Methods for the refinement of genome-scale metabolic networks

by

Rodrigo Liberal Fernandes

A thesis submitted for the degree of
Doctor of Philosophy of Imperial College London

Division of Molecular Biosciences
Imperial College London
London SW7 2AZ, England

Typeset in Times by LaTeX

This report is the result of my own work.

No part of this dissertation has already been, or is currently being submitted by the author for any other degree or diploma or other qualification.

This dissertation does not exceed 100,000 words, excluding appendices, bibliography, footnotes, tables and equations. It does not contain more than 150 figures.

This work is supported by a Portuguese Foundation for Science and Technology (FCT) grant and completed in the Division of Molecular Biosciences at Imperial College, London.

All trademarks used in this dissertation are acknowledged to be the property of their respective owners.

Abstract

More accurate metabolic networks of pathogens and parasites are required to support the identification of important enzymes or transporters that could be potential targets for new drugs. The overall aim of this thesis is to contribute towards a new level of quality for metabolic network reconstruction, through the application of several different approaches.

After building a draft metabolic network using an automated method, a large amount of manual curation effort is still necessary before an accurate model can be reached. PathwayBooster, a standalone software package, which I developed in Python, supports the first steps of model curation, providing easy access to enzymatic function information and a visual pathway display to enable the rapid identification of inaccuracies in the model.

A major current problem in model refinement is the identification of genes encoding enzymes which are believed to be present but cannot be found using standard methods. Current searches for enzymes are mainly based on strong sequence similarity to proteins of known function, although in some cases it may be appropriate to consider more distant relatives as candidates for filling these pathway holes. With this objective in mind, a protocol was devised to search a proteome for superfamily relatives of a given enzymatic function, returning candidate enzymes to perform this function.

Another, related approach tackles the problem of misannotation errors in public gene databases and their influence on metabolic models through the propagation of erroneous annotations. I show that the topological properties of metabolic networks contains useful

information about annotation quality and can therefore play a role in methods for gene function assignment.

An evolutionary perspective into functional changes within homologous domains opens up the possibility of integrating information from multiple genomes to support the reconstruction of metabolic models. I have therefore developed a methodology to predict functional change within a gene superfamily phylogeny.

Contents

List of Figures

List of Tables

Acknowledgements

Doing a PhD has required a great deal of hard work, and these past few years would have been much harder if it were not for the help of some people who I would like to thank.

First of all, I would like to thank Francisca for her constant support and for believing that this part of my life would finish one day. And of course, for all the cakes and awesome food with which I have been regularly presented.

To my family for always being present whenever I needed them, and with a special thanks to my parents who managed to ask me less than once a week when I would submit my PhD. I truly believe it must have been hard for them. By the way, I am submitting it now!

To all my friends in the UK and in Portugal for making me feel at home in both places.

To all PDBCs, the PDBC07 in particular, who have walked a parallel path over the past few years. A special thanks to my “paizinho” brother Hugo for his help and support. Hopefully now he will not have to worry about my thesis and can focus on finishing his.

To all the members of the Theoretical Systems Biology group that I have met during my PhD for creating an enjoyable working environment, for all the discussions (mostly Maxime), group meetings, runs, synchronised swimming, badminton and table tennis matches. Special thanks to Beata with whom I have built PathwayBooster.

For all those that have spent some time proof reading this thesis, with a special thanks to Paul. This has greatly improved its readability.

To end this weeping, I would like to thank both my supervisor, John Pinney, and co-supervisor, Michael Stumpf, for accepting me within their groups. In particular, I would like to thank Michael for doing amazing work that motivated me to come to London, and John for his support, patience and scientific enthusiasm that have been invaluable to me.

Chapter 1

Introduction

1.1 Background

Systems biology is an inter-disciplinary field of study that integrates mathematical and computing approaches with biological systems information. Systems Biology focuses on the interaction of the individual components, aiming to understand the system as a whole. One of these components is the group of biochemical reactions and their interactions responsible for the metabolic processes that determine the cell's functions: the metabolic network. Only understanding both genetic and metabolic organisation, we will be able to forecast phenotypic traits caused by alterations of the genome or metabolic network. This may enable us to identify important enzymes or transporters that could be potential targets for new drugs, and support the optimisation of pathways responsible for the consumption/production of certain molecules in biotechnological applications.

This knowledge can be acquired through the modelling of metabolic networks, a process known as metabolic reconstruction. More precisely, metabolic reconstruction is the process of building a map of an organism's metabolic network, using evidence from its genome sequence.

A possible approach for network reconstruction is based on manual curation. An expert in a specific organism uses several available resources together with experimental studies and available literature and then manually inspects all the metabolic network annotations. This includes steps such as literature search, performing lab experiments, etc, in order to find evidence that supports or rejects each of the annotations. This process just by itself is very time consuming. MPMP (Ginsburg (2000)) is an example of a metabolic database that relies heavily on manual reconstruction from a world expert on malaria parasite biochemistry. It uses several publicly available resources and presents the information together with genomic annotations from other databases. Presently, MPMP has the most current and informed pathways for the human malaria parasite *Plasmodium falciparum*. The increasing speed with which we are able to sequence an organism's genome makes bioinformatics an increasingly important source of information.

