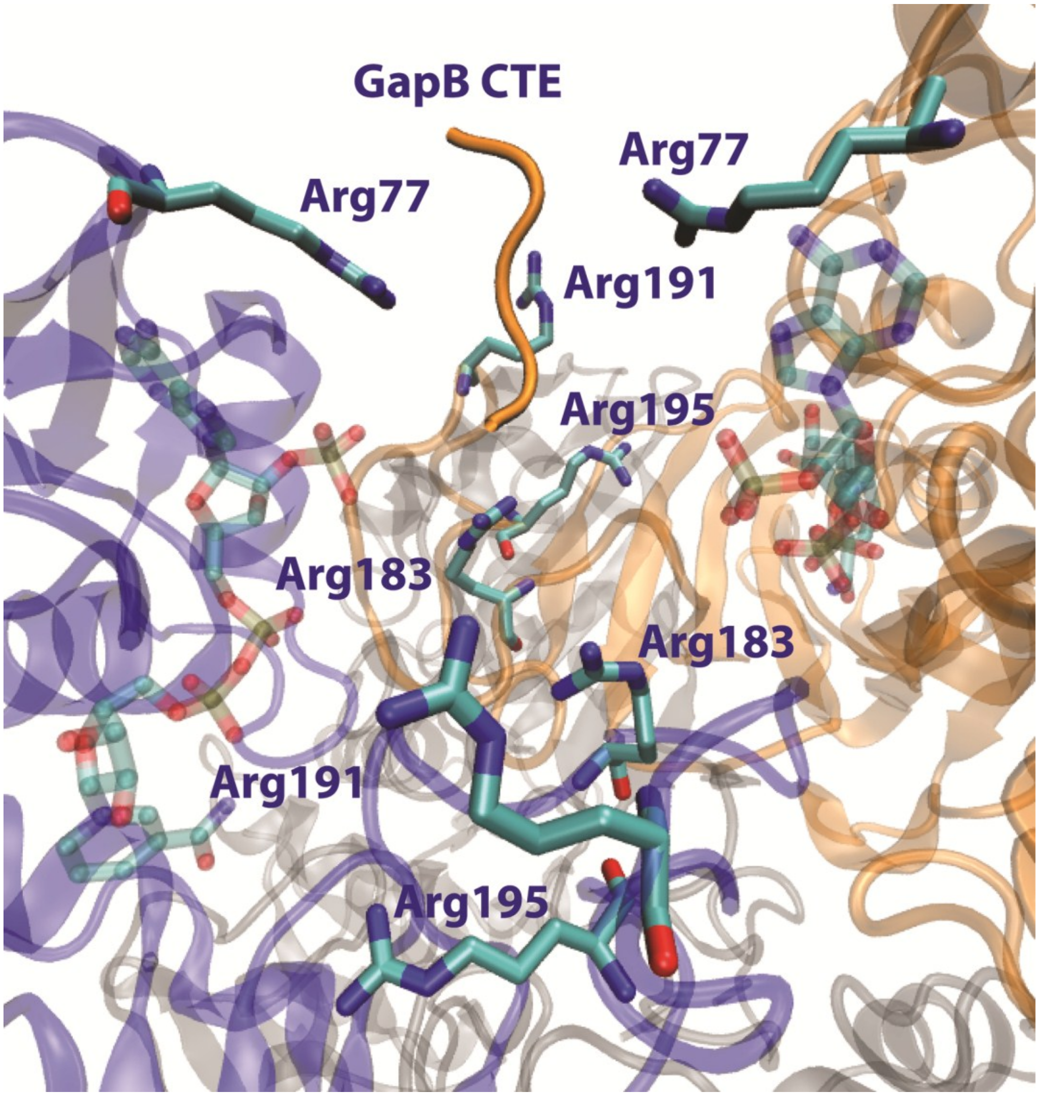


FIGURE 4.5 THE "CLEFT" AND CP12 BINDING REGION BETWEEN MONOMERS IN THE A2B2 TETRAMER OF GAPDH FROM SPINACH (2PKQ). The C-terminal extension of GapB, homologous to C-terminal regulatory region of CP12 and important conserved arginine residues are indicated. Bound NADP+ is coloured by atom type but is shown as transparent as are the P and R monomers. Conserved arginines, important for CP12 interaction are shown from both monomers.

4.3.1.5 Other Sites in the Coenyzme Binding Domain

There are other sites within the coenzyme domain predicted to be functionally divergent, including sites predicted to be divergent in both the chromalveolate GapC1 sequences and the plastid-targeted GAPDH sequences of the green group. While we cannot conclusively assign functional roles to all of these residues based on experimental studies, some are located in close proximity to the residues described above, including four residues all located on the same loop as arginine 77 (Val74, Ser75, Asp 76, and Asn79). Because Arg77 plays an important role in both coenzyme binding and the thioredoxin-mediated regulation of GAPDH during light/dark cycles, residues nearby may also play direct or indirect roles in regulation, or are coevolving. With the exception of Asn79, which was only predicted to be functionally divergent by the Difference Evolutionary-Trace method, these residues were only predicted by FunDi to be functionally divergent in both groups, Val74 regardless of which program is used for site-likelihood calculation and Ser75 and Asp76 only with QmmRAxML. Asp76 is further predicted to be functionally divergent only within the chromalveolates by FunDi using RAxML and Ser75 is predicted to be functionally divergent only within the greens by Difference Evolutionary-Trace and FunDi using FastTree or RAxML. FunDi with QmmRAxML is the only method to consistently identify these residues as functionally divergent in both groups. Residues 74, 75, and 76 are almost universally conserved within the archeaplastids and cyanobacterial NADPH-dependent GAPDHs with a V-S-[NDT] motif, while these three positions are not conserved among cytosolic GAPDH sequences. The chromalveolate GapC1 sequences are less strongly conserved, with a tendency for a [ST]-[HA]-T motif, although there are some exceptions. It is possible that these positions may influence GapC1 regulation. The motif for *Odontella* for instance, which lacks CP12 regulation (Michels *et al.* 2005), is a very divergent S-R-C while for *Ascophyllum*, which does appear to have CP12 regulation, it is the more canonical GapC1 motif of S-A-T. Based on the sequence conservation patterns in these residues it does not appear as if there is a strong argument for convergence between the green group and GapC1 sequences, although the region may still be important for CP12 mediated regulation, and the differences observed within the chromalveolates may partially explain some of the observed regulatory differences. The accumulation of more regulatory and sequence data will be needed to test these speculations.

4.3.1.6 Other Catalytic Domain Residues

Several residues in the region from 206 to 211 are also predicted as being functionally divergent within the green group alone, within the chromalveolates alone, or in both. This region plays a role in the difference between the P_i and “New” P_i site (Falini *et al.* 2003; Fermani *et al.* 2007) which is an important structural difference between NAD+- and NADPH-dependent activity of catalysis. Down-regulated A_4 or A_2B_2 bound to NAD+ instead of NADPH has a slightly altered conformation where the inorganic phosphate is located closer to the catalytic residues in what has been called the “New” Pi site instead of the “classic” Pi site where it is usually located in NADPH-bound GAPDH.

There are other amino acids in this region, predicted to be functionally divergent, whose exact functional roles cannot be so easily quantified. In the chromalveolate GapC1 sequences two sites stand out that correspond to asparagine 146 and cysteine 153 in the 2PKQ crystal structure (Cys149 is the catalytic cysteine in this numbering). In the chromalveolates, the conserved Asn146 has been substituted for a cysteine residue, while the conserved Cys153 has been substituted by a glycine. While the first was predicted to be functionally divergent by all three prediction methods (and using all three site-likelihood calculation programs with FunDi), the latter was not predicted to be functionally divergent by the Difference-Evolutionary Trace Method.

4.3.2 Functionally Divergent Residues and Intrinsic Structural Disorder

Previous studies have reported that functionally divergent residues are preferentially located in or near regions of structural disorder (Aharoni *et al.* 2005; Chakrabarti *et al.* 2007; Capra & Singh 2008; Chakrabarti & Panchenko 2009). To investigate whether intrinsic disorder is related to functional divergence in GAPDH we used sequence-based predictors of disordered regions on the GapA sequence of spinach, the GapC1 sequence of the brown alga *Ascophyllum nodosum* used for homology model construction, and the cytosolic GAPDH sequences of *Ascophyllum* and *Homo sapiens*. In all cases, only residues retained in the masked sequence alignment were classified as structured/unstructured. For all four sequences, predictions of structural disorder were made by PONDR, DisEmbl, and isUnstruct. DisEmbl contains three different predictors of structural disorder, so for these analysis we selected the "Hot Coils" method as providing the most information for putative disordered regions while minimizing the false positive rate (Linding *et al.* 2003) compared to considering all loop regions as disordered or predicting missing coordinates. Disordered regions were mapped on to the reference alignment used for the prediction of functional divergence and checked for overlap.

IsUnstruct predicted the fewest number of unstructured residues, mostly limited to the N- and C-terminal portions of the sequence (not including targeting peptides, which were removed from the alignment), with the exception of the *Ascophyllum* and human cytosolic GAPDH sequences with a short internal region that was predicted to be inherently disordered. IsUnstruct predictions are not shown, as they provided no significant information compared to PONDR and DisEmbl. PONDR and DisEmble overlapped in some of their predictions but differ in others (Table 4.2). For example in the spinach sequence, PONDR and DisEmbl only overlap by five residues while they overlap by 32 residues in the *Ascophyllum* GapC1 sequence. Larger regions of structural disorder (by number of residues) were observed in both of the plastid-targeted forms of GAPDH than in either of the two cytosolic sequences.

TABLE 4.2 REGIONS PREDICTED TO BE INTRINSICALLY UNSTRUCTURED BY DISEMBLE AND PONDR. Numbers are according to the position numbers; not including gaps, in the respective sequences from the masked multiple sequence alignment.

	DisEmbl	PONDR
2PKQ	1-24, 65-89, 118-130, 176-186, 233-242	17-19, 117,118, 190-198, 211-233, 256-275
GapC1 *Ascophyllum nodosum*	1-25, 58-75, 108-117, 178-203, 216-230, 251-259, 285-294	9-11,13,63-80,120-130,136-139,185-212,250-273
Ascophyllum nodosum	1-12, 20-37, 172-205	176-187,239-240,262-268,322,325
Homo sapiens	1-27, 68-79, 174-185, 241-250	182-185,198-205,244-248,321-326

Regions of predicted structural disorder by either method were mapped on to the NADPH-dependent GAPDH structures for spinach (2PKQ) and the *Ascophyllum nodosum* homology model (Figure 4.6 A and B). In both structures, when both DisEmble and PONDR predictions are considered, disordered regions are extensive and include regions in both the co-enzyme binding and catalytic domains as well as the majority of the s-loop. In addition, these regions, which were predicted from sequence information alone, also include regions of known secondary structure such as alpha helices and beta-

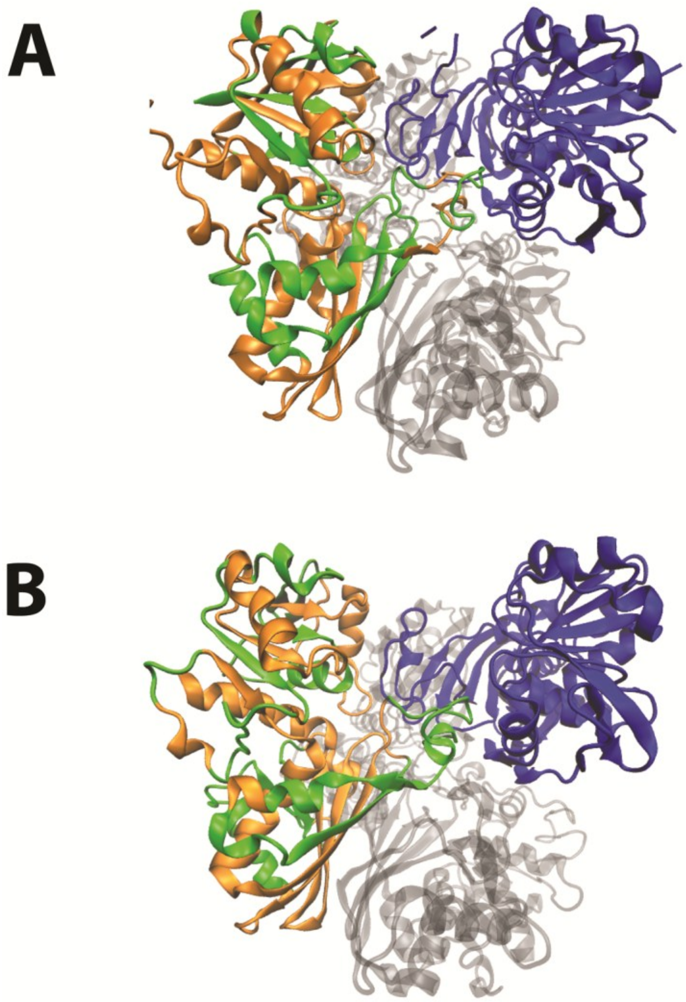


FIGURE 4.6 REGIONS PREDICTED TO BE STRUCTURALLY DISORDERED BY BOTH DISEMBL AND PONDR IN THE SEQUENCE OF SPINACH GAPA (A) AND ASCOPHYLLUM NODOSUM GAPC1 (B) COLOURED IN GREEN. Regions of predicted disorder (by sequence alone) contain regions of defined secondary structure in the x-ray crystal structure as well as the homology model. The majority of long loop regions in both structures are also predicted to contain significant intrinsic disorder.

sheets. In the spinach sequence, arginine 77 and serine 188, along with the important arginine residues on the S-loop are contained within regions of predicted disorder, the former in a region predicted only by DisEmble while Ser188 and the arginines were all in a region predicted only by PONDR. In the *Ascophyllum nodosum* homology model, whose sequence contains even more regions of predicted structural disorder than that of spinach GapA, arginine 77 and serine 188 are also located in disordered regions. Both lie in regions predicted by DisEmble and serine 188 lies in a PONDR predicted region. As with the spinach structure, disordered regions are located in both domains as well as the S-loop, and include a substantial amount of area with defined secondary structure.

The residues located on the loop which contains aspartate 32, which is functionally divergent in the chromalveolates, does not lie within a region of predicted structural disorder by either method, and in either group of sequences. Sites predicted to be functionally divergent lie throughout the resolved crystal structure of 2PKQ or the homology model of *Ascophyllum* GapC1 with no indication of clustering or a preference for location in regions predicted to be disordered. To determine whether these observations were statistically significant, two-sided fisher's exact tests were carried out on predictions for the 2PKQ sequence using either the PONDR or DisEmble predictions and either all sites predicted to be functionally divergent in both the green group and chromalveolates, or only those sites predicted to be divergent in both groups by FunDi using QmmRAxML. Hypothesizing that those sites predicted to be functionally divergent in both the greens and chromalveolates are more likely to represent sites involved in coenzyme discrimination/binding and regulation via protein-protein interactions we first limited our analysis to those sites. Fisher's exact tests could not reject independence of the predictions in these cases (P-values all >> 0.05).

Given the large degree of inconsistency between the two methods used to predict intrinsic disorder, which both use the primary sequence information alone, it is difficult to determine the true regions of disorder. Combining the two prediction methods results in very little overlap and extremely large regions of the secondary structure to be considered disordered. We also performed a Fisher’s exact test for independence between PONDR and DisEmble. These two predictors overlapped in only 5 disordered predictions while PONDR predicted 78 sites to be disordered that were not by DisEmble and there were 52 sites where the reverse was true. The majority of sites (216) were not predicted to be disordered by either program. Expected counts under independence were 13.48, 69.52, 43.52, and 224.48 respectively. Although independence was rejected (P-value = 0.0034), the dataset was highly skewed relative to the number of predicted disordered sites.

4.4 Discussion

While there is substantial overlap in the sites predicted to be functionally divergent in NADPH-dependent GAPDH sequences by all of the predictors used in this work, especially among sites that have been previously linked with functional differences, there are also clear discrepancies between them. FunDi predicted the most sites, especially when using QmmRAxML for background calculation of site likelihoods. FunDi with QmmRAxML also consistently predicted more known functional sites as being functionally divergent. Serine 188 and arginine 77 were both predicted by FunDi with QmmRAxML while from the other predictors only GroupSim predicted the former and the Difference Evolutionary Trace method the latter. The only known functionally divergent residues missed by FunDi, with any likelihood calculation method, were two of the potentially regulatory arginines (191 and 195) on the S-loop.

There has also been a high degree of convergent evolution between cyanobacterial derived NADPH-dependent GAPDHs and GapC1 of chromalveolates. Not only were 28 residues predicted to be functionally divergent in both groups between all of the prediction methods tested here, but also many key positions converged on the same or physico-chemically similar amino acid residues. Arginine 77 and serine 188 for instance, previously identified in crystallographic and mutagenesis studies as being two key residues for the discrimination between NADPH and NADH, converged on the same residue. Serine 188 in particular is striking, as the homologous position in cytosolic GAPDHs is generally a proline. While positions 32, 187, and 188 have been identified as fingerprints for NADPH-binding, and were thus used to characterize function in chromalveolate GapC1 sequences (Fagan et al. 1998; Fast et al. 2001; Harper and Keeling 2003), to date there have not been systematic analyses of functional divergence and convergent evolution between these and other plastid-targeted GAPDH sequences. Additionally, aspartate 32 has incorrectly been described as non-conserved in plastid-targeted GAPDHs. Mutagenesis experiments had typically substituted this residue, as it is not conserved in the very distantly related archaeal GAPDH that has dual co-enzyme specificity. However, in cyanobacterial Gap2 and plastid-targeted GAPDH position 32 is still a conserved aspartate but the side-chain rotates upon NADPH-binding to a different conformation to prevent steric clashes with the 2’-phosphate of this coenzyme. The substitution to alanine at this position in the chromalveolate GapC1 sequences represents a unique adaptation to their role in the Calvin cycle.

