Université de Montréal

Molecular protein function prediction using sequence similarity-based and similarity-free approaches

par

Sivakumar Kannan

Département de Biochimie

Faculté de Médecine

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Doctorat

en Bio-informatique

September, 2007

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Molecular protein function prediction using sequence similarity-based and similarity-free approaches

présentée par :

Sivakumar Kannan

a été évaluée par un jury composé des personnes suivantes :

Normand Brisson, président-rapporteur

Gertraud Burger, directeur de recherche

Mathieu Blanchette, membre du jury

Marcel Turcotte, examinateur externe

François Major, représentant du doyen de la FES

Résumé

Le séquençage de génomes en tant que tel ne révèle aucune information : il génère des données brutes. C'est plutôt l'analyse des séquences *in silico* qui découvre la fonction des gènes. Or, en général, la fonction de seulement 50% des gènes codant de protéines peut être déchiffré avec les méthodes courantes d'annotation à grande échelle, - toutes étant basées sur la similarité des séquences. Cet état de fait accentue le besoin urgent de nouvelles méthodes d'annotation qui exploitent, au lieu d'un seul attribut (la similarité des séquences), de multiples caractéristiques des séquences biologiques. Ainsi, l'objectif principal de ma thèse est d'améliorer l'annotation fonctionnelle en exploitant d'abord les méthodes bio-informatiques de pointe existantes et ensuite, en développant une nouvelle méthode prédictive. Le but de cette nouvelle méthode est de détecter les signatures et les patrons cachés dans les données biologiques intégrées et d'utiliser ces nouvelles connaissances pour déchiffrer à grande échelle les séquences génomiques.

La recherche de patrons dans des grands volumes de données intégrées ('data mining') permet de déduire des modèles prédictifs robustes. Ceci se fait en trois étapes. Dans un premier temps, un ensemble de séquences protéiques est décrit en utilisant différents attributs calculés directement à partir de la séquence (par exemple, propriétés physico-chimiques, structure secondaire prédite, etc.). Dans un second temps, un algorithme de recherche de données décèle des patrons dans les données décrites et apprend des règles à partir des patrons observés. Comme dernière étape, les règles apprises seront

vérifiées avec des données connues et les règles donnant la meilleure performance seront utilisées pour prédire la fonction des protéines inconnues.

Un objectif important de ce projet est l'étude des différentes représentations des séquences permettant une recherche efficace de données. Démonstration de faisabilité, un ensemble de données mitochondriales de bonne qualité a été traité exhaustivement. Ensuite, nous avons développé une nouvelle façon de valider des prédictions *in silico* à large échelle. Ceci permet la formulation d'hypothèses de travail prometteuses pour fins de vérifications expérimentales. Finalement, la fonction prédite d'une protéine particulière a été scrutée en profondeur en utilisant des méthodes bio-informatiques de pointe ainsi que diverses connaissances enzymatiques, physiologiques et génétiques décrites dans la littérature, qui attestent de la bonne performance de l'approche d'annotation fonctionnelle que nous avons développé.

Mots-clés : Annotation de génomes; prédiction de la fonction protéique; exploration de donneés; apprentissage machine; classification; arbre de décision; validation par connaissance expert.

Abstract

The blueprint of a living organism is specified in its genome whose coding segments are called genes. Whole-genome sequencing projects have produced an enormous amount of genomic data, and function prediction—the process of assigning function to these genes or their inferred protein sequences *in silico*—is crucial for making sense of this data. Large-scale function prediction using sequence-similarity-based methods such as BLAST can only assign function to ~50% of a typical genome while profile based methods such as PSI-BLAST and HMMER are more sensitive but still leave a substantial portion of the sequences with unassigned function. Therefore, we sought to develop a large-scale method for annotating the 'left-over' proteins that cannot be assigned function by current methods.

Our sequence-similarity-free method involves data mining i.e., searching for strong patterns and relationships in the data and inducing generalized rules. Protein sequences were represented by their physicochemical properties, and amino acid composition. A decision tree machine learning algorithm was used for inducing rules and predicting the function of the function-*unknown* sequences. As a proof of concept, we applied this method for predicting the function of the mitochondrion encoded function-unknown proteins across eukaryotes.

The prediction accuracy of our method, when tested on the function-*known* data, exceeds 80%. Using this method, we assigned function to more than 1,000 function-

unknown mitochondrial proteins. By our new validation procedure that assesses the predictions using domain-specific knowledge, about half of them received positive support, making these proteins candidates for targeted experimental validation. For one of the predictions that received positive support, we employed sensitive *in silico* methods together with the most recent domain-knowledge from the literature corroborating our prediction beyond doubt.

Keywords : Genome annotation; protein function prediction; data mining; machine learning; classification; decision trees; domain-specific validation.

Table of contents

List of Tables

Chapter 2

Chapter 3

List of Supplementary Tables

Chapter 2 :

Chapter 3 :

List of Figures

INTRODUCTION

Chapter 2 :

Chapter 3 :

List of Supplementary Figures

Chapter 2 :

To my beloved parents :

My mother, Chandra

My father, Kannan

Sine quibus non of my life

Acknowledgements

My research and thesis would not have been possible without many people who have supported me in different ways during this adventurous journey.

First of all, I would like to thank my research director and mentor, Dr. Gertraud Burger without whom I would never have finished this work. She has been very supportive and provided valuable supervision of my project, cheering me up whenever I lost my spirits and gently cracking the whip when I slowed down. Apart from teaching me academic skills, Dr. Burger has also been an example for me of a fine human being.

I would like to thank my pre-doctoral committee members, Dr. Pascal Chartrand, Dr. Michel Bouvier and Dr. Bálazs Kégl for their advice in the earlier stages of my project.

My special thanks to Dr. Franz Lang for his valuable feedback and discussions during the lab meetings, which he made a lot more fun with his intellectual humour.

A big thanks to my collaborator and friend Dr. Amy Hauth who taught me some of the most valuable work practices.

Of all my lab mates, the always cheerful Shen deserves special thanks for critically reading my manuscripts and for being a good friend.

I cannot thank enough all the wonderful people in our group, Liisa, David, Veronique, Claudia, William, Naiara, Prem, Yun, Jung Hwa, Lise, Shona, Henner, Yu, Pasha, Rachid, Pierre and Dorothee. They were among the many others who made the whole experience very enjoyable.

My special thanks to all the GOBASE group members, Veronique, Emmet, Eric, and Yue, the summer student Mathieu and our system administrator, Allan.

I would like to express my sincere gratitude to the secretaries, Elaine Meunier, Marie Robichaud, Denise Lessard and Sylvie Beauchemin for helping me breeze through the administrative tasks.

I would like to thank the Faculté des études supérieures for granting me permission to write my thesis in English and for providing me with the tuition fee exoneration.

The financial support from Canadian Institutes of Health Research (CIHR) Strategic Training Program Grant in Bioinformatics is gratefully acknowledged.

My loving thanks to my best friends, Ramkiran, Mallika and Shantha.

I am very grateful to my Canadian mother, Tove, who took me under her wing helping me integrate into Western culture, and for her unconditional love and support.

The wonderful woman Karen, my number1 fan, deserves special mention of all here for her love, understanding and caring and for everything.

I owe my gratitude to my family back home in India for their constant support that has kept me going. My loving thanks to my grandmother, my brother Ramesh, my sister-in-law Saraswathi and my two little nieces, Jayasri and Manjusri and my younger brother Ashok who will always remain the smartest in the family no matter what letters come after my name.

Above all, I am very grateful to my parents for their love and their absolute confidence in me. To them, in gratitude, I dedicate this thesis.

Overview of the thesis

The central theme of this thesis is how to predict protein function without recurring to sequence similarity. The thesis is organized as follows.

In the **Introduction**, we define the problem of function prediction, critically review the existing state-of-the-art methods and their limitations, and state the objectives of this work as well as our approach.

Chapter 1 is a published journal article on predicting the function of mitochondrial proteins from the protist *Seculamonas ecaudoriensis* using the state-of-the-art methods.

Chapter 2 is a submitted manuscript, in which we describe the methodology of our newly developed large-scale function prediction system, and a validation procedure, which assigns different levels of support.

Chapter 3 is a submitted manuscript that reports the verification of one of our function predictions with positive support, using various sensitive *in silico* methods and diverse biological evidence.

Finally, we close with a general **Conclusion** and future research directions.

INTRODUCTION

Life originated on Earth at least 4 billion years ago (Zimmer, 2005). Since then evolution by natural selection has produced millions of different living organisms with enormous diversity as to morphology, habitat, life-style, metabolism, and modes of reproduction. However, all these species have evolved from a universal common ancestor, and share many characteristics of their genetic information and housekeeping systems.

The information required for the development, functioning and reproduction of a species is encoded in its genome. Hence, scientists have started sequencing whole genomes to better understand the biology of species. The first complete genome to be sequenced was that of *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and since then, the advancements of high-throughput sequencing technologies lead to an explosion of published genomic data. As of now, complete bacterial and nuclear genome sequences are available for ~700 species and >2,000 are underway (Liolios *et al.*, 2006). In addition, there are >1,300 complete organellar (from mitochondria and chloroplasts), and >1,400 viral genomes. Further, expressed sequence tags (EST) sequencing projects are contributing enormous amount of data on their part. With this astronomical quantity of genomic data, the way biology research is done has been transformed fundamentally.

Gene function prediction

In the genomics era, one of the greatest challenges facing biologists is to make sense of this sea of data. The primary step is finding the function of genes before we can

answer questions about development, metabolism, and evolution of an organism. A gene is defined as "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions" (Pearson, 2006). In this study, we are focussed only on the protein coding genes. Function annotation—the process of assigning function to genes or their products—is a central problem in life sciences.

Although function annotation is an active research area, a large proportion of published sequences is still unassigned. For example, the best studied model organisms *Escherichia coli* and *Caenorhabditis elegans* have, respectively, 50% and 88% of genes with no or ambiguous and uninformative annotation (Hawkins and Kihara, 2007). With millions of sequences already available and the tremendous rate at which new data are being added, it is obvious that there is a great need for large-scale function prediction (Figure 1).

Protein function

The function of a protein can be described at different levels or dimensions such as molecular, cellular, physiological, and organismal. Further, various features or aspects of the protein can describe protein function such as the sub-cellular localization, secondary structure, and post-translational modifications.

Traditionally, protein function has been annotated by describing all available information with free text such as in SWISS-PROT (Boeckmann *et al.*, 2003), making it

machine-unreadable. The resulting inconsistency and incoherence is a notorious problem in genome analysis (Dobson *et al.*, 2004).

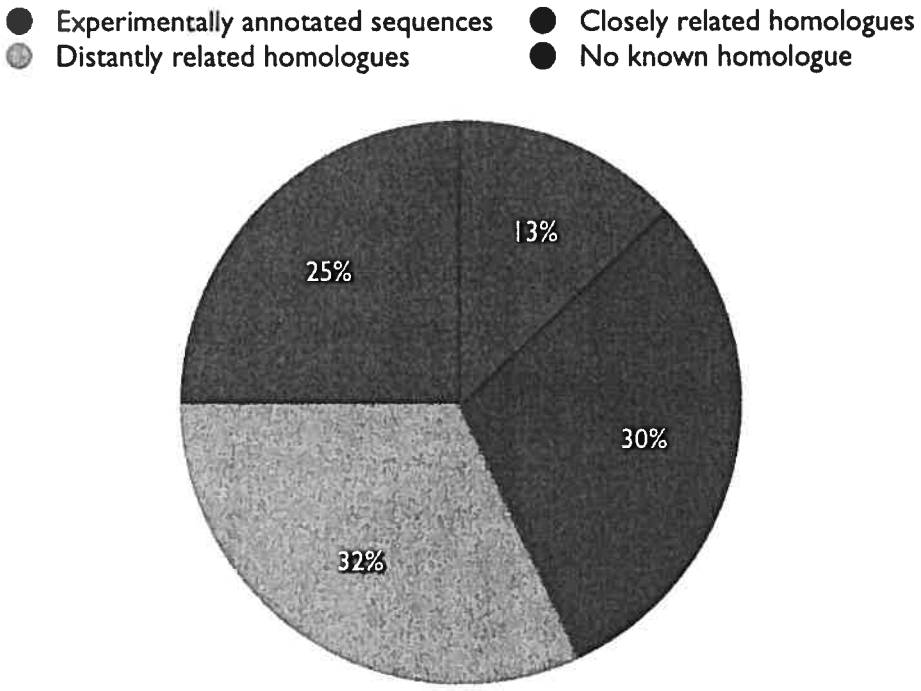

Figure 1. Annotation status of all published protein sequences (~2 million) (Modified from: Ofran, 2005).

To automate function prediction, many attempts have been made to standardize the description of protein function. Generally, protein function is classified in a hierarchical fashion starting with generic function and progressing toward more specific function. Species-specific hierarchical categorization of protein function using controlled vocabulary was established for *E. coli* (Riley, 1993) and *Saccharomyces cerevisiae* (Mewes *et al.*, 1997); the latter was then extended to an organism-independent classification system called Functional Catalogue (FunCat) (Ruepp *et al.*, 2004). Enzyme Commission (E.C.) is a four-level numerical hierarchy to classify enzymes based on the biochemical reactions they catalyze (Enzyme_Commission, 1999). The E.C. system is commonly used, but obviously it is unsuited for non-enzyme proteins.

Recently, biologists have elaborated ontologies to describe protein function systematically and consistently. An ontology is a formal description of concepts and relationships between concepts in a given domain employing a controlled vocabulary. The need for and applications of ontology-based function classification have been discussed in detail by others (e.g., Karp, 2000). The most widely used is Gene Ontology (Ashburner *et al.*, 2000a), which describes protein function in terms of 'cellular component' (e.g., nucleus), 'biological process' (e.g., signal transduction) and 'molecular function' (e.g., adenylate cyclase). The comparison of different functional annotation schemas is presented in Rison *et al.* (2000), and Soldatova and King (2005) critically assess the effectiveness of the current bio-ontologies.

Function prediction using experimental methods

Traditionally, protein function has been determined for individual proteins, one at a time, by biochemical and molecular biology experiments (Whisstock and Lesk, 2003). This approach is still considered the gold standard.

Recent functional genomics and proteomics methods infer structure and function of genes or their products by applying high-throughput experimental methods, usually combined with extensive statistical or computational analyses of the results (Hieter and Boguski, 1997). The simultaneous study of numerous genes or gene products provides a window on how they function together at the molecular, cellular or even at the systems level.

DNA microarrays (or Gene or Genome Chips) allow gene expression profiling, i.e., monitoring the expression levels of thousands of genes simultaneously (Schena *et al.*, 1995). Comparison of gene expression patterns in different cell types and tissues, or under normal, diseased and stress conditions, determines groups of genes having similar profiles, which are considered functionally related or as part of the same cellular process. For example, by using this technique, hundreds of previously unknown genes were identified to be associated with cancer (Walker *et al.*, 1999).

Similar to the large-scale study of genes, proteomics examines hundreds of proteins simultaneously. Frequently used proteomics methods are mass spectrometry (Andersen and Mann, 2000), two-dimensional gel electrophoresis (Encarnación *et al.*, 2005; O'Farrell,

1975), yeast two-hybrid screening for studying protein-protein interactions (Fields and Song, 1989; Giot *et al.*, 2003; Uetz *et al.*, 2000), and protein microarrays (Jones *et al.*, 2006). Marcotte (1999b) used data generated from high-throughput methods along with predicted information from computational methods and to assign function to 1,600 previously function-unknown ORFs in *S. cerevisiae*.

Finally, reverse genetics methods such as mutagenesis or gene disruption (knock-out) are also effectively used at a large scale for inferring function. When the expression of the gene of interest is abolished, the resulting phenotype can hint at the functional role of the gene. Such deletion studies were carried out at a large-scale for *S. cerevisiae* as part of the EUROFAN (European Functional Analysis Network) (Dujon, 1998; Schomburg *et al.*, 2004).

One of the problems is the inconsistency in function predictions made by different methods (Bader and Hogue, 2002; von Mering *et al.*, 2002). This is because gene expression and protein-protein interactions are affected by a multitude of factors that cannot be easily controlled and the displayed phenotype of a gene disruption depends on the conditions tested. The best way to make accurate high-throughput function inference is by combining evidences from different methods.

Function prediction using *in silico* methods

It is apparent that experimental characterization of protein function cannot scale up to the rate at which the genomic data is being produced. According to a recent estimate, out

of approximately two million inferred protein sequences available in public databases, only <15% has experimentally characterized function (Ofran *et al.*, 2005). This situation has set the stage for developing computational methods for function annotation. The basic underlying assumptions in using *in silico* methods for function assignment are that:

1. sequence determines structure and structure ultimately determines function; hence all the information necessary to infer the function of a protein is present in its sequence. Therefore, it should be possible to predict some aspects of protein function directly from the sequence.
2. homologous proteins (i.e., proteins derived from a common ancestor) have similar function and share many common features.

If, based on these assumptions, a function-*unknown* protein is similar (in sequence, secondary structure, or physicochemical properties, etc.) to a function-*known* protein, then it should also have similar if not the same function. Thus, the function of the known protein can be transferred to the function-*unknown* homolog. This is the basic concept behind *in silico* function prediction methods.

Biological databases

An important prerequisite for carrying out *in silico* analyses effectively is that, all the available information about the function-*known* sequences is stored in databases. The most important publicly accessible databases that organize information about DNA or protein sequences are GenBank of NCBI—mostly a data repository (Bilofsky and Burks, 1988), and SWISS-PROT—a curated protein sequence database (Boeckmann *et al.*, 2003).

Function prediction using sequence-similarity based methods

Sequence-sequence comparison

Protein function prediction by sequence-sequence comparison involves comparing the function-unknown sequence with all function-known sequences. If some kind of resemblance is found with statistical significance, the function of the function-known sequence can then be transferred to the function-unknown sequence.

Sequence-similarity based function transfer

Among the many features that are common between two homologous proteins, sequence similarity is the strongest one and hence the most exploited for function annotation. Using sequence similarity to assign function is referred to as sequence-similarity based function transfer. (Note: Though the term 'homology-based' function transfer is also used in the literature, it is incorrect because programs like BLAST detects only sequence similarity. Similarity does not imply homology; homology detection, sensu strictu, requires phylogenetic analysis).

The most commonly used tools to compare two protein sequences are BLAST (Altschul *et al.*, 1990a) and FASTA (Pearson, 1990), which compare the function-unknown sequences against sequence databases and report statistically significant hits.

The advantages of sequence-similarity-based function transfer are that it is fast allowing large-scale automated function annotation, and that it assigns the molecular function rather than the broad functional class or any other aspect of protein function.

Limitations of sequence similarity-based function annotation

While similarity-based function annotation methods are easy to employ, require relatively little resources and scale very well to the sea of genomic data, there are also many limitations:

1. function can be assigned only if a similar annotated sequence is found in the database, which is not always the case. For example in a recent study, only 35% of all proteins from 105 entire proteomes could be annotated with a <5% error rate (Carter et al., 2003).

2. it is not obvious to quantify the level of similarity required between the known and the unknown to transfer function confidently. It has been suggested that at least 40% global sequence similarity is required (Devos and Valencia, 2000; Todd et al., 2001; Wilson et al., 2000). However, in a study by Rost (2002) one third of the enzymes having more than 70% sequence similarity did not even share the first E.C. number, meaning these enzymes have entirely different functions.

3. similarity-based function annotation can be challenging for proteins with multiple domains, which is quite common in eukaryotes. If only one domain is matching to a

function-known protein, then function transfer will be incorrect (Galperin and Koonin, 1998; Smith and Zhang, 1997).

4. there are many cases where two proteins have statistically significant sequence similarity but have completely different functions (Smith and Zhang, 1997). In fact, divergent evolution of paralogs is one of Nature's main strategies to 'invent' new function (Whisstock and Lesk, 2003). For example, sequence similarity is very high among the various nucleotidyl cyclases as well as among the different protein kinases, but only few residues determine their functional specificity (Hannenhalli and Russell, 2000).

5. function is not always determined by the gene itself, but may depend on the context such as tissue, sub-cellular localization, developmental stages and life-style. For instance in birds, crystallins in the eye lens and lactate dehydrogenase and enolase enzymes in other tissues are identical in sequence but exert completely unrelated functions (Wistow and Piatigorsky, 1987). Another example is the heat shock protein DegP that functions as a chaperone at low temperatures and as a proteinase at high temperatures (Spiess et al., 1999). In these cases, *in silico* annotation is bound to fail.

Annotation errors using similarity-based function transfer are abound in the literature (Ichikawa *et al.*, 1997; Keller *et al.*, 2002) and hence this approach warrants critical assessment of the results.

Intermediate Sequence Search (ISS)

If two homologous sequences are too derived to be identified by simple pairwise alignment, these two sequences can be related by using an intermediate or a third sequence that is homologous to both of the initial two sequences (Park *et al.*, 1997). This method is called Intermediate Sequence Search (ISS) (John and Sali, 2004; Salamov *et al.*, 1999). Using a single intermediate to relate two remote homologs performs better than simple pairwise alignment methods such as BLAST in identifying the remote homologs (Park *et al.*, 1998); and using multiple intermediates is even more effective than using a single intermediate (Salamov *et al.*, 1999).

The inherent risk of ISS methods is that they can relate two non-homologous proteins, especially if these proteins carry multiple domains.