Also of interest are residues located near to arginine 77 (positions 74, 75, and 76) predicted by FunDi with QmmRAxML as being functionally divergent in both groups of NADPH-dependent GAPDH sequences, along with position 79, which was predicted to be divergent in both groups only by GroupSim. To our knowledge, no experimental work has suggested a functional role related to the differences between NADH- and NADPH-dependent GAPDH sequences for these sites. However, their level of conservation in these groups suggests that, at the very least, they contribute to the physical and chemical environment at arginine 77, or may potentially be involved in interactions with CP12. These residues, while functionally divergent in both groups, are not an example of convergent evolution as they feature very different sequence motifs. In the Chromalveolata, fingerprint regions such as this where there is some variability among lineages may help explain the apparent differences observed in GapC1 regulation between chromalveolate taxa, although more functional and sequence data is required.

While previous studies (See 4.3.4) have indicated a connection between functional divergence and instrinsic structural disorder, we could find no statistically significant overlap here between the two. Given the minimal overlap between the two disorder prediction methods used, and the large area of coverage if both sets of predictions are combined, it is apparent that predicting disorder also remains a difficult task. Additionally previous examples rely on only a small amount of data from biological datasets, an issue we have already addressed in chapter 3. However, secondary structure information and predictions of disorder may still prove useful when placing predictions of functional divergence within their proper biological and structural context.

The foregoing analyses serve as an excellent example of the utility of FunDi for the identification of key residues involved in functional divergence, especially when analysing large and phylogenetically complex datasets. While no single classifier identified all of the important specificity-determining sites with experimental validations, FunDi predicted the most successfully and was markedly better at identifying cases of convergent evolution. Based on previous simulation results (Chapter 3), we can be confident that this improved true positive prediction rate is not generally achieved at the expense of a drastic increase in false positives. It is also probable that many of the sites predicted to be functionally divergent without experimental validation, are in fact functionally divergent and/or are co-evolving with functionally divergent residues or are responsible for maintaining appropriate protein folding dynamics, stability, or function.

Chapter 5 Conclusions

Phylogenomic methods have been used to tackle a variety of biological problems. In this thesis, I implement two new methods that use a phylogenomic approach for two different bioinformatic problems. PhyloPred-HMM is a general phylogenomic framework for assigning functional annotations to unknown sequences of interest. Although it was designed to predict the localization of nucleus-encoded proteins to subcellular compartments such as mitochondria or related organelles (hydrogenosomes and mitosomes), it can be used to assign any kind of annotation on the basis of phylogeny. FunDi, the second tool I have developed, uses a maximum-likelihood mixture modeling approach that has a phylogenetic "independence" component that models sites in protein multiple sequence alignments that are contributing to functional divergence across a split of interest in protein families. Both of these problems relate to the evolution of diverse protein families, and more importantly how protein functions change over evolutionary time. By explicitly using phylogenetic techniques and by using specific evolutionary models, I have shown that these methods can improve upon predictive methods that do not use phylogenetic information.

5.1 PhyloPred-HMM

PhyloPred-HMM combines large scale clustering of biological sequence data, along with automated multiple sequence alignment and phylogenetic tree reconstruction; hidden markov models for sequence comparison; and a phylogenomic method for assigning subcellular localizations to unknown sequences of interest. We also developed a simple BLAST-based approach (CBOrg) for quickly filtering data from transcriptome sequencing projects to identify putative MRO localized sequences. Both of these programs were compared to several other widely used and previously validated prediction methods that can be both installed on local computers and are capable of performing predictions at the 'genomic scale' (hundreds to tens of thousands of predictions) in reasonable amounts of time. In addition, we compared the performance of several *de novo* clustering techniques on the reference biological sequence data used by PhyloPred-HMM and three different phylogenetically-based distance measures for assigning localization data.

The type of clustering method used had a small but noticeable effect on the performance of PhyloPred-HMM, with larger differences observed when performing predictions on whole proteomes versus evaluating phylogenetic distance measures on clusters with sequences from both MRO-localized and non-MRO-localized sequences. These results are due to the “tightness” and size of the resulting clusters. The hierarchical clustering method, the simplest tested in this work, resulted in the largest number of clusters but these clusters on average were smaller. The Markov Clustering algorithm with default parameters resulted in fewer clusters that were larger on average, while increasing the inflation parameter to two produced results intermediate between the two extremes. The MCL method, with an inflation parameter of two, was selected as optimal based on these properties as well as its performance.

Surprisingly, of the three distance measures tested in this work, the nearest distance (ND) metric had the best performance compared to more complicated averaging (SAD and T-SAD) methods. It was hypothesized that this was due primarily to cases where there were large differences in the percentage of MRO-localized proteins in a sequence cluster; although there were certainly cases where this was true, which we examined in more detail (Chapter 2), no consistent trend could be observed in the overall data. The average distance measures tested had comparable performances, and were characterized by their increased recall of MRO-localized sequences, but at the expense of an accompanying increase in false positive predictions. However, the two SAD and ND measures do largely overlap. The ND measure also has an advantage in not explicitly depending on the validity of the ‘Ortholog Conjecture’ and neither this method, nor the SAD/T-SAD measures, require the identification of orthology/paralogy relationships in order to make functional assignments. However, in cases where subcellular localization switches frequently over the phylogenetic tree, all measures (and indeed any phylogeny-based method) will have difficulty making accurate classifications.

The PhyloPred-HMM method performed comparably, or better, to other published prediction methods on sequences from the Metazoa, Fungi, or Archeaplastida that localized to canonical mitochondria. Unlike previous prediction methods, which use only sequences with an experimentally determined subcellular locations, the performance of PhyloPred-HMM on similar datasets did not appear to suffer by using sequences annotated at any confidence level. It is likely that because the three groups of eukaryotes mentioned above are so well studied, localization data for known sequences has accumulated such that measures of sequence similarity alone (e.g. BLAST scores) are likely to be good predictors in the majority of cases. However, for sequences from more diverse microbial eukaryotes, there has been very little investigation of the performance of subcellular localization prediction. PhyloPred-HMM was specifically constructed to include as much taxonomic and sequence diversity as possible, especially since several microbial eukaryotes have recently had the contents of their MROs determined proteomically. When PhyloPred-HMM's performance was compared to that of CBOrg and several other published prediction methods on the complete genomes of two microbial eukaryotes, *Tetrahymena thermophila* and *Trichomonas vaginalis*, it was clearly superior; however, there still remains a great deal of room for improvement as all methods performed relatively poorly. The performance of PhyloPred-HMM would improve with the inclusion of an N-terminal targeting prediction algorithm. The presence of an N-terminal targeting sequence is the most reliable indication of MRO localization, although not all organellar sequences have one. To date the majority of targeting prediction programs, like localization classifiers in general, have been trained only on a small subset of taxonomic and organellar diversity, as has been previously shown there are differences, particularly in length, between the localization peptides found for the mitosomes and hydrogenosomes of other microbial eukaryotes. Building a suitable training set should now be possible for these taxa, and would greatly enhance localization prediction in related organisms.

Despite these limitations, a phylogenomic approach to MRO targeting prediction clearly performs at least as well as more complex machine-learning approaches, which do not explicitly use phylogenetic information, on sequences from well-studied groups where a wealth of genomic data is available. However, as we move to less well-studied organisms where data from closely related taxa is sparse, performance decreases, but is still superior to that of other methods, including homology-based methods that use only BLAST or other sequence comparison methods and not a full phylogenomic framework.

5.2 Functional Divergence

Our FunDi program detects functionally divergent sites between two or more monophyletic groups within protein families and, like PhyloPred-HMM was developed with scalability to larger scale phylogenomic analyses in mind. FunDi allows for the modular use of any maximum-likelihood phylogenetic program capable of calculating and outputting site-likelihoods given a user supplied phylogenetic tree in order to compare a multiple sequence alignment under the standard ‘dependent’ model of protein evolution and our ‘independent’ model which is designed to capture significant signal of functional divergence. This approach, in contrast to some other phylogeny-based methods that have been developed such as DIVERGE, allows FunDi to make use of improvements in maximum-likelihood based phylogenetic programs and to be rapidly adapted to improved models of protein evolution.

Because no true ‘gold standard’ examples exist for comparing the performance of new functional divergence prediction methods, we constructed two new simulation strategies for creating datasets containing functionally divergent protein residues and also compared the performance of FunDi to other classifiers on several reasonably well characterized biological datasets. We then studied the performance of FunDi, compared to GroupSim and the Difference Evolutionary-Trace method, by studying a case of convergent evolution and functional divergence between the plastid-targeted NADPH-dependent GAPDHs of the Archeaplastida and the Chromalveolata.

Comparisons of FunDi to other classifiers on nine biological datasets in Chapter **3** showed wide variation within individual classifiers on different datasets as measured by both the area under the Precision-Recall (AUC-PR) and Receiver Operator Characteristic curves (AUC-ROC). FunDi performed comparably to several other predictors on this dataset, along with GroupSim, Sequence Harmony, and Multi-RELIEF. Because there are unknown, truly functionally divergent sites in these real datasets that will distort all measures of performance, we constructed two different methods for simulating alignments that contained both functionally divergent and non-divergent sites. Under both sets of simulation conditions, we showed that FunDi using QmmRAxML for site-likelihood calculations and with the ConsWin windowing method consistently outperformed other methods, while GroupSim came a close second.

We also evaluated performance differences under different phylogenetic tree shape and size scenarios to identify situations where we could expect better or worse performance from individual classifiers.Two phylogenetic tree shapes were identified with similar numbers of taxa and the same normalized tree length, but with radically different performance characteristics as measured by AUC-PR. These trees, were quite different in the relationship of their subtrees. The tree on which FunDi performed well in terms of prediction was more balanced, possessing subtrees with similar branch lengths while the poorly performing tree was more unbalanced. This unbalanced tree had a large difference in the normalized tree length of the subtrees and relatively few taxa in the short subtree compared to the larger. This large difference in normalized tree lengths between subtrees was compounded by the relatively small normalized tree length of both, particularly the short subtree.

Branch length re-scaling also impacted the performance of FunDi in two different ways depending on which branches were being re-scaled. When the branch separating the two subtrees was re-scaled performance tended to decrease with increasing length, with the effect most noticeable on the tree we had identified as poorly performing. As this branch lengthens the amount of overall divergence between the two subtrees increases, including for sites not contributing to any sort of functional divergence. This long branch closely resembles the alternative 'independent' model of functional divergence of FunDi making it more difficult to distinguish between truly functionally divergent and non-divergent sites. In contrast as the branches within subtrees are lengthened the amount of divergence time within subtrees is increased. Because this lengthening does not reflect the component of the mixture model used to model functional divergence, and functionally divergent sites are being modeled under altered substitution models, there is more time to "lock in" those evolutionary differences compared to the neutral substitutions in non-divergent sites.

It is clear that the overall properties of the phylogenetic tree in question must be carefully considered when conducting any sort of analysis of functional divergence. Future work to quantify the impact these factors may have on determining suitable cut-offs for functional divergence scoring would lead to improved predictive performance on biological datasets.

5.3 Functional Divergence in GAPDH

Additional analyses on the GAPDH dataset in Chapter 4 showed that compared to GroupSim and the Difference Evolutionary-Trace method, FunDi was better able to identify residues of convergent functional divergence between the two groups of plastid-targeted sequences that had been verified by previous experimental work. Given the complex evolutionary history of GAPDH in this dataset, and the fact that we maximized sequence diversity by including paralogs from as many eukaryotic taxa as possible, the use of a robust phylogenetic framework with appropriate models of sequence evolution is key to distinguishing between sites undergoing functional divergence and those we might expect under purely neutral evolutionary conditions.

Phylogenomic approaches and the use of maximal sequence and taxonomic diversity are key to improving predictions of protein function. While it possible to create information theoretic or machine-learning based models to detect sequence composition patterns associated with a function, these are always based on an incomplete picture of biological reality. Phylogenetic models of sequence evolution are necessary to distinguish between meaningful evolutionary signal and biological noise.

Bibliography

Abhiman, S., & Sonnhammer, E. L. L. (2005). Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins, 60(4), 758-68.

Aguileta, G., Bielawski, J. P., & Yang, Z. (2004). Gene conversion and functional divergence in the beta-globin gene family. J. Mol Evol., 59(2), 177-89.

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., & Tawfik, D. S. (2005). The "evolvability" of promiscuous protein functions. Nat. Gen., 37(1), 73-6.

Altenhoff, A. M., Schneider, A., Gonnet, G. H., & Dessimoz, C. (2011). OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Res., 39(Database issue), D289-94.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25(17), 3389-402.

Andersson, S. G., Zomorodipour, A, Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C., Podowski, R. M., Näslund, a K., et al. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature, 396(6707), 133-40.

Andersson, S. G. E., Karlberg, O., Canbäck, B., & Kurland, C. G. (2003). On the origin of mitochondria: a genomics perspective. Phil. Trans. Roy. Soc. Lon. B. Biol. Sci., 358(1429), 165-79

Anisimova M., Bielawski J.P., Yang Z. (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19:950-8

Anisimova, M., & Liberles, D. A. (2007). The quest for natural selection in the age of comparative genomics. Heredity, 99(6), 567-79

Archibald, J.M., Cavalier-Smith, T., Maier, U., & Douglas, S. (2001). Molecular chaperones encoded by a reduced nucleus: the cryptomonad nucleomorph. J. Mol Evol., 52(6), 490-501

Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics, 22(2), 195-201

Atteia, A., Adrait, A., Brugière, S., Tardif, M., van Lis, R., Deusch, O., Dagan, T., et al. (2009). A proteomic survey of Chlamydomonas reinhardtii mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. Mol. Biol. Evol., 26(7), 1533-48

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Carlton, J. M., Dommer, J., Fischer, S., Gajria, B., et al. (2009). GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*. Nucleic Acids Res., 37(Database issue), D526-30

Aurrecoechea, C., Brestelli, J., Brunk, B. P., Fischer, S., Gajria, B., Gao, X., Gingle, A., et al. (2010). EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res., 38(Database issue), D415-9

Baalmann, E., Scheibe, R., Cerff, R., & Martin, W. (1996). Functional studies of chloroplast glyceraldehyde-3-phosphate dehydrogenase subunits A and B expressed in Escherichia coli: formation of highly active A4 and B4 homotetramers and evidence that aggregation of the B4 complex is mediated by the B subunit carb. Plant Mol. Biol., 32(3), 505-13

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics, 16(5), 412-424.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18, 298–305.

Barberà, M. J., Ruiz-Trillio, I., Leigh, J., Hug, L. A., and Roger, A. J. (2007). The diversity of mitochondrion-related organelles amongst eukaryotic microbes. In ‘‘Origin of Mitochondria and Hydrogenosomes’’ (W. F. Martin and M. Müller, eds), pp. 239–275. Springer, Berlin.

Barkman, T.J. et al. (2007) Positive Selection for Single Amino Acid Change Promotes Substrate Discrimination of a Plant Volatile-Producing Enzyme. Mol. Biol. Evol. 24(6): 1320-9

Baurain, D., Brinkmann, H., Petersen, J., Rodríguez-Ezpeleta, N., Stechmann, A., Demoulin, V., Roger, A. J., et al. (2010). Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol. Biol. Evol., 27(7), 1698-709

Ben-Menachem, R., Regev-Rudzki, N., & Pines, O. (2011). The aconitase C-terminal domain is an independent dual targeting element. Jour. Mol. Biol., 409(2), 113-23

Berglund, A.-C., Sjölund, E., Ostlund, G., & Sonnhammer, E. L. L. (2008). InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res., 36(Database issue), D263-6

Betrán, E., & Long, M. (2003). Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics, 164(3), 977-88.

Bielawski, J. P., & Yang, Z. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol Evol., 59(1), 121-32

Biesecker, G., Harris, J. I., Thierry, J. C., Walker, J. E., & Wonacott, A. J. (1977). Sequence and structure of D-glyceraldehyde 3-phosphate dehydrogenase from Bacillus stearothermophilus. Nature, 266(5600), 328-33.

Bloom, J. D., Raval, A., & Wilke, C. O. (2007). Thermodynamics of neutral protein evolution. Genetics, 175(1), 255-66

Blouin, C., Boucher, Y., & Roger, A. J. (2003). Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res., 31(2), 790-797

Blouin, C., Butt, D., and Roger A.J. 2005. Impact of taxon sampling on the estimation of rates of evolution at sites. Mol. Biol. Evol. 22(3):784-91

Blum, T., Briesemeister, S., & Kohlbacher, O. (2009). MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. BMC Bioinformatics, 10, 274

Bodén, M., and Hawkins, J. (2005). Prediction of subcellular localisation using sequence biased recurrent networks. Bioinformatics 21, 2279–2286.

Boeckmann, B., Robinson-Rechavi, M., Xenarios, I., & Dessimoz, C. (2011). Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. Brief. Bioinformatics, 12(5)

Boggetto, N., Gontero, B., & Maberly, S. C. (2007). Regulation of phosphoribulokinase and glyceraldehyde 3-phosphate dehydrogenase in a freshwater diatom, Asterionella formosa 1. Journal of Phycology, 43(6), 1227-1235

Bolender, N., Sickmann, A., Wagner, R., Meisinger, C., and Pfanner, N. (2008). Multiple pathways for sorting mitochondrial precursor proteins. EMBO Rep. 9, 42–49.

Boxma, B., de Graaf, R. M., van der Staay, G. W., van Alen, T. A., Ricard, G., Gabaldón, T., van Hoek, A. H., Moon-van der Staay, S. Y., Koopman, W. J., Bui, E. T., Bradley, P. J., and Johnson, P. J. (1996). A common evolutionary origin for mitochondria and hydrogenosomes. Proc. Natl. Acad. Sci. USA 93, 9651–9656

Bradley, P. J., Lahti, C. J., Plümper, E., & Johnson, P. J. (1997). Targeting and translocation of proteins into the hydrogenosome of the protist Trichomonas: similarities with mitochondrial protein import. The EMBO journal, 16(12), 3484-93

Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., et al. (2009). Fast statistical alignment. PLoS Comp. Biol., 5(5), e1000392

Brandt, B.W., Feenstra, K.A., Heringa, J. 2010. Multi-Harmony: detecting functional specificity from sequence alignment. Nucl. Acids Res. 38:W35-40

Briesemeister, S., Blum, T., Brady, S., Lam, Y., Kohlbacher, O., & Shatkay, H. (2009). SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. Journal of Proteome Research, 8(11), 5363-6

Brindefalk, B., Ettema, T. J. G., Viklund, J., Thollesson, M., & Andersson, S. G. E. (2011). A Phylometagenomic Exploration of Oceanic Alphaproteobacteria Reveals Mitochondrial Relatives Unrelated to the SAR11 Clade. PLoS ONE, 6(9), e24457. doi:10.1371/journal.pone.0024457

Brinkmann, H, Martinez, P., Quigley, F., Martin, W., & Cerff, R. (1987). Endosymbiotic origin and codon bias of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. J. Mol Evol., 26(4), 320-8.

Buchanan, B B. (1980). Role of Light in the Regulation of Chloroplast Enzymes. Annual Review of Plant Physiology, 31(1), 341-374. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA

Buchanan, Bob B, & Balmer, Y. (2005). Redox regulation: a broadening horizon. Ann. Rev. Plant Biol., 56, 187-220

Buchanan, Bob B. (1984). The Ferredoxin/Thioredoxin System: A Key Element in the Regulatory Function of Light in Photosynthesis. BioScience, 34(6), 378-383

Burki, F., Inagaki, Y., Bråte, J., Archibald, J. M., Keeling, P. J., Cavalier-Smith, T., Sakaguchi, M., et al. (2009). Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, telonemia and centroheliozoa, are related to photosynthetic chromalveolates. Gen. Biol. Evol., 1, 231-8

Burki, F., Shalchian-Tabrizi, K., & Pawlowski, J. (2008). Phylogenomics reveals a new “megagroup” including most photosynthetic eukaryotes. Biology letters, 4(4), 366-9

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjaeveland, A., Nikolaev, S. I., Jakobsen, K. S., & Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE, 2(8), e790

Butterfield, D. A., Hardas, S. S., & Lange, M. L. B. (2010). Oxidatively modified glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and Alzheimer’s disease: many pathways to neurodegeneration. Journal of Alzheimer’s disease : JAD, 20(2), 369-93

Caffrey, D.R., Lunney, E. Moshinsky, D.J. (2008). Prediction of specificity-determining residues for small-molecule kinase inhibitors. BMC Bioinformatics, 9: 49

Cao, S., Kumimoto, R. W., Siriwardana, C. L., Risinger, J. R., & Holt, B. F. (2011). Identification and characterization of NF-Y transcription factor families in the monocot model plant Brachypodium distachyon. PLoS ONE, 6(6), e21805

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972-3

Capra, J. A, & Singh, M. (2008). Characterization and prediction of residues determining protein functional specificity. Bioinformatics, 24(13), 1473-80

Capra, J.A, et al (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comp. Biol., 5(12): e1000585

Capra, J.A. and Singh, M. 2007. Predicting functionally important residues from sequence conservation. Bioinformatics, 23(15):1875-82.

Carlton, Jane M, Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zhao, Q., Wortman, J. R., et al. (2007). Draft Genome Sequence of the Sexually Transmitted Pathogen Trichomonas vaginalis. Science, 315(5809), 207-212.

Carrie, C., Giraud, E., & Whelan, J. (2009). Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. The FEBS journal, 276(5), 1187-95

Carugo, O, & Argos, P. (1997). NADP-dependent enzymes. I: Conserved stereochemistry of cofactor binding. Proteins, 28(1), 10-28

Carugo, Oliviero. (2007). Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots. BMC Bioinformatics, 8, 380

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol., 17(4), 540-52.

Cavalier-Smith, T. (1999). Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. Jour. Euk. Microbiol., 46(4), 347-66

Cavalier-Smith, T. (2002). Chloroplast evolution: secondary symbiogenesis and multiple losses. Curr. Biol., 12(2), R62-4.

Cavalier-Smith, T. (2003). Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). Phil. Trans. Roy. Soc. Lon. B. Biol. Sci., 358(1429), 109-33; discussion 133-4

Cerff, R, & Chambers, S. E. (1978). Glyceraldehyde-3-phosphate dehydrogenase (NADP) from Sinapis alba L. Isolation and electrophoretic characterization of isoenzymes. Hoppe-Seyler's Zeitschrift für physiologische Chemie, 359(6), 769-72

Cerff, R. (1979). Quaternary structure of higher plant glyceraldehyde-3-phosphate dehydrogenases. European journal of biochemistry / FEBS, 94(1), 243-7

Cerff, Rüdiger, & Chambers, S. E. (1979). Subunit Structure of Higher Plant Glyceraldehyde-3-phosphate Dehydrogenases (EC 1.2.1.12 and EC 1.2.1.13)*. The Jour. Biol. Chem., 254(13), 6094-6098

Cerkasovová, A., Lukasová, G., Cerkasòv, J., and Kulda, J. (1973). Biochemical characterization of large granule fraction of *Tritrichomonas foetus* (strain KV1). J. Protozool. 20, 525

Chakrabarti, S. and Panchenko, A.R. (2009) Ensemble approach to predict specificity determinants: benchmarking and validation. BMC Bioinformatics. 10:207

Chakrabarti, S., Bryant, S. H., & Panchenko, A. R. (2007). Functional specificity lies within the properties and evolutionary changes of amino acids. Jour. Mol. Biol., 373(3), 801-10.

Chen, F., Mackey, A.J., Stoeckert, C. J., & Roos, D.S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res., 34(Database issue), D363-8

Chen, F., Mackey, A. J., Vermunt, J. K., & Roos, D. S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE, 2(4), e383

Chiu, J. C., Lee, E. K., Egan, M. G., Sarkar, I. N., Coruzzi, G. M., & DeSalle, R. (2006). OrthologID: automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics, 22(6), 699-707

Chou, K.-C., & Shen, H.B. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE, 5(4), e9931

Chou, K.-C., Wu, Z.-C., & Xiao, X. (2011). iLoc-Euk: A Multi-Label Classifier for Predicting the Subcellular Localization of Singleplex and Multiplex Eukaryotic Proteins. (C. Schönbach, Ed.)PLoS ONE, 6(3), e18258

Chou, K.-chen, & Shen, H.-bin. (2007). Euk-mPLoc : A Fusion Classifier for Large-Scale Eukaryotic Protein Subcellular Location Prediction by Incorporating Multiple Sites research articles. Journal of Proteome Research, 6, 1728 – 1734

Claros, M. G., and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur. J. Biochem. 241, 779–786

Clermont, S., Corbier, C., Mely, Y., Gerard, D., Wonacott, A., & Branlant, G. (1993). Determinants of coenzyme specificity in glyceraldehyde-3-phosphate dehydrogenase: Role of the acidic residue in the fingerprint region of the nucleotide binding fold. Biochemistry, 32(38), 10178-10184

Colell, A., Green, D., & Ricci, J.E. (2009). Novel roles for GAPDH in cell death and carcinogenesis. Cell death and differentiation, 16(12), 1573-81

Consortium, T. U. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res., 39(Database issue), D214-9

Corbier, C., Mougin, A., Mely, Y., Adolph, H. W., Zeppezauer, M., Gerard, D., Wonacott, A., et al. (1990). The nicotinamide subsite of glyceraldehyde-3-phosphate dehydrogenase studied by site-directed mutagenesis. Biochimie, 72(8), 545-554

Courtiade, J., Pauchet, Y., Vogel, H., & Heckel, D. G. (2011). A comprehensive characterization of the caspase gene family in insects from the order Lepidoptera. BMC Genomics, 12(1), 357

Cristianini, N., and Shawe-Taylor, J. (2000). ‘‘An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods’’ Cambridge University Press, Cambridge.

Cuervo, A. M., Wong, E. S. P., & Martinez-Vicente, M. (2010). Protein degradation, aggregation, and misfolding. Movement Disorders, 25 Suppl 1, S49-54

Danne, J. C., & Waller, R. F. (2011). Analysis of dinoflagellate mitochondrial protein sorting signals indicates a highly stable protein targeting system across eukaryotic diversity. Jour. Mol. Biol., 408(4), 643-53

Datta, R. S., Meacham, C., Samad, B., Neyer, C., & Sjolander, K. (2009). Berkeley PHOG : PhyloFacts orthology group prediction web server. Nucleic Acids Res, 37, 84-89

Daugbjerg, N., & Andersen, R. a. (1997). Phylogenetic analyses of the rbcL sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. Mol. Biol. Evol., 14(12), 1242-51

Davis, J. and Goadrich, D. (2006) The relationship between precision-recall and ROC curves. 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA. June 26th-28th.

Dayhoff, M. O., Schwartz, R. M., & Orcutt, B. C. (1978). A model of evolutionary change in proteins. (M. O. Dayhoff, Ed.)Atlas of protein sequence and structure, 5(Suppl 3), 345-352

De Grassi, A., Lanave, C., & Saccone, C. (2008). Genome duplication and gene-family evolution: the case of three OXPHOS gene families. Gene, 421(1-2), 1-6

Delage, L., Leblanc, C., Nyvall Collén, P., Gschloessl, B., Oudot, M.-P., Sterck, L., Poulain, J., et al. (2011). In silico survey of the mitochondrial protein uptake and maturation systems in the brown alga Ectocarpus siliculosus. PLoS ONE, 6(5), e19540

DePristo, M. A, Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet., 6(9), 678-87

Didierjean, C., Rahuel-Clermont, S., Vitoux, B., Dideberg, O., Branlant, G., & Aubry, A. (1997). A crystallographic comparison between mutated glyceraldehyde-3-phosphate dehydrogenases from Bacillus stearothermophilus complexed with either NAD+ or NADP+. Jour. Mol. Biol., 268(4), 739-59

Dinur-Mills, M., Tal, M., and Pines, O. (2008). Dual targeted mitochondrial proteins are characterized by lower MTS parameters and total Net charge. PLoS ONE 3, e2161

Dolezal, P., Likic, V., Tachezy, J., and Lithgow, T. (2006). Evolution of the molecular machines for protein import into mitochondria. Science 313, 314–318.

Dolezal, P., Smíd, O., Rada, P., Zubácová, Z., Bursać, D., Suták, R., Nebesárová, J., et al. (2005). Giardia mitosomes and trichomonad hydrogenosomes share a common mode of protein targeting. Proc. Natl. Acad. Sci. USA, 102(31), 10924-9

Douglas, S. E., & Penny, S. L. (1999). The plastid genome of the cryptophyte alga, Guillardia theta: complete sequence and conserved synteny groups confirm its common ancestry with red algae. J. Mol Evol., 48(2), 236-44

Dumke, R., Hausner, M., & Jacobs, E. (2011). Role of Mycoplasma pneumoniae glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in mediating interactions with the human extracellular matrix. Microbiology (Reading, England), 157(Pt 8), 2328-38

Dunker, A.K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., Vacic, V., et al. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. BMC genomics, 9 Suppl 2, S1

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics, 14(9), 755-763

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comp. Biol., 7(10), e1002195

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-1

Eisen, J.A. (1998a). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research, 8(3), 163-167

Eisen, J.A. (1998b). A phylogenomic study of the MutS family of proteins. Nucleic Acids Res., 26(18), 4291-4300

Eisen, J.A., Sweder, K. S., & Hanawalt, P. C. (1995). Evolution of the SNF2 family of proteins : subfamilies with distinct sequences and functions. Nucleic Acids Res., 23(14), 2715-2723

Eisen, J.A. (1998). Phylogenomics : Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Research, 8, 163-167

Eisen, J.A., Kaiser, D., & Myers, R. A. (1997). Gastrogenomic delights : A movable feast. Nat. Med., 3(10), 1076-1078

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300, 1005–1016.

Embley, T. M. (2006). Multiple secondary origins of the anaerobic lifestyle in eukaryotes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361, 1055–1067.

Embley, T. M., and Martin, W. (2006). Eukaryotic evolution, changes and challenges. Nature 440, 623–630

Engelhardt B.E., Jordan M.I., Muratore K.E., Brenner S.E. 2005. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comp. Biol. 1:e45

Engelhardt, B. E., Jordan, M. I., Repo, S. T., & Steven, E. (2009). Phylogenetic molecular function annotation. J Phys, 180(1), 1-7

Engelhardt, B. E., Jordan, M. I., Srouji, J. R., & Brenner, S. E. (2011). Genome-scale phylogenetic function annotation of large and diverse protein families. Gen. Res. Epub Ahead of Print

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res., 30(7), 1575-1584

Erales, J., Gontero, B., & Maberly, S. C. (2008). Specificity and Function of Glyceraldehyde-3-Phosphate Dehydrogenase in a Freshwater Diatom, *Asterionella Formosa* (Bacillariophyceae) 1. Journal of Phycology, 44(6), 1455-1464

Erales, J., Lignon, S., & Gontero, B. (2009). CP12 from Chlamydomonas reinhardtii, a permanent specific “chaperone-like” protein of glyceraldehyde-3-phosphate dehydrogenase. Jour. Biol. Chem., 284(19), 12735-44

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., et al. (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol., 21(9), 1643-60

Eyschen, J., Vitoux, B., Rahuel-Clermont, S., Marraud, M., Branlant, G., & Cung, M. T. (1996). Phosphorus-31 nuclear magnetic resonance studies on coenzyme binding and specificity in glyceraldehyde-3-phosphate dehydrogenase. Biochemistry, 35(19), 6064-6072

Fagan, T., Woodland Hastings, J., & Morse, D. (1998). The phylogeny of glyceraldehyde-3-phosphate dehydrogenase indicates lateral gene transfer from cryptomonads to dinoflagellates. J. Mol Evol., 47(6), 633-9

Falini, Giuseppe, Fermani, S., Ripamonti, A., Sabatino, P., Sparla, F., Pupillo, P., & Trost, P. (2003). Dual coenzyme specificity of photosynthetic glyceraldehyde-3-phosphate dehydrogenase interpreted by the crystal structure of A4 isoform complexed with NAD. Biochemistry, 42(16), 4631-9

Fast, N. M., Kissinger, J. C., Roos, D. S., & Keeling, P. J. (2001). Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol. Biol. Evol., 18(3), 418-26

Feenstra K.A., Pirovano W., Krab K., Heringa J. (2007). Sequence harmony: detecting functional specificity from alignments. Nuc. Acid. Res. 35:W495-8

Fermani S. et al. 2007. Molecular mechanism of thioredoxin regulation in photosynthetic A2B2-glyceraldehyde-3-phosphate dehydrogenase. Proc. Natl. Acad. Sci. USA. 104:11109-14

Fermani, S, Ripamonti, A., Sabatino, P., Zanotti, G., Scagliarini, S., Sparla, F., Trost, P., et al. (2001). Crystal structure of the non-regulatory A(4)isoform of spinach chloroplast glyceraldehyde-3-phosphate dehydrogenase complexed with NADP. Jour. Mol. Biol., 314(3), 527-42

Ferri, G., Comerio, G., Iadarola, P., Zapponi, M. C., & Speranza, M. L. (1978). Subunit structure and activity of glyceraldehyde-3-phosphate dehydrogenase from spinach chloroplasts. Biochimica et biophysica acta, 522(1), 19-31

Finlay, B. J., Span, A. S. W., and Harman, J. M. P. (1983). Nitrate respiration in primitive eukaryotes. Nature 303, 333–336

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Systematic Zoology, 19(2), 99-113

Fitch, W.M. (2000). Homology a personal view on some of the problems. Trends in Genetics, 16, 227-231

Fitzpatrick, D.A, Creevey, C. J., & McInerney, J. O. (2006). Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. Mol. Biol. Evol., 23(1), 74-85

Fletcher, W. and Yang, Z. (2009) INDELible: a flexible simulator of biological sequence evolution. Mol. Biol. Evol 26(8):1879-88

Fong, J. H., & Panchenko, A. R. (2010). Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. Molecular BioSystems, 6(10), 1821-8

Fong, J. H., Shoemaker, B. A, Garbuzynskiy, S. O., Lobanov, M. Y., Galzitskaya, O. V., & Panchenko, A. R. (2009). Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. PLoS Comp. Biol., 5(3), e1000316

Force, S.A., Lynch, M., Pickett, F. B., Amores, A, Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics, 151(4), 1531-45

Forsberg, R., & Christiansen, F. B. (2003). A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol. Biol. Evol., 20(8), 1252-1259

Frech, C., & Chen, N. (2010). Genome-wide comparative gene family classification. PLoS ONE, 5(10), e13409

Gabriel, K., and Pfanner, N. (2007). The mitochondrial machinery for import of precursor proteins. Methods Mol. Biol. 390, 99–117

Gardebien, F., Thangudu, R. R., Gontero, B., & Offmann, B. (2006). Construction of a 3D model of CP12, a protein linker. Journal of molecular graphics & modelling, 25(2), 186-95

Gaston, D., Tsaousis, A. D., & Roger, A. J. (2009). Predicting Proteomes of Mitochondria and Related Organelles from Genomic and Expressed Sequence Tag Data. Mitochondrial Function, Part B: Mitochondrial Protein Kinases, Protein Phosphatases and Mitochondrial Diseases, 457(09), 21-47

Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Brief. Bioinformatics, bbr042

Georgiades, K., Madoui, M.-A., Le, P., Robert, C., & Raoult, D. (2011). Phylogenomic Analysis of Odyssella thessalonicensis Fortifies the Common Origin of Rickettsiales, Pelagibacter ubique and Reclimonas americana Mitochondrion. (Wenjun Li, Ed.)PLoS ONE, 6(9), e24857

Gerlt, J.A. and Babbitt, P.C. (2000) Can sequence determine function? Genome Biol. 1(5): reviews0005.1-0005.10

Germot, A., Phillipe, H., and Le Guyader, H. (1996). Presence of a mitochondrial-type 70-kDa heat shock protein in Trichomonas vaginalis suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc. Natl. Acad. Sci. USA 93: 14614–14617

Gharib, W. H., & Robinson-Rechavi, M. (2011). When orthologs diverge between human and mouse. Brief. Bioinformatics, 12(5), 436-441

Gill, E. E., Diaz-Triviño, S., Barberà, M. J., Silberman, J. D., Stechmann, A., Gaston, D., Tamas, I., et al. (2007). Novel mitochondrion-related organelles in the anaerobic amoeba Mastigamoeba balamuthi. Molecular microbiology, 66(6), 1306-20