Sequence-pattern based function transfer

Two proteins with the same function generally have a common sequence pattern (motif) or a group of motifs (fingerprint) since the functional sites consist of several residues. Many protein annotation methods make use of such motifs, usually represented as regular expressions (Attwood, 2000). These motifs are compiled and stored in public databases (for example, PROSITE) against which the function-unknown sequences can be searched (Attwood, 2002; Henikoff *et al.*, 2000; Hulo *et al.*, 2006).

Yet, inferring function using sequence-pattern based methods requires caution since there are many conserved sequence patterns without any functional significance such as the signals for post-translational modifications (Whisstock and Lesk, 2003).

Profile-sequence comparison

Using motifs for identifying distant protein homologues has been employed with limited success. A better representation of shared characteristics between related sequences is achieved by profiles, i.e. position-specific scoring matrices generated from a multiple alignment of sequences with same function. Function-unknown sequences can then be searched against these profiles. Profile - sequence comparison methods are more sensitive in identifying distant homologues compared to simple sequence-sequence comparison methods (Aravind and Koonin, 1999). PSI-BLAST is a commonly used tool for profile-sequence comparison (Altschul *et al.*, 1997). We employed profile-sequence comparison extensively in one study that aimed at identifying the composition of enzyme complexes in primitive eukaryotes (see Chapter 1). Profile Hidden Markov Models (Profile HMMs) are even more effective (Eddy, 1998) because they contain, in addition to the amino acid frequency for each position, position-specific probabilistic scores for insertions and deletions along the alignment. Among the most widely employed profile HMM-sequence comparison tools, SAM (Hughey and Krogh, 1996) outperforms in most cases HMMER (Eddy, 1998) and PSI-BLAST, but SAM is significantly slower (Madera and Gough, 2002). Profile HMMs, as well as profile-profile comparison discussed below, served us for a detailed scrutiny of the *in silico* predicted function of the MURF1 protein (see Chapter 3).

Profile-profile comparison

Instead of searching a single sequence against a database of profiles, recently developed methods build a profile from a set of related unknown sequences and this profile is then searched against the profiles built from function-known proteins. Profile-profile comparison methods are up to 30% more sensitive than profile-sequence comparison methods in identifying distantly related homologues (Panchenko, 2003). Readily available tools include prof_sim (Yona and Levitt, 2002), COMPASS (Sadreyev and Grishin, 2003) and HHSearch (Soding, 2005); the latter was shown to outperform earlier developed methods (sequence-sequence, profile-sequence and profile-profile) (Soding, 2005). While these profile-profile comparison tools are promising, they require a profile generated from a multiple alignment of related unknown sequences. Yet, many unknown sequences are 'orphans' and hence this method cannot be applied to them.

Function prediction using sequence-similarity-free methods

As mentioned above, similarity-based function annotation methods leave us with about 50% unassigned proteins for a newly sequenced genome even for well-studied organisms.

Obviously, this low success rate generated much interest in developing methods that capture other similarities than sequence similarity between the unknown and the known protein sequences. Such methods are collectively called similarity-free or *ab initio* methods and they are by no means a substitute for similarity-based methods but rather

complementary to the latter. Since similarity-free methods typically use multiple features, they are less prone to incorrect assignments for multi-domain proteins.

Comparative genomics based methods

Gene function can be inferred by comparing two or more genomes of related species or different strains of the same species, and this approach is called comparative genomics (Wei *et al.*, 2002). Comparative studies can be conducted at various levels such as genome structure, gene order, and co-evolution of the genes, coding regions, and non-coding regions. The extensive experimental data available for model organisms shed light on related organisms by comparing their genomes.

Phylogenetic profiles

Proteins involved in the same metabolic pathway or a structural complex are likely to evolve in a concerted fashion. For example, the eight proteins involved in glycolysis, have a similar pattern of presence/absence across genomes (phylogenetic profile). So, if two given proteins share similar phylogenetic profiles, they are most likely functionally linked. Using this method, phylogenetic profiles of ~4,000 proteins in *E.coli* were compared with those from 16 other genomes and an estimated half of these proteins can be assigned a general function (Pellegrini *et al.*, 1999).

Domain fusions

Two different but interacting proteins in one species may be fused into a single protein chain in other species. Based on this observation, several thousands of protein-protein interactions were predicted in *E. coli* and *S. cerevisiae* (Marcotte *et al.*, 1999a).

Gene neighbours & gene order

If gene order and gene neighbours are conserved across genomes, it is likely that these genes are functionally linked. This applies to prokaryotes most of whose genes are organized in operons. Function annotation methods that exploit this situation search for pairs of topologically close bi-directional best hits in different genomes (Overbeek *et al.*, 1999).

Limitations of comparative genomics methods

There are several limitations to protein function assignment by comparative genomics:

1. these methods are based on the occurrence of particular observations (e.g., domain fusion or prokaryotic operons), which are not common to all organismal groups.

2. the chance for false positives is high because the observation (e.g., a given gene order across different genomes) can be due to mere chance without any functional significance.

3. most of these methods cannot predict the specific molecular function but only a broad function such as being involved in a certain pathway or interacting with some other protein.

4. although sequence similarity is not required for assigning function to the unknown, it is needed for identifying the presence of homologs in other genomes; hence, these methods are not completely similarity-free.

Similarity-free function prediction from sequences by feature extraction

As explained above, sequence-similarity based and comparative genomics methods fail to predict the function for a certain portion of proteins. Therefore, methods are needed that recognize homologues by other features such as secondary structure, and physicochemical properties, i.e., features that can be extracted directly from the sequence.

Prediction of protein function using extracted features has been modelled as a data mining/machine-learning problem. Data mining is the search for implicit global patterns and subtle relationships in an integrated data and the construction of predictive models based on them. Machine learning algorithms such as decision trees can be used for inducing the rules that map the data to classes and for predicting the class of the unknown data based on these rules. The advantage of decision tree algorithms such as C4.5 (Quinlan, 1993a) is that the rules are human-readable, providing insights into the underlying biology. For protein function prediction, protein sequence data are represented as extracted features that are mapped to a functional class (Figure 2).

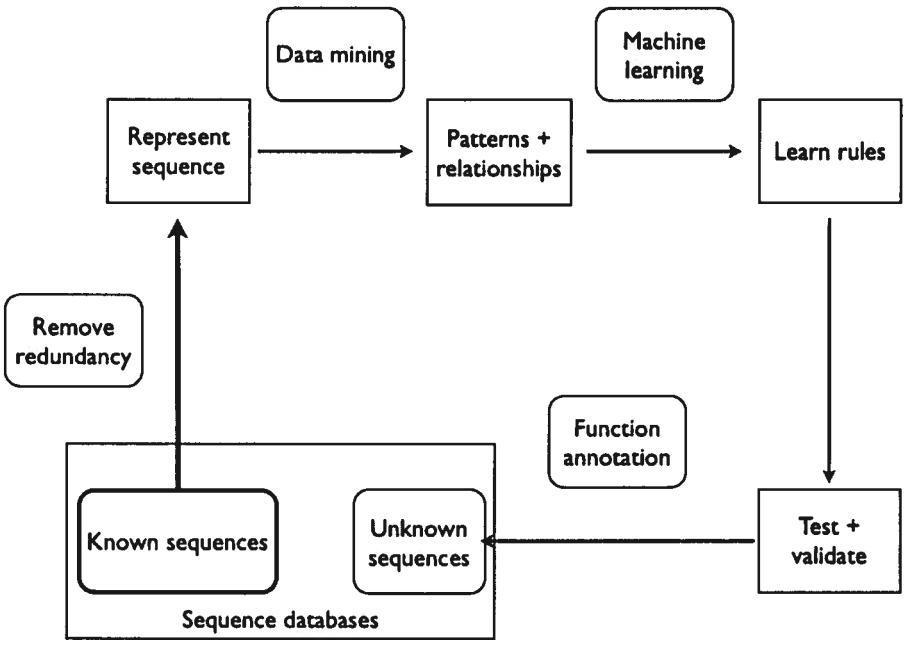

Figure 2. General methodology for function annotation using machine learning algorithms.

One of the earliest applications of feature extraction for predicting protein function was a study to discriminate whether a given protein is an enzyme or not (des Jardins *et al.*, 1997). Since then, numerous machine-learning based methods have been developed that use extracted features for predicting various aspects of protein function, in particular subcellular localization.

The first machine learning application to predict molecular protein function using extracted sequence features was carried out for *S. cerevisiae*, *E. coli* and *A. thaliana* (Clare *et al.*, 2006a; Clare and King, 2003; King *et al.*, 2001a). A latter retrospective study showed that some of the predictions made earlier were validated by experimental studies, thus testifying to the power of this approach (King *et al.*, 2004a). Still, this approach predicts only broad functional categories rather than fine-grained function. Second, the predictor is not entirely sequence similarity free as it depends on similarity for calculating some of the features such as phylogenetic attributes. Finally, it is species-specific, i.e., the method can predict the function of ORFs only from the same species used for training. Though the species-specificity is due to the lack of protein function schemas that encompass different organisms, it was not clear whether feature-extraction based machine learning algorithms could generalize function across taxa since their orthologues evolve at a different pace. However, a recent work has shown that the algorithms do learn and perform well across the species (Al-Shahib *et al.*, 2007).

How to solve the problem?

The above cited studies show the great potential of machine learning methods using extracted sequence features for function annotation, and this approach has become an active research area.

Our contribution to the field was to overcome the limitations described above, by developing a similarity-free method to predict the 'fine-grained' molecular function—the most specific function in the Gene Ontology hierarchy—across taxon boundaries (see Chapter 2).

Evaluating function predictions

For *in silico* function prediction, and this applies particularly to similarity-free methods, one of the major challenges is to demonstrate that the predictions are correct.

Experimental validation

Obviously, the ultimate way of validating a predicted function is by biochemical or molecular biology experiments. One example is the study by King and co-workers mentioned above (2004a), and there are many other instances where single predictions were confirmed by experimental studies. One case was exceptional where experiment-based function assignments were shown to be wrong and alternative functions were proposed with strong support using *in silico* methods (Iyer *et al.*, 2001).

Validating predicted function experimentally is not feasible at a large scale, thus calling for *in silico* approaches. Computational validation is of great value for building promising working hypothesis, directing the design of targeted experiments for ultimate validation (see Chapter 3). In addition, *in silico* validation will facilitate to compare the performances of the various function prediction tools and thus advance the field of automated function prediction in general (Godzik *et al.*, 2007).

Validating machine learning based predictions

Currently, the performance of machine learning based function prediction programs is evaluated by their performance on the training (known) data. But there is no guarantee that programs that performed well on the training data will do so on the test data or even unknown data. This is especially true if the training data and the test data are dissimilar. The good prediction performance of the classifier on the training data can be due to overfitting or class imbalance in the training data. Hence, validation methods that are independent from the training data are needed.

Since 1994, the protein structure prediction community in collaboration with structural biologists conducts a bi-annual competition called CASP (Critical Assessment of techniques for protein Structure Prediction). The objective is to assess the performance of prediction programs by comparing the predicted structure of proteins with their unpublished experimentally solved structures. This has resulted in a great progress in computational structure prediction and serves as a gold standard for evaluating new approaches.

A similar initiative is emerging in molecular function prediction with an annual competition called AFP (Automated Function Prediction) (Rodrigues *et al.*, 2007). However gold standards for function prediction programs need yet to be established.

Predictions have also been evaluated by comparing the results from different methods for the same unknown protein (Brenner, 1999) and then relying on the consensus (which still may be wrong). Our approach to the problem was to develop an evaluation procedure that takes advantage of the domain knowledge about the data (see Chapter 2). For a proof of principle, we chose molecular function prediction of hypothetical mitochondrial proteins to assign external support values to each prediction.

As an extension to this work, we demonstrated the merits of this validation procedure in one instance, MURF1 of kinetoplastids (see Chapter 3).

Objectives

The specific objectives of the thesis are

- to explore the strengths and limitations of protein function prediction methods that are based on sequence similarity
- to develop a sequence similarity-free method for fine-grained prediction of molecular protein function
- to design an *in silico* procedure that evaluates predicted molecular functions as a means to generate working hypotheses for experimentation
- to verify biologically relevant predictions by using a combination of state-of-the-art *in silico* methods together with in-depth literature searches.

Chapter 1 : ARTICLE

Subunits of complex III were identified by 2D-gel electrophoresis and ESI-MS/MS protein sequencing. Peptide identification was not straight forward, because the nuclear genome sequences of *Seculomonas ecuadoriensis* is not known. Therefore, we used sensitive sequence similarity methods to find homologues in other species. My contribution to this work was to conduct all the function prediction analyses and evaluation using state-of-the-art sequence similarity-based methods and writing up the methodology and results section for these analyses.

Structure of the bc_1 Complex from *Seculamonas ecuadoriensis*, a Jakobid Flagellate with an Ancestral Mitochondrial Genome

Stefanie Marx,‡*[1] *Maja Baumgärtner,†* *Sivakumar Kunnan,‡* *Hans-Peter Braun,** *B. Franz Lang,‡* *and Gertraud Burger‡*

*Institut für Angewandte Genetik, Universität Hannover, Hannover, Germany; †Gesellschaft für Biotechnologische Forschung, Braunschweig, Germany; and ‡Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, Montréal, Québec, Canada

In eubacteria, the respiratory bc_1 complex (complex III) consists of three or four different subunits, whereas that of mitochondria, which have descended from an α-proteobacterial endosymbiont, contains about seven additional subunits. To understand better how mitochondrial protein complexes evolved from their simpler bacterial predecessors, we purified complex III of *Seculamonas ecuadoriensis*, a member of the jakobid protists, which possess the most bacteria-like mitochondrial genomes known. The *S. ecuadoriensis* complex III has an apparent molecular mass of 460 kDa and exhibits antimycin-sensitive quinol:cytochrome *c* oxidoreductase activity. It is composed of at least eight subunits between 6 and 46 kDa in size, including two large "core" subunits and the three "respiratory" subunits. The molecular mass of the *S. ecuadoriensis* bc_1 complex is slightly lower than that reported for other eukaryotes, but about 2× as large as complex III in bacteria. This indicates that the departure from the small bacteria-like complex III took place at an early stage in mitochondrial evolution, prior to the divergence of jakobids. We posit that the recruitment of additional subunits in mitochondrial respiratory complexes is a consequence of the migration of originally α-proteobacterial genes to the nucleus.

Introduction

The bc_1 complex (also termed complex III or ubiquinol:cytochrome *c* reductase) and the structurally and functionally similar b_6f complex, are integral components of respiratory and photosynthetic electron transfer chains across all domains of life. The bc_1 complex of both mitochondria and bacteria is a dimeric enzyme and is present in aerobic as well as anaerobic energy-transducing respiratory chains. It catalyzes reduction of cytochrome *c* by oxidation of ubiquinol, with concomitant generation of a proton gradient that is utilized by the F_0F_1 ATP synthase to generate ATP. Electron transport is carried out by three redox center-bearing subunits: cytochromes *b* and c_1, and the "Rieske" iron-sulfur protein (Trumpower 1990*b*; Crofts and Berry 1998). In most bacteria, the bc_1 complex consists solely of these three so-called respiratory subunits and, in some cases, one additional low-molecular-mass subunit, which is noncatalytic (Gennis et al. 1993; Darrouzet et al. 1999; Montoya et al. 1999).

The evolutionary origin of mitochondria has been traced back to the α-subdivision of Proteobacteria (John and Whatley 1975; Yang et al. 1985; reviewed by Gray, Burger, and Lang 2001). Testifying to this ancestry are, among other features, three subunits of the mitochondrial bc_1 complex that are clearly homologs of the α-proteobacterial respiratory subunits (for a review, see Schütz et al. 2000). However, complex III from very different eukaryotes, i.e., potato, yeast, and bovine, contains seven additional subunits: two relatively large subunits designated "core" proteins, and five small polypeptides. The core proteins as well as the small subunits do not carry redox centers and are not directly involved in electron transfer (Brandt et al. 1994; Braun and Schmitz 1995*b*; Xia et al. 1997; Iwata et al. 1998; Zhang et al. 1998, 2000). In bovine, and probably other animals as well, the presequence of the "Rieske" iron-sulfur protein is retained in the bc_1 complex after proteolytic cleavage of the precursor protein, constituting an 11th subunit. To date, the bc_1 complex of only three protists has been studied in detail, *Euglena gracilis*, *Crithidia fasciculatum*, and *Leishmania tarentolae*, which all belong to the Euglenozoa lineage. The subunit composition of the bc_1 complexes from these organisms is much the same as in fungal, plant, and animal species (Mukai et al. 1989; Priest and Hajduk 1992; Horváth et al. 2000).

For a long time, the functions of the seven additional subunits of mitochondrial complex III were unknown. Only the core subunits of the plant mitochondrial bc_1 complex were found to possess processing peptidase activity (Braun et al. 1992; Eriksson, Sjoling, and Glaser 1996; Brumme et al. 1998). In contrast, the core subunits of animals and yeast are proteolytically inactive, and the mitochondrial processing peptidase (MPP) of these organisms is a soluble enzyme localized in the mitochondrial matrix. The core proteins of the bc_1 complex and MPP exhibit sequence similarity and share typical features with metalloendoproteinases of the pitrilysin family, indicating a common phylogenetic origin (Braun and Schmitz 1995*a*). Gene deletions and complementation experiments suggest that the core subunits of yeast are involved in the assembly of the bc_1 complex (Tzagoloff, Wu, and Crivellone 1986; Oudshoorn et al. 1987). The function and origin of the five small subunits of the bc_1 complex remain unknown.

To understand better how the mitochondrial bc_1 complex evolved from its much simpler α-proteobacterial predecessor, and how it diversified in the various eukaryotic lineages, we characterized the bc_1 complex from *Seculamonas ecuadoriensis*. This protist belongs to the

[1]Present address: Institut für Pflanzenphysiologie, Martin-Luther-Universität Halle-Wittenberg, Halle, Germany.

Key words: jakobid flagellates, *Seculamonas ecuadoriensis*, mitochondria, bc_1 complex, evolution.

[redacted]

Mol. Biol. Evol. 20(1):145–153. 2003
DOI: 10.1093/molbev/msg016

jakobids, a group of unicellular, heterotrophic flagellates that comprise the better-known species *Reclinomonas americana* (Flavin and Nerad 1993; Lang et al. 1997). Jakobids are assumed to have a very basal position in molecular phylogenies and include five aerobic genera: Seculamonas, Reclinomonas, Histiona, Jakoba, and Malawimonas. The four first genera share more morphological and ultrastructural features with one another than with Malawimonas and are therefore referred to as "core" jakobids (Edgcomb et al. 2001; O'Kelly, unpublished data). Up to now, six mitochondrial genomes of jakobids were completely sequenced, among these five core jakobids (*R. americana* NZ, *R. americana* 284, *S. ecuadoriensis*, *Jakoba libera*, and *J. bahamensis*) and the non-core jakobid, *Malawimonas jakobiformis* (Lang et al. 1997, http://megasun.bch.umontreal.ca/ogmp/projects/sumprog.html). Core jakobid mitochondrial genomes display an astonishing number of bacterial features more closely resembling the genome of the ancestral α-proteobacterial symbiont than any other mtDNA investigated today (Gray 1998; Gray et al. 1998; Lang, Gray, and Burger 1999). Some of the 18 or so extra genes present in most jakobid mitochondrial genomes have apparently been lost during evolution in all other eukaryotic lineages and functionally replaced by genes of other origin (e.g., $\alpha_2\beta,\beta'\sigma$ RNA polymerase, which has been replaced by a T3/T7-type enzyme; Cermakian et al. 1996, 1997). Most of the extra genes, however, are believed to have migrated to the nucleus in more derived eukaryotes. Among these are mostly genes coding for ribosomal proteins, but respiratory chain components migrate as well. One well-documented example is succinate-ubiquinone oxidoreductase (respiratory complex II), whose subunits 2, 3, and 4 (Sdh2 to Sdh4) are mitochondrially encoded in core jakobids (and also some plants and some protists like *Chondrus crispus*, *Porphyra purpurea*, *Rhodomonas salina*). Nucleus-encoded genes specifying Sdh2 have been identified in a number of fungi and animals. Phylogenetic analysis of bacterial, mitochondrial, and nuclear DNA-encoded Sdh2 sequence strongly suggest that the nuclear sdh2 genes originated by transfer from a mitochondrial genome in which it was originally resident (Burger et al. 1996).

As former studies on the mitochondrial respiratory chain, and the bc_1 complex in particular, have been conducted exclusively with derived eukaryotic taxa, jakobids are the organisms of choice to address the above evolutionary questions. Instead of *R. americana*, whose mtDNA sequence has been published previously (Lang et al. 1997), we have chosen the sister taxon *S. ecuadoriensis* for the protein-chemical experiments described here. The latter species is better amenable to biochemical studies that require a substantial amount of cell material, while it displays the same ancestral features and an almost identical mitochondrial gene set as *R. americana* (Burger and Lang, unpublished data).

Materials and Methods

Cell Culture

Seculamonas ecuadoriensis ATCC 50688 was grown in 2.5-liter culture flasks with gentle shaking at 24°C in WCL medium (http://megasun.bch.umontreal.ca/People/lang/FMGP/FMGP.html). The protists were fed with live *Enterobacter aerogenes* (ATCC 13048). A 600 ml culture yielded about 0.5 g of *S. ecuadoriensis* cells after 8 days.