Goldberg, A. V., Molik, S., Tsaousis, A. D., Neumann, K., Kuhnke, G., Delbac, F., Vivares, C. P., Hirt, R. P., Lill, R., and Embley, T.M.(2008). Localization and functionality of microsporidian iron–sulphur cluster assembly proteins. Nature 452, 624–628

Golding, G.B. and Dean, A.M. (1998) The structural basis of molecular adaptation. Mol. Biol. Evol. 15:355-69

Gould, S. B., Waller, R. F., & McFadden, G. I. (2008). Plastid evolution. Ann. Rev. Plant Biol., 59, 491-517

Graciet, E., Gans, P., Wedel, N., Lebreton, S., Camadro, J.-M., & Gontero, B. (2003). The small protein CP12: a protein linker for supramolecular complex assembly. Biochemistry, 42(27), 8163-70

Graciet, E., Lebreton, S., & Gontero, B. (2004). Emergence of new regulatory mechanisms in the Benson-Calvin pathway via protein-protein interactions: a glyceraldehyde-3-phosphate dehydrogenase/CP12/phosphoribulokinase complex. Journal of experimental botany, 55(400), 1245-54

Gray, M. W. (1999). Mitochondrial Evolution. Science, 283(5407), 1476-1481

Gray, M. W., Burger, G., and Lang, B. F. (1999). Mitochondrial evolution. Science 283, 1476–1481

Gray, M. W., Lang, B. F., and Burger, G. (2004). Mitochondria of protists. Annu. Rev. Genet. 38, 477–524

Gray, M. W., Lang, B. F., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T. G., Plante, I., Rioux, P., et al. (1998). Genome structure and gene content in protist mitochondrial DNAs. Nucl. Acids Res. 26, 865–878

Gray, M.W., Burger, G., & Lang, B. (2001). The origin and early evolution of mitochondria. Genome Biology, 2(6), 1018.1-1018.5

Gregersen, N., Bolund, L., & Bross, P. (2005). Protein misfolding, aggregation, and degradation in disease. Molecular biotechnology, 31(2), 141-50

Gribaldo, S., Casane, D., Lopez, P., & Philippe, H. (2003). Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. Mol. Biol. Evol., 20(11), 1754-9

Groben, R., Kaloudas, D., Raines, C.A., Offmann, B., Maberly, S. C., & Gontero, B. (2010). Comparative sequence analysis of CP12, a small protein involved in the formation of a Calvin cycle complex in photosynthetic organisms. Photosynthesis research, 103(3), 183-94

Gruber, A., Weber, T., Bártulos, C. R., Vugrinec, S., & Kroth, P. G. (2009). Intracellular distribution of the reductive and oxidative pentose phosphate pathways in two diatoms. Journal of basic microbiology, 49(1), 58-72

Gschloessl, B., Guermeur, Y., & Cock, J. M. (2008). HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinformatics, 9, 393

Gsponer, J., & Babu, M. M. (2009). The rules of disorder or why disorder rules. Progress in biophysics and molecular biology, 99(2-3), 94-103

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol., 16(12), 1664-74

Gu, X. (2001). Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol., 18(4), 453-64

Gu, X. and Vander Velden, K. (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18(3): 500-1

Gu, X. (2006). A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol. Biol. Evol., 23(10), 1937-45

Guda, C., Fahy, E., and Subramaniam, S. (2004a). MITOPRED: A genome-scale method for prediction of nuclear-encoded mitochondrial proteins. Bioinformatics 20, 1785–1794

Guda, C., Guda, P., Fahy, E., and Subramaniam, S. (2004b). MITOPRED: A web server for the prediction of mitochondrial proteins. Nucleic Acids Res. 32, W372–W374

Gurney, K. (1997). ‘‘An Introduction to Neural Networks’’ Taylor & Francis, Bristol, PA.

Hackett, J. D., Yoon, H. S., Li, S., Reyes-Prieto, A., Rümmele, S. E., & Bhattacharya, D. (2007). Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Mol. Biol. Evol., 24(8), 1702-13

Hampl, V., Silbermann, J. D., Stechmann, A., Diaz-Trivino, S., Johnson, P. J., and Roger, A. J. (2008). Genetic evidence for a mitochondriate ancestry in the 'amitochondriate' flagellate *Trimastix pyriformis*. PLoS ONE 3, e1383

Hampl, V., Stairs, C. W., & Roger, A. J. (2011). The tangled past of eukaryotic enzymes involved in anaerobic metabolism. Mobile genetic elements, 1(1), 71-74

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., & Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. Genome research, 19(5), 859-67

Hannenhalli, S. S., & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. Jour. Mol. Biol., 303(1), 61-76

Harper, J. T., & Keeling, P. J. (2003). Nucleus-Encoded, Plastid-Targeted Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) Indicates a Single Origin for Chromalveolate Plastids. Mol. Biol. Evol., 20(10), 1730-1735

Hawkins, J., and Bodén, M. (2006). Detecting and sorting targeting peptides with recurrent networks and support vector machines. J. Bioinform. Comput. Biol. 4, 1–18

Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., Uversky, V. N., et al. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comp. Biol., 2(8), e100

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics, 169(2), 1157-64

Heazlewood, J. L., Tonti-Fillippini, J. S., Gout, A. M., Day, D. A., Whelan, J., and Millar, A. H. (2004). Experimental analysis of the Arabidopsis mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. Plant Cell 16, 241–256

Henikoff, S. et al. (1997) Gene Families: The Taxonomy of Protein Paralogs and Chimeras. Science. 278(5338):609-14

Henze, K, Badr, a, Wettern, M., Cerff, R., & Martin, W. (1995). A nuclear gene of eubacterial origin in Euglena gracilis reflects cryptic endosymbioses during protist evolution. Proc. Natl. Acad. Sci. USA, 92(20), 9122-6

Höglund, A., Blum, T., Brady, S., Dönnes, P., Miguel, J. S., Rocheford, M., Kohlbacher, O., and Shatkay, H. (2006a). Significantly improved prediction of subcellular localization by integrating text and protein sequence data. Proceedings of the Pacific Symposium on Biocomputing (PSB 2006), pp. 16–27

Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W., & Kohlbacher, O. (2006b). MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics, 22(10), 1158-65

Horner, D. S., Hirt, R. P., Kilvington, S., Lloyd, D., and Embley, T. M. (1996). Molecular data suggest an early acquisition of the mitochondrion endosymbiont. Proc. R. Soc. B Biol. Sci. 263, 1053–1059.

Horton, P., Park, K. J., Obayashi, T., and Nakai, K. (2006). Protein subcellular localization prediction with WoLF PSORT. Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference APBC06, Taipei, Taiwan pp. 39–48

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., and Nakai, K. (2007). WoLF PSORT: Protein localization predictor. Nucl. Acids Res. 35, w585–w587

Howe, C. J. (2008). Cellular evolution: What's in a mitochondrion? Curr. Biol. 18, R429–R431.

Hrdy I., Hirt R.P., Dolezal P., Bardonova L., Foster P.G., Tachezy J., Embley T.M. (2004) Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432 (7017), 618-622.

Huang, W.L., Tung, C.W., Ho, S.W., Hwang, S.F., & Ho, S.Y. (2008). ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics, 9, 80

Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., et al. (2009). Ensembl 2009. Nucleic Acids Res., 37(Database issue), D690-7

Huerta-Cepas, J., Capella-gutierrez, S., Pryszcz, L. P., Denisov, I., Kormes, D., Marcet-houben, M., & Gabaldon, T. (2011). PhylomeDB v3 . 0 : an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Database, 39(November 2010), 556-560

Hug, L. a, Stechmann, A., & Roger, A. J. (2010). Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes. Mol. Biol. Evol., 27(2), 311-24

Hughes, T., & Liberles, D. A. (2007). The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. J. Mol Evol., 65(5), 574-88

Inagaki, Y. uj., Blouin, C., Susko, E., & Roger, A. J. (2003). Assessing functional divergence in EF-1 and its paralogs in eukaryotes and archaebacteria. Nucleic Acids Res., 31(14), 4227-4237

Jackson, A. P. (2007). Evolutionary consequences of a large duplication event in Trypanosoma brucei: chromosomes 4 and 8 are partial duplicons. BMC Genomics, 8, 432

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J., et al. (2009). Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics, 10, 136

Jedelský, P. L., Doležal, P., Rada, P., Pyrih, J., Smíd, O., Hrdý, I., Sedinová, M., et al. (2011). The minimal proteome in the reduced mitochondrion of the parasitic protist *Giardia intestinalis*. PLoS ONE, 6(2), e17285

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. Genome Biology, 2(8), interactions1002-interactions1002.3

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics, 11(1), 431

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences, 8(3), 275-82

Kalinina, O. V., Mironov, A. A., Gelfand, M. S., & Rakhmaninova, A. B. (2004). Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Science, 13, 443-456

Kalinina, O. V., Novichkov, P. S., Mironov, A. A, Gelfand, M. S., & Rakhmaninova, A. B. (2004). SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. Nucleic Acids Res., 32(Web Server issue), W424-8

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res., 30(14), 3059-3066

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinformatics, 9(4), 286-98

Katoh, K., & Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics, 26(15), 1899-900

Katoh, K., Kuma, K., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res., 33(2), 511-8

Kecman, V. (2001). ‘‘Learning and Soft Computing—Support Vector Machines, Neural Networks, Fuzzy Logic Systems’’ The MIT Press, Cambridge, MA.

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Res., 37(Database issue), D387-92

Knudsen, B., & Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. USA, 98(25), 14512-7. doi:10.1073/Proc. Natl. Acad. Sci. USA.251526398

Knudsen, B., Miyamoto, M. M., Laipis, P. J., & Silverman, D. N. (2003). Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases. Genetics, 164(4), 1261-9.

Kobayashi, M., Matsuo, Y., Takimoto, A., Suzuki, S., Maruo, F., and Shoun, H. (1996). Denitrification, a novel type of respiratory metabolism in fungal mitochondrion. J. Biol. Chem. 271, 16263–16267

Kondrashov, F. A, Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. Genome biology, 3(2), RESEARCH0008

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. Annual review of genetics, 39, 309-38

Koski, L. B., & Golding, G. B. (2001). The Closest BLAST Hit Is Often Not the Nearest Neighbor. J. Mol Evol., 540-542

Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. Annals of Mathematical Statistics 22 (1): 79–86

Lang, B F, Gray, M. W., & Burger, G. (1999). Mitochondrial genome evolution and the origin of eukaryotes. Annual review of genetics, 33, 351-97

Lartillot, N. and Phillipe, H. (2004) A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino Acid Replacement Process. Mol. Biol. Evol 21(6):1095-1109

Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol., 21(6), 1095-109

Le, S.Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. Mol. Biol. Evol. 25(7): 1307-20

Lebreton, S., Andreescu, S., Graciet, E., & Gontero, B. (2006). Mapping of the interaction site of CP12 with glyceraldehyde-3-phosphate dehydrogenase from *Chlamydomonas reinhardtii*. Functional consequences for glyceraldehyde-3-phosphate dehydrogenase. The FEBS journal, 273(14), 3358-69

Li W., Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics . 22:1658-9

Li W., Jaroszewski L. Godzik A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 17:282-283.

Li W., Jaroszewski .L, Godzik A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics. 18:77-82

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178-89

Li, W.H. (1983) Evolution of duplicated genes. In: Nei M, Koehn RK, editors. Evolution of genes and proteins. Sunderland, Mass: Sinauer Associates. p. 14-37

Li, X., Romero, P., Rani, M., Dunker, A., & Obradovic, Z. (1999). Predicting Protein Disorder for N-, C-, and Internal Regions. Genome informatics. Workshop on Genome Informatics, 10(I), 30-40

Liang, H., Plazonic, K. R., Chen, J., Li, W.-H., & Fernández, A. (2008). Protein underwrapping causes dosage sensitivity and decreases gene duplicability. PLoS genetics, 4(1), e11

Liaud, M. F., Brandt, U., Scherzinger, M., & Cerff, R. (1997). Evolutionary origin of cryptomonad microalgae: two novel chloroplast/cytosol-specific GAPDH genes as potential markers of ancestral endosymbiont and host cell components. J. Mol Evol., 44 Suppl 1, S28-37

Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. Jour. Mol. Biol., 257(2), 342-58

Lin, J. (1991) Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory. 37(1): 145–151

Linard, B., Thompson, J. D., Poch, O., & Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics, 12(1), 11

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein Disorder Prediction. Structure, 11, 1453-1459

Lindmark, D. G., and Müller, M. (1973). Hydrogenosome, a cytoplasmic organelle of the anaerobic flagellate *Tritrichomonas foetus*, and its role in pyruvate metabolism. J. Biol. Chem. 248, 7724–7728.

Lopez, P, Casane, D., & Philippe, H. (2002). Heterotachy, an important process of protein evolution. Mol. Biol. Evol., 19(1), 1-7

Loughran, N. B., O'Connor, B., O'Fágáin, C., & O'Connell, M. J. (2008). The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions. BMC Evol. Biol., 8, 101. doi:10.1186/1471-2148-8-101

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D. S., Poulin, B., Anvik, J., Macdonell, C., and Eisner, R. (2004). Predicting protein subcellular localization of proteins using machine-learned classifiers. Bioinformatics 20, 547–556

Lynch, M., & Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science, 290(5494), 1151-1155

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics, 154(1), 459-73.

Maberly, S. C., Courcelle, C., Groben, R., & Gontero, B. (2010). Phylogenetically-based variation in the regulation of the Calvin cycle enzymes, phosphoribulokinase and glyceraldehyde-3-phosphate dehydrogenase, in algae. Jour. Exp. Bot., 61(3), 735-45

Madabushi, S., Gross, A. K., Philippi, A., Meng, E. C., Wensel, T. G., & Lichtarge, O. (2004). Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. The Jour. Biol. Chem., 279(9), 8126-32. doi:10.1074/jbc.M312671200

Mai, Z., Ghosh, S., Frisardi, M., Rosenthal, B., Rogers, R., and Samuelson, J. (1999). Hsp60 is targeted to a cryptic mitochondrion-derived organelle (''crypton'') in the microaerophilic protozoan parasite Entamoeba histolytica. Mol. Cell Biol. 19, 2198–2205

Marande, W., and Burger, G. (2007). Mitochondrial DNA as a genomic jigsaw puzzle. Science 318, 415

Marcotte, E. M., Xenarios, I., van Der Bliek, A. M., and Eisenberg, D. (2000). Localizing proteins in the cell from their phylogenetic profiles. Proc. Natl. Acad. Sci. USA 97, 12115–12120

Marri, L., Sparla, F., Pupillo, P., & Trost, P. (2005). Co-ordinated gene expression of photosynthetic glyceraldehyde-3-phosphate dehydrogenase, phosphoribulokinase, and CP12 in *Arabidopsis thaliana*. Journal of experimental botany, 56(409), 73-80

Marri, L., Trost, P., Trivelli, X., Gonnelli, L., Pupillo, P., & Sparla, F. (2008). Spontaneous assembly of photosynthetic supramolecular complexes as mediated by the intrinsically unstructured protein CP12. The Jour. Biol. Chem., 283(4), 1831-8

Martens, C., & Van de Peer, Y. (2010). The hidden duplication past of the plant pathogen Phytophthora and its consequences for infection. BMC genomics, 11, 353

Martin, W., Scheibe, R., Schnarrenberger, C., Leegood, R., Sharkey, T., & Caemmerer, S. (2004). The Calvin Cycle and Its Regulation. In R. C. Leegood, T. D. Sharkey, & S. Caemmerer (Eds.), Photosynthesis (Vol. 9, pp. 9-51). Dordrecht: Kluwer Academic Publishers

McCarroll, S. A, & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. Nat. Gen., 39(7 Suppl), S37-42

McClellan, A. J., Tam, S., Kaganovich, D., & Frydman, J. (2005). Protein quality control: chaperones culling corrupt conformations. Nat. Cell Biol., 7(8), 736-41

Mei, S., Wang, F., & Zhou, S. (2011). Gene ontology based transfer learning for protein subcellular localization. BMC Bioinformatics, 12(1), 44

Melo-Minardi, R.C. de, Bastard, K., Artiguenave, F. (2010). Identification of Subfamily-specific Sites based on Active Sites Modeling and Clustering. Bioinformatics 26(24): 3075-3082

Mentel, M., Zimorski, V., Haferkamp, P., Martin, W., & Henze, K. (2008). Protein import into hydrogenosomes of Trichomonas vaginalis involves both N-terminal and internal targeting signals: a case study of thioredoxin reductases. Eukaryotic Cell, 7(10), 1750-7

Michels AK, Wedel N, Kroth PG. 2005. Diatom Plastids Possess a Phosphoribulokinase with an Altered Regulation and No Oxidative Pentose Phosphate Pathway. Plant physiology. 137:911-920

Mihalek, I, Res, I., & Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. Jour. Mol. Biol., 336(5), 1265-82

Mi-ichi, F., Abu Yousuf, M., Nakada-Tsukui, K., & Nozaki, T. (2009). Mitosomes in Entamoeba histolytica contain a sulfate activation pathway. Proc. Natl. Acad. Sci. USA, 106(51), 21731-6

Mirny, L. (2002). Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors. Jour. Mol. Biol., 321(1), 7-20

Montanari, F., Shields, D. C., & Khaldi, N. (2011). Differences in the Number of Intrinsically Disordered Regions between Yeast Duplicated Proteins, and Their Relationship with Functional Divergence. PLoS ONE, 6(9), e24989

Moras, D., Olsen, K. W., Sabesan, M. N., Buehner, M., Ford, G. C., & Rossmann, M. G. (1975). Studies of asymmetry in the three-dimensional structure of lobster D-glyceraldehyde-3-phosphate dehydrogenase. The Jour. Biol. Chem., 250(23), 9137-62

Müller, M. (1993). The hydrogenosome. J. Gen. Microbiol. 139, 2879–2889

Nakai, K., and Horton, P. (1999). PSORT: A program for detecting the sorting signals of proteins and predicting their subcellular localization. Trends Biochem. Sci. 24, 34–35

Nehrt, N. L., Clark, W. T., Radivojac, P., & Hahn, M. W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comp. Biol., 7(6), e1002073

Nepusz, T., Sasidharan, R., & Paccanaro, A. (2010). SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. BMC Bioinformatics, 11, 120

Neuwald, A. F. (2011). Surveying the Manifold Divergence of an Entire Protein Class for Statistical Clues to Underlying Biochemical Mechanisms. Statistical Applications in Genetics and Molecular Biology, 10(1)

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6

Nowak, M. a, Boerlijst, M. C., Cooke, J., & Smith, J. M. (1997). Evolution of genetic redundancy. Nature, 388(6638), 167-71

Ohno, S. (1970). Evolution by gene duplication (p. 160). Berlin: Springer.

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., & Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC genomics, 9 Suppl 1, S1

Oliviera, M.C., & Bhattacharya, D. (2000). Phylogeny of the Bangiophycidae (Rhodophyta) and the Secondary Endosymbiotic Origin of Algal Plastids. American journal of botany, 87(4), 482-492

Opitz, D., & Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research, 11, 169-198

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., Frings, O., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res., 38(Database issue), D196-203

Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., et al. (2008). A mitochondrial protein compendium elucidates complex I disease biology. Cell 134, 112–123.

Palmer, J. D. (1997). Organelle genomes: Going, going, gone! Science 275, 790–791

Palmer, J. D. (2003). THE SYMBIOTIC BIRTH AND SPREAD OF PLASTIDS : HOW MANY TIMES AND WHODUNIT ? Journal of Phycology, 39, 4-11

Patil, A., Kinoshita, K., & Nakamura, H. (2010a). Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. Protein Science, 19(8), 1461-8

Patil, A., Kinoshita, K., & Nakamura, H. (2010b). Hub promiscuity in protein-protein interaction networks. International journal of molecular sciences, 11(4), 1930-43

Patron, N. J., Inagaki, Y., & Keeling, P. J. (2007). Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. Curr. Biol., 17(10), 887-91

Patron, N. J., Rogers, M. B., & Keeling, P. J. (2004). Gene Replacement of Fructose-1, 6-Bisphosphate Aldolase Supports the Hypothesis of a Single Photosynthetic Ancestor of Chromalveolates †. Eukaryotic Cell, 3(5), 1169-1175

Pawlowski, K. and Godzik, A. (2001). Surface map comparison: studying function diversity of homologous proteins. Jour. Mol. Biol., 309(3), 793-806

Peeters, N., and Small, I. (2001). Dual targeting to mitochondria and chloroplasts. Biochem. Biophys. Acta 1541, 54–63

Pei, J., Cai, W., Kinch, L. N., & Grishin, N. V. (2006). Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics, 22(2), 164-71

Peitsch, M. C. (1995). Protein Modeling by E-mail. Biotechnology, 13(7), 658-660

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., & Pupko, T. (2010). GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res., (4), 1-6

Petersen, J., Teich, R., Becker, B., Cerff, R., & Brinkmann, H. (2006). The GapA/B gene duplication marks the origin of Streptophyta (charophytes and land plants). Mol. Biol. Evol., 23(6), 1109-18

Petersen, J., Teich, R., Brinkmann, H., & Cerff, R. (2006). A “green” phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. J. Mol Evol., 62(2), 143-57

Petsalaki, E. I., Bagos, P. G., Litou, Z. I., & Hamodrakas, S. J. (2006). PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics, proteomics & bioinformatics / Beijing Genomics Institute, 4(1), 48-55

Pfanner, N, & Geissler, A. (2001). Versatility of the mitochondrial protein import machinery. Nature reviews. Mol. Cell Biol., 2(5), 339-49

Pfanner, N., Wiedemann, N., Meisinger, C., & Lithgow, T. (2004). Assembling the mitochondrial outer membrane. Nature structural & molecular biology, 11(11), 1044-8

Philippe, H, & Lopez, P. (2001). On the conservation of protein sequences in evolution. Trends in biochemical sciences, 26(7), 414-6

Philippe, H., Casane, D., Gribaldo, S., Lopez, P., & Meunier, J. (2003). Heterotachy and functional shift in protein evolution. IUBMB life, 55(4-5), 257-65

Pierleoni, A., Martelli, P. L., Fariselli, P., & Casadio, R. (2006). BaCelLo: a balanced subcellular localization predictor. Bioinformatics, 22(14), e408-16

Pirovano, W., Feenstra, K. A., & Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. Nucleic Acids Res., 34(22), 6540-8

Plaxton, W. C. (1996). the Organization and Regulation of Plant Glycolysis. Ann. Rev. Plant Phys. and Plant Mol. Biol., 47, 185-214

Pohlmeyer, K., Paap, B. K., Soll, J., & Wedel, N. (1996). CP12: a small nuclear-encoded chloroplast protein provides novel insights into higher-plant GAPDH evolution. Plant Mol. Biol., 32(5), 969-78

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol., 26(7), 1641-50

Price, M.N., Dehail, P.S., Arkin, A.P. (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. PLoS ONE. 5(3):e9490

Prokisch, H., Scharfe, C., Camp, D. G. 2nd, Xiao, W., David, L., Andreoli, C., Monroe, M. E., Moore, R. J., Gritsenko, M. A., Kozany, C., Hixson, K. K., Mottaz, H. M., et al. (2004). Integrative analysis of the mitochondrial proteome in yeast. PLoS Biol. 2, e160.

Pryszcz, L. P., Huerta-Cepas, J., & Gabaldón, T. (2011). MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res., 39(5), e32

Qian, W., & Zhang, J. (2008). Gene dosage and gene duplicability. Genetics, 179(4), 2319-24

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 77, 257–286

Rada, P., Jedelsky, P. L., Bursac, D., Perry, A. J., Miroslava, S., Dolez, P., Hrdy, I., et al. (2011). The Core Components of Organelle Biogenesis and Membrane Transport in the Hydrogenosomes of *Trichomonas vaginalis*. PLoS ONE, 6(9), e24428

Rastogi, S., & Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol. Biol., 5(1), 28

Raviscioni, M., He, Q., Salicru, E. M., Smith, C. L., & Lichtarge, O. (2006). Evolutionary Identification of a Subtype Specific Functional Site in the Ligand Binding Domain of Steroid Receptors. Bioinformatics, 1057(October 2005), 1046 -1057

Regev-Rudzki, N., & Pines, O. (2007). Eclipsed distribution: a phenomenon of dual targeting of protein and its significance. BioEssays : news and reviews in molecular, cellular and developmental biology, 29(8), 772-82

Reinders, J., Zahedi, R. P., Pfanner, N., Meisinger, C., and Sickmann, A. (2006). Toward the complete yeast mitochondrial proteome: Multidimensional separation techniques for mitochondrial proteomics. J. Proteome Res. 5, 1543–1554

Reyes-Prieto, A., Weber, A. P. M., & Bhattacharya, D. (2007). The origin and establishment of the plastid in algae and plants. Annual review of genetics, 41, 147-68

Ricard, G., Molina, J., Chrast, J., Gu, W., Gheldof, N., Pradervand, S., Schütz, F., et al. (2010). Phenotypic consequences of copy number variation: insights from Smith-Magenis and Potocki-Lupski syndrome mouse models. PLoS Biology, 8(11), e1000543

Ridout, K. E., Dixon, C. J., & Filatov, D. A. (2010). Positive selection differs between protein secondary structure elements in Drosophila. Gen. Biol. Evol., 2, 166-79

Rodriguez, G. J., Yao, R., Lichtarge, O., & Wensel, T. G. (2010). Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. Proc. Natl. Acad. Sci. USA, 107(17), 7787-92

Rodríguez-Trelles, F., Tarrío, R., & Ayala, F. J. (2003). Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. Proc. Natl. Acad. Sci. USA, 100(23), 13413-7

Roger, A. J., and Silberman, J. D. (2002). Cell evolution: Mitochondria in hiding. Nature 418, 827–829

Roger, A. J., Clark, C. G., and Doolittle, W. F. (1996). A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis*. Proc. Natl. Acad. Sci. USA 93, 14618–14622

Rokach, L. (2010). Ensemble-based classifiers. Artificial Intelligence Review, 33, 1-39

Romero, O., & Dunker, K. (1997). Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. Genome informatics. Workshop on Genome Informatics, 8, 110-124

Romero, P, Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, a K. (2001). Sequence complexity of disordered protein. Proteins: Structure, Function, and Bioinformatics, 42, 38-48

Roth, A. C., Gonnet, G. H., & Dessimoz, C. (2008). Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics, 9(1), 518

Sael, L. et al. (2008). Rapid comparison of properties on protein surface. Proteins: Structure, Function, and Bioinformatics, 73(1): 1–10

Safi, A., Wallace, K.A., Rusche, L.N. (2008) Evolution of New Function through a Single Amino Acid Change in the Yeast Repressor Sum1p. Mol and Cell Biol. 28(8): 2567-2578

Sankararaman, S. et al. (2010). Active site prediction using evolutionary and structural information. Bioinformatics 26(5), 617-24

Schmidt, H.A., et al.(2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502-4

Schneider, R. E., Brown, M. T., Shiflett, A. M., Dyall, S. D., Hayes, R. D., Xie, Y., Loo, J. A, et al. (2011). *The Trichomonas vaginalis* hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. International journal for parasitology doi:10.1016/j.ijpara.2011.10.001

Schreiber, F., Pick, K., Erpenbeck, D., Wörheide, G., & Morgenstern, B. (2009). OrthoSelect : a protocol for selecting orthologous groups in phylogenomics. BMC Bioinformatics, 12, 1-12

Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., Jovanovic, M., Malmström, J., Brunner, E., et al. (2009). Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS biology, 7(3), e48

Schwartz, R. S., & Mueller, R. L. (2010). Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. BMC Evol. Biol., 10, 5

Shannon, C. E. (1948). A mathematical theory of communication. Bell Systems Technical Journal, 27, 379-423

Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnes, P., & Kohlbacher, O. (2007). SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics, 23(11), 1410-7

Shen, Y. Q., and Burger, G. (2007). 'Unite and conquer': Enhanced prediction of protein subcellular localization by integrating multiple specialized tools. BMC Bioinformatics 8, 420–430

Shin, C. J., Wong, S., Davis, M. J., & Ragan, M. A. (2009). Protein-protein interaction as a predictor of subcellular location. BMC systems biology, 3, 28. doi:10.1186/1752-0509-3-28

Sickmann, A., Reinders, J., Wagner, Y., Joppich, C., Zahedi, R., Meyer, H. E., Schönfisch, B., Perschil, I., Chacinska, A., Guiard, B., Rehling, P., Pfanner, N., et al. (2003).. Proc. Natl. Acad. Sci. USA 100, 13207–13212.

Siltberg-Liberles, J. (2011). Evolution of structurally disordered proteins promotes neostructuralization. Mol. Biol. Evol., 28(1), 59-62. doi:10.1093/molbev/msq291

Simpson, A. G. B., & Roger, A. J. (2004). The real "kingdoms" of eukaryotes. Curr. Biol., 14(17), R693-6

Sjölander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics, 20(2), 170-179

Sjölander, K.et al. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci.Aug;12(4):327-45

Skarzyński, T., Moody, P. C., & Wonacott, A. J. (1987). Structure of holo-glyceraldehyde-3-phosphate dehydrogenase from Bacillus stearothermophilus at 1.8 A resolution. Jour. Mol. Biol., 193(1), 171-87

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4, 1581–1590

Smíd, Ondrej, Matusková, A., Harris, S. R., Kucera, T., Novotný, M., Horváthová, L., Hrdý, I., et al. (2008). Reductive evolution of the mitochondrial processing peptidases of the unicellular parasites trichomonas vaginalis and giardia intestinalis. PLoS pathogens, 4(12), e1000243

Smith, D. G. S., Gawryluk, R. M. R., Spencer, D. F., Pearlman, R. E., Siu, K. W. M., & Gray, M. W. (2007). Exploring the mitochondrial proteome of the ciliate protozoon Tetrahymena thermophila: direct analysis by tandem mass spectrometry. Jour. Mol. Biol., 374(3), 837-63

Sonnhammer, E. L. L., & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. Trends in genetics , 18(12), 619-20

Sparla F et al. 2005. Regulation of Photosynthetic GAPDH Dissected by Mutants. Plant physiology. 138:2210-2219

Sparla, F., Fermani, S., Falini, G., Zaffagnini, M., Ripamonti, A., Sabatino, P., Pupillo, P., et al. (2004). Coenzyme site-directed mutants of photosynthetic A4-GAPDH show selectively reduced NADPH-dependent catalysis, similar to regulatory AB-GAPDH inhibited by oxidized thioredoxin. Jour. Mol. Biol., 340(5), 1025-37

Stairs, C. W., Roger, A. J., & Hampl, V. (2011). Eukaryotic pyruvate formate lyase and its activating enzyme were acquired laterally from a firmicute. Mol. Biol. Evol., 28(7), 2087-99

Stamatakis A. (2006a.) Phylogenetic Models of Rate Heterogeneity : A High Performance Computing Perspective. In: Proceedings of the 20th International Parallel and Distributed Processing Symposium. Rhodes, Greece.

Stamatakis A. (2006b). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics . 22:2688-90

Stechmann, A., Hamblin, K., Pérez-Brocal, V., Gaston, D., Richmond, G. S., van der Giezen, M., Clark, C. G., et al. (2008). Organelles in Blastocystis that blur the distinction between mitochondria and hydrogenosomes. Curr. Biol., 18(8), 580-5

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., et al. (2003). The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. PLoS Biology, 1(2), E45

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. J. Mol Evol., 49(2), 169-81

Storm, C. E., & Sonnhammer, E. L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics, 18(1), 92-99

Strope, C.L., Scott, S.D., Moriyama, E.N. (2007) indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. Mol. Biol. Evol 24(3): 640-9

Strope, C.L.,et al. (2009) Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Mol. Biol. Evol 26(11): 2581-93

Studer, R. A, & Robinson-Rechavi, M. (2010). Large-scale analysis of orthologs and paralogs under covarion-like and constant-but-different models of amino acid evolution. Mol. Biol. Evol., 27(11), 2618-2627

Studer, R. A., & Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? Trends in genetics, 25(5), 210-6

Studer, R. A., Penel, S., Duret, L., & Robinson-Rechavi, M. (2008). Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome Research, 18, 1393-1402

Susko, E. et al. (2002) Testing for differences in rates-across-sites distributions in phylogenetic trees. Mol. Biol. Evol. 19(9): 1514-23

Susko, E., Spencer, M., Roger, A.J. (2005) Biases in phylogenetic estimation can be caused by random sequence segments. J. Mol. Evol. 61(3):351-9.

Sutak, R., Dolezal, P., Fiumera, H. L., Hrdy, I., Dancis, A., Delgadillo-Correa, M., Johnson, P. J., Müller, M., and Tachezy, J. (2004). Mitochondrial-type assembly of Fe-S centers in the hydrogenosomes of the amitochondriate eukaryote *Trichomonas vaginalis*. Proc. Natl. Acad. Sci. USA 101, 10368–10373.

Tachezy, J., Sanchez, L. B., and Müller, M. (2001). Mitochondrial type iron–sulfur cluster assembly in the amitochondriate eukaryotes *Trichomonas* vaginalis and *Giardia intestinalis*, as indicated by the phylogeny of IscS. Mol. Biol. Evol. 18, 1919–1928

Takaya, N., Suzuki, S., Kuwazaki, S., Shoun, H., Maruo, F., Yamaguchi, M., and Takeo, K. (1999). Cytochrome P450nor, a novel class of mitochondrial cytochrome P450 involved in nitrate respiration in the fungus Fusarium oxysporum. Arch. Biochem. Biophys. 372, 340–346

Takishita, K., Yamaguchi, H., Maruyama, T., & Inagaki, Y. (2009). A hypothesis for the evolution of nuclear-encoded, plastid-targeted glyceraldehyde-3-phosphate dehydrogenase genes in “chromalveolate” members. PLoS ONE, 4(3), e4737

Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic biology, 56(4), 564-77

Tekaia, F., & Latgé, J.P. (2005). *Aspergillus fumigatus*: saprophyte or pathogen? Current opinion in microbiology, 8(4), 385-92

Thrash, J. C., Boyd, A., Huggett, M. J., Grote, J., Carini, P., Yoder, R. J., Robbertse, B., et al. (2011). Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Scientific Reports, 1, 1-9

Tielens, A. G., and Van Hellemond, J. J. (1998). The electron transport chain in anaerobically functioning eukaryotes. Biochim. Biophys. Acta 1365, 71–78

Tielens, A. G., Rotte, C., Van Hellemond, J. J., and Martin, W. (2002). Mitochondria as we don’t know them. Trends Biochem. Sci. 27, 564–572

Timmis, J. N., Ayliffe, M. a, Huang, C. Y., & Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet., 5(2), 123-35

Tovar, J., Fischer, A., and Clark, C. G. (1999). The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica. Mol. Microbiol. 32, 1013–1021

Tovar, J., Leon-Avila, G., Sanchez, L. B., Sutak, R., Tachezy, J., van der Giezen, M., Hernandez, M., Mü ller, M., and Lucocq, J. M. (2003). Mitochondrial remnant organelles of Giardia function in iron–sulphur protein maturation. Nature 426, 172–176

Trost, P, Fermani, S., Marri, L., Zaffagnini, M., Falini, G., Scagliarini, S., Pupillo, P., et al. (2006). Thioredoxin-dependent regulation of photosynthetic glyceraldehyde-3-phosphate dehydrogenase: autonomous vs. CP12-dependent mechanisms. Photo. Res., 89(2-3), 263-75

Tsaousis, A. D., Kunji, E. R., Goldberg, A. V., Lucocq, J. M., Hirt, R. P., and Embley, T. M. (2008). A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. Nature 453, 553–556

Tsaousis, A. D., Leger, M. M., Stairs, C. A. W., & Roger, A. J. (2012). The biochemical adaptations of mitochondrion-related organelles of parasitic and free-living microbial eukaryotes to low oxygen environment. In Anoxia (Vol. 21, pp. 51-81) (A. V. Altenbach, J. M. Bernhard, & J. Seckbach Eds.), Dordrecht: Springer Netherlands