Isolation of Membranes from *S. ecuadoriensis*

The following steps were carried out at 4°C, unless specified otherwise. For isolation of membranes, cells were pelleted by centrifugation, suspended in 0.2 M Na-phosphate buffer, pH 7.2, and disrupted by sonication. Cellular debris was removed by centrifugation at 12,000×*g* for 8 min. Membranes were then separated from the supernatant by centrifugation through sucrose step gradients (60%, 32%, 15% sucrose in 1 mM EDTA, 1 mM PMSF, and 10 mM MOPS/KOH, pH 7.2) at 92,000×*g* and 2°C for 1 h. The membrane fraction from the 15%/32% interphase was collected and diluted with 1 mM EDTA, 1 mM PMSF, and 10 mM MOPS/KOH, pH 7.2. Membranes were pelleted by centrifugation at 100,000×*g* for 90 min. The enrichment of mitochondrial membranes was monitored by cytochrome *c* oxidase activity measurements according to Hodges and Leonard (1974).

Isolation of Mitochondria from *Solanum tuberosum*

Mitochondria from potato tubers were isolated as described by Braun and Schmitz (1995*c*). The organelles were suspended in 0.4 M mannitol, 0.1% BSA, 1 mM EGTA, 0.2 mM PMSF, and 10 mM KH_2PO_4, pH 7.2, at a concentration of 10 mg of mitochondrial protein per milliliter.

Cytochrome *c* Affinity Chromatography

In preparation for affinity chromatography, 1.5 g membranes from *S. ecuadoriensis* were suspended in 2 ml ice-cold water and solubilized by slow addition of 10% Triton X100, to a final concentration of 3.3%. To remove membrane fragments and lipids, the suspension was centrifuged for 10 min at 60,000×*g*. A detailed protocol for the subsequent cytochrome *c* affinity chromatography is given in Linke and Weiss (1986). Proteins bound to the cytochrome *c* column were eluted using a Tris-acetate gradient (20–200 mM Tris-acetate [pH 7.0]/0.04% Triton, 5% sucrose, 0.2 mM phenylmethylsulfonyl fluoride). Fractions containing subunits of the bc_1 complex were identified by immunoblotting, pooled, and subsequently concentrated by ultrafiltration through filters with an exclusion limit of 300 kDa. Finally, the concentrate was analyzed by two-dimensional Blue-Native gel electrophoresis as described in the following section.

Separation of Mitochondrial Protein Complexes by Blue-Native and Tricine-SDS-PAGE

To determine the apparent molecular mass of mitochondrial protein complexes from *S. ecuadoriensis*, Blue-Native polyacrylamide gel electrophoresis (BN-PAGE) was carried out (Schägger, Cramer, and von Jagow 1994). Protein complexes of potato mitochondria were loaded onto the Blue-Native gels as a size standard. The BN

gels consisted of a separation gel (4.95% to 12.6% acrylamide) and a stacking gel (4% acrylamide). Sample preparation and electrophoresis was carried out as described by Jänsch et al. (1996). To separate the subunits of the protein complexes resolved by BN-PAGE, entire stripes of the BN gel were transferred horizontally on Tricine-SDS-PAGE gels. A protocol for this second-dimension Tricine-SDS gel is given in Schägger, Cramer, and von Jagow (1994). Tricine-SDS gels were either stained with Coomassie blue or silver nitrate or blotted onto filter membranes for immunological identification of proteins.

Identification of Proteins by Amino Acid Sequencing and Immunostaining

For internal sequence analysis, protein spots were cut from Tricine-SDS gels and digested with trypsin as outlined by Kruft et al. (2001). The resulting peptides were analyzed by Electrospray Ionization Tandem Mass Spectrometry (ESI-MS/MS). For immunostaining, Tricine-SDS gels were blotted onto nitrocellulose membranes. Blots were incubated with antibodies directed against the core II protein from *N. crassa* (dilution 1:1000). Visualization of immunopositive bands was performed using the Vectastain ABC-Kit (Vector Laboratories, Burlingame, CA) according to the manufacturer's instructions.

Quinol:Ferricytochrome *c* Activity Measurement

The quinol:ferricytochrome *c* activity assay was essentially carried out as described by Linke and Weiss (1986). Cytochrome *c* reduction was monitored at 24°C in a dual-wavelength photometer at 550 nm and 580 nm, using the extinction coefficient 20 mM^{-1} cm^{-1}. The test solution contained 50 mM LiMOPS, pH 6.8, 100 mM K_2SO_4, 40 μM cytochrome *c* from horse heart (Sigma, type III), 40 μM KCN, and 100 μM decylquinone (kindly provided by Dr. U. Schulte, Düsseldorf University, Germany). Antimycin, an inhibitor of bc_1 activity, was added to a final concentration of 2 μM. The turnover number is determined by extrapolating the rates of enzymatic reaction corrected for nonenzymatic rates to the infinite quinol concentration.

Sequence Similarity Searches

The peptide sequences determined in this study were searched against the following sequence repositories: the local jakobid database of the Organelle Genome Megasequencing Unit (OGMP), the nonredundant database (nrdb) of the National Centre for Biotechnology Information (NCBI); MITOP of the Munich Information Center for Protein Sequences (MIPS), a database for mitochondria-related genes, proteins, and diseases (Scharfe et al. 2000); dbEST, the division of GenBank that contains sequence data and other information on "single pass" cDNA sequences and Expressed Sequence Tags (ESTs) from a number of organisms (Boguski, Lowe, and Tolstoshev 1993); the motif databases Pfam (Bateman et al. 2002) and Prosite (Falquet et al. 2002); as well as the collection of HMM protein families (Eddy 1998, 2001). Downloaded and formatted for local searches were MITOP, dbEST, the Pfam families of Rieske (Fe-S) proteins, 14.5 kDa and 9.5 kDa subunits of complex III (UCR_14.5kD, UCR_9.5kD), and mitochondrial processing peptidases.

As search tools, we used FASTA, BLAST, PSI-BLAST and "Search for short sequences," available on the NCBI Blast homepage (Altschul et al. 1990, 1997). A very lenient e-value (1000) was used to account for the short length of the input peptide sequences (<30 residues). The word size of 2 and PAM 30 matrix were used as advised by the BLAST "Search for short sequences." The search against the HMM protein families was performed using HMMER (online version) (Eddy 2001). We also employed the online version of Mascot, which is a powerful search engine which uses mass spectrometry data to identify proteins from primary sequence databases (Perkins et al. 1999).

Results

Purification of the bc_1 Complex of *S. ecuadoriensis*

The bc_1 complex of *S. ecuadoriensis* was purified using cytochrome *c* affinity chromatography as published for *N. crassa* and plants (Linke and Weiss, 1986; Braun and Schmitz, 1992). According to the original protocols, the starting material for affinity chromatography should be isolated mitochondria. Owing to the slow growth of jakobid cultures, however, the preparation of pure mitochondria from quite limiting amounts of *S. ecuadoriensis* cells proved difficult. Therefore fractions of enriched mitochondrial membranes were generated from *S. ecuadoriensis* as described in the *Experimental Procedures* section. The fractions were fivefold to eightfold enriched in mitochondrial membrane proteins as monitored by cytochrome *c* oxidase measurements. The main purification step for the *S. ecuadoriensis* bc_1 complex, cytochrome *c* affinity chromatography, takes advantage of the specific interaction of the cytochrome c_1 subunit of the bc_1 complex and cytochrome *c*, the natural binding partner during respiratory electron transport. Proteins bound to the cytochrome *c* column were eluted by a salt gradient of 20–200 mM Tris-acetate, and the fractions obtained were analyzed by Tricine-SDS-PAGE and by immunoblotting. A 46 kDa band of fractions 19–23 eluted by 90–100 mM Tris-acetate strongly cross reacted with an antiserum directed against the core II protein from *N. crassa* (fig. 1). These peak fractions were pooled and concentrated by ultrafiltration using filters with an exclusion limit of 300 kDa. Finally the concentrate was analyzed by two-dimensional BN-PAGE/Tricine-SDS-PAGE. One protein complex was visible on the one-dimensional gels that could be resolved into seven protein bands of 46 kDa, 33 kDa, 29 kDa, 28 kDa, 14.5 kDa, 10 kDa, and 6 kDa upon separation on a second gel dimension (fig. 2).

That the isolated complex is indeed complex III of the respiratory chain was confirmed by two experiments. First, the native complex exhibits quinol:cytochrome *c* oxidoreductase activity (turnover number: 9.5 s-1), which is antimycin-sensitive. It should be noted that the electron

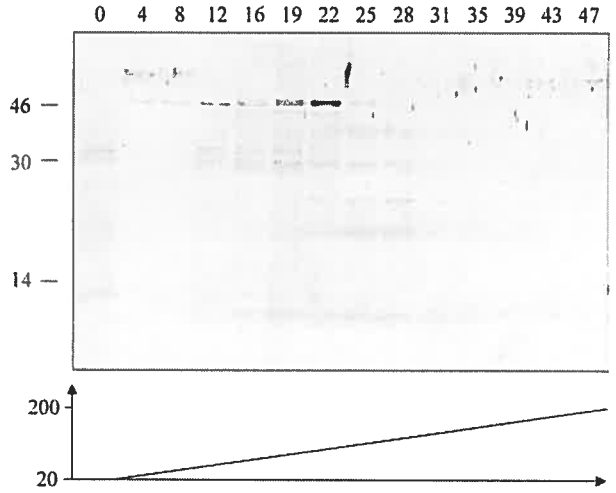

FIG. 1.—Purification of the bc_1 complex from *Seculamonas ecuadoriensis* by cytochrome *c* affinity chromatography. Fractions eluted from the affinity column were separated by SDS-PAGE, blotted, and probed with an antibody directed against the core II protein from *N. crassa*. The numbers of the fractions are indicated above the gel, and the molecular masses of standard proteins are shown to the left of the gel (in kDa). A graphical illustration of the Tris-acetate gradient (20–200 mM) used for elution of proteins bound to the cytochrome *c* column is given below the gel.

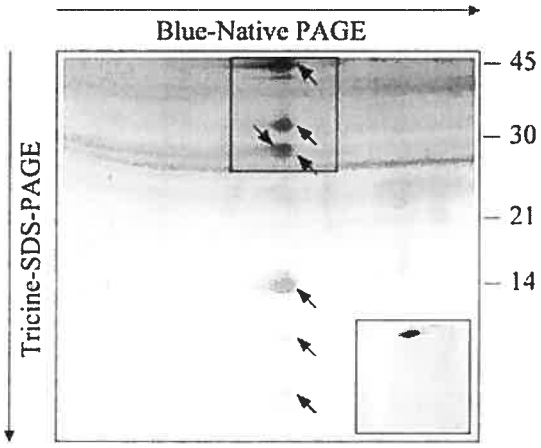

FIG. 2.—Characterization of the purified bc_1 complex from *S. ecuadoriensis* by BN-PAGE/Tricine-SDS-PAGE. Sizes of standard proteins are given on the right; protein spots of the bc_1 complex are marked with arrows. The spot at 46 kDa cross reacts with an antiserum directed against the core II protein from *N. crassa* (Inset: Western blot).

transfer activity is relatively low compared to preparations from other eukaryotes (Linke and Weiss 1986; Trumpower 1990*a*; Braun and Schmitz 1992), which might be due to the use of heterologous cytochrome *c* (from horse heart) as electron acceptor. Second, in an immunoblotting experiment, the 46 kDa spot on the 2D gels was shown to cross-react with the antiserum directed against the core II subunit of the bc_1 complex from *N. crassa* (fig. 2, *inset*).

Analysis of Protein Complexes from *S. ecuadoriensis* by BN PAGE/Tricine-SDS-PAGE and Determination of the Molecular Mass of the bc_1 Complex

To obtain further information on size and subunit composition of the bc_1 complex from *S. ecuadoriensis*, protein complexes from fractions enriched in mitochondrial membranes were analyzed directly by Blue-Native and Tricine-SDS-PAGE. BN-PAGE is a very reliable method for molecular mass determination of protein complexes (Schägger et al. 1994). Apparent molecular masses were estimated by co-electrophoresis of protein complexes from *S. ecuadoriensis* and mitochondrial protein complexes from potato, which served as size reference (fig. 3*A*). The membrane fraction of *S. ecuadoriensis* contains five protein complexes in the size range of 100 to 700 kDa and additional minor bands. Analysis of the separated protein complexes on a second gel dimension by Tricine-SDS-PAGE (fig. 3*B*) allowed to identify the protein complexes from potato on the basis of subunit compositions (Jänsch et al. 1996). One of the separated protein complexes of *S. ecuadoriensis* comprises an identical subunit composition like the bc_1 complex purified by affinity chromatography (fig. 2). Direct comparison of the protein complexes of potato mitochondria and *S. ecuadoriensis* mitochondrial fractions on Blue-Native gels revealed an apparent molecular mass of 460 kDa for the bc_1 complex from *S. ecuadoriensis* (fig. 3*A*). A further protein complex of *S. ecuadoriensis* runs at 550 kDa on Blue-Native gels and can be separated into 10 subunits upon analysis on a second gel dimension. The subunit composition of this 550 kDa complex resembles the one reported for mitochondrial ATP synthase complexes from other organisms (Jänsch et al., 1996; Boyer, 1997). The identity of the dominant protein complex at about 150 kDa could not be determined on the basis of subunit composition.

Identification of Individual Subunits of Protein Complexes Separated by BN-PAGE/Tricine-SDS-PAGE

To confirm the identity of the protein complexes and to obtain data on individual subunits, selected protein spots were subjected to peptide sequencing by Electrospray Ionization/Tandem Mass Spectrometry (ESI-MS/MS). We determined 1 to 3 peptide sequences of 7 different proteins (table 1), which form part of three different protein complexes and which are indicated and numbered on the gel in figure 3*B*. The sequence of peptide 1 of the 46 kDa protein forming part of the bc_1 complex exhibits significant similarity to a conserved stretch of the β MPP/core I subunit of the bc_1 complex from other eukaryotes. Notably, this peptide covers one of the few regions that distinguish α and β MPP paralogs (fig. 4). Peptide 2 of the same protein spot exhibits some weak similarities to the core II proteins from *N. crassa*. As shown above, this protein spot also cross reacts with an antiserum directed against the core II protein from *N. crassa* (fig. 1). These data strongly suggest that the bc_1 complex from *S. ecuadoriensis* comprises two core proteins with identical apparent molecular masses of 46 kDa.

Peptides 1 and 3 of the 29 kDa protein of the *S. ecuadoriensis* bc_1 complex exhibit low sequence identity

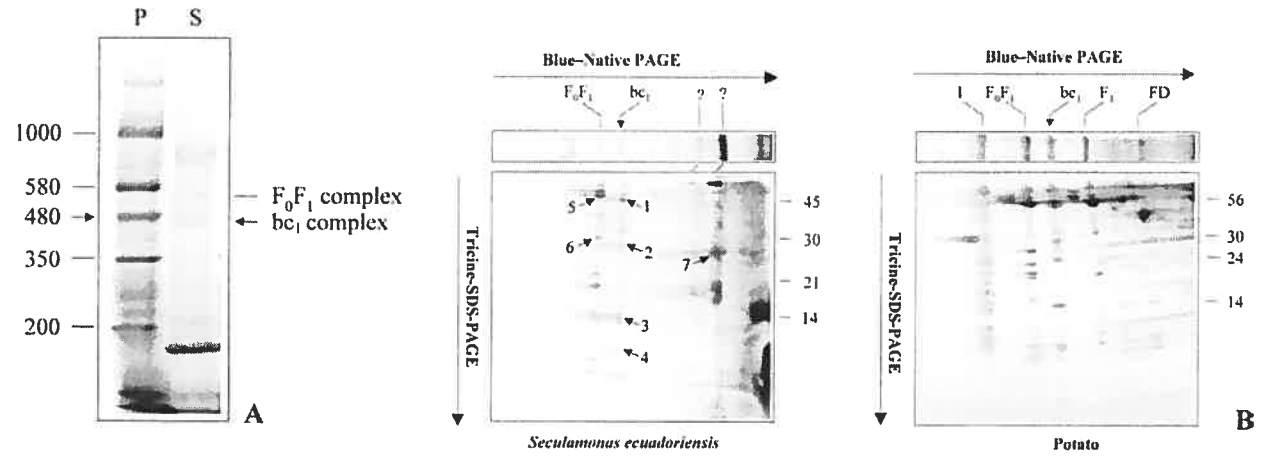

FIG. 3.—Determination of the molecular mass of the bc_1 complex from *S. ecuadoriensis* by Blue-Native gel electrophoresis (BN-PAGE). *A*, One-dimensional resolution of protein complexes from potato (P) and *S. ecuadoriensis* (S) by BN-PAGE. The molecular masses of the mitochondrial protein complexes from potato are taken from Jänsch et al. (1996) (the band at 480 kDa corresponds to the potato bc_1 complex). The bc_1 complex from *S. ecuadoriensis* runs at about 460 kDa. *B*, Two-dimensional resolution of protein complexes from *S. ecuadoriensis* and mitochondrial protein complexes from potato by BN-PAGE/Tricine-SDS-PAGE. Sizes of standard proteins are given on the right. Protein complexes from potato were identified by their characteristic subunit composition (Jänsch et al. 1996); protein complexes from *S. ecuadoriensis* were identified by sequence analysis of individual subunits by mass spectrometry (analyzed subunits are indicated by arrows and numbered according to table 1). I, NADH dehydrogenase; F_0F_1, F_0F_1 ATP synthase; bc_1, bc_1 complex; F_1, F_1 part of the F_0F_1 ATP synthase; FD, formate dehydrogenase; question marks indicate protein complexes that could not unambiguously be identified.

to cytochrome c_1 (table 1). However, the stretches of similarity represent unconserved regions of the protein and contain numerous insertions/deletions across the taxa, and therefore they did not withstand more rigorous statistical tests.

Also, neither peptides of the 14.5 kDa subunit nor those of the 10 kDa subunit of the *S. ecuadoriensis* bc_1 complex displayed significant sequence similarity to known proteins. Sequence conservation of the small subunits of the mitochondrial bc_1 complex is notoriously low (Braun and Schmitz 1995*b*). Much longer peptide sequences would be required than those determined in this study to identify phylogenetically distant, and, even more so, weakly conserved homologs.

The 48 and 30 kDa proteins of the F_0F_1 ATP synthase complex were unambiguously identified as the β and the γ subunits of this protein complex. Both peptide sequences obtained for 48 kDa protein exhibit high sequence identity to an internal sequence stretch of β subunits of the F_0F_1 ATP synthase complex from other organisms, and the two

Table 1
Peptide Sequences of Subunits of Protein Complexes from *Seculamonas ecuadoriensis*

No.[a]	Apparent Molecular Mass[b]	Peptide[c]	Sequence[d]	Sequences Producing Significant Alignments[e]	Source[f]	Subunit[g]
1.....	46 kDa	1–1	KTALLMDLDGSTPVA	KTSLLLALDGTTPVA	*B. emersonii* (gi1145777)	core I, bc_1 complex
		1–2	KFAPAPASPVSFEPAK			?
		1–3	KGEFSPALLQVPATVETSLK			?
2.....	29 kDa	2–1	LFKENGGLAVMQQFVK			?
		2–2	SEFQF			?
		2–3	HNLDVLNDLVDLN			?
3.....	14.5 kDa	3–1	ALQAASASLGATLPK			?
		3–2	ASAVDEDKSNN			?
4.....	10 kDa	4–1	LASNFAPNK			?
5.....	48 kDa	5–1	PSAVGYQPTLSEEMGILQ	PSAVGYQPTLNELQY	*B. aphidicola* (gi8977804)	β subunit, F_0F_1 ATPase
		5–2	TIAMDATEGLV	TIAMDATEGLV	*R. sphaeroides* (gi4633072)	β subunit, F_0F_1 ATPase
6.....	30 kDa	6–1	KIFSALLENATSEQGAR	KIFSALLENATSEQGAR	*S. ecuadoriensis* (OGMP)	γ subunit, F_0F_1 ATPase
		6–2	ELIEIISCASAVSSK	ELIEIISCASAVSSK	*S. ecuadoriensis* (OGMP)	γ subunit, F_0F_1 ATPase
7.....	26 kDa	7–1	RTTQLALPVLVLLFMGPGK			?

[a] The numbers correspond to the protein numbers as indicated in figure 3*B*.
[b] Apparent molecular masses as determined by Tricine-SDS-PAGE (fig. 3*B*).
[c] The first numbers correspond to the protein numbers as indicated in figure 1*B*. The second numbers indicate different peptides of these proteins.
[d] Peptide sequences as determined by Electrospray Ionization/Tandem Mass Spectrometry (ESI-MS/MS). Amino acids are given in the one-letter code.
[e] Sequences producing significant alignments as identified by using BLAST at NCBI or FASTA at the local jakobid database of the OGMP.
[f] Source of sequences producing significant alignments.
[g] Identity as determined by sequence similarity search. Question marks indicate that there is no significant identity to published sequences.