Tung, T. Q., & Lee, D. (2009). A method to improve protein subcellular localization prediction by integrating various biological data sources. BMC Bioinformatics, 10 Suppl 1, S43

Turoverov, K. K., Kuznetsova, I. M., & Uversky, V. N. (2010). The protein kingdom extended: ordered and intrinsically disordered proteins, their folding, supramolecular complex formation, and aggregation. Progress in biophysics and molecular biology, 102(2-3), 73-84

Van de Peer, Y, Taylor, J. S., Braasch, I., & Meyer, a. (2001). The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. J. Mol Evol., 53(4-5), 436-46

van der Giezen, M., and Tovar, J. (2005). Degenerate mitochondria. EMBO Rep. 6, 525–530

van der Giezen, M., Slotboom, D. J., Horner, D. S., Dyal, P. L., Harding, M., Xue, G. P., Embley, T. M., and Kunji, E. R. (2002). Conserved properties of hydrogenosomal and mitochondrial ADP/ATP carriers: A common origin for both organelles. EMBO J. 21, 572–579

van der Heijden, R. T. J. M., Snel, B., van Noort, V., & Huynen, M. a. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics, 8, 83. doi:10.1186/1471-2105-8-83

van Hellemond, J. J., Opperdoes, F. R., and Tielens, A. G. (1998). Trypanosomatidae produce acetate via a mitochondrial acetate:succinate CoA transferase. Proc. Natl. Acad. Sci. USA 95, 3036–3041

van Hellemond, J. J., Tielens, A. G., Friedrich, T., et al. (2005). An anaerobic mitochondrion that produces hydrogen. Nature 434, 74–79

Veitia, R. A. (2002). Exploring the etiology of haploinsufficiency. BioEssays : news and reviews in molecular, cellular and developmental biology, 24(2), 175-84

Wall, P. K., Leebens-Mack, J., Müller, K. F., Field, D., Altman, N. S., & dePamphilis, C. W. (2008). PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Res., 36(Database issue), D970-6

Wang, H.-C., Li, K., Susko, E., & Roger, A. J. (2008). A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol. Biol., 8(1), 331

Ward, R. M., Venner, E., Daines, B., Murray, S., Erdin, S., Kristensen, D. M., & Lichtarge, O. (2009). Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics, 25(11), 1426-7

Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J., & Kriventseva, E. V. (2011). OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res., 39(Database issue), D283-8

Weber, T., Gruber, A., & Kroth, P. G. (2009). The presence and localization of thioredoxins in diatoms, unicellular algae of secondary endosymbiotic origin. Molecular plant, 2(3), 468-77

Wedel, N, & Soll, J. (1998). Evolutionary conserved light regulation of Calvin cycle activity by NADPH-mediated reversible phosphoribulokinase/CP12/ glyceraldehyde-3-phosphate dehydrogenase complex dissociation. Proc. Natl. Acad. Sci. USA, 95(16), 9699-704

Wedel, N, Soll, J., & Paap, B. K. (1997). CP12 provides a new mode of light regulation of Calvin cycle activity in higher plants. Proc. Natl. Acad. Sci. USA, 94(19), 10479-84

Whelan, S. and Goldman, N. (2001) A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. Mol. Biol. Evol. 18(5):691-9

Wierenga, R. (1986). Prediction of the occurrence of the ADP-binding-fold in proteins, using an amino acid sequence fingerprint. Jour. Mol. Biol., 187(1), 101-107

Wierenga, R. K., De Maeyer, M. C. H., & Hol, W. G. J. (1985). Interaction of pyrophosphate moieties with alpha-helixes in dinucleotide-binding proteins. Biochemistry, 24(6), 1346-1357

Wilke, C. O., Bloom, J. D., Drummond, D. A., & Raval, A. (2005). Predicting the tolerance of proteins to random amino acid substitution. Biophysical journal, 89(6), 3714-20

Williams, B. A., Hirt, R. P., Lucocq, J. M., and Embley, T. M. (2002). A mitochondrial remnant in the microsporidian Trachipleistophora hominis. Nature 418, 865–869

Williams, G. J. (2009). Rattle : A Data Mining GUI for R. The R Journal, 1(2), 45-55

Wolosiuk, R. a, & Buchanan, B. B. (1978). Activation of Chloroplast NADP-linked Glyceraldehyde-3-Phosphate Dehydrogenase by the Ferredoxin/Thioredoxin System. Plant physiology, 61(4), 669-71

Yang, Z. (1994). Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites : Approximate Methods. J. Mol Evol., 39(3), 306-314

Yang, Z. and Rannala, B. (1997) Bayesian Phylogenetic Inferences Using DNA Sequences: A Markov Chain Monte Carlo Method. Mol. Biol. Evol. 14(7): 717-24

Yao, H., Mihalek, I., & Lichtarge, O. (2006). Rank Information : A Structure-Independent Measure of Evolutionary Trace Quality That Improves Identification of Protein Functional Sites. Bioinformatics, 123: 111-123

Ye, K., Feenstra, K. A., Heringa, J., Ijzerman, A. P., & Marchiori, E. (2008). Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. Bioinformatics, 24(1), 18-25

Ye, K., Lameijer, E.-wubbo M., Beukers, M. W., & Ijzerman, A. P. (2006). A Two-Entropies Analysis to Identify Functional Positions in the Transmembrane Region of Class A G Protein-Coupled Receptors. Proteins: Structure, Function, and Bioinformatics, 63, 1018 - 1030

Yoon, H. S., Hackett, J. D., Pinto, G., & Bhattacharya, D. (2002). The single, ancient origin of chromist plastids. Proc. Natl. Acad. Sci. USA, 99(24), 15507-12

Yu, C. S., Chen, Y. C., Lu, C. H., and Hwang, J. K. (2006). Prediction of protein subcellular localization. Proteins. 64, 643–651

Zapponi, M. C., Iadarola, P., Stoppini, M., & Ferri, G. (1993). Limited proteolysis of chloroplast glyceraldehyde-3-phosphate dehydrogenase (NADP) from Spinacia oleracea. Biological chemistry Hoppe-Seyler, 374(6), 395-402

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. Ann. Rev. Gen. Hum. Gen., 10, 451-81

Zhang, J., Rosenberg, H. F., & Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proceedings of the National Academy of Sciences, 95(7), 3708-3713

Zhang, Z., Green, B. R., & Cavalier-Smith, T. (2000). Phylogeny of Ultra-Rapidly Evolving Dinoflagellate Chloroplast Genes : A Possible Common Origin for Sporozoan and Dinoflagellate Plastids. J. Mol Evol., 51, 26-40

Zheng, L., Roeder, R. G., & Luo, Y. (2003). S phase activation of the histone H2B promoter by OCA-S, a coactivator complex that contains GAPDH as a key component. Cell, 114(2), 255-66

Zmasek, C. M., & Eddy, S. R. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics, 3, 14

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary Divergence and Convergence, in Proteins . In V. Bryson & H. . Vogel (Eds.), Evolving Genes and Proteins (pp. 97-166). New York: Academic Press.

Zwickl, D.J. and Hillis, D.M. (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51(4):588-9

Appendix A Copyright Agreement

Dear Mr. Daniel Gaston,

We hereby grant you permission to reprint the material detailed below at no charge in your thesis subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

“This article was published in Publication title, Vol number, Author(s), Title of article, Page Nos, Copyright Elsevier (or appropriate Society name) (Year).”

3. Your thesis may be submitted to your institution in either print or electronic form.

4. Reproduction of this material is confined to the purpose for which permission is hereby given.

5. This permission is granted for non-exclusive world English rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.

6. This includes permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely,

Kayleigh Harris

Kayleigh Harris :: Rights Associate :: Global Rights :: ELSEVIER

T: +44 (0)1865 843467 :: F: +44 (0)1865 853333

E: k.harris@elsevier.com

Working hours: Tues-Fri, 8.30am-5.30pm

To use the following material:

Title: Predicting proteomes of mitochondria and related organelles from genomic and expressed sequence tag data.

Author(s): D Gaston, AD Tsaousis, AJ. Roger

Volume: 457

Issue: 1

Year: 2009

Pages: 21 - 47

Article title: Predicting proteomes of mitochondra...

How much of the requested material is to be used:

Figure 2.1, Table 2.2

Section 1 (Introduction)

Section 3 (Program Descriptions)

Section 4 (Program Descriptions)

Are you the author: Yes

Author at institute: Yes

How/where will the requested material be used: [how_used]

Details:

Part of this material is to be used in the introduction of my PhD thesis and the introduction of a particular chapter of my thesis. In addition Table 2.2 is to be used in the results section of that chapter. The work on CBOrg presented in this review forms a part of a chapter in my thesis, on predicting subcellular localization along with additional material.

Rights retained by ALL Oxford Journal Authors

The right, after publication by Oxford Journals, to use all or part of the Article and abstract, for their own personal use, including their own classroom teaching purposes;

The right, after publication by Oxford Journals, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;

The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgment is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article and to Oxford University Press and/or the learned society.

The only exception to this is for the re-use of material for commercial purposes, as defined in the information available via the above url. Permission for this kind of re-use is required and can be obtained by using Rightslink:

With Copyright Clearance Center's Rightslink ® service it's faster and easier than ever before to secure permission from OUP titles to be republished in a coursepack, book, CD-ROM/DVD, brochure or pamphlet, journal or magazine, newsletter, newspaper, make a photocopy, or translate.

Simply visit: www.oxfordjournals.org and locate your desired content.

Click on (Order Permissions) within the table of contents and/ or at the bottom article's abstract to open the following page:

Select the way you would like to reuse the content

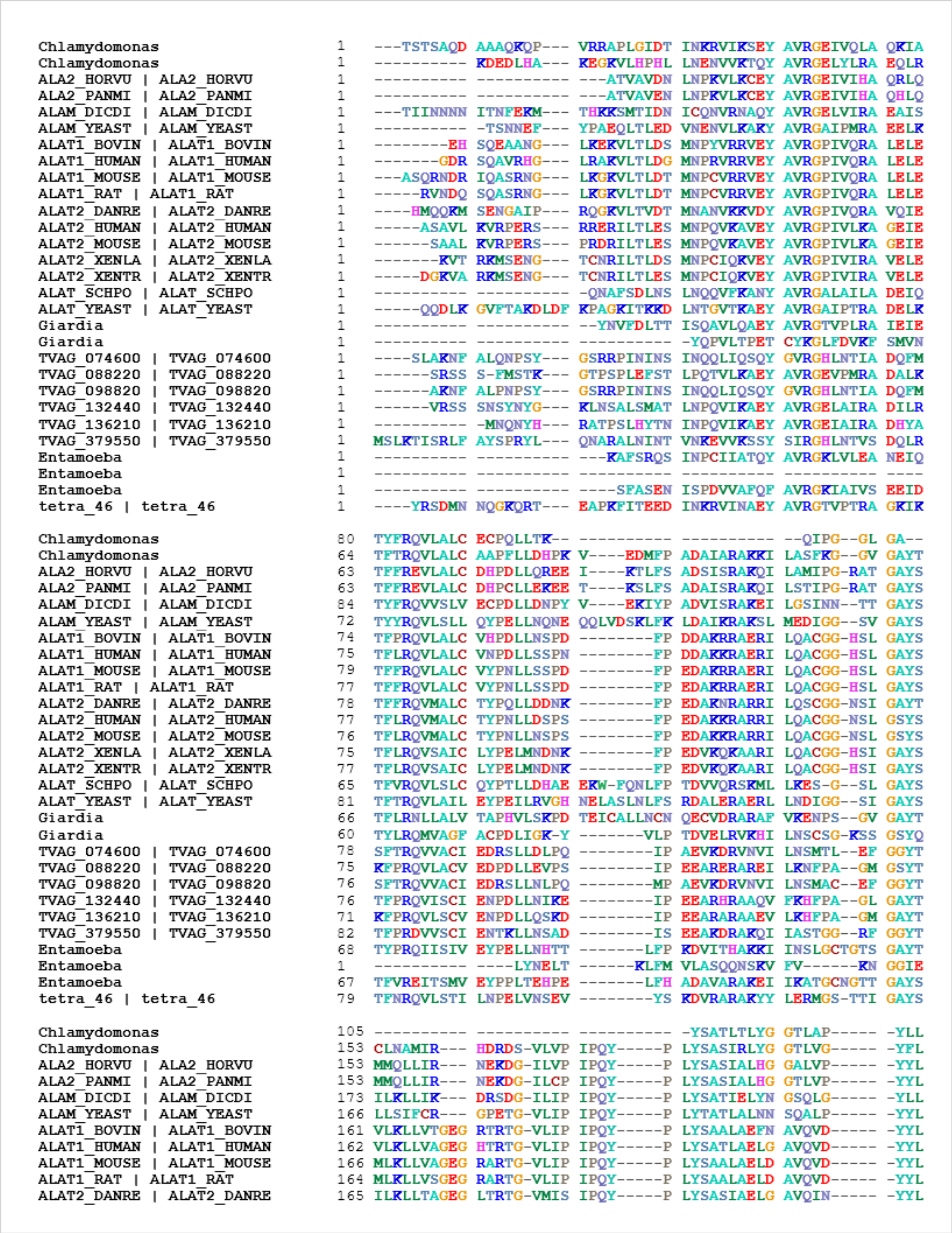
Create an account or login to your existing account

Accept the terms and conditions and permission is granted

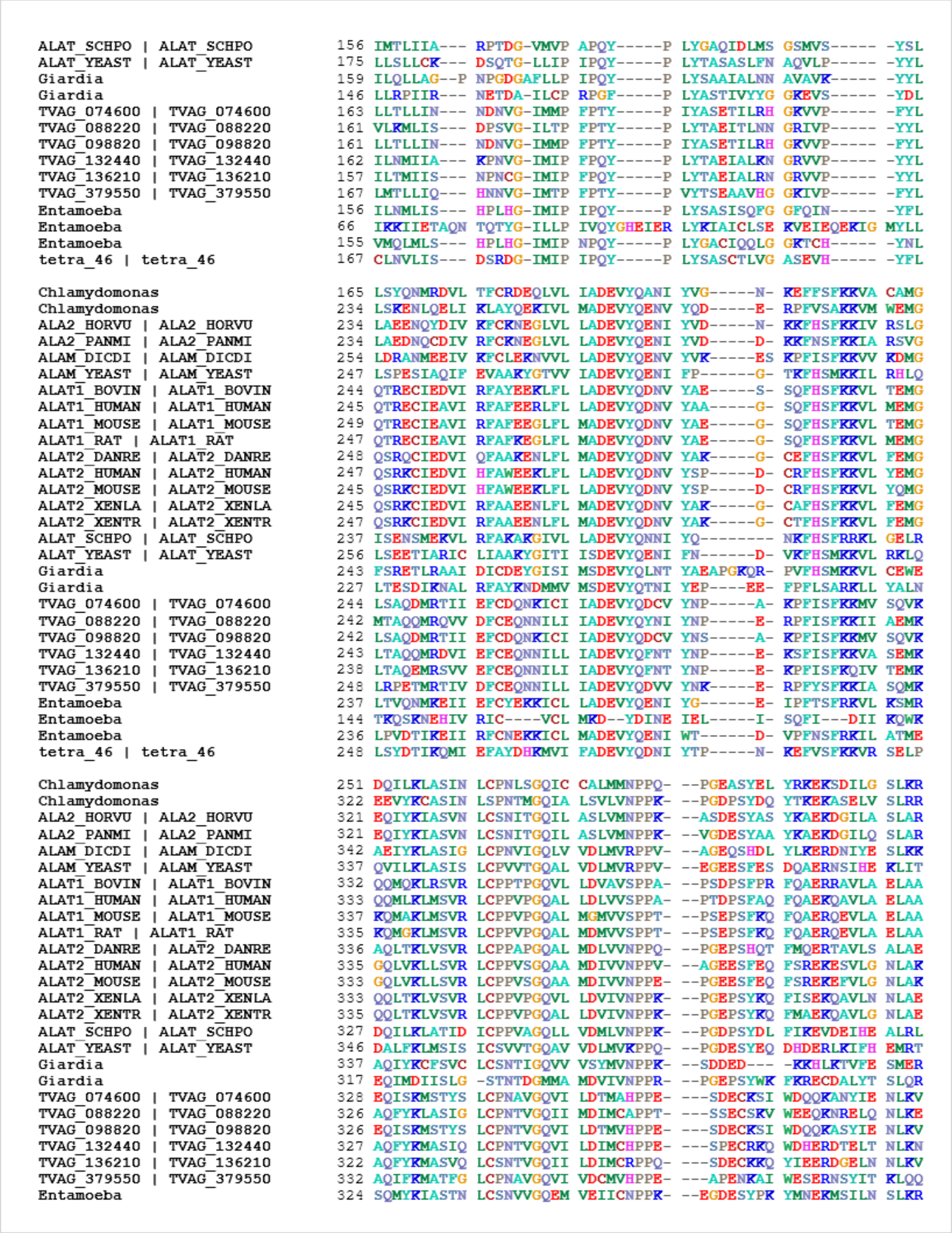
For questions about using the Rightslink service, please contact Customer Support via phone 877/622-5543 (toll free) or 978/777-9929, or email Rightslink customer care.

Appendix B

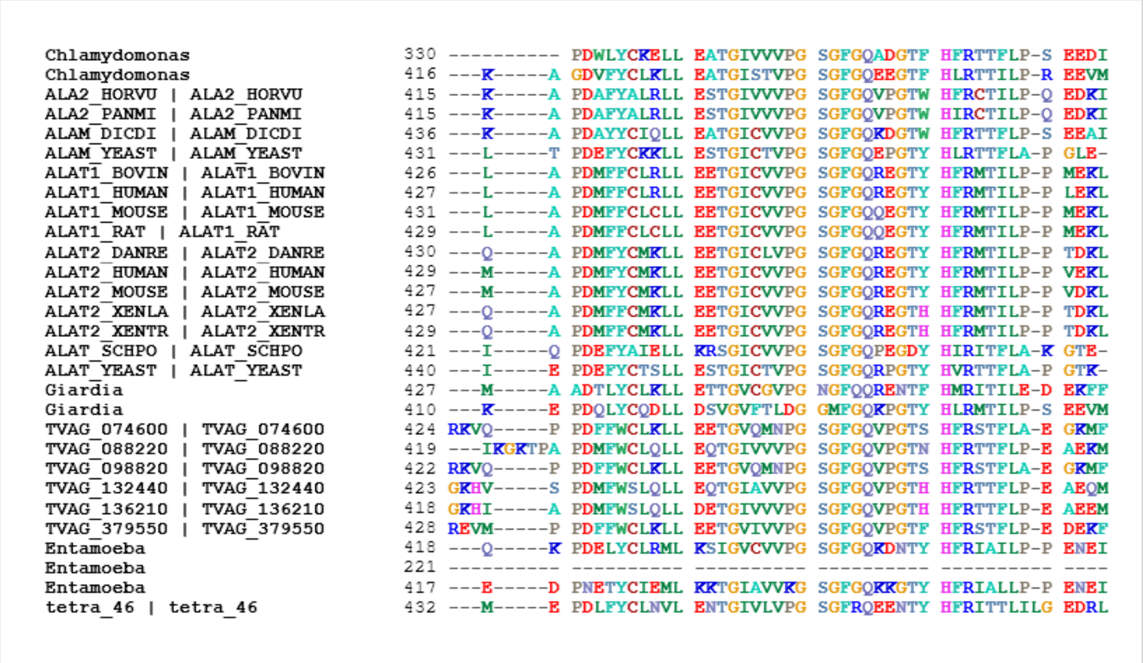
B2 Classification of Subcellular Localization



```
Chlamydomonas                  1   ---TSTSAQD AAAQKQP--- VRRAPLGIDT INKRVIKSEY AVRGEIVQLA QKIA
Chlamydomonas                  1   ---------- KDEDLHA--- KEGKVLHPHL LNENVVKTQY AVRGELYLRA EQLR
ALA2_HORVU | ALA2_HORVU        1   ---------- ---------- ---ATVAVDN LNPKVLKCEY AVRGEIVIHA QRLQ
ALA2_PANMI | ALA2_PANMI        1   ---------- ---------- ---ATVAVEN LNPKVLKCEY AVRGEIVIHA QHLQ
ALAM_DICDI | ALAM_DICDI        1   ---TIINNNN ITNFEKM--- THKKSMTIDN ICQNVRNAQY AVRGELVIRA EAIS
ALAM_YEAST | ALAM_YEAST        1   ---------- -TSNNEF--- YPAEQLTLED VNENVLKAKY AVRGAIPMRA EELK
ALAT1_BOVIN | ALAT1_BOVIN      1   --------EH SQEAANG--- LKEKVLTLDS MNPYVRRVEY AVRGPIVQRA LELE
ALAT1_HUMAN | ALAT1_HUMAN      1   -------GDR SQAVRHG--- LRAKVLTLDG MNPRVRRVEY AVRGPIVQRA LELE
ALAT1_MOUSE | ALAT1_MOUSE      1   ---ASQRNDR IQASRNG--- LKGKVLTLDT MNPCVRRVEY AVRGPIVQRA LELE
ALAT1_RAT | ALAT1_RAT          1   -----RVNDQ SQASRNG--- LKGKVLTLDT MNPCVRRVEY AVRGPIVQRA LELE
ALAT2_DANRE | ALAT2_DANRE      1   ----HMQQKM SENGAIP--- RQGKVLTVDT MNANVKKVDY AVRGPIVQRA VQIE
ALAT2_HUMAN | ALAT2_HUMAN      1   -----ASAVL KVRPERS--- RRERILTLES MNPQVKAVEY AVRGPIVLKA GEIE
ALAT2_MOUSE | ALAT2_MOUSE      1   ------SAAL KVRPERS--- PRDRILTLES MNPQVKAVEY AVRGPIVLKA GEIE
ALAT2_XENLA | ALAT2_XENLA      1   -------KVT RKMSENG--- TCNRILTLDS MNPCIQKVEY AVRGPIVIRA VELE
ALAT2_XENTR | ALAT2_XENTR      1   -----DGKVA RKMSENG--- TCNRILTLES MNPCIQKVEY AVRGPIVIRA VELE
ALAT_SCHPO | ALAT_SCHPO        1   ---------- ---------- -QNAFSDLNS LNQQVFKANY AVRGALAILA DEIQ
ALAT_YEAST | ALAT_YEAST        1   -----QQDLK GVFTAKDLDF KPAGKITKKD LNTGVTKAEY AVRGAIPTRA DELK
Giardia                        1   ---------- ---------- --YNVFDLTT ISQAVLQAEY AVRGTVPLRA IEIE
Giardia                        1   ---------- ---------- ---------- -YQPVLTPET CYKGLFDVKF SMVN
TVAG_074600 | TVAG_074600      1   ----SLAKNF ALQNPSY--- GSRRPININS INQQLIQSQY GVRGHLNTIA DQFM
TVAG_088220 | TVAG_088220      1   ------SRSS S-FMSTK--- GTPSPLEFST LPQTVLKAEY AVRGEVPMRA DALK
TVAG_098820 | TVAG_098820      1   ------AKNF ALPNPSY--- GSRRPININS INQQLIQSQY GVRGHLNTIA DQFM
TVAG_132440 | TVAG_132440      1   ------VRSS SNSYNYG--- KLNSALSMAT LNPQVIKAEY AVRGELAIRA DILR
TVAG_136210 | TVAG_136210      1   ---------- -MNQNYH--- RATPSLHYTN INPQVIKAEY AVRGEIAIRA DHYA
TVAG_379550 | TVAG_379550      1   MSLKTISRLF AYSPRYL--- QNARALNINT VNKEVVKSSY SIRGHLNTVS DQLR
Entamoeba                      1   ---------- ---------- ---KAFSRQS INPCIIATQY AVRGKLVLEA NEIQ
Entamoeba                      1   ---------- ---------- ---------- ---------- ---------- ----
Entamoeba                      1   ---------- ---------- ----SFASEN ISPDVVAFQF AVRGKIAIVS EEID
tetra_46 | tetra_46            1   ----YRSDMN NQGKQRT--- EAPKFITEED INKRVINAEY AVRGTVPTRA GKIK

Chlamydomonas                  80  TYFRQVLALC ECPQLLTK-- ---------- ---------- --QIPG--GL GA--
Chlamydomonas                  64  TFTRQVLALC AAPFLLDHPK V----EDMFP ADAIARAKKI LASFKG--GV GAYT
ALA2_HORVU | ALA2_HORVU        63  TFFREVLALC DHPDLLQREE I----KTLFS ADSISRAKQI LAMIPG-RAT GAYS
ALA2_PANMI | ALA2_PANMI        63  TFFREVLALC DHPCLLEKEE T----KSLFS ADAISRAKQI LSTIPG-RAT GAYS
ALAM_DICDI | ALAM_DICDI        84  TYFRQVVSLV ECPDLLDNPY V----EKIYP ADVISRAKEI LGSINN--TT GAYS
ALAM_YEAST | ALAM_YEAST        72  TYYRQVLSLL QYPELLNQNE QQLVDSKLFK LDAIKRAKSL MEDIGG--SV GAYS
ALAT1_BOVIN | ALAT1_BOVIN      74  TFPRQVLALC VHPDLLNSPD --------FP DDAKRRAERI LQACGG-HSL GAYS
ALAT1_HUMAN | ALAT1_HUMAN      75  TFLRQVLALC VNPDLLSSPN --------FP DDAKKRAERI LQACGG-HSL GAYS
ALAT1_MOUSE | ALAT1_MOUSE      79  TFFRQVLALC VYPNLLSSPD --------FP EDAKRRAERI LQACGG-HSL GAYS
ALAT1_RAT | ALAT1_RAT          77  TFFRQVLALC VYPNLLSSPD --------FP EDAKRRAERI LQACGG-HSL GAYS
ALAT2_DANRE | ALAT2_DANRE      78  TFFRQVMALC TYPQLLDDNK --------FP EDAKNRARRI LQSCGG-NSI GAYT
ALAT2_HUMAN | ALAT2_HUMAN      77  TFLRQVMALC TYPNLLDSPS --------FP EDAKKRARRI LQACGG-NSL GSYS
ALAT2_MOUSE | ALAT2_MOUSE      76  TFLRQVMALC TYPNLLNSPS --------FP EDAKKRARRI LQACGG-NSL GSYS
ALAT2_XENLA | ALAT2_XENLA      75  TFLRQVSAIC LYPELMNDNK --------FP EDVKQKAARI LQACGG-HSI GAYS
ALAT2_XENTR | ALAT2_XENTR      77  TFLRQVSAIC LYPELMNDNK --------FP EDVKQKAARI LQACGG-HSI GAYS
ALAT_SCHPO | ALAT_SCHPO        65  TFVRQVLSLC QYPTLLDHAE EKW-FQNLFP TDVVQRSKML LKES-G--SL GAYS
ALAT_YEAST | ALAT_YEAST        81  TFTRQVLAIL EYPEILRVGH NELASLNLFS RDALERAERL LNDIGG--SI GAYS
Giardia                        66  TFLRNLLALV TAPHVLSKPD TEICALLNCN QECVDRARAF VKENPS--GV GAYT
Giardia                        60  TYLRQMVAGF ACPDLIGK-Y -------VLP TDVELRVKHI LNSCSG-KSS GSYQ
TVAG_074600 | TVAG_074600      78  SFTRQVVACI EDRSLLDLPQ --------IP AEVKDRVNVI LNSMTL--EF GGYT
TVAG_088220 | TVAG_088220      75  KFPRQVLACV EDPDLLEVPS --------IP EEARERAREI LKNFPA--GM GSYT
TVAG_098820 | TVAG_098820      76  SFTRQVVACI EDRSLLNLPQ --------MP AEVKDRVNVI LNSMAC--EF GGYT
TVAG_132440 | TVAG_132440      76  TFPRQVISCI ENPDLLNIKE --------IP EEARHRAAQV FKHFPA--GL GAYT
TVAG_136210 | TVAG_136210      71  KFPRQVLSCV ENPDLLQSKD --------IP EEARARAAEV LKHFPA--GM GAYT
TVAG_379550 | TVAG_379550      82  TFPRDVVSCI ENTKLLNSAD --------IS EEAKDRAKQI IASTGG--RF GGYT
Entamoeba                      68  TYPRQIISIV EYPELLNHTT -------LFP KDVITHAKKI INSLGCTGTS GAYT
Entamoeba                      1   ---------- ----LYNELT ------KLFM VLASQQNSKV FV------KN GGIE
Entamoeba                      67  TFVREITSMV EYPPLTEHPE -------LFH ADAVARAKEI IKATGCNGTT GAYS
tetra_46 | tetra_46            79  TFNRQVLSTI LNPELVNSEV --------YS KDVRARAKYY LERMGS-TTI GAYS

Chlamydomonas                  105 ---------- ---------- ---------- -YSATLTLYG GTLAP----- -YLL
Chlamydomonas                  153 CLNAMIR--- HDRDS-VLVP IPQY-----P LYSASIRLYG GTLVG----- -YFL
ALA2_HORVU | ALA2_HORVU        153 MMQLLIR--- NEKDG-ILVP IPQY-----P LYSASIALHG GALVP----- -YYL
ALA2_PANMI | ALA2_PANMI        153 MMQLLIR--- NEKDG-ILCP IPQY-----P LYSASIALHG GTLVP----- -YYL
ALAM_DICDI | ALAM_DICDI        173 ILKLLIK--- DRSDG-ILIP IPQY-----P LYSATIELYN GSQLG----- -YLL
ALAM_YEAST | ALAM_YEAST        166 LLSIFCR--- GPETG-VLIP IPQY-----P LYTATLALNN SQALP----- -YYL
ALAT1_BOVIN | ALAT1_BOVIN      161 VLKLLVTGEG RTRTG-VLIP IPQY-----P LYSAALAEFN AVQVD----- -YYL
ALAT1_HUMAN | ALAT1_HUMAN      162 VLKLLVAGEG HTRTG-VLIP IPQY-----P LYSATLAELG AVQVD----- -YYL
ALAT1_MOUSE | ALAT1_MOUSE      166 MLKLLVAGEG RARTG-VLIP IPQY-----P LYSAALAELD AVQVD----- -YYL
ALAT1_RAT | ALAT1_RAT          164 MLKLLVSGEG RARTG-VLIP IPQY-----P LYSAALAELD AVQVD----- -YYL
ALAT2_DANRE | ALAT2_DANRE      165 ILKLLTAGEG LTRTG-VMIS IPQY-----P LYSASIAELG AVQIN----- -YYL
```