Neurospora	β MPP	- V S E A E V E R A K A Q L K A S I L I S L D G T T A V A E D - - I G R Q I V T	415	A29881
yeast	β MPP	K I S D A E V N R A K A Q L K A A L L I S L D G S T A I V E D - - I G R Q V V T	404	P10507
man	β MPP	- V T E S E V A R A R N L L K T N M L L Q L D G S T P I C E D - - I G R Q M L C	428	O75439
man	Core I	- A T E S E V A R G K N I L R N A L V S H L D G T T P V C E D - - I G R S L L T	419	I52367
S. ecuadoriensis		- - - - - - - - - - - - - - - K T A L L M D L D G S T P V A - - - - - - - - - -	15	
B. emersonii	β MPP	- P S L G E V A I A K Q Q L K T S L L I A L D G T T P V A E E - - I G R Q M L A	404	U41300
potato	Core I	- V S D A D V T H A C N Q L K S S L M T H I D G T S P V A E D - - I G R H V L T	469	B48529
yeast	Core I	- V T D T E V E R A K S L L K L Q L G Q L Y E S G N P V N D A N L L G A E V L I	396	P07256

FIG. 4.—Identification of the core I/β MPP subunit of *S. ecuadoriensis* by sequence comparison with core I and β MPP proteins of other organisms. Residues identical in at least six organisms are underlayed in black; other residues conserved in at least 4 organisms are underlayed in gray. Positions of the sequence stretches and accession numbers of the proteins are given on the right.

peptide sequences of the 30 kDa protein correspond exactly to amino acid stretches of the mitochondrial encoded γ subunit of the F_1 part of *S. ecuadoriensis*.

The identity of the dominant protein complex at 150 kDa could not be resolved on the basis of the peptide sequence of a 26 kDa subunit.

Discussion

This article reports the identification of two protein complexes of the respiratory chain from the jakobid flagellate *S. ecuadoriensis*, the F_0F_1 ATP synthase and the bc_1 complex. The experiments described here focus on the purification and characterization of the bc_1 complex. To our knowledge this is the first report on a biochemical preparation of an enzyme from jakobid flagellates. The *S. ecuadoriensis* bc_1 complex was purified on the basis of its affinity to the natural binding partner during respiratory electron transport, cytochrome *c*. The purified bc_1 complex retains both quinol:cytochrome *c* oxidoreductase activity and antimycin sensivity.

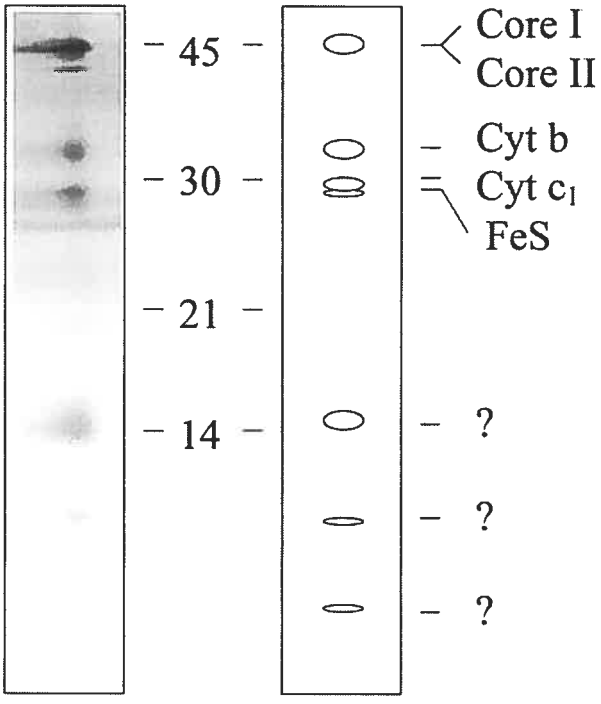

FIG. 5.—Identities of the subunits of the bc_1 from *S. ecuadoriensis*. A scheme of the gel is given on the right, and the sizes of standard proteins are given in the middle. Cyt, cytochrome; FeS, iron-sulfur protein; question marks indicate subunits of unknown identity.

The molecular mass of the bc_1 complex from *S. ecuadoriensis* lies at 460 kDa as determined by BN-PAGE. Under denaturing electrophoresis conditions the complex was resolved into seven protein bands with apparent molecular masses of 46 kDa, 33 kDa, 29 kDa, 28 kDa, 14.5 kDa, 10 kDa, and 6 kDa (fig. 2). Three lines of evidence strongly suggest that the largest protein spot encompasses two proteins, the core I and core II subunits. First, the 46 kDa spot cross reacts with an antiserum directed against the core II protein of the bc_1 complex from *N. crassa*, while a peptide derived from this spot also exhibits significant sequence identity to the core I/β-MPP subunit from different eukaryotes. Second, the core subunits of all mitochondrial bc_1 complexes characterized to date closely comigrate in gel electrophoresis in the size range of 45 to 55 kDa. Third, the apparent molecular mass of the *S. ecuadoriensis* bc_1 complex (460 kDa) can only be explained by assuming a dimeric holoenzyme that includes two core subunits per monomer as further discussed below.

In all bacteria and mitochondria characterized up to now, quinol:cytochrome *c* oxidoreductase activity is based on electron transfer reactions between the prosthetic groups of cytochrome *b*, cytochrome c_1, and the iron-sulfur protein. Given that the purified bc_1 complex of *S. ecuadoriensis* displays electron transport activity, it must include these three subunits. The molecular masses of the respiratory subunits are quite conserved in potato, bovine, and yeast, with 42–44 kDa for cytochrome *b*, 27–28 kDa for cytochrome c_1, and 20–23 kDa for the iron-sulfur protein (Braun and Schmitz, 1995*b*). A considerable degree of conservation across three different eukaryotic phyla allowed us to assign the following three protein species of the purified bc_1 complex of *S. ecuadoriensis* to respiratory subunits. First, the 33 kDa protein of *S. ecuadoriensis* is most likely cytochrome *b*. This is consistent with the fact that cytochrome *b*, because of its highly hydrophobic properties, typically displays a migration behavior that makes its molecular mass appear ~25% smaller than it really is (Mendel-Hartvig and Nelson 1983; Berry, Huang, and Derose 1991; Priest and Hajduk 1992; Braun and Schmitz 1995*b*). Second, the 29 kDa protein of *S. ecuadoriensis* is probably cytochrome c_1 and is thus only slightly larger than the proteins of its well-characterized counterparts. Finally, the 28 kDa subunit

of *S. ecuadoriensis* represents, very likely, the iron-sulfur protein. Although the *S. ecuadoriensis* protein is approximately 30% larger than its homologs of the model organisms, such a size deviation is not without precedent. A comparatively large iron-sulfur subunit was also reported for the bc_1 complex from *L. tarentolae* (Horváth et al. 2000). Figure 5 summarizes the demonstrated and inferred subunit assignments of the bc_1 complex from *S. ecuadoriensis*.

As already mentioned, bacterial and mitochondrial bc_1 complexes alike are dimers. Assuming the presence of two core proteins, the molecular masses of the eight separated subunits of the *S. ecuadoriensis* bc_1 complex sums up to 423 kDa for the dimeric complex, a value that is somewhat smaller than the experimentally determined molecular mass of the complex (460 kDa). The difference of 27 kDa could be due to the presence of one or two further low molecular mass subunits that may not have been detected in our experiments. These small proteins are difficult to spot because they migrate closely together and are poorly stainable.

While the bacterial bc_1 complex comprises three or four subunits at most, with a molecular mass of the dimeric complex of ~220 kDa (Yu et al. 1999), the mitochondrial counterpart is at least twice as large. At present, complex III has been characterized from highly diverse eukaryotes, including plants, fungi, animals, and euglenoid and kinetoplastid protists (Mukai et al. 1989; Priest and Hajduk 1992; Brandt et al. 1994; Gutiérrez-Cirlos et al. 1994; Braun and Schmitz 1995*b*; Xia et al. 1997; Iwata et al. 1998; Zhang et al. 1998; Horváth et al. 2000). All these mitochondrial bc_1 complexes have an apparent molecular mass of 470–495 kDa and consist of three respiratory subunits, two large core proteins and, at least in higher eukaryotes (but probably in all eukaryotic taxa investigated up to now), five small subunits. As we show here, the bc_1 complex from the jakobid protistan *S. ecuadoriensis* has only a slightly smaller molecular mass (460 kDa) than that from fungi, mammals, and plants (470–495 kDa) and includes at least eight different subunits. Because jakobids are believed to be a primitive eukaryotic lineage, this finding was unexpected.

The view that jakobids are minimally derived eukaryotes is based on their ultrastructural similarities with the retortamonads, an amitochondriate group considered to have diverged close to the eukaryotic origin (O'Kelly 1993). *S. ecuadoriensis* is a typical member of jakobids and of the core jakobids in particular, as first established by analysis of the basal body ultrastructure and other cellular characters (O'Kelly 1993). Furthermore, phylogenetic analyses using mitochondrion-encoded protein genes clearly affiliate *S. ecuadoriensis* with *R. americana* (not shown). In global eukaryotic trees, however, available multiple mitochondrial protein data fail to place the jakobids relative to the other eukaryotic lineages with confidence, which is most likely a result of the low sampling of jakobid taxa (Lang, Gray, and Burger 1999). Similarly uncertain topologies are obtained with single nuclear genes (α- and β-tubulin) (Edgcomb et al. 2001). Nevertheless, and irrespective of their exact phylogenetic position relative to the other eukaryotes, mitochondrial genomes of core jakobids, such as *R. americana* and *S. ecuadoriensis*, display an astonishing number of bacterial features, more closely resembling the genome of the ancestral α-proteobacterial symbiont than any other mtDNA investigated today (Gray 1998; Gray et al. 1998; Lang, Gray, and Burger 1999).

Our initial hypothesis posited that jakobid mitochondrial complexes are evolutionary intermediates between those of α-proteobacteria and mitochondria from higher eukaryotes. Although the finding of much the same complex III structure in *S. ecuadoriensis* and plants or fungi does not corroborate this view, the particular complex investigated may not be a suitable choice for detecting a more bacteria-like structure. Indeed, considering the number of mitochondrion-encoded subunits, complex III of the jakobids is as much derived as that of all other eukaryotes studied. Among the three genes coding for respiratory subunits of α-proteobacterial origin, only the apoprotein b gene still resides in mtDNA, whereas the other two, i.e., the cytochrome c_1 and iron-sulfur protein, are nucleus-encoded.

It is conceivable that the number of originally α-proteobacterial genes that have migrated to the nucleus is correlated with the number of secondarily acquired subunits in a respiratory complex. In fact, the migration of mitochondrial genes to the nucleus may actually be the underlying cause for the recruitment of additional subunits the crucial role of which may involve mediation of protein complex assembly from components that are now synthesized in different cellular compartments and synchronization of the expression of genes now located in different genomes.

If this hypothesis is correct, one should expect a more primitive structure in those mitochondrial membrane complexes that have retained a larger number of mitochondrion-encoded subunits in jakobids than in other eukaryotes. The mitochondrial NADH dehydrogenase (complex I of the respiratory chain) and the ATP synthase are two such cases. Additional subunits of these protein complexes are encoded by mitochondrial genes in *R. americana* (as well as *S. ecuadoriensis*, unpublished data) if compared to the mitochondrial genomes of other eukaryotes (Lang et al. 1997). Like the situation in complex III, all mitochondrial ATP synthases characterized in eukaryotes other than jakobids have approximately five more subunits over and above the eight that are typically present in the bacterial enzyme (Boyer 1997). To test the hypothesis that subunit recruitment is the consequence of gene migration from the mitochondrial genome to the nuclear genome, investigation of the complex composition of *S. ecuadoriensis* ATP synthase is under way.

Acknowledgments

We thank Dr. U. Schulte, Düsseldorf, for instructing us on quinol:cytochrome c reductase activity measurements in his laboratory. This work was supported by the Deutsche Forschungsgemeinschaft, the Fonds der Chemischen Industrie, and the Canadian Institute for Health Research (CIHR). G.B. is a Canadian National Associate,

and B.F.L is an Imasco fellow in the program of Evolutionary Biology of the Canadian Institute of Advanced Research (CIAR), which we thank for salary and interaction support.

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer. 2002. The Pfam protein families database. Nucleic Acids Res. **30**:276–280.

Berry, E. A., L. S. Huang, and V. J. Derose. 1991. Ubiquinol-cytochrome c oxidoreductase of higher plants. Isolation and characterization of the *bc*1 complex from potato tuber mitochondria. J. Biol. Chem. **266**:9064–9077.

Boguski, M. S., T. M. Lowe, and C. M. Tolstoshev. 1993. dbEST-database for "expressed sequence tags." Nat. Genet. **4**:332–333.

Boyer, P. D. 1997. The ATP synthase—a splendid molecular machine. Annu. Rev. Biochem. **66**:717–749.

Brandt, U., S. Uribe, H. Schägger, and B. L. Trumpower. 1994. Isolation and characterization of QCR10, the nuclear gene encoding the 8.5-kDa subunit 10 of the *Saccharomyces cerevisiae* cytochrome *bc*1 complex. J. Biol. Chem. **269**:12947–12953.

Braun, H. P., M. Emmermann, V. Kruft, and U. K. Schmitz. 1992. The general mitochondrial processing peptidase from potato is an integral part of cytochrome c reductase of the respiratory chain. EMBO J. **11**:3219–3227.

Braun, H. P., and U. K. Schmitz. 1992. Affinity purification of cytochrome c reductase from potato mitochondria. Eur. J. Biochem. **208**:761–767.

———. 1995*a*. Are the 'core' proteins of the mitochondrial *bc*1 complex evolutionary relics of a processing protease? Trends Biochem. Sci. **20**:171–175.

———. 1995*b*. The bifunctional cytochrome c reductase/processing peptidase complex from plant mitochondria. J. Bioenerg. Biomembr. **27**:423–436.

———. 1995*c*. Molecular structure of the 8 kDa subunit of the cytochrome c reductase from potato and its $\Delta\Psi$-dependent import into isolated mitochondria. Biochem. Biophys. Acta **1229**:181–186.

Brumme, S., V. Kruft, U. K. Schmitz, and H. P. Braun. 1998. New insights into the co-evolution of cytochrome c reductase and the mitochondrial processing peptidase. J. Biol. Chem. **273**:13143–13249.

Burger, G., B. F. Lang, M. Reith, and M. W. Gray. 1996. Genes encoding the same three subunits of respiratory complex II are present in the mitochondrial DNA of two phylogenetically distant eukaryotes. Proc. Natl. Acad. Sci. USA **93**:2328–2332.

Cermakian, N., T. M. Ikeda, R. Cedergren, and M. W. Gray. 1996. Sequence homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucleic Acids Res. **24**:648–654.

Cermakian, N., T. M. Ikeda, P. Miramontes, B. F. Lang, M. W. Gray, and R. Cedergren. 1997. On the evolution of the single-subunit RNA polymerase. J. Mol. Evol. **45**:671–681.

Crofts, A. R., and E. A. Berry. 1998. Structure and function of the cytochrome *bc*1 complex of mitochondria and photosynthetic bacteria. Curr. Opin. Struct. Biol. **8**:501–509.

Darrouzet, E., M. Valkova-Valchanova, T. Ohnishi, and F. Daldal. 1999. Structure and function of the bacterial *bc*1 complex: domain movement, subunit interactions, and emerging rationale engineering attempts. J. Bioenerg. Biomembr. **31**:275–288.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics **14**:755–763.

———. 2001. HMMER: Profile hidden Markov models for biological sequence analysis. http://hmmer.wustl.edu/.

Edgcomb, V. P., A. J. Roger, A. G. Simpson, D. T. Kysela, and M. L. Sogin. 2001. Evolutionary relationships among "jakobid" flagellates as indicated by alpha- and beta-tubulin phylogenies. Mol. Biol. Evol. **18**:514–522.

Eriksson, A. C., S. Sjoling, and E. Glaser. 1996. Characterization of the bifunctional mitochondrial processing peptidase (MPP)/*bc*1 complex in *Spinacia oleracea*. J. Bioenerg. Biomembr. **28**:285–292.

Falquet, L., M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann, and A. Bairoch. 2002. The PROSITE database, its status in 2002. Nucleic Acids Res. **30**:235–238.

Flavin, M., and T. A. Nerad. 1993. *Reclinomonas americana* N. G., N. Sp., a new freshwater heterotrophic flagellate. J. Eukaryot. Microbiol. **40**:172–179.

Gennis, R. B., B. Barquera, B. Hacker, S. R. Van Doren, S. Arnaud, A. R. Crofts, E. Davidson, K. A. Gray, and F. Daldal. 1993. The *bc*1 complexes of *Rhodobacter sphaeroides* and *Rhodobacter capsulatus*. J. Bioenerg. Biomembr. **25**:195–209.

Gray, M. W. 1998. *Rickettsia*, typhus and the mitochondrial connection. Nature **396**:109–110.

Gray, M. W., G. Burger, and B. F. Lang. 2001. The origin and early evolution of mitochondria. Genome Biol. **2**:1018.1–1018.5.

Gray, M. W., B. F. Lang, R. Cedergren et al. (15 co-authors). 1998. Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res. **26**:865–878.

Gutiérrez-Cirlos, E. B., A. Antaramian, M. Vázquez-Acevedo, R. Coria, and D. González-Halphen. 1994. A highly active ubiquinol-cytochrome c reductase (bc_1 complex) from the colorless algae *Polytomella* ssp., a close relative of *Chlamydomonas*. J. Biol. Chem. **269**:9147–9154.

Hodges, T. K., and R. T. Leonard. 1974. Purification of a plasma membrane-bound adenosine triphosphate from plant roots. Meth. Enzymol. **32**:392–406.

Horváth, A., E. A. Berry, L. S. Huang, and D. A. Maslov. 2000. *Leishmania tarentolae*: a parallel isolation of cytochrome *bc*(1) and cytochrome c oxidase. Exp. Parasitol. **96**:160–167.

Iwata, S., J. W. Lee, K. Okada, J. K. Lee, M. Iwata, B. Rasmussen, T. A. Link, S. Ramaswamy, and B. K. Jap. 1998. Complete structure of the 11-subunit bovine mitochondrial cytochrome *bc*1 complex. Science **281**:64–71.

Jänsch, L., V. Kruft, U. K. Schmitz, and H. P. Braun. 1996. New insights into the composition, molecular mass and stoichiometry of the protein complexes of plant mitochondria. Plant J. **9**:357–368.

John, P., and F. R. Whatley. 1975. *Paracoccus denitrificans* and the evolutionary origin of the mitochondrion. Nature **254**:495–498.

Kruft, V., H. Eubel, W. Werhahn, L. Jänsch, and H. P. Braun. 2001. Proteomic approach to identify novel mitochondrial functions in *Arabidopsis thaliana*. Plant Physiol. **127**:1694–1710.

Lang, B. F., G. Burger, C. J. O'Kelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, and M. W. Gray. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature **387**:493–497.

Lang, B. F., M. W. Gray, and G. Burger. 1999. Mitochondrial

genome evolution and the origin of eukaryotes. Annu. Rev. Genet. **33**:351–397.

Linke, P., and H. Weiss. 1986. Reconstitution of ubiquinol-cytochrome-c reductase from *Neurospora* mitochondria with regard to subunits I and II. Methods Enzymol. **126**:201–210.

Mendel-Hatvig, I., and B. D. Nelson. 1983. Studies on beef heart ubiquinol-cytochrome c reductase. Topological studies on the core proteins using proteolytic digestion and immunoreplication. J. Bioenerg. Biomembr. **15**:27–36.

Montoya, G., K. te Kaat, S. Rodgers, W. Nitschke, and I. Sinnig. 1999. The cytochrome *bc*1 complex from *Rhodovulum sulfidophilum* is a dimer with six quinones per monomer and an additional 6-kDa component. Eur. J. Biochem. **259**:709–718.

Mukai, K., M. Yoshida, H. Toyosaki, Y. Yao, S. Wakabayashi, and H. Matsubara. 1989. An atypical heme-binding structure of cytochrome c1 of *Euglena gracilis* mitochondrial complex III. Eur. J. Biochem. **178**:649–656.

O'Kelly, C. J. 1993. The jakobid flagellates: structural features of *Jakoba*, *Reclinomonas* and *Histiona* and implications for the early diversification of eukaryotes. J. Euk. Microbiol. **40**:627–636.

Oudshoorn, P., H. Van Steeg, B. W. Swinkels, P. Schoppink, and L. A. Grivell. 1987. Subunit II of yeast QH2:cytochrome-c oxidoreductase. Nucleotide sequence of the gene and features of the protein. Eur. J. Biochem. **163**:97–103.

Perkins, D. N., D. J. Pappin, D. M. Creasy, and J. S. Cottrell. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20**:3551–3567.

Priest, J. W., and S. L. Hajduk. 1992. Cytochrome c reductase purified from *Crithidia fasciculata* contains an atypical cytochrome c1. J. Biol. Chem. **267**:20188–20195.

Schägger, H., W. A. Cramer, and G. von Jagow. 1994. Analysis of molecular masses and oligomeric states of protein complexes by blue native electrophoresis and isolation of membrane protein complexes by two-dimensional native electrophoresis. Anal. Biochem. **217**:220–230.

Scharfe, C., P. Zaccaria, K. Hoertnagel et al. (12 co-authors). 2000. MITOP, the mitochondrial proteome database: 2000 update. Nucleic Acids Res. **28**:155–158.

Schütz, M., M. Brugna, E. Lebrun et al. (12 co-authors). 2000. Early evolution of cytochrome *bc* complexes. J. Mol. Biol. **300**:663–675.

Trumpower, B. L. 1990*a*. Cytochrome *bc*1 complexes of microorganisms. Microbiol. Rev. **54**:101–129.

———. 1990*b*. The protonmotive Q cycle. Energy transduction by coupling of proton translocation to electron transfer by the cytochrome *bc*1 complex. J. Biol. Chem. **265**:11409–11412.

Tzagoloff, A., M. A. Wu, and M. Crivellone. 1986. Assembly of the mitochondrial membrane system. Characterization of COR1, the structural gene for the 44-kilodalton core protein of yeast coenzyme QH2-cytochrome c reductase. J. Biol. Chem. **261**:17163–17169.

Xia, D., C. A. Yu, H. Kim, J. Z. Xia, A. M. Kachurin, L. Zhang, L. Yu, and J. Deisenhofer. 1997. Crystal structure of the cytochrome *bc*1 complex from bovine heart mitochondria. Science **277**:60–66.

Yang, D., Y. Oyaizu, H. Oyaizu, G. J. Olsen, and C. R. Woese. 1985. Mitochondrial origins. Proc. Natl. Acad. Sci. USA **82**:4443–4447.

Yu, L., S.-C. Tso, S. K. Shenoy, B. N. Quinn, and D. Xia. 1999. The role of the supernumerary subunit of the *Rhodobacter sphaeroides* cytochrome bc_1 complex. J. Bioenerg. Biomembr. **31**:251–257.

Zhang, Z., E. A. Berry, L. S. Huang, and S. H. Kim. 2000. Mitochondrial cytochrome *bc*1 complex. Subcell. Biochem. **35**:541–580.

Zhang, Z., L. Huang, V. M. Shulmeister, Y. I. Chi, K. K. Kim, L. W. Huang, A. R. Crofts, E. A. Berry, and S. H. Kim. 1998. Electron transfer by domain movement in cytochrome *bc*1. Nature **392**:677–684.

Geoffrey McFadden, Associate Editor

Accepted September 24, 2002

Chapter 2 : ARTICLE

Sequence similarity-based methods leave a large portion of genes with unannotated function. In the case of mitochondrion-encoded genes whose sequences are all strored in the organelle genome database GOBASE, ~9% are of unknown function. As a part of GOBASE's data curation efforts, we developed a similarity-free method called MOPS.

I have designed, developed and implemented the Mitochondrial ORF function Prediction System (MOPS) and I wrote the manuscript.

Function prediction of hypothetical proteins without sequence similarity to proteins of known function

Sivakumar Kannan*, Amy M. Hauth and Gertraud Burger

Canadian Institute for Advanced Research

Robert Cedergren Research Center for Bioinformatics and Genomics

Département de Biochimie, Université de Montréal

2900 Boulevard Edouard-Montpetit

Montréal, Québec H3T 1J4, Canada

*Corresponding author

[redacted] : [redacted]

Abstract

Background

Function prediction by sequence-similarity based methods identifies only ~50% of the proteins deduced from newly sequenced genomes. Therefore, we set out to develop an approach to annotate the 'leftover proteins' i.e., those which cannot be assigned function using sequence similarity. Our goal is to perform pan-taxonomic prediction specifying fine-grained molecular function (rather than a broad functional category).

Results

Our sequence-similarity free approach to function annotation involves representation of proteins by a host of calculated attributes such as physicochemical properties and amino acid composition. We employed a decision tree algorithm and addressed both data redundancy and class imbalance in an innovative manner. To demonstrate the merits of this approach, we developed MOPS (Mitochondrial ORF function Prediction System) whose accuracy exceeds 82% when tested with mitochondrion-encoded proteins of known function. In addition, we developed a validation scheme that assesses predictions using domain-specific knowledge. Based on our results, we critically discuss current performance measures and validation methods employed for protein function prediction.

Background

Large-scale genome and EST (Expressed Sequence Tag) sequencing projects are producing data at an ever-increasing rate. Assigning function to the high volume of inferred proteins makes *in silico* methods indispensable. Most function annotation methods use sequence similarity by first finding a similar protein of known function and then extrapolating the known function to the inferred protein. Generally, annotation tools employ BLAST (Altschul *et al.*, 1990b), but applications using machine learning algorithms are available as well (e.g., Vinayagam *et al.*, 2004). However, sequence-similarity approaches have limits; they typically annotate less than 50% of the inferred proteins deduced from a newly sequenced genome.

To increase the number of proteins annotated, recent research explores sequence-similarity free ('similarity-free') function prediction, as a complement to similarity-based methods such as BLAST. Rather than directly using amino-acid sequence, similarity-free function prediction methods use protein features (e.g., physicochemical properties and protein secondary structure) calculated from it and employ machine learning algorithms. This approach has been used successfully in predicting protein structure, functional sites and subcellular location, etc. (Dobson *et al.*, 2004; Szilagyi *et al.*, 2005). King and co-workers were the first to employ this approach for assigning the molecular function to hypothetical proteins from *M. tuberculosis*, *E. coli* and *S. cerevisiae* (King *et al.*, 2000; King *et al.*, 2001b); subsequent experimental validation (King *et al.*, 2004b) testifies to the power of similarity-free prediction of molecular protein function. Still, this work has

limitations. First, the predictor identifies only a broad functional category (e.g., transport/binding proteins) instead of a specific function (e.g., SecY-type transporter protein). Second, the predictor still depends to a certain extent on sequence similarity, since PSI-BLAST serves for calculating phylogenetic attributes. Third, the predictor is species-specific, meaning that proteins only from the species used for training can be annotated.

Our research seeks to build an efficient automated system to annotate inferred proteins that lack recognizable sequence similarity (referred to in the following as hypothetical proteins). In this study we focus on the molecular function of a protein as defined by Gene Ontology (Ashburner *et al.*, 2000b). The approach we present here addresses shortcomings discussed above: (i) it predicts a fine-grained molecular function (i.e., most specific in the Gene Ontology hierarchy) rather than a broad functional category, and (ii) the predictor is applicable to any taxon, not just the species included in the training set. Data are represented by a host of multiple similarity-free protein features. The predictor is trained using a decision tree algorithm (Quinlan, 1993b), which has the distinct advantage of producing rules that can be interpreted by humans, and thus provides a window on the underlying biology.

To test the effectiveness of this similarity-free annotation system, we built a classifier that predicts the molecular function of mitochondrion-encoded hypothetical proteins. Mitochondrial genomes (mtDNA) harbor up to ~70 protein-coding genes that specify components involved in respiration, ATP synthesis, protein synthesis, transcription, protein import and maturation (Table 1). In contrast to nuclear genomes, gene families are

absent from mtDNA (for a review, see Gray *et al.*, 2004). As of now, around 2,000 hypothetical protein-coding genes of unknown function (ORFs) have been reported in mitochondria. Further, mtDNA-encoded ORFs are assumed to originate from 'regular' mitochondrial genes that are too derived to be recognized using sequence similarity. Predicting the likely function of mitochondrion-encoded ORFs is an important part of the expert data curation task of GOBASE – the publicly accessible, taxonomically-broad organelle genome database that organizes, integrates and validates all available data on mitochondria and chloroplasts (O'Brien *et al.*, 2006b).

Our new approach has been implemented as an automated system for the annotation of mitochondrion-encoded hypothetical proteins, named MOPS (Mitochondrial ORF function Prediction System). The performance of MOPS is assessed in two ways: by standard machine-learning specific measures and by a new procedure based on expert knowledge. This new validation procedure is honed for function prediction of mitochondrion-encoded hypothetical proteins, but the principle is generally applicable to other biological problems involving *in silico* predictions.

Results

Protein sequence representation

Open reading frames annotated as hypothetical proteins lack significant sequence similarity with known proteins (See Methods section 'Dataset'). To determine function, proteins must be represented by attributes other than sequence similarity.

A major challenge in machine learning-based function annotation is the selection of effective attributes that is, knowledge representation of proteins. First, we investigated the predictive power of attributes used in earlier studies (King *et al.*, 2001b; Lu *et al.*, 2004), notably physicochemical properties (molecular weight, isoelectrical point, etc.,) and atomic and residue composition of proteins alone and in combination (Table 2, top). Then, we explored pairs of amino acids (dipeptides; Table 2, bottom) in an attempt to capture local amino acid correlations, e.g., regularly spaced hydrophilic and hydrophobic amino acids, three or four residues apart in amphipathic helices. Dipeptides, in particular ungapped ones, have been shown previously to be effective attributes for machine-learning-based inference of cellular location and broad functional classes (King *et al.*, 2001b; Park and Kanehisa, 2003). (Note that dipeptides do not contain enough context information to be exploited by similarity-based methods.) When testing various attribute combinations, we found that inclusion of gapped dipeptides did not improve and, in some cases, decreased prediction performance. The best prediction performance was obtained by the combination of physiochemical properties, atomic and residual composition and ungapped dipeptides (Table 3). Hence in all further studies, protein sequences were represented in this optimal way.

Sequence similarity in dataset

Classifiers built with the above mentioned attributes displayed an unexpectedly high overall precision and specificity (>90%) when trained with the basic dataset of proteins (Table 4, top). This could indicate that sequence similarity might permeate to some extent

physicochemical and compositional attributes, despite the abstraction from all contextual information. Since our aim is to eliminate reliance on sequence similarity, a predictor should be trained and evaluated using sequences that exhibit only moderate similarity to one another. Therefore, we formed new datasets clustered at sequence identity thresholds down to 50%. Remarkably, classifiers trained with each of these datasets maintained high prediction specificity. Yet, precision and sensitivity dropped significantly, due to an increase of false negatives (Table 4). As we show in the next section, the underlying reason for the growing number of false negatives lies in the declining number of instances.

Class size, class imbalance, undersampling and performance

In the basic data set, the number of instances per functional class (i.e., the class size) of mitochondrion-encoded proteins ranges from one to 3,476 instances (see Supplementary information, Table S1). Such an imbalance is well known to reduce the performance of machine learning algorithms. In order to visualize in how far the predictor performance depends on the class size, we plotted the number of instances versus precision for each class. Figure 1 shows that class sizes of >30 instances on average yield satisfying prediction.

Undersampling is a widely employed method to address class imbalance and generally involves removing randomly-selected instances from the majority classes (Drummond and Holte, 2003). In fact, forming classes of equal size (fully balanced) has been shown to yield the best prediction performance compared to using data with class imbalance (Al-Shahib *et al.*, 2005). Our primary undersampling method reduced class size

based on sequence identity. We explored several ways of clustering the basic dataset of known proteins (denoted DS-1000 because of the 1000-fold difference in class size) to form new datasets having less class imbalance. First, all classes were clustered at the same sequence identity thresholds (see Methods section 'Sequence clustering'). This substantially reduced the size of classes containing many instances, but it also reduced class size for those having few instances. Second, class-specific thresholds were chosen to decrease the number of instances in a class as much as possible while retaining at least 30 instances, a threshold based on the analysis presented in Figure 1. The resultant dataset (denoted DS-100) had 100-fold overall class imbalance. Finally, to reduce the imbalance further, a dataset was clustered first on a class-by-class basis and then instances in classes larger than 40 were removed randomly. This created a dataset having only a 10-fold imbalance (denoted DS-10; See Supplementary information Figure S1 for the class size distribution of various datasets).

Using the datasets DS-1000, DS-100 and DS-10, we trained three classifiers (denoted C-1000, C-100 and C-10, respectively) and evaluated them by ten-fold cross-validation. The overall prediction performance indicates that C-1000 is the best of the three, while C-10 performs least well (Table 5). Closer inspection revealed that different classifiers have predictive strength at different class sizes. For large classes (>30 instances), C-1000 performed substantially better than the other predictors (Table 5). In contrast, for small classes (<=30 instances), classifiers trained on the more balanced datasets performed better than that based on the highly imbalanced DS-1000 (note that training and test sets for all three classifiers are identical for small classes). Finally, we compared the confidence

factors assigned to individual function predictions for known proteins. Since there was no clear correlation between the confidence factor and correctness (data not shown), this measure was not considered further in the context of this study.

Function prediction of hypothetical proteins

For predicting the function of the 1,336 hypothetical proteins, we constructed three classifiers (MOPS-1000, -100 and -10) based on the full datasets of known proteins DS-1000, DS-100 and DS-10. The reason for three classifiers is that, as we showed above, classifiers trained with the different datasets displayed complementary performance in the prediction of small and large functional classes for known proteins. A full list of ORF function predictions is available in the supplementary information (See Supplementary information Table S5).

We compared the prediction performance of MOPS with BLAST, even though this comparison is not fair because as mentioned earlier, our method is complementary to sequence-similarity-based methods rather than an alternative. We stated in the Methods section that a small portion of the hypothetical proteins (361) has significant BLAST hits when searched against the known proteins. About 100 MOPS predictions concur with assignments made by BLAST. BLAST and MOPS disagree on 260 protein annotations, most of which are annotated as "endonucleases" by BLAST. Yet this disagreement is not surprising, since "endonuclease" is an ill-defined and divergent class having at least eight families in the Pfam database (Finn *et al.*, 2006b). In the next section, we show that ~700

MOPS predictions receive positive support testifying to the merits of MOPS compared to BLAST.

Discussion

We developed an effective automated method for molecular function prediction of 'left-over' proteins i.e., those that do not display significant sequence similarity to known proteins. This new method employs a host of physicochemical and compositional protein properties, addresses class imbalance and overcomes several limitations seen in other predictors. First, our method is able to predict fine-grained functional classes (e.g., 'succinate dehydrogenase subunit 2'), rather than broad functional categories (e.g., 'component involved in mitochondrial electron transport'). Second, our method is taxonomically independent in that, unlike other methods (e.g., King *et al.*, 2001b), it does not rely on data from the same species for training. The following is a critical discussion regarding the high performance of our predictors (for known proteins) and the relevance of constructing meaningful and effective test and training sets. We also point to potential pitfalls in the interpretation of machine-learning-based function classification of proteins. Finally, we provide a new approach for validating individual function predictions of hypothetical proteins.

Reasons for high predictor performance

The C-1000 classifier (trained with the dataset DS-1000 having a 1000-fold class imbalance) exhibited an exceptionally high performance (see Table 4). A reason therefore

might be that the sequences within large classes are quite similar, introducing overfitting. Indeed, after clustering the sequences at a threshold of 75%, false negatives increased and the overall precision decreased to 82% – a performance level comparable to that of other methods (e.g.,Clare *et al.*, 2006b).

When dissecting performance of classifiers by class size, we observed that for proteins from large classes, function is best predicted by C-1000 (see Table 5). In contrast, for proteins from small classes, function is best predicted by classifiers trained with a dataset having reduced class imbalance. This observation can be exploited for function prediction of unassigned proteins, by selecting a classifier based on the distribution of proteins within the dataset under investigation.

The conundrum of leave-one-taxon-out

The leave-one-out cross-validation in machine learning involves the removal of randomly chosen instances, one at a time, from the training set and using them for testing the predictor. For multi-taxon datasets as in the present case, one taxon can be left out, that is, all instances from a given species. This approach was used previously for evaluating sub-cellular localization predictions (e.g., Lu *et al.*, 2004), but has not been exploited so far for protein function prediction. In fact, the leave-one-taxon-out method simulates the ultimate task of a classifier—to predict the function of hypothetical proteins derived from a newly sequenced genome.

Obviously, when the selected organism to be left out has numerous closely related relatives in the dataset, classifier performance will be very good. To assess to which extent this test is meaningful, we studied the effect of leave-one-taxon-out based on the 'population density' of clades (groups of organisms related by descent). Clade circumscriptions were obtained from published phylogenies (for a recent review, see Keeling *et al.*, 2005) and organisms were selected from large, medium and small clades. Note that 'population density' here relates to the number of organisms in a clade for which mitochondrion-encoded protein data are available and not to the total number of recognized species in a clade (see Supplementary information Tables S2, S3 for the taxonomic distribution of instances in the dataset).

The leave-one-taxon-out procedure was conducted for several clades of roughly comparable evolutionary diversity, with numbers of member species ranging from 2,030 (*Drosophila*) to 2 (Jakobid flagellates). For a given clade, we removed a single species from DS-1000 and used the residual set for training; then proteins from the removed species were used for testing the classifier. Figure 2 shows that, for the leave-one-taxon-out test, the performance of the predictor varies from as high as 90% for clades with large population density, to below 50% in the poorly sampled jakobids. Therefore, leave-out cross validation should be conducted not only with a densely populated clade as often seen in the literature (e.g., Lu *et al.*, 2004), but also with poorly and moderately populated clades, in order to avoid over-estimation of predictor performance.

Evaluating function predictions of hypothetical proteins

We chose to assign molecular function to hypothetical proteins using the three classifiers MOPS-1000, MOPS-100 and MOPS-10, because, as we showed above, they have different predictive strength depending on the class size. Yet, evaluating correctness of such predictions is intrinsically impossible. While typical publications on machine learning-based function prediction leave us with untested hypotheses, we have taken a step ahead. We developed domain-specific criteria for assessing individual function predictions, as described in more detail below.

Assessing *in silico* function prediction using domain-specific knowledge

To identify which ORF function assignments are likely to be correct or incorrect, a taxonomically comprehensive review of mitochondrion-encoded proteins across eukaryotes (see GOBASE, 'Gene distribution' http://gobase.bcm.umontreal.ca/searches/compilations.php) yielded helpful insights: (i) in contrast to nuclear genomes, mitochondrial genomes do not contain gene families; and (ii) mtDNAs of closely related species often encode the same set of proteins (Gray *et al.*, 2004).

A combination of three domain-specific criteria assisted in evaluating the correctness of a predicted function, which we describe in the order employed in our 'biological' evaluation scheme (Figure 3). The 'uniqueness criterion' limits a mitochondrial genome to contain only a single gene for each function. For example, a prediction is likely

incorrect ('negative support') if for a given genome the predicted function of an ORF conflicts with a gene known to reside in this mtDNA and that has this function. For remaining predictions, the 'completeness criterion' evaluates whether a complete mitochondrial genome sequence for this species is available or not. The strongest evidence for a correct prediction satisfies both the uniqueness and completeness rules ('strong positive support'). Finally, for ORFs from incompletely sequenced mtDNAs, the 'solidarity criterion' corroborates a prediction if the predicted function is present in a closely related species ('weak positive support'). Predictions that remain after application of these three criteria cannot be evaluated regarding their correctness ('no support').

The effectiveness of domain-specific criteria is demonstrated by their good performance on known proteins: 99.9% of the correct predictions given by C-1000, C-100 and C-10 received positive support. Evaluating all function predictions given by C-1000 using these three criteria, 96% of the predictions received 'positive support' (strong plus weak, see above) with 94% of these being correct and 2% incorrect. Among the predictions obtained with C-100 and C-10, 65% and 56% of those with positive support are correct. A detailed breakdown of the evaluation is available in Table 6.

"Validating" function predictions of hypothetical proteins

Of the 1,336 ORFs, 672 have function predictions with positive support (569 strong positive and 103 weak positive) from at least one of the three MOPS classifiers. Considering each classifier independently, 31-34% of the ORFs have functions with positive support, 14-16% with no support and 50-55% with negative support (Table 7).

Note that only 126 predictions with strong positive support are cross-confirmed by two or more classifiers. But since each classifier has complementary predictive power depending on class size, we expect that function predictions from the individual classifiers will differ.

At first glance, the number of most likely incorrect predictions (negative support) appears considerable. One reason could be class imbalance, a known problem in machine learning where unknown instances are assigned more often to classes overrepresented in the training set than to underrepresented ones (Weiss and Provost, 2001). Another reason may reflect the expectation that some of the mitochondrial ORFs are not expressed ('spurious'), being a relic of gene migration to the nucleus (van den Boogaart *et al.*, 1982) or a product of recent segmental genome duplication and reshuffling, for example as documented in plant mitochondria (Hanson, 1991). If these spurious ORFs still carry remnants of (one or several) functional genes, then this will necessarily lead to functional misidentification, especially when using highly sensitive approaches.

Our analysis defines a core of 126 highly trusted ORF function predictions, assigned consistently by at least two MOPS classifiers and positively supported by the biological evaluation schema (Figure 3). In addition, for over half of the ORFs, the function predicted by one of the classifiers received positive support. These assignments provide powerful working hypotheses, on which the function of an ORF may be confirmed. For instance, computational confirmation could be achieved through multiple sequence alignment of an ORF with proteins known to have the same molecular function, but with varying intermediate degrees of derivation. Such a meticulous exercise was conducted for

ORF150 of an alga (Chesnick *et al.*, 2000) and Var1 of fungi (Bullerwell *et al.*, 2000), based on clues inferred from gene order and other biological considerations. However, these bioinformatics analyses require extensive expert intervention, as they have not been automated.

The ultimate way of validating the prediction of protein function is by biochemical and molecular biology methods, yet, current techniques do not lend themselves to high throughput processing. Therefore, *in silico* predictions are of great value for wet-lab experimentalists, especially if individual predictions are validated by biological criteria as shown above.

Outlook

The annotation system presented here could be enhanced in several aspects. For example, additional protein attributes should be explored, such as gapped dipeptides for localized regions of the proteins that have the propensity to form helices. In this manner, amphipathic helices could be captured and exploited for function prediction.

A second aspect worth improvement is the clustering technique. CD-HIT clusters the sequences based on sequence identity using short-word filters (deca-, penta-peptides etc.,) thereby avoiding time-consuming full pair-wise alignment. A more sensitive approach for comparing two proteins (for clustering) would employ amino acid substitution matrices and would attempt to align the proteins fully, e.g., by using MUSCLE (Edgar, 2004b).

Another issue is that a few functional classes, especially intron maturase and endonuclease, are not well defined as to their molecular function, likely introducing noise in the training process. A thorough classification based on multiple sequence alignments and phylogenetic analysis will be required to better circumscribe these classes.