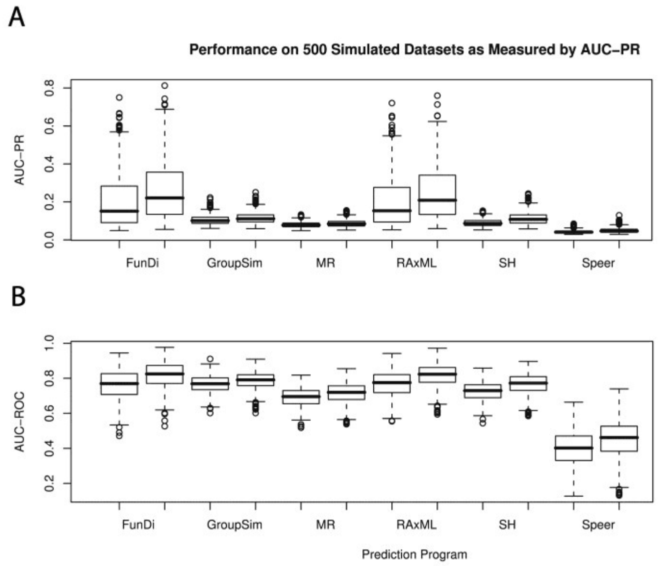
```
ALAT_SCHPO | ALAT_SCHPO      156 IMTLIIA--- RPTDG-VMVP APQY-----P LYGAQIDLMS GSMVS----- -YSL
ALAT_YEAST | ALAT_YEAST      175 LLSLLCK--- DSQTG-LLIP IPQY-----P LYTASASLFN AQVLP----- -YYL
Giardia                      159 ILQLLAG--P NPGDGAFLLP IPQY-----P LYSAAIALNN AVAVK----- -YYL
Giardia                      146 LLRPIIR--- NETDA-ILCP RPGF-----P LYASTIVYYG GKEVS----- -YDL
TVAG_074600 | TVAG_074600    163 LLTLLIN--- NDNVG-IMMP FPTY-----P IYASETILRH GKVVP----- -FYL
TVAG_088220 | TVAG_088220    161 VLKMLIS--- DPSVG-ILTP FPTY-----P LYTAEITLNN GRIVP----- -YYL
TVAG_098820 | TVAG_098820    161 LLTLLIN--- NDNVG-IMMP FPTY-----P IYASETILRH GKVVP----- -FYL
TVAG_132440 | TVAG_132440    162 ILNMIIA--- KPNVG-IMIP FPQY-----P LYTAEIALKN GRVVP----- -YYL
TVAG_136210 | TVAG_136210    157 ILTMIIS--- NPNCG-IMIP FPQY-----P LYTAEIALRN GRVVP----- -YYL
TVAG_379550 | TVAG_379550    167 LMTLLIQ--- HNNVG-IMTP FPTY-----P VYTSEAAVHG GKIVP----- -FYL
Entamoeba                    156 ILNMLIS--- HPLHG-IMIP IPQY-----P LYSASISQFG GFQIN----- -YFL
Entamoeba                    66  IKKIIETAQN TQTYG-ILLP IVQYGHEIER LYKIAICLSE KVEIEQEKIG MYLL
Entamoeba                    155 VMQLMLS--- HPLHG-IMIP NPQY-----P LYGACIQQLG GKTCH----- -YNL
tetra_46 | tetra_46          167 CLNVLIS--- DSRDG-IMIP IPQY-----P LYSASCTLVG ASEVH----- -YFL

Chlamydomonas                165 LSYQNMRDVL TFCRDEQLVL IADEVYQANI YVG-----N- KEFFSFKKVA CAMG
Chlamydomonas                234 LSKENLQELI KLAYQEKIVL MADEVYQENV YQD-----E- RPFVSAKKVM WEMG
ALA2_HORVU | ALA2_HORVU      234 LAEENQYDIV KFCKNEGLVL LADEVYQENI YVD-----N- KKFHSFKKIV RSLG
ALA2_PANMI | ALA2_PANMI      234 LAEDNQCDIV RFCKNEGLVL LADEVYQENI YVD-----D- KKFNSFKKIA RSVG
ALAM_DICDI | ALAM_DICDI      254 LDRANMEEIV KFCLEKNVVL LADEVYQENV YVK-----ES KPFISFKKVV KDMG
ALAM_YEAST | ALAM_YEAST      247 LSPESIAQIF EVAAKYGTVV IADEVYQENI FP------G- TKFHSMKKIL RHLQ
ALAT1_BOVIN | ALAT1_BOVIN    244 QTRECIEDVI RFAYEEKLFL LADEVYQDNV YAE-----S- SQFHSFKKVL TEMG
ALAT1_HUMAN | ALAT1_HUMAN    245 QTRECIEAVI RFAFEERLFL LADEVYQDNV YAA-----G- SQFHSFKKVL MEMG
ALAT1_MOUSE | ALAT1_MOUSE    249 QTRECIEAVI RFAFEEGLFL MADEVYQDNV YAE-----G- SQFHSFKKVL TEMG
ALAT1_RAT | ALAT1_RAT        247 QTRECIEAVI RFAFKEGLFL MADEVYQDNV YAE-----G- SQFHSFKKVL MEMG
ALAT2_DANRE | ALAT2_DANRE    248 QSRQCIEDVI QFAAKENLFL MADEVYQDNV YAK-----G- CEFHSFKKVL FEMG
ALAT2_HUMAN | ALAT2_HUMAN    247 QSRKCIEDVI HFAWEEKLFL LADEVYQDNV YSP-----D- CRFHSFKKVL YEMG
ALAT2_MOUSE | ALAT2_MOUSE    245 QSRKCIEDVI HFAWEEKLFL LADEVYQDNV YSP-----D- CRFHSFKKVL YQMG
ALAT2_XENLA | ALAT2_XENLA    245 QSRKCIEDVI RFAAEENLFL MADEVYQDNV YAK-----G- CAFHSFKKVL FEMG
ALAT2_XENTR | ALAT2_XENTR    247 QSRKCIEDVI RFAAEENLFL MADEVYQDNV YAK-----G- CTFHSFKKVL FEMG
ALAT_SCHPO | ALAT_SCHPO      237 ISENSMEKVL RFAKAKGIVL LADEVYQNNI YQ-------- NKFHSFRRKL GELR
ALAT_YEAST | ALAT_YEAST      256 LSEETIARIC LIAAKYGITI ISDEVYQENI FN------D- VKFHSMKKVL RKLQ
Giardia                      243 FSRETLRAAI DICDEYGISI MSDEVYQLNT YAEAPGKQR- PVFHSMKKVL CEWE
Giardia                      227 LTESDIKNAL RFAYKNDMMV MSDEVYQTNI YEP----EE- FPFLSARKLL YALN
TVAG_074600 | TVAG_074600    244 LSAQDMRTII EFCDQNKICI IADEVYQDCV YNP-----A- KPFISFKKMV SQVK
TVAG_088220 | TVAG_088220    242 MTAQQMRQVV DFCEQNNILI IADEVYQYNI YNP-----E- RPFISFKKII AEMK
TVAG_098820 | TVAG_098820    242 LSAQDMRTII EFCDQNKICI IADEVYQDCV YNS-----A- KPFISFKKMV SQVK
TVAG_132440 | TVAG_132440    243 LTAQQMRDVI EFCEQNNILL IADEVYQFNT YNP-----E- KSFISFKKVA SEMK
TVAG_136210 | TVAG_136210    238 LTAQEMRSVV EFCEQNNILI IADEVYQFNT YNP-----E- KPFISFKQIV TEMK
TVAG_379550 | TVAG_379550    248 LRPETMRTIV DFCEQNNILL IADEVYQDVV YNK-----E- RPFYSFKKIA SQMK
Entamoeba                    237 LTVQNMKEII EFCYEKKICL LADEVYQENI YG------E- IPFTSFRKVL KSMR
Entamoeba                    144 TKQSKNEHIV RIC----VCL MKD--YDINE IEL-----I- SQFI---DII KQWK
Entamoeba                    236 LPVDTIKEII RFCNEKKICL MADEVYQENI WT------D- VPFNSFRKIL ATME
tetra_46 | tetra_46          248 LSYDTIKQMI EFAYDHKMVI FADEVYQDNI YTP-----N- KEFVSFKKVR SELP

Chlamydomonas                251 DQILKLASIN LCPNLSGQIC CALMMNPPQ- --PGEASYEL YRKEKSDILG SLKR
Chlamydomonas                322 EEVYKCASIN LSPNTMGQIA LSVLVNPPK- --PGDPSYDQ YTKEKASELV SLRR
ALA2_HORVU | ALA2_HORVU      321 EQIYKIASVN LCSNITGQIL ASLVMNPPK- --ASDESYAS YKAEKDGILA SLAR
ALA2_PANMI | ALA2_PANMI      321 EQIYKIASVN LCSNITGQIL ASLVMNPPK- --VGDESYAA YKAEKDGILQ SLAR
ALAM_DICDI | ALAM_DICDI      342 AEIYKLASIG LCPNVIGQLV VDLMVRPPV- --AGEQSHDL YLKERDNIYE SLKK
ALAM_YEAST | ALAM_YEAST      337 QVILKLASIS LCPVVTGQAL VDLMVRPPV- --EGEESFES DQAERNSIHE KLIT
ALAT1_BOVIN | ALAT1_BOVIN    332 QQMQKLRSVR LCPPTPGQVL LDVAVSPPA- --PSDPSFPR FQAERRAVLA ELAA
ALAT1_HUMAN | ALAT1_HUMAN    333 QQMLKLMSVR LCPPVPGQAL LDLVVSPPA- --PTDPSFAQ FQAEKQAVLA ELAA
ALAT1_MOUSE | ALAT1_MOUSE    337 KQMAKLMSVR LCPPVPGQAL MGMVVSPPT- --PSEPSFKQ FQAERQEVLA ELAA
ALAT1_RAT | ALAT1_RAT        335 KQMGKLMSVR LCPPVPGQAL MDMVVSPPT- --PSEPSFKQ FQAERQEVLA ELAA
ALAT2_DANRE | ALAT2_DANRE    336 AQLTKLVSVR LCPPAPGQAL MDLVVNPPQ- --PGEPSHQT FMQERTAVLS ALAE
ALAT2_HUMAN | ALAT2_HUMAN    335 GQLVKLLSVR LCPPVSGQAA MDIVVNPPV- --AGEESFEQ FSREKESVLG NLAK
ALAT2_MOUSE | ALAT2_MOUSE    333 GQLVKLLSVR LCPPVSGQAA MDIVVNPPE- --PGEESFEQ FSREKEFVLG NLAK
ALAT2_XENLA | ALAT2_XENLA    333 QQLTKLVSVR LCPPVPGQVL LDVIVNPPK- --PGEPSYKQ FISEKQAVLN NLAE
ALAT2_XENTR | ALAT2_XENTR    335 QQLTKLVSVR LCPPVPGQAL LDVIVNPPK- --PGEPSYKQ FMAEKQAVLG NLAE
ALAT_SCHPO | ALAT_SCHPO      327 DQILKLATID ICPPVAGQLL VDMLVNPPK- --PGDPSYDL FIKEVDEIHE ALRL
ALAT_YEAST | ALAT_YEAST      346 DALFKLMSIS ICSVVTGQAV VDLMVKPPQ- --PGDESYEQ DHDERLKIFH EMRT
Giardia                      337 AQIYKCFSVC LCSNTIGQVV VSYMVNPPK- --SDDED--- -KKHLKTVFE SMER
Giardia                      317 EQIMDIISLG -STNTDGMMA MDVIVNPPR- --PGEPSYWK FKRECDALYT SLQR
TVAG_074600 | TVAG_074600    328 EQISKMSTYS LCPNAVGQVI LDTMAHPPE- ---SDECKSI WDQQKANYIE NLKV
TVAG_088220 | TVAG_088220    326 AQFYKLASIG LCPNTVGQII MDIMCAPPT- ---SSECSKV WEEQKNRELQ NLKE
TVAG_098820 | TVAG_098820    326 EQISKMSTYS LCPNTVGQVI LDTMVHPPE- ---SDECKSI WDQQKASYIE NLKV
TVAG_132440 | TVAG_132440    327 AQFYKMASIQ LCPNTVGQVI LDIMCHPPE- ---SPECRKQ WDHERDTELT NLKN
TVAG_136210 | TVAG_136210    322 AQFYKMASVQ LCSNTVGQII LDIMCRPPQ- ---SDECKKQ YIEERDGELN NLKV
TVAG_379550 | TVAG_379550    332 AQIFKMATFG LCPNAVGQVI VDCMVHPPE- ---APENKAI WESERNSYIT KLQQ
Entamoeba                    324 SQMYKIASTN LCSNVVGQEM VEIICNPPK- --EGDESYPK YMNEKMSILN SLKR
```



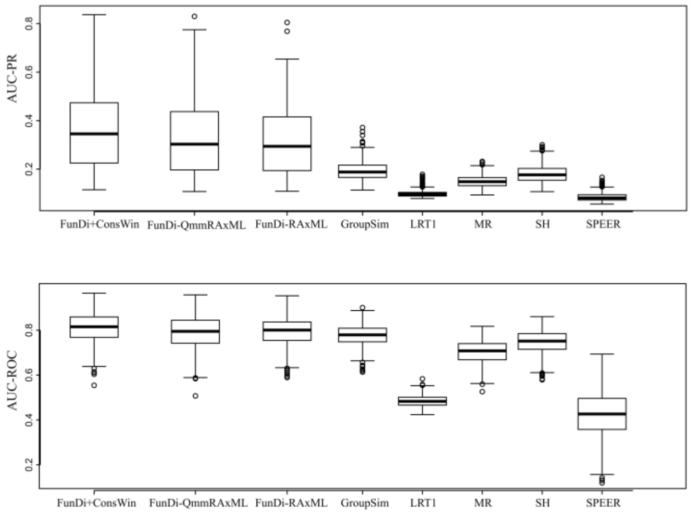
```
Chlamydomonas                 330 ---------- PDWLYCKELL EATGIVVVPG SGFGQADGTF HFRTTFLP-S EEDI
Chlamydomonas                 416 ---K-----A GDVFYCLKLL EATGISTVPG SGFGQEEGTF HLRTTILP-R EEVM
ALA2_HORVU | ALA2_HORVU       415 ---K-----A PDAFYALRLL ESTGIVVVPG SGFGQVPGTW HFRCTILP-Q EDKI
ALA2_PANMI | ALA2_PANMI       415 ---K-----A PDAFYALRLL ESTGIVVVPG SGFGQVPGTW HIRCTILP-Q EDKI
ALAM_DICDI | ALAM_DICDI       436 ---K-----A PDAYYCIQLL EATGICVVPG SGFGQKDGTW HFRTTFLP-S EEAI
ALAM_YEAST | ALAM_YEAST       431 ---L-----T PDEFYCKKLL ESTGICTVPG SGFGQEPGTY HLRTTFLA-P GLE-
ALAT1_BOVIN | ALAT1_BOVIN     426 ---L-----A PDMFFCLRLL EETGICVVPG SGFGQREGTY HFRMTILP-P MEKL
ALAT1_HUMAN | ALAT1_HUMAN     427 ---L-----A PDMFFCLRLL EETGICVVPG SGFGQREGTY HFRMTILP-P LEKL
ALAT1_MOUSE | ALAT1_MOUSE     431 ---L-----A PDMFFCLCLL EETGICVVPG SGFGQQEGTY HFRMTILP-P MEKL
ALAT1_RAT | ALAT1_RAT         429 ---L-----A PDMFFCLCLL EETGICVVPG SGFGQQEGTY HFRMTILP-P MEKL
ALAT2_DANRE | ALAT2_DANRE     430 ---Q-----A PDMFYCMKLL EETGICLVPG SGFGQREGTY HFRMTILP-P TDKL
ALAT2_HUMAN | ALAT2_HUMAN     429 ---M-----A PDMFYCMKLL EETGICVVPG SGFGQREGTY HFRMTILP-P VEKL
ALAT2_MOUSE | ALAT2_MOUSE     427 ---M-----A PDMFYCMKLL EETGICVVPG SGFGQREGTY HFRMTILP-P VDKL
ALAT2_XENLA | ALAT2_XENLA     427 ---Q-----A PDMFFCMKLL EETGICVVPG SGFGQREGTH HFRMTILP-P TDKL
ALAT2_XENTR | ALAT2_XENTR     429 ---Q-----A PDMFFCMKLL EETGICVVPG SGFGQREGTH HFRMTILP-P TDKL
ALAT_SCHPO | ALAT_SCHPO       421 ---I-----Q PDEFYAIELL KRSGICVVPG SGFGQPEGDY HIRITFLA-K GTE-
ALAT_YEAST | ALAT_YEAST       440 ---I-----E PDEFYCTSLL ESTGICTVPG SGFGQRPGTY HVRTTFLA-P GTK-
Giardia                       427 ---M-----A ADTLYCLKLL ETTGVCGVPG NGFQQRENTF HMRITILE-D EKFF
Giardia                       410 ---K-----E PDQLYCQDLL DSVGVFTLDG GMFGQKPGTY HLRMTILP-S EEVM
TVAG_074600 | TVAG_074600     424 RKVQ-----P PDFFWCLKLL EETGVQMNPG SGFGQVPGTS HFRSTFLA-E GKMF
TVAG_088220 | TVAG_088220     419 ---IKGKTPA PDMFWCLQLL EQTGIVVVPG SGFGQVPGTN HFRTTFLP-E AEKM
TVAG_098820 | TVAG_098820     422 RKVQ-----P PDFFWCLKLL EETGVQMNPG SGFGQVPGTS HFRSTFLA-E GKMF
TVAG_132440 | TVAG_132440     423 GKHV-----S PDMFWSLQLL EQTGIAVVPG SGFGQVPGTH HFRTTFLP-E AEQM
TVAG_136210 | TVAG_136210     418 GKHI-----A PDMFWSLQLL DETGIVVVPG SGFGQVPGTH HFRTTFLP-E AEEM
TVAG_379550 | TVAG_379550     428 REVM-----P PDFFWCLKLL EETGVIVVPG SGFGQVPGTF HFRSTFLP-E DEKF
Entamoeba                     418 ---Q-----K PDELYCLRML KSIGVCVVPG SGFGQKDNTY HFRIAILP-P ENEI
Entamoeba                     221 ---------- ---------- ---------- ---------- ---------- ----
Entamoeba                     417 ---E-----D PNETYCIEML KKTGIAVVKG SGFGQKKGTY HFRIALLP-P ENEI
tetra_46 | tetra_46           432 ---M-----E PDLFYCLNVL ENTGIVLVPG SGFRQEENTY HFRITTLILG EDRL
```

SUPPLEMENTARY FIGURE 2.1 ALIGNMENT OF MEMBERS OF PUTATIVE AMINOTRANSFERASE CLASS I AND II FAMILY CLUSTER. Tetra_46 is the Tetrahymena thermophila query sequence sequences were aligned and automatically trimmed as described in 2.2.

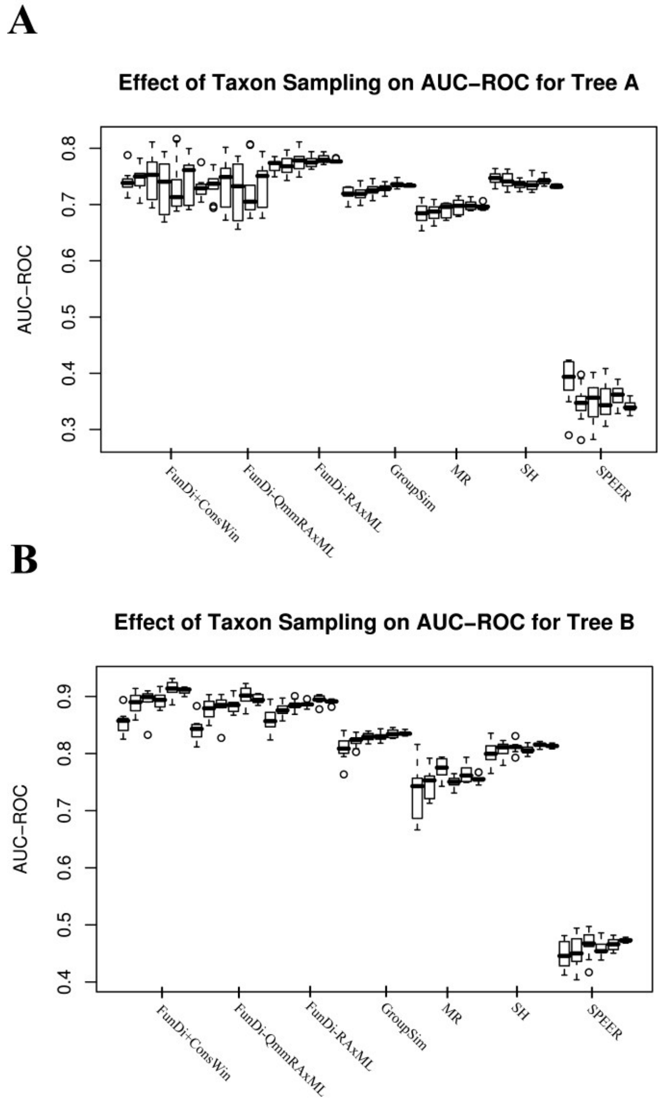
B3 Predicting Functionally Divergent Protein Residues



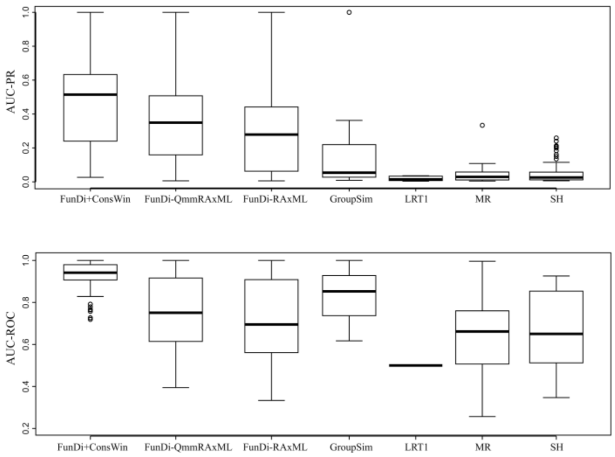
SUPPLEMENTARY FIGURE 3.1 PERFORMANCE DIFFERENCES BETWEEN TYPE I (LEFT) AND TYPE II (RIGHT) SITES FOR EACH OF THE TESTED PREDICTION PROGRAMS ACROSS 500 RANDOMLY SIMULATED DATASETS AS MEASURED BY AUC-PR (A) AND AUC-ROC (B).



SUPPLEMENTARY FIGURE 3.2 PERFORMANCE AS MEASURED BY AUC-PR AND AUC-ROC ON 500 SIMULATED DATASETS AS IN FIGURE 1 OF MANUSCRIPT WITH THE ADDITION OF THE LIKELIHOOD RATIO TEST OF KNUDSEN ET AL (2002,2003)



SUPPLEMENTARY FIGURE 3.3 PERFORMANCE UNDER TAXON SAMPLING CONSTRAINTS AS MEASURED BY ROC CURVE ON TREES A AND B.



SUPPLEMENTARY FIGURE 3.4 PERFORMANCE AS MEASURED BY AUC-PR AND AUC-ROC ON 70 SIMULATED DATASETS AS IN FIGURE 6 OF MANUSCRIPT WITH THE ADDITION OF THE LIKELIHOOD RATIO TEST OF KNUDSEN ET AL (2002,2003)