Finally, we showed that taxonomic information is useful for classifier evaluation, but it could be exploited for molecular function prediction as well. We also contemplate to extend our approach to chloroplast-encoded and ultimately bacterial proteins.

Methods

Dataset

The dataset spans all publicly available mitochondrion-encoded protein sequences as available in the curated database GOBASE (http://megasun.bch.umontreal.ca/gobase/gobase.html). These proteins are conceptual translations mostly from genomic sequences, with a small portion from EST sequences. From GOBASE release 12.0, we retrieved 52,360 complete (non-partial) mitochondrion-encoded protein sequences of known function that do not contain the amino acid 'x' (unknown amino acid, which would disable the calculation of sequence composition). A total of 1,754 ORFs (product name 'hypothetical protein') were retrieved from the GOBASE release 15.0. (At the time of retrieval, the function of all these ORFs were unknown. But recently when these ORFs were searched against the set of function known proteins using BLAST, 26% of them found hits with a BLAST score greater than 65. The

threshold based on our experience; note that the BLAST e-value depends on the database size and therefore is not meaningful for comparison with searches against other databases such as NCBI's non-redundant (nr) database). The 74 functional categories or 'classes' (e.g., cytochrome c oxidase 1) are listed in Table 1; 'unknown' is not used as a class.

Sequence clustering

Identical or similar sequences in the dataset can introduce bias both in the training of a predictor and in performance evaluation. To minimize this bias, known proteins are placed in clusters using CD-HIT (Li *et al.*, 2001) based on sequence identity greater than or equal to a fixed-percentage cutoff threshold. A new dataset is constructed by retaining a single representative from each cluster, namely the longest sequence. For instance, clustering at 99% yielded the 'basic' dataset of known proteins comprising 18,871 sequences. For the various experiments described in Results, clustering was performed at thresholds from 99% down to 50%. The 1,754 ORFs are also clustered at 99%, which resulted in 1,336 ORFs. In the clustered dataset, 27 % of ORFs have BLAST hits with scores greater than 65.

Protein representation / attributes

We experimented with different types of attributes (see Table 2). Physicochemical properties and atomic and residue composition of proteins were calculated by using the ProtParam web server (http://www.ca.expasy.org/tools/protparam.html).

Given 20 different amino acids, there are theoretically 400 different dipeptides. By sliding a window, two positions wide, along the entire protein sequence, all occurring dipeptides are counted. Dipeptide ratio is calculated as number of occurrences of a dipeptide XY / total number of dipeptides in the protein sequence. Ungapped dipeptides are two consecutive amino acids; one-gapped dipeptides are two amino acids in a distance of one residue, etc. Ungapped and gapped dipeptide ratios were calculated directly by using in-house Perl scripts. In order to compare the performance of classifiers constructed with various combinations of protein features, we clustered all the function-known mitochondrial-encoded protein sequences at 99% and 50% sequence identity and used them for training the classifier. The classifier performance was evaluated by ten-fold cross validation (see next section).

Classifier algorithm and performance evaluation

Building the Classifier

For machine-learning based classification, we chose a decision-tree algorithm for two reasons: (i) it infers which attributes are informative for each class (King *et al.*, 2004b), and (ii) it accommodates both attributes with continuous and with categorical values. We used the algorithm C4.5 (Quinlan, 1993b) implemented in the Weka data mining software (Witten and Frank, 2005). Sequences of known molecular function from GOBASE were clustered using CD-HIT at sequence identity thresholds 99%, 75% and 50% denoted DS-99, DS-75 and DS-50, respectively. Using these data sets as training data, we built three classifiers denoted C-1000, C-100 and C-10 were built, respectively.

Classifier performance evaluation

Classifier performance was evaluated by training and testing with proteins of known function. The standard procedure is ten-fold cross validation, which involves subdivision of the data into ten randomly chosen, equally-sized subsets, each of which is used as a test set while the remaining nine serve to build the classifier. This procedure was performed ten times with the overall performance reported as the average across all 100 classifiers. We also employed the 'leave-one-out' cross validation by removing individual instances from the training set and using them for testing only. In both cases, performance was determined by measuring precision (PR), specificity (SP), sensitivity (SE) and accuracy (AC) for each class, i: $PR_i = TP_i / (TP_i + FP_i)$, $SP_i = TN_i / (TN_i + FP_i)$, $SE_i = TP_i / (TP_i + FN_i)$ and $AC_i = TP_i / (TP_i + FP_i + TN_i + FN_i)$ where TP is the number of true positives, FP is false positives, TN is true negatives, and FN is false negatives. Here, we report performance over all classes, n, as a weighted average of the number of instances in each class, Ki. For instance, the weighted average for precision is $PR = \sum_{i=1...n} PR_i / K_i$. Finally, C4.5 assigns a confidence factor to each unknown prediction, based on the rules describing a given class a prediction satisfies.

Authors' contributions

SK designed, developed and implemented MOPS and drafted the manuscript. AMH carried out the class imbalance analysis and participated in the implementation of MOPS. GB conceived the study, participated in its design and in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Veronique Marie, Eric Wang and Emmet O'Brien for help in accessing the GOBASE database, Mathieu Courcelle for assistance in programming, and Balazs Kegl (Computer science department) for discussions. SK and AMH are Canadian Institute for Health Research (CIHR) Strategic Training Fellows in Bioinformatics (Institute for Genetics grant STG-63292). This work was supported by a grant from the CIHR (Institute for Genetics grant MOP-15331). The Canadian Institute for Advanced Research (CIFAR) is acknowledged for travel and interaction support provided to GB.

References

Al-Shahib, A., Breitling, R., and Gilbert, D. (2005) Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl Bioinformatics* **4:** 195-203.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *J Mol Biol* **215:** 403-410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* **25:** 25-29.

Bullerwell, C.E., Burger, G., and Lang, B.F. (2000) A novel motif for identifying Rps3 homologs in fungal mitochondrial genomes. *Trends Biochem Sci* **25:** 363-365.

Chesnick, J.M., Goff, M., Graham, J., Ocampo, C., Lang, B.F., Seif, E., and Burger, G. (2000) The mitochondrial genome of the stramenopile alga *Chrysodidymus synuroideus*. Complete sequence, gene content and genome organization. *Nucl. Acids Res.* **28:** 2512-2518.

Clare, A., Karwath, A., Ougham, H., and King, R.D. (2006) Functional bioinformatics for *Arabidopsis thaliana*. *Bioinformatics* **22:** 1130-1136.

Dobson, P., Cai, Y., Stapley, B., and Doig, A. (2004) Prediction of protein function in the absence of significant sequence similarity. *Curr Med Chem* **11:** 2135-2142.

Drummond, C. and Holte, R.C. (2003) C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. In *Workshop on Learning from Imbalanced Data Sets II*, Washington, DC, USA.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32:** 1792-1797.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. et al. (2006) Pfam: clans, web tools and services. *Nucl. Acids Res.* **34:** D247-251.

Gray, M.W., Lang, B.F., and Burger, G. (2004) Mitochondria of Protists. *Annu Rev Genet* **38:** 477-524.

Hanson, M.R. (1991) Plant mitochondrial mutations and male sterility. *Ann Rev Genet* **25:** 461-486.

Keeling, P.J., Burger, G., Durnford, D.G., Lang, B.F., Lee, R.W., Pearlman, R.E., Roger, A.J., and Gray, M.W. (2005) The tree of eukaryotes. *Trends Ecol Evol* **20:** 670-676.

King, R.D., Karwath, A., Clare, A., and Dehaspe, L. (2000) Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis* and *Escherichia coli* genomes using data mining. *Yeast* **17:** 283-293.

King, R.D., Karwath, A., Clare, A., and Dehaspe, L. (2001) The utility of different representations of protein sequence for predicting functional class. *Bioinformatics* **17:** 445-454.

King, R.D., Wise, P.H., and Clare, A. (2004) Confirmation of data mining based predictions of protein function. *Bioinformatics* **20:** 1110-1118.

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

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

O'Brien, E.A., Zhang, Y., Yang, L., Wang, E., Marie, V., Lang, B.F., and Burger, G. (2006) GOBASE--a database of organelle and bacterial genome information. *Nucl. Acids Res.* **34:** D697-699.

Park, K.-J. and Kanehisa, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* **19:** 1656-1663.

Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann publishers, San Francisco.

Szilagyi, A., Grimm, V., Arakaki, A.K., and Skolnick, J. (2005) Prediction of physical protein-protein interactions. *Phys Biol* **2:** S1-16.

van den Boogaart, P., Samallo, J., and Agsteribbe, E. (1982) Similar genes for a mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of *Neurospora crassa*. *Nature* **298:** 187-189.

Vinayagam, A., König, R., Moormann, J., Schubert, F., Eils, R., Glatting, K.-H., and Suhai, S. (2004) Applying Support Vector Machines for Gene ontology based gene function prediction. *BMC Bioinformatics* **5**: 116.

Weiss, G.M. and Provost, F. (2001) The effect of class distribution on classifier learning: An empricial study. Dept. of Computer Science, Rutgers University.

Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann publishers, San Francisco.

FIGURES

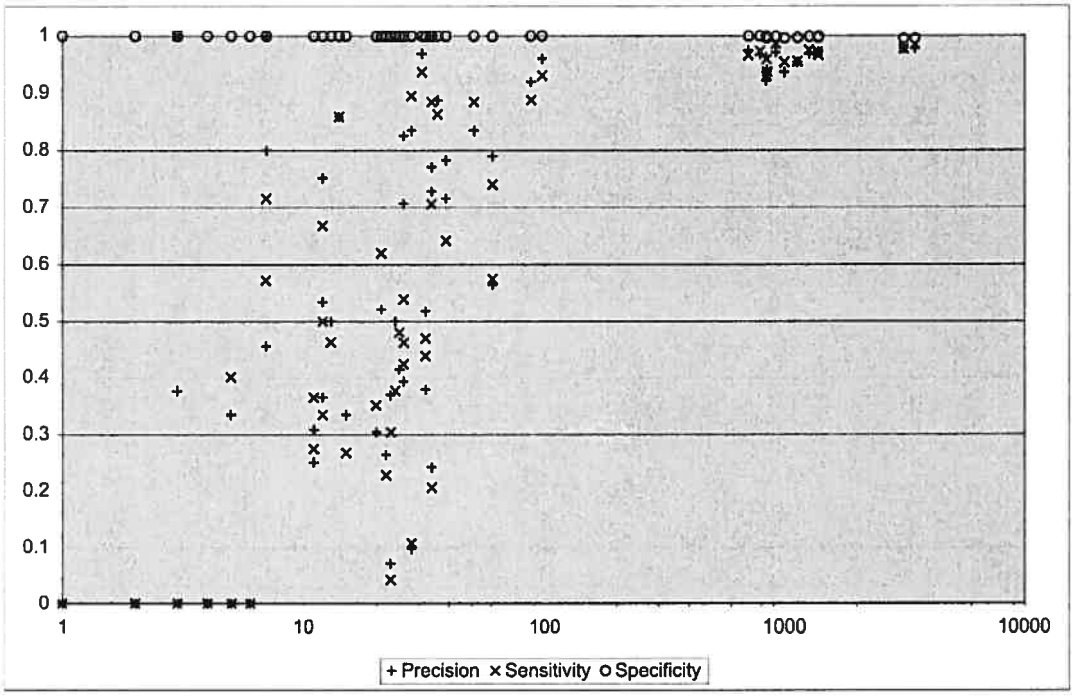

Figure 1. The effect of the number of instances per functional class (horizontal axis in logarithmic scale) on the prediction performance for the class (vertical axis). The sequences are clustered at 99% sequence identity threshold.

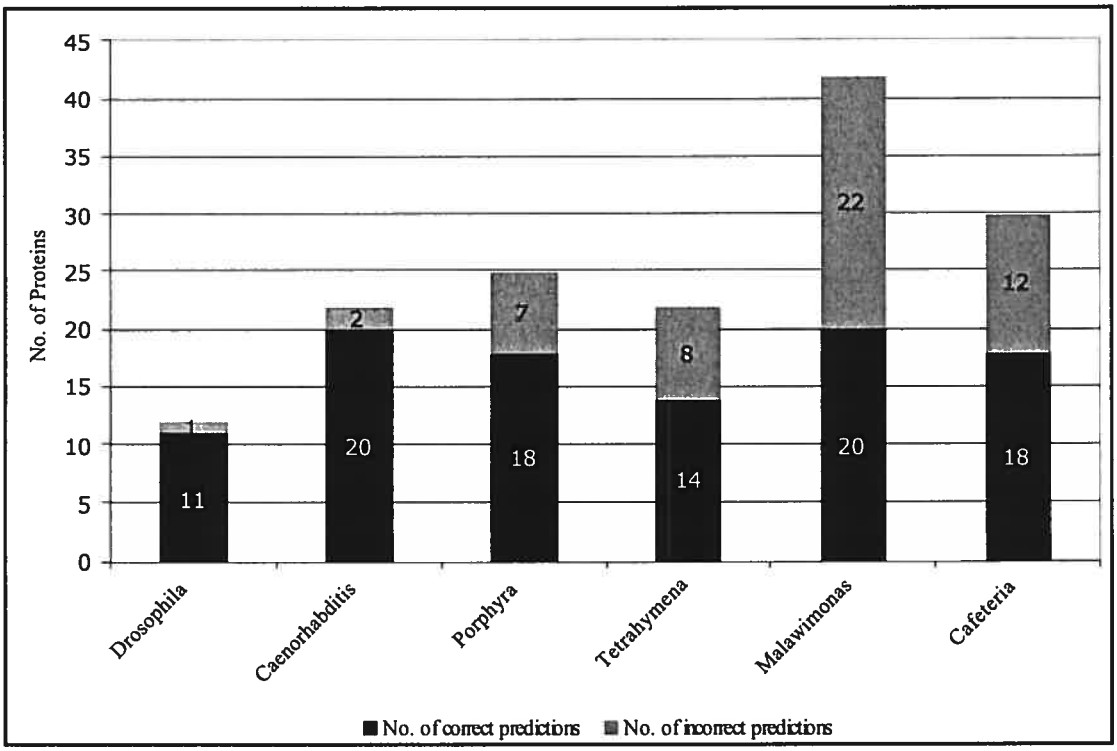

Figure 2. Leave-one-taxon-out validation. All sequences from a particular taxon are removed from the training set and used as test set. The species removed are (the name of the clade and the number of species contained in it are shown in parentheses): *Drosophila melanogaster* (Diptera; 2030), *Caenorhabditis elegans* (Rhabditida; 97), *Porphyra purpurea* (Bangiales; 54), *Tetrahymena pyriformis* (Hymenostomatida; 2), *Malawimonas jakobiformis* (Malawimonadidae; 2), *Cafeteria roenbergensis* (Bicosoecida; 1).

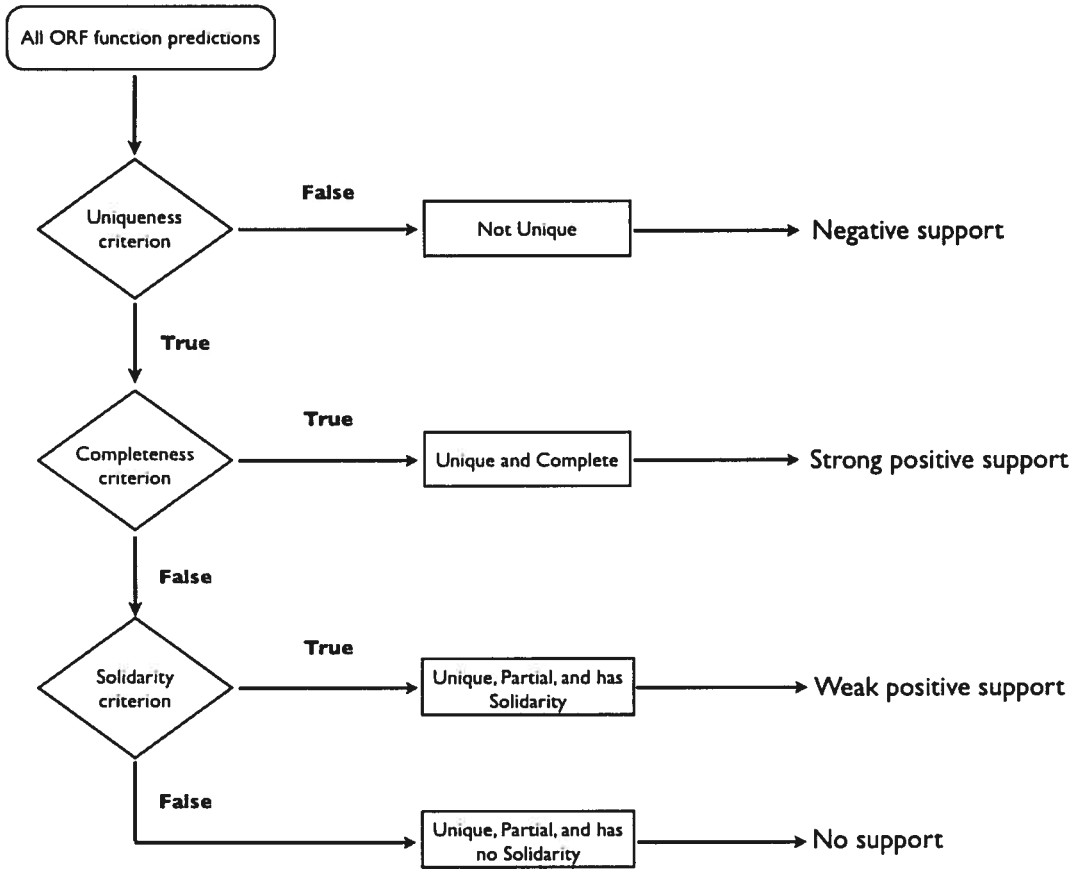

Figure 3. Evaluation of MOPS function predictions based on domain-specific knowledge. Uniqueness criterion: the predicted function is not reported for the genome of the same species. Completeness criterion: a complete genome is available for the species. Solidarity criterion: a gene with the predicted function is present in the genome of the neighboring species.

Tables

Table 1. List of functional classes for mitochondrial proteins

Functional classes
ABC transporter subunits: ATP-binding, channel, C
Apocytochrome b
ATP synthase F0 subunits: 8, a- c
ATP synthase F1 subunits: alpha, gamma
Cytochrome c oxidase subunits: 1-3
DNA adenine methyltransferase
DNA polymerase
Elongation Factor Tu
Endonuclease
Intron maturase
Large subunit ribosomal proteins: L1, L2, L5, L6, L10, L11, L14, L16, L18-L20, L27, L31, L32, L34, L36
mutS-like protein
NADH dehydrogenase subunits: 1-4, 4L, 5-11
Protein involved in haem biosynthesis, haem lyase
Reverse transcriptase
RNA polymerase subunits: alpha, beta, beta'
RNA polymerase (T3/T7 type)
Sec-independent protein translocase components: TatA, TatC
SecY-type transporter protein
Sigma-like factor
Small subunit ribosomal proteins: S1-S4, S7, S8, S10-S14, S16, S19
Succinate dehydrogenase subunits: 2-4

Table 2. Features used to represent protein sequences

Attributes	Description	No. of Attributes
Physicochemical		
Sequence Length	No. of amino acids	1
Molecular Weight	Computed molecular weight	1
Theoretical pI	Theoretical isoelectric point (pI)	1
Aliphatic Index	Computed aliphatic index	1
Hydropathy Index	Grand average of hydropathy index (GRAVY)	1
Composition		
Atomic Composition	Composition of elements: C, H, O, S	5
Residue Composition	Individual amino acid content	20
Dipeptide Ratios		
Ungapped dipeptide ratio	Residue pair at positions N, N+1	400
1-gapped dipeptide ratio	Residue pair at positions N, N+2	400
2-gapped dipeptide ratio	Residue pair at positions N, N+3	400
3-gapped dipeptide ratio	Residue pair at positions N, N+4	400

Table 3. Effect[a] of gapped and ungapped dipeptide attributes on classifier performance

	Basic[b]	Basic + Ungapped dipeptides	Basic + one-gapped dipeptides	Basic + two-gapped dipeptides	Basic + three-gapped dipeptides	Basic + four-gapped dipeptides
99%[c]						
Precision	0.941	0.944	0.938	0.940	0.938	0.942
Specificity	0.997	0.997	0.996	0.997	0.996	0.997
Sensitivity	0.942	0.945	0.939	0.943	0.940	0.944
50%[c]						
Precision	0.627	0.634	0.619	0.607	0.612	0.630
Specificity	0.986	0.986	0.986	0.985	0.985	0.985
Sensitivity	0.632	0.638	0.623	0.610	0.625	0.639

[a] Evaluated by 10-fold cross validation.

[b] Combination of physiochemical features, atomic and residual composition.

[c] Set of known proteins clustered at different sequence identity thresholds.

Table 4. Classifier performance[a] on known data clustered at different sequence identity thresholds[b]

Clustering Threshold	Instances in Dataset[c]		Precision	Specificity	Sensitivity
99%	18,871	(3,476 / 1)	0.944	0.997	0.945
75%	4,743	(580 / 1)	0.819	0.993	0.823
50%	1,932	(222 / 1)	0.634	0.986	0.638

[a] Evaluated by 10-fold cross validation. Attributes used: physicochemical properties, residue and atomic composition, ungapped dipeptides.

[b] More complete table is provided in the supplementary information (Table S4).

[c] Maximum and minimum class size in the dataset is given in parentheses.

Table 5. Effect of class imbalance on classifier performance

		All classes	Class size > 30	Class size[a] ≤ 30
C-1000				
	Precision	0.944	0.958	0.419
	Specificity	0.997	0.997	0.999
	Sensitivity	0.945	0.960	0.389
C-100				
	Precision	0.651	0.705	0.485
	Specificity	0.989	0.987	0.996
	Sensitivity	0.659	0.710	0.496
C-10				
	Precision	0.567	0.662	0.553
	Specificity	0.991	0.991	0.991
	Sensitivity	0.566	0.673	0.550

[a] Number of instances per class in the training set is identical for all the three classifiers

Table 6. Domain-knowledge-based evaluation of MOPS predictions on the function-known proteins (treated as unknown)

Support	C-1000		C-100		C-10	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Strong positive	4,127	133	788	132	401	119
Weak positive	13,693	123	724	41	325	38
No	3	120	0	106	0	83
Negative	0	672	0	535	0	338
Total	17,823	1,048	1,512	814	726	578

Table 7. Domain-knowledge-based evaluation of MOPS predictions on ORFs

Support	MOPS-1000 Predictions	MOPS-100 Predictions	MOPS-10 Predictions
Strong Positive	342	344	391
Weak Positive	70	65	67
No	193	183	209
Negative	731	744	669
Total	1,336	1,336	1,336

A total of 672 unique proteins were predicted with positive support.

Supplementary Information

Figures

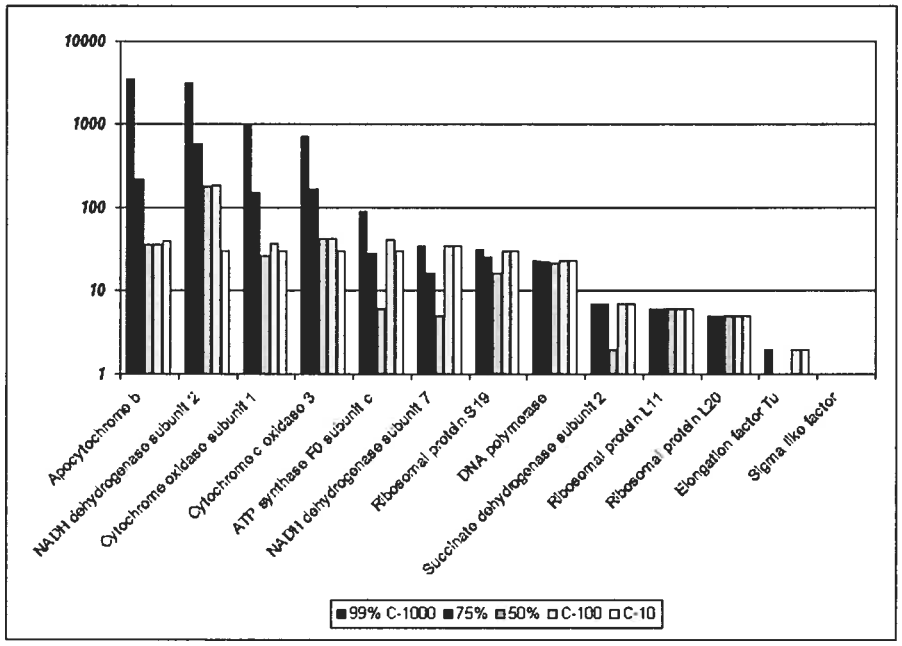

Figure S1. Number of instances per class (vertical axis) in datasets with different degrees of class imbalance. Only 13 of the 74 functional classes are shown (horizontal axis).

Tables

Table S1. Number of instances per protein functional class in the DS-1000 dataset (sequences clustered at 99% sequence identity)

Protein Functional class	No. of instances
ABC transporter ATP-binding subunit	3
ABC transporter channel subunit	12
ABC transporter subunit C	12
Apocytochrome b	3,476
ATP synthase F0 subunit 8	1,388
ATP synthase F0 subunit a	1,135
ATP synthase F0 subunit b	26
ATP synthase F0 subunit c	88
ATP synthase F1 subunit alpha	34
ATP synthase F1 subunit gamma	3
Cytochrome c oxidase subunit 1	1,002
Cytochrome c oxidase subunit 2	1,382
Cytochrome c oxidase subunit 3	712
DNA adenine methyltransferase	1
DNA polymerase	23
Elongation factor Tu	2
Endonuclease	34
Haem lyase	26
Intron maturase	98
mutS-like protein	4
NADH dehydrogenase subunit 1	1,265
NADH dehydrogenase subunit 2	3,123
NADH dehydrogenase subunit 3	920
NADH dehydrogenase subunit 4	842

Protein Functional class	No. of instances
NADH dehydrogenase subunit 4L	789
NADH dehydrogenase subunit 5	843
NADH dehydrogenase subunit 6	848
NADH dehydrogenase subunit 7	34
NADH dehydrogenase subunit 8	7
NADH dehydrogenase subunit 9	36
NADH dehydrogenase subunit 10	7
NADH dehydrogenase subunit 11	14
Protein involved in haem biosynthesis	3
Reverse transcriptase	11
Ribosomal protein L1	2
Ribosomal protein L2	28
Ribosomal protein L5	32
Ribosomal protein L6	22
Ribosomal protein L10	2
Ribosomal protein L11	6
Ribosomal protein L14	26
Ribosomal protein L16	39
Ribosomal protein L18	2
Ribosomal protein L19	2
Ribosomal protein L20	5
Ribosomal protein L27	2
Ribosomal protein L31	5
Ribosomal protein L32	1
Ribosomal protein L34	2
Ribosomal protein L36	1
Ribosomal protein S1	11

Protein Functional class	No. of instances
Ribosomal protein S2	28
Ribosomal protein S3	64
Ribosomal protein S4	32
Ribosomal protein S7	25
Ribosomal protein S8	20
Ribosomal protein S10	21
Ribosomal protein S11	24
Ribosomal protein S12	51
Ribosomal protein S13	64
Ribosomal protein S14	39
Ribosomal protein S16	1
Ribosomal protein S19	31
RNA polymerase subunit alpha	1
RNA polymerase subunit beta'	2
RNA polymerase subunit beta	2
RNA polymerase (T3/T7 type)	15
Sec-independent protein translocase component TatA	1
Sec-independent protein translocase component TatC	23
SecY-type transporter protein	2
Sigma-like factor	1
Succinate dehydrogenase subunit 2	7
Succinate dehydrogenase subunit 3	12
Succinate dehydrogenase subunit 4	13

Table S2. Taxonomic distribution of training data at different sequence identity clustering thresholds

Clustering threshold	Number of Sequences									
	Total	I*	II	III	IV	V	VI	VII	VIII	IX
99%	18,871	16,378	706	1,004	307	100	54	1	170	151
95%	12,997	10,827	643	780	292	85	51	1	169	149
90%	9,561	7,606	595	643	273	77	48	1	169	149
85%	7,156	5,396	518	563	251	67	46	1	167	147
80%	5,822	4,176	470	514	246	63	42	1	164	146
75%	4,743	3,223	424	472	232	56	36	1	156	143
70%	3,846	2461	380	425	209	54	31	1	150	135
65%	3,260	1,970	345	399	199	53	25	1	140	128
60%	2,700	1,524	313	358	181	48	23	1	128	124
55%	2,283	1,219	279	328	161	45	22	1	117	111
50%	1,932	973	235	302	144	44	22	0	110	102

* Taxonomic group names listed below

I Animals

II Unikonts without animals

III Plantae

IV Chromalveolates

V Alveolates

VI Euglenozoa

VII Rhizaria

VIII Excavates

IX Other

Table S3. Taxonomic distribution in the largest functional class (apocytochrome b; DS-1000)

Taxonomic Group	No. of Instances
Animals	3,303
Unikonts without Animals	101
Plantae	36
Chromalveolates	9
Alveolates	14
Rhizaria	0
Euglenozoa	5
Excavates	3
Other	5

Table S4. Classifier performance[a] using training data clustered with different sequence identity thresholds

Clustering threshold	No. of instances in dataset[b]		Precision	Specificity	Sensitivity
99%	18,871	(3,476 / 1)	0.944	0.997	0.945
95%	12,997	(2,134 / 1)	0.923	0.996	0.925
90%	9,561	(1,541 / 1)	0.898	0.995	0.899
85%	7,156	(1,064 / 1)	0.872	0.994	0.874
80%	5,822	(778 / 1)	0.850	0.994	0.855
75%	4,743	(580 / 1)	0.819	0.993	0.823
70%	3,846	(439 / 1)	0.787	0.991	0.792
65%	3,260	(375 / 1)	0.755	0.990	0.759
60%	2,700	(304 / 1)	0.710	0.989	0.719
55%	2,283	(259 / 1)	0.666	0.986	0.672
50%	1,932	(222 / 1)	0.634	0.986	0.638

[a] Evaluated by 10-fold cross validation. Attributes used: physico-chemical properties, residue and atomic composition, ungapped dipeptides.

[b] Maximum and minimum size of the class in the dataset is given in parentheses.

Chapter 3 : ARTICLE

The best way to validate *in silico* function prediction is by experimentation. However, provided that when several homologues of the unknown are available, computational validation may be possible. Kinetoplastid MURF1, predicted by MOPS as NAD2, has seven members and therefore *in silico* validation has been attempted by using the most sensitive methods such as Profile HMM-Profile HMM comparison.

I have carried out the sequence analyses, detailed literature survey of life sciences literature and wrote the manuscript.

Unassigned MURF1 of kinetoplastids codes for NADH dehydrogenase subunit 2

Sivakumar Kannan* and Gertraud Burger

Canadian Institute for Advanced Research

Robert Cedergren Research Center for Bioinformatics and Genomics

Département de Biochimie, Université de Montréal

2900 Boulevard Edouard-Montpetit

Montréal, Québec H3T 1J4, Canada

*Corresponding author

[redacted]

Abstract

In a previous study, we conducted a large-scale similarity-free function prediction of mitochondrion-encoded hypothetical proteins, by which the hypothetical gene *murf1* (maxicircle unidentified reading frame 1) was assigned as *nad2*, encoding subunit 2 of NADH dehydrogenase (Complex I of the respiratory chain). This hypothetical gene occurs in the mitochondrial genome of kinetoplastids, a group of unicellular eukaryotes including the causative agents of African sleeping sickness and leishmaniasis. In the present study, we test this assignment by using bioinformatics methods that are highly sensitive in identifying remote homologs and confront the prediction with available biological knowledge.

Comparison of MURF1 profile Hidden Markov Model (HMM) against function-known profile HMMs in Pfam, Panther and TIGR shows that MURF1 is a Complex I protein, but without specifying the exact subunit. Therefore, we constructed profile HMMs for each individual subunit, using all available sequences clustered at various identity thresholds. HMM-HMM comparison of these of individual NADH subunits against MURF1 clearly identifies this hypothetical protein as NAD2. Further, we collected the relevant experimental information about kinetoplastids, which provides additional evidence for the *in silico* assignment of MURF1 being a highly divergent member of NAD2.

Introduction

The single-celled flagellated eukaryotes of the group kinetoplastids include notorious human pathogens such as *Trypanosoma* and *Leishmania.* Mitochondrial (mt) genomes of numerous trypanosomatids have been sequenced, with complete and nearly complete mtDNA sequences available for five species: *Leishmania tarentolae* (GenBank Accession No: NC_000894), *Trypanosoma brucei* (M94286), *T. cruzi* (DQ343645), *Crithidia oncopelti* (X56015), *Leptomonas seymouri* (DQ239758), and major portions of mtDNA for two other members of the group: *Leishmania major* (AH015294), *Leptomonas collosoma* (AH015822). For a review, see (Feagin, 2000).

The unassigned open reading frame (ORF) *murf1* in *T. brucei* mtDNA has been known for 25 years (Eperon *et al.*, 1983), but until today, there is no protein of known function that shares significant sequence similarity with this ORF. In a recent study, we conducted a comprehensive function prediction of all hypothetical mitochondrion-encoded proteins using the machine-learning-based classifier MOPS (S.Kannan, AM.Hauth, G.Burger, under review). This classifier does not rely on sequence similarity but rather on a host of other features including physico-chemical properties of proteins, and hence should be able to detect remote homologs. MOPS predicted MURF1 of the kinetoplastid *Phytomonas serpens* as subunit 2 (NAD2) of the NADH dehydrogenase Complex (NADHdh), but only with moderate support. We chose to scrutinize this function assignment in detail, motivated by several reasons: the long-standing controversy

surrounding MURF1, the large available body of related biological knowledge, and the significance of this organismal group for human health.

Results

As mentioned in the Introduction, the hypothetical protein MURF1 was predicted by the automated similarity-free classifier MOPS to be a divergent NADHdh subunit 2 (NAD2). To test this prediction, we conducted the following analyses.

Sequence - Sequence Comparison

BLAST searches of *Phytomonas* MURF1 sequence against NRDB or UniProt did not result in any informative hits, but identified all the MURF1 homologs from other kinetoplastids such as *T. brucei, L. tarentolae,* etc. In contrast, FASTA searches against UniProt returned, after MURF1 homologs, NADHdh subunit 5 from the kinetoplastid *Crithidia* as top informative hit with an e-value of $6.5e^{-09}$, followed by NAD2 from the red alga *Chondrus crispus* with an e-value of $8.8e^{-07}$. A list of all hits and their corresponding e-values is compiled in Table 1.

Profile - Sequence Comparison

For the identification of distantly related sequences, methods that exploit profiles (i.e., position-specific descriptions of the consensus of a multiple sequence alignment) are more sensitive than those based on pairwise alignment such as BLAST and FASTA. Here,

we used PSI-BLAST to generate a MURF1 profile and searched it against NRDB, but no other proteins beyond kinetoplastids MURF1 were found.

Profile HMM - Profile HMM Comparison

In contrast to simple sequence profiles, Profile Hidden Markov Models (HMMs) contain extra information about insertions/deletions and gap scores. Profile HMM – Profile HMM comparison is more sensitive than the profile – sequence comparison in identifying distant homologs. HHsearch (the first implementation of this approach), was shown to outperform profile – sequence comparison methods such as PSI-BLAST and HMMER, profile - profile comparison tools such as PROF_SIM and COMPASS and the other HMM - HMM comparison tool PRC (Soding, 2005).

We built a profile HMM for MURF1 from the multiple alignment of seven kinetoplastids MURF1 sequences. Using HHsearch, we searched this profile HMM against the profile HMMs available in Pfam, PANTHER, COG and TIGR. In most cases, the top hit was the "NADH-Ubiquinone/plastoquinone (Complex I)", which is a multi-subunit protein complex. Only the search against the COG database returned a specific subunit as top hit, i.e., NAD2. HHsearch results are summarized in Table 2.

To narrow down the exact function of MURF1, we generated profile HMMs for all 12 subunits of NADHdh. We clustered the protein sequences of all function-known proteins of NADHdh subunits at different sequence identity thresholds ranging from 40% to 75%, constructed a multiple sequence alignment for each of the subunits at each

threshold, and generated profile HMMs. In order to choose the best profile HMM for each NADHdh subunit, we evaluated these profile HMMs by searching them against a database of all function-known proteins encoded by mitochondria. The best profile HMMs are those that are able to identify all instances of their corresponding subunits with minimum error. Finally, we searched the MURF1 profile HMM against the 12 optimal profiles of NADHdh subunits. The top hit is NAD2 with an e-value of $1e^{-15}$. The e-value of the second best hit is 3 orders of magnitude worse (Table 3).

Discussion

While sequence – sequence comparison and profile HMM – profile HMM comparison point to MURF1 being a subunit of NADHdh, profile – profile comparison against the profile HMMs of individual subunits of NADHdh is able to clearly assign MURF1 to NAD2. In the following, we will confront this *in silico* prediction with the available biological knowledge. If the MURF1 protein of trypanosomes is indeed NAD2, then the following criteria must apply :

1. **There should be no previously annotated *nad2* gene in either mitochondrial or nuclear genomes of kinetoplastids.** A *nad2* gene has not been reported in any mitochondrial genome of kinetoplastids. Recently, the sequence of the nuclear genome became available for the *P. serpens* (Pappas *et al.*, 2005). Neither genome nor EST data (2,190 ESTs) indicate the presence of this gene.

2. **There should be numerous precedents for *nad2* being encoded by mtDNA.** The *nad2* gene is mtDNA-encoded by the large majority of eukaryotes (see GOBASE, 'Gene Distribution' http://gobase.bcm.umontreal.ca/searches/compilations.php). The rare species that lack this mitochondrial gene also lack other NADH subunits (Apicomplexa, yeast).

3. **The *murf1* gene should be transcribed.** Evidence for *murf1* being expressed rather than being a spurious ORF is provided by several observations. First, the deduced amino acid sequence is conserved across trypanosomes, despite considerable divergence at the nucleotide level. Second, transcription of this gene has been demonstrated in *P. serpens* (Maslov *et al.*, 1999).

4. **Rotenone-sensitive NADH dehydrogenase Complex I should be present in kinetoplastids.** The presence of Complex I has been biochemically confirmed in *Trypanosoma* and *Phytomonas* (Fang *et al.*, 2001; González-Halphen and Maslov, 2004).

On all accounts enumerated above, the biological knowledge reinforces the *in silico* prediction. Thus, MURF1 is identified beyond doubt as a highly derived homolog of NAD2. For illustration purpose, Figure. 1 depicts the multiple protein sequence alignment of the most conserved block of known NAD2 proteins and kinetoplastid MURF1 sequences.

Outlook

Notably, a functional NADHdh is crucial to the survival of trypanosomes. Under aerobic conditions (procyclic, insect stage), NADHdh is required as a component of the respiratory chain, to catalyze electron transport toward complex IV. The thus generated proton gradient is utilized for ATP synthesis. Under anaerobic conditions (bloodstream form), a functional NADHdh is equally essential. In the blood stream of mammals, NADHdh provides electrons for the alternative oxidase, a pathway required for maintaining the balance of NADH/NAD+ in the cell. This confirms that trypanosomes depend on a functional NADHdh. In fact, Atovaquone, an anti-leishmanial drug, inhibits the NADHdh activity in *P. serpens* and this inhibition was suggested to underlie the anti-leishmanial activity of that drug (González-Halphen and Maslov, 2004). In this context, the identification of MURF1 as a divergent NAD2 could offer new avenues to the prevention or treatment of trypanosomatid-caused diseases.

Methods

Dataset

All function-known protein sequences used in this study were retrieved from the organelle genome database GOBASE release 12.0 (O'Brien *et al.*, 2006a). The homologs for MURF1 were retrieved from Entrez (Ostell, 2002), and their accession numbers are given in Table S1.

Sequence - sequence comparison

A BLAST (blastp) (Altschul *et al.*, 1990c) search was conducted for the MURF1 protein sequence against NCBI's NRDB (non-redundant protein database) (October, 2006; 4,565,699 sequences), with default parameters. In addition, a FASTA (Pearson, 1990) search was conducted for the MURF 1 protein sequence against UniProt (release 10.4) with default parameters, at the EBI website (http://www.ebi.ac.uk/fasta33).

Profile - sequence comparison

This comparison was conducted in two different ways. First, PSI-BLAST (Altschul *et al.*, 1997) was employed to search MURF1 remotely against NCBI's NRDB, with four iterations. Second, we per performed profile HMM - sequence comparison using profiles from Pfam version 21.0 (Finn *et al.*, 2006a), executed at the Pfam website (http://www.sanger.ac.uk/Software/Pfam).

Profile HMM - profile HMM comparison

For Profile HMM - profile HMM comparison, we used HHsearch of the HHpred package (Soding, 2005), which takes the MURF1 sequence as input and searches against NRDB using PSI-BLAST. The MURF1 homologs obtained from the PSI-BLAST search are then used to generate a profile HMM. As a next step, this MURF1 profile HMM is searched against all profile HMMs of function-known proteins available from the public databases Pfam, PANTHER, SMART, COG, PDB and SCOP. In addition, we generated our own profile HMMs for each of the 12 NADHdh subunits (1-11 and 4L) from all known

sequences of these protein classes. These sequences were clustered at different identity thresholds using CD-HIT (Li and Godzik, 2006), followed by multiple sequence alignment performed with MUSCLE (Edgar, 2004a). The multiple alignment served as input for generating profiles using hmmbuild of HMMER version 2.3.2, 2003 package (Eddy, 1998). The efficiency of the profile HMMs was assessed by searching the profile HMMs against the known NADH dehydrogenase subunits and calculating how many known sequences can by identified by these profile HMMs. Finally, the MURF1 profile HMM was searched against all these profiles using HHsearch.

Acknowledgments

We thank Yaoqing Shen for critically reading the manuscript. SK is Canadian Institute for Health Research (CIHR) Strategic Training Fellow in Bioinformatics (Genetics Institute grant STG-63292). This work was supported by grants from the CIHR Genetics Institute (grants MOP-15331 and MOP-79303). The Canadian Institute for Advanced Research (CIAR) is acknowledged for travel and interaction support provided to GB.

References

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25:** 3389-3402.

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

Edgar, R. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32:** 1792-1797.

Eperon, I., Janssen, J., Hoeijmakers, J., and Borst, P. (1983) The major transcripts of the kinetoplast DNA of *Trypanosoma brucei* are very small ribosomal RNAs. *Nucl. Acids Res.* **11:** 105-125.

Fang, J., Wang, Y., and Beattie, D. (2001) Isolation and characterization of complex I, rotenone-sensitive NADH: ubiquinone oxidoreductase, from the procyclic forms of *Trypanosoma brucei. Eur J Biochem* **268:** 3075-3082.

Feagin, J. (2000) Mitochondrial genome diversity in parasites. *Int J Parasitol* **30:** 371-390.

Finn, R., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* **34:** D247-251.

González-Halphen, D. and Maslov, D. (2004) NADH-ubiquinone oxidoreductase activity in the kinetoplasts of the plant trypanosomatid *Phytomonas serpens. Parasitol Res* **92:** 341-346.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658-1659.

Maslov, D., Nawathean, P., and Scheel, J. (1999) Partial kinetoplast-mitochondrial gene organization and expression in the respiratory deficient plant trypanosomatid *Phytomonas serpens. Mol Biochem Parasitol* **99:** 207-221.

O'Brien, E., Zhang, Y., Yang, L., Wang, E., Marie, V., Lang, B., and Burger, G. (2006) GOBASE--a database of organelle and bacterial genome information. *Nucl. Acids Res.* **34:** D697-699.

Ostell, J. (2002) The Entrez search and retrieval system. In *The NCBI Handbook.*

Pappas, G., Benabdellah, K., Zingales, B., and González, A. (2005) Expressed sequence tags from the plant trypanosomatid *Phytomonas serpens*. *Mol Biochem Parasitol* **142:** 149-157.

Pearson, W. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth Enzymol* **183:** 63-98.

Soding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21:** 951-960.

Figures

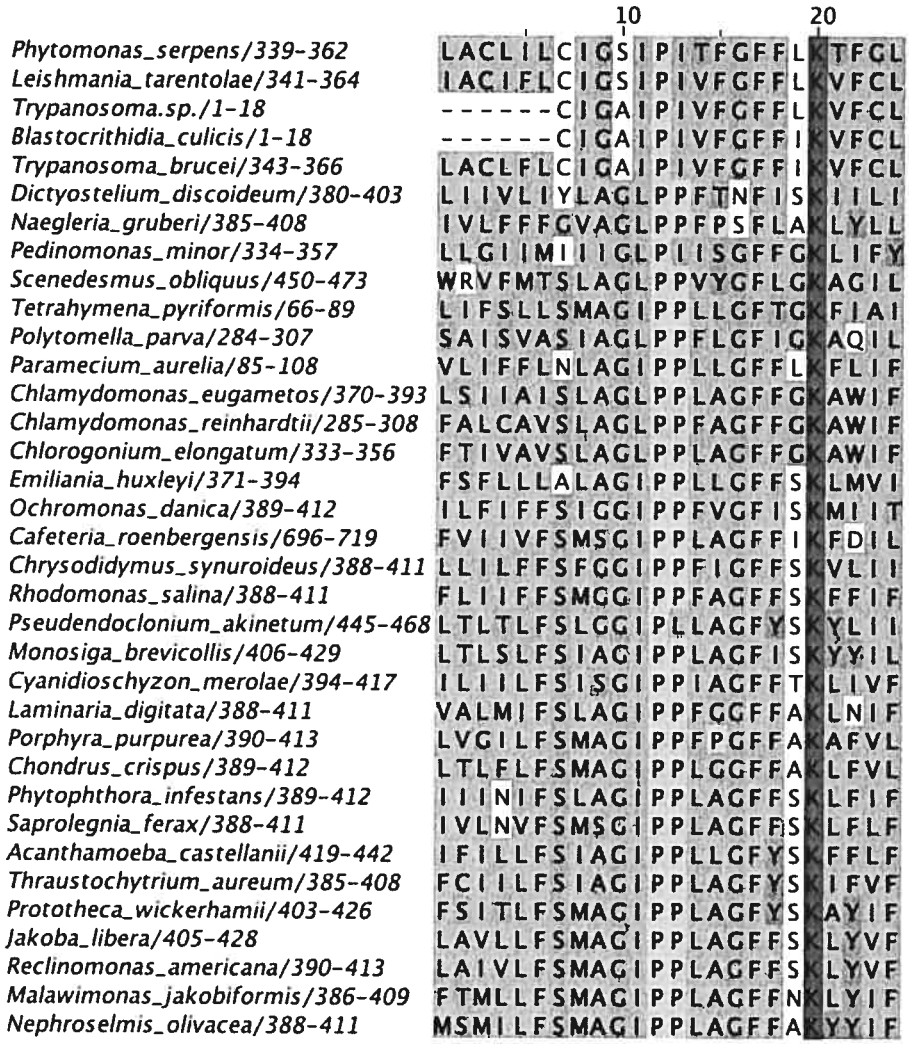

Figure 1. Multiple sequence alignment of kinetoplastid MURF1 sequences (first five sequences in the alignment above) with NAD2 sequences from other eukaryotes.

TABLES

Table 1. List of FASTA hits for *P. serpens* MURF1 searched against UniProt

UniProt ID	Species Name	Protein Name	e-value	Similarity
Q9XKY50	*Phytomonas serpens*	MURF1	$5.3e^{-148}$	100.0%
Q33559	*Leishmania tarentolae*	MURF1	$2.7e^{-109}$	90.9%
Q8HE85	*Trypanosoma sp.*	MURF1	$3.1e^{-17}$	87.6%
Q33547	*Blastocrithidia culicis*	MURF1	$1.2e^{-16}$	86.5%
Q33552	*Crithidia fasciculata*	MURF1	$6e^{-16}$	88.4%
Q33556	*Herpetomonas muscarum*	MURF1	$5.1e^{-13}$	85.0%
Q34937	*Leishmania. tarentolae*	MURF2	$2e^{-09}$	60.4%
Q34096	*Crithidia fasciculata*	MURF2	$3.2e^{-09}$	56.3%
Q34192	*Crithidia oncopelti*	NAD5	$3.8e^{-09}$	54.5%
P48903	*Chondrus crispus*	NAD2	$5.4e^{-07}$	57.9%
Q5LRX2	*Silicibacter pomeroyi*	Putative membrane protein	$1.2e^{-06}$	58.0%
Q6E773	*Saprolegnia ferax*	NAD2	$1.5e^{-06}$	53.8%
Q6SKY5	*Speleonectes tulumensis*	NAD5	$2.3e^{-06}$	55.2%
Q5AG49	*Candida albicans*	Hypothetical protein	$3.3e^{-06}$	67.7%
Q5AGI5	*Candida albicans*	Hypothetical protein	$7.1e^{-06}$	67.9%
Q8SKS6	*Ancylostoma duodenale*	NAD4	$7.4e^{-06}$	57.3%
Q85TH7	*Melipona bicolor*	NAD4	$7.7e^{-06}$	58.1%
Q33575	*Trypanosoma brucei*	NAD4	$8.7e^{-06}$	57.3%
P24499	*Trypanosoma brucei brucei*	ATP6	$1.1e^{-05}$	55.4%
Q70NW4	*Strongyloides stercoralis*	NAD4	$1.2e^{-05}$	56.7%
Q33570	*Trypanosoma cruzi*	ATP6	$1.5e^{-05}$	56.9%
Q5CV17	*Cryptosporidium parvum*	Hypothetical protein	$1.5e^{-05}$	61.5%
Q057W5	*Buchnera aphidicola*	NADH dehydrogenase I chain L	$1.9e^{-05}$	54.9%
Q8IBJ6	*Plasmodium falciparum*	Hypothetical protein	$2.9e^{-05}$	58.4%

Table 2. Best informative hits for the MURF1 profile HMM when searched against profile HMMs from various databases

	Best informative hit	**e-value**	**Identity**	**Probability**
Pfam	NADH-Ubiquinone/plastoquinone (Complex I), various subunits	1.6e-08	21%	96.80
PANTHER	NADH dehydrogenase	4.3e-09	16%	99.20
COG	NADH:Ubiquinone oxidoreductase subunit 2	3.8e-03	19%	39.65
TIGR	NDH_I_N Proton-translocating NADH-Quinone oxidoreductase	91	19%	75.95

Table 3. Best informative hits for the MURF1 profile HMM when searched against the profile HMMs of all NADH dehydrogenase subunits. The hits are ranked based on e-values

Profile HMMs*	e-value	Identity	Probability
NAD2_0.45	$1.0e^{-15}$	26%	96.6
NAD4_0.4	$7.8e^{-12}$	21%	76.0
NAD6_0.4	$2.4e^{-09}$	24%	91.8
NAD5_0.4	$7.9e^{-09}$	21%	87.8
NAD1_0.5	$4e^{-08}$	18%	87.0
NAD3_0.4	$5.7e^{-06}$	30%	62.1
NAD4L_0.4	$1.6e^{-04}$	25%	34.2
NAD4L_0.4	$1.6e^{-04}$	25%	34.2

*The number following the subunit name is the sequence identity threshold used for clustering the sequences from which we generate the profile HMM. For example, NAD2_0.45 profile HMM is generated by clustering all known NAD2 sequences at 45% sequence identity threshold using CD-HIT.

Supplementary Information

Table S1. List of kinetoplastid MURF1 sequences with GenBank Accession Numbers

Species Name	GenBank Accession
Phytomonas serpens	AAD28358
Leishmania tarentolae	NP_050068
Trypanosoma brucei	E22845
Trypanosoma sp.	AAN86606
Blastocrithidia culicis	AAA73417
Crithidia fasciculata	AAA73421
Herpetomonas muscarum	AAA73415

CONCLUSION

Function prediction remains a central challenge in genomics research, with the gap between the numbers of function-known and unknown sequences widening progressively. Our newly developed function annotation approach complements the existing state-of-the-art methods in closing that gap, since it overcomes species boundaries (pan-taxonomic) and provides expert knowledge-based *in silico* validation of the predicted functions. Further, this work led us to identify the function of a previously unknown protein involved in infectious disease.

Original contributions to Knowledge

This thesis makes the following original contributions to our current knowledge:

1. Function assignment for the Complex III subunits of a primitive eukaryote *Seculomonas ecuadoriensis* using state-of-the-art sequence similarity-based methods.

2. Development of a similarity-free function annotation method MOPS and annotation of all the mitochondrion-encoded proteins of unknown function.

3. Development of a prediction evaluation procedure that is independent of the machine learning classifier and training data by using domain-specific knowledge. More than half of the predicted functions received positive support.

4. Identification of the function of a controversial Kinetoplastid gene MURF1 as NADHdh Subunit 2, providing multiple lines of support from *in silico* analyses and from life sciences literature.

The experience we gained during this study allows us to formulate several recommendations for future work in this area:

- Validation of classifier predictions should be done with methods that are independent of the training data and the classifier. The assumption that the good performance of the classifier on the known data implies good performance on the unknown data is not always true.
- In order to avoid over-estimation of predictor performance, leave-one-taxon-out cross validation should be conducted not only with a densely populated clade but also with poorly and moderately populated clades.
- Class imbalance in the training data should be addressed in a biological meaningful way. Large classes should be undersampled by clustering sequences at different sequence identity thresholds rather than by random removal; oversampling of small classes, which is usually done by duplication, is biologically not sensible.
- Sensitive methods such as Profile HMM – Profile HMM comparison has lot of potential in identifying the distant homologues. However, it requires several homologues of unknown sequences for generating the multiple sequence alignment and Profile HMM.

- Mitochondrial genomes are an ideal data set for machine learning based function annotation due to their reasonable number of well-defined protein functional classes.

Future Developments

My work opens several new areas of research for machine learning based molecular protein function prediction. For example, there are many other factors besides amino acid sequence that are essential for protein function. One of these factors is post-translational modifications (e.g. phosphorylation or glycosylation), yet none of the current methods take modifications into account. It would be worthwhile to explore whether the inclusion of features such as these will enhance function prediction accuracy.

In addition, the role of proteins in physiological and cellular processes rely on temporal and spatial regulation of gene expression, with regulatory signals located not in the coding regions but in the untranslated portions of a gene. The integration of regulatory and structural information in protein function prediction appears to be a promising yet challenging direction for future research.

In this study, we used physicochemical properties and amino acid frequencies to represent the protein sequences. But it would be interesting to test more attributes. For example, 4-gapped dipeptides should capture amphipathic helices, which are present in many mitochondrial proteins. However, the ratio of 4-gapped dipeptides should be calculated only for the corresponding domains rather than for the full-length sequences, in order not to dilute the signal. Further promising attributes to include are predicted

secondary structure such as the number of alpha helices, beta sheets and coils, which are believed to be conserved between distant homologues where the shared sequence similarity is very low. Finally, information about gene neighbors can also be exploited as features, especially since mitochondrial DNAs of many taxa have maintained vestiges of the eubacterial operon structure.

Another avenue for future studies would be to analyse the rules generated by the decision tree classifiers to see whether biological knowledge can be extracted. If the tree is too complex with numerous leaf nodes, feature selection and aggressive pruning can be used to reduce the tree size and hence potentially allow inference of biological insights.

Finally, another area worth exploring is the use of other machine learning methods such as Support Vector Machines (SVM), which are less sensitive to class imbalance but have at least two drawbacks : they are computationally expensive and extraction of biological knowledge is not straightforward.

REFERENCES

Al-Shahib, A., Breitling, R., and Gilbert, D. (2007) Predicting protein function by machine learning on amino acid sequences--a critical evaluation. *BMC Genomics* **8:** 78.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Andersen, J. and Mann, M. (2000) Functional genomics by mass spectrometry. *FEBS Lett* **480:** 25-31.

Aravind, L. and Koonin, E. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *J Mol Biol* **287:** 1023-1040.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Attwood, T. (2000) The quest to deduce protein function from sequence: the role of pattern databases. *The International Journal of Biochemistry & Cell Biology* **32:** 139-155.

Attwood, T. (2002) The PRINTS database: a resource for identification of protein families. *Brief Bioinformatics* **3:** 252-263.

Bader, G. and Hogue, C. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20:** 991-997.

Bilofsky, H. and Burks, C. (1988) The GenBank genetic sequence data bank. *Nucleic Acids Res* **16:** 1861-1863.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., O'Donovan, C., Phan, I. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31:** 365-370.

Brenner, S. (1999) Errors in genome annotation. *Trends Genet* **15:** 132-133.

Carter, P., Liu, J., and Rost, B. (2003) PEP: Predictions for Entire Proteomes. *Nucleic Acids Res* **31:** 410-413.

Clare, A., Karwath, A., Ougham, H., and King, R. (2006) Functional bioinformatics for Arabidopsis thaliana. *Bioinformatics* **22:** 1130-1136.

Clare, A. and King, R. (2003) Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics* **19 Suppl 2:** II42-II49.

des Jardins, M., Karp, P., Krummenacker, M., Lee, T., and Ouzounis, C. (1997) Prediction of enzyme classification from protein sequence without the use of sequence similarity. In *Proc Int Conf Intell Syst Mol Biol.*, pp. 92-99.

Devos, D. and Valencia, A. (2000) Practical limits of function prediction. *Proteins: Structure, Function, and Genetics* **41:** 98-107.

Dobson, P., Cai, Y., Stapley, B., and Doig, A. (2004) Prediction of protein function in the absence of significant sequence similarity. *Curr Med Chem* **11:** 2135-2142.

Dujon, B. (1998) European Functional Analysis Network (EUROFAN) and the functional analysis of the Saccharomyces cerevisiae genome. *Electrophoresis* **19:** 617-624.

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

Encarnación, S., Hernández, M., Martínez-Batallar, G., Contreras, S., Vargas, M.C., and Mora, J. (2005) Comparative proteomics using 2-D gel electrophoresis and mass spectrometry as tools to dissect stimulons and regulons in bacteria with sequenced or partially sequenced genomes. *Biological procedures online* **7:** 117-135.

Enzyme_Commission. (1999) Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). *Eur J Biochem* **264:** 610-650.

Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340:** 245-246.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., and Merrick, J. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496-512.

Galperin, M. and Koonin, E. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol (Gedrukt)* **1:** 55-67.

Giot, L., Bader, J., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y., Ooi, C., Godwin, B., Vitols, E. et al. (2003) A protein interaction map of Drosophila melanogaster. *Science* **302:** 1727-1736.

Godzik, A., Jambon, M., and Friedberg, I. (2007) Computational protein function prediction: Are we making progress? *Cell Mol Life Sci.*

Hannenhalli, S. and Russell, R. (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* **303:** 61-76.

Hawkins, T. and Kihara, D. (2007) Function prediction of uncharacterized proteins. *Journal of bioinformatics and computational biology* **5:** 1-30.

Henikoff, J., Greene, E., Pietrokovski, S., and Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* **28:** 228-230.

Hieter, P. and Boguski, M. (1997) Functional genomics: it's all how you read it. *Science* **278:** 601-602.

Hughey, R. and Krogh, A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci* **12:** 95-107.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P., Pagni, M., and Sigrist, C. (2006) The PROSITE database. *Nucleic Acids Res* **34:** D227-230.

Ichikawa, T., Suzuki, Y., Czaja, I., Schommer, C., Lessnick, A., Schell, J., and Walden, R. (1997) Identification and role of adenylyl cyclase in auxin signalling in higher plants. *Nature* **390:** 698-701.

Iyer, L., Aravind, L., Bork, P., Hofmann, K., Mushegian, A., Zhulin, I., and Koonin, E. (2001) *Quoderat demonstrandum?* The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol* **2:** RESEARCH0051.

John, B. and Sali, A. (2004) Detection of homologous proteins by an intermediate sequence search. *Protein Sci* **13:** 54-62.

Jones, R., Gordus, A., Krall, J., and MacBeath, G. (2006) A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* **439:** 168-174.

Karp, P. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics* **16:** 269-285.

Keller, J., Smith, P., Benach, J., Christendat, D., deTitta, G., and Hunt, J. (2002) The crystal structure of MT0146/CbiT suggests that the putative precorrin-8w decarboxylase is a methyltransferase. *Structure* **10:** 1475-1487.

King, R., Karwath, A., Clare, A., and Dehaspe, L. (2001) The utility of different representations of protein sequence for predicting functional class. *Bioinformatics* **17:** 445-454.

King, R., Wise, P., and Clare, A. (2004) Confirmation of data mining based predictions of protein function. *Bioinformatics* **20:** 1110-1118.

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

Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* **30:** 4321-4328.

Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T., and Eisenberg, D. (1999a) Detecting protein function and protein-protein interactions from genome sequences. *Science* **285:** 751-753.

Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T., and Eisenberg, D. (1999b) A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83-86.

Mewes, H., Albermann, K., Heumann, K., Liebl, S., and Pfeiffer, F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* **25:** 28-30.

O'Farrell, P. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250:** 4007-4021.

Ofran, Y., Punta, M., Schneider, R., and Rost, B. (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov Today* **10:** 1475-1482.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G., and Maltsev, N. (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol (Gedrukt)* **1:** 93-108.

Panchenko, A. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res* **31:** 683-689.

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

Park, J., Teichmann, S., Hubbard, T., and Chothia, C. (1997) Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* **273:** 349-354.

Pearson, H. (2006) Genetics: what is a gene? *Nature* **441:** 398-401.

Pearson, W. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth Enzymol* **183:** 63-98.

Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., and Yeates, T. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *PNAS* **96:** 4285-4288.

Quinlan, J. (1993) C4.5 Programs for Machine Learning. Morgan Kaufmann publishers, San Francisco.

Riley, M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol Rev* **57:** 862-952.

Rison, S., Hodgman, T., and Thornton, J. (2000) Comparison of functional annotation schemes for genomes. *Funct Integr Genomics* **1:** 56-69.

Rodrigues, A.P., Grant, B.J., Godzik, A., and Friedberg, I. (2007) The 2006 automated function prediction meeting. *BMC Bioinformatics* **8 Suppl 4:** S1-4.

Rost, B. (2002) Enzyme function less conserved than anticipated. *J Mol Biol* **318:** 595-608.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. et al. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* **32:** 5539-5545.

Sadreyev, R. and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* **326:** 317-336.

Salamov, A., Suwa, M., Orengo, C., and Swindells, M. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng* **12:** 95-100.

Schena, M., Shalon, D., Davis, R., and Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467-470.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* **32:** D431-433.

Smith, T. and Zhang, X. (1997) The challenges of genome sequence annotation or "the devil is in the details". *Nat Biotechnol* **15:** 1222-1223.

Soding, J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21:** 951-960.

Soldatova, L. and King, R. (2005) Are the current ontologies in biology good ontologies? *Nat Biotechnol* **23:** 1095-1098.

Spiess, C., Beil, A., and Ehrmann, M. (1999) A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. *Cell* **97:** 339-347.

Todd, A., Orengo, C., and Thornton, J. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307:** 1113-1143.

Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403:** 623-627.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417:** 399-403.

Walker, M., Volkmuth, W., Sprinzak, E., Hodgson, D., and Klingler, T. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res* **9:** 1198-1203.

Wei, L., Liu, Y., Dubchak, I., Shon, J., and Park, J. (2002) Comparative genomics approaches to study organism similarities and differences. *Journal of biomedical informatics* **35:** 142-150.

Whisstock, J. and Lesk, A. (2003) Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36:** 307-340.

Wilson, C.A., Kreychman, J., and Gerstein, M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297:** 233-249.

Wistow, G. and Piatigorsky, J. (1987) Recruitment of enzymes as lens structural proteins. *Science* **236:** 1554-1556.

Yona, G. and Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol* **315:** 1257-1275.

Zimmer, C. (2005) How and where did life on Earth arise? *Science* **309:** 89.

Online supplementary information

Chapter 2

Table S5. List of ORF predictions by MOPS classifiers with domain-specific support information

http://megasun.bcm.umontreal.ca/~siva/MOPS_ORF_Predictions.pdf