Currently, automatic function assignments are still mainly performed using sequence similarity methodologies. The use of sequence similarity searches works well as long as an annotation of a closely related organism is already available, but it will not be sensitive enough to detect all the enzymes in many species, as is the case with *Plasmodium*. Enzymes may have functions that are not represented in sequence databases, may be functional analogs to other unrelated proteins catalysing the same reaction, or may simply have diverged too far to be recognisable. Briefly, for more distantly related proteins, where only certain sequence features or structural motifs are conserved, the similarity between two proteins cannot easily be recognised by pairwise alignment methods, nor even by the more sensitive profile based methods (Pinney *et al.* (2007)).

Owing to this difficulty in assigning specific enzymatic functions, initial metabolic reconstruction usually produces networks with many *holes*: reactions essential for a complete biochemical pathway, but for which no enzyme has been annotated in the genome. The existence of a hole in a pathway can be due to several reasons. This can be caused by a fault of the method used. It can be that the gene is not yet or incorrectly annotated. In parasites, another possible explanation might be that the metabolites produced by the reaction are obtained from the host.

A previous study of the human metabolic network identified 203 pathway holes, for which putative genes were found for 25 (Romero *et al.* (2004)). In the microbial genomes, between 200 and 300 pathway holes are expected, where the majority of the holes are believed to be the result of a failure to identify the correct gene (Karp *et al.* (2010)).

To identify the missing enzymes that catalyse reactions thought to be present, some techniques have been developed. Some use a comparative genomic strategy (Osterman and Overbeek (2003)) where information from closely related genomes is used. Still within comparative genomics, there are examples of studies that try to find functionally analogous genes (Morett *et al.* (2003)). Other approaches use Machine Learning techniques in order to evaluate the candidate gene using homology, genomic context and pathway-based evidence (Green and Karp (2004)). As yet, however, such methods have had little impact on the overall quality of automatically generated metabolic maps.

New and more accurate methods for the metabolic reconstruction of pathogens and parasites are needed if we are to make full use of systems biology in identifying enzymes or transporters that may be viable targets for the development of more effective drugs.

This thesis has the objective to contribute towards a new level of quality for metabolic network reconstruction, through the application of several different approaches in the metabolic reconstruction steps.

1.1.1 Background

Enzymes

The genetic information passed on in cell division is contained in the deoxyribonucleic acid (DNA). DNA consists of two 2 chains of nucleotides (adenine (A), guanine (G), cytosine (C), and thymine (T)) grouped in the form of a double helix connected by hydrogen bonds between the complementary bases (A-G and C-T).

Some parts of the sequence, called *genes*, can be transcribed to a ribonucleic acid (RNA) sequence and decoded to *proteins*, polymer chains made of amino acids linked together by peptide bonds. The rest of the DNA sequence is not yet completely understood (Kapranov and Laurent (2012)).

Proteins may have several different functions making them one of the bases for all cell functions. There are specialised proteins, the antibodies, that make part of our immune system and help defend the organism against external objects such as viruses. Others are called structural proteins and can be found in the extracellular matrix in our tissues or for example our hair. The proteins are responsible virtually for all of the cell's functions.

One class of proteins is called enzymes. Enzymes are proteins with the task to catalyse chemical reactions. Almost all cellular reactions need an enzyme in order to be catalysed. Those reactions that do not need an enzyme are called *spontaneous*. The enzyme works by decreasing the minimum energy necessary to start the reaction, consequently increasing the rate of the reaction.

When discussing metabolic networks, enzymes and their function play a crucial role.

Enzymatic function labeling

Identification of an enzyme to a specific function can sometimes be a very difficult process. To help with this task it is crucial to have a structured enzymatic classification system. Currently the two most used ones are EC Numbers and GO terms.

EC Numbers

The Enzyme Commission number (EC number) is a hierarchical numerical classification scheme for enzymes, based on the chemical reactions they catalyze. Every enzyme code consists of the letters EC followed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzyme. The first number divides the enzymatic functions into six groups:

1. *Oxidoreductases*: in this group are included all the oxidation or reduction reactions: These reactions are characterised by the transfer of an oxygen or hydrogen atom as well as electrons within the molecules.

2. *Transferases*: these reactions are responsible for the transfer of a functional group: Example of functional groups are methyl, phosphate, etc.

3. *Hydrolases*: this group contains all the hydrolysis reactions: reactions responsible for the cleavage of a molecule by adding water.

4. *Lyases*: These enzymes are responsible for cleaving non-hydrolytic chemical bonds (such as: Carbon-Carbon (C-C), Carbon-Nitrogen(C-N), Carbon-Oxygen (C-O), Carbon- Sulphur (C-S)).

5. *Isomerases*: These reactions transform one molecule into another molecule which has exactly the same set of atoms.

6. *Ligases*: These enzymes are responsible for the synthesis of chemical bonds by breaking down ATP.

The numbers to the right in the EC number notation divides the reactions into finer groups untill the 4th number, which specifies the reaction at the substrate level.

Gene Ontology

GO terms are part of a project with the objective of standardising the gene and gene product annotations across species and different data sources (Ashburner *et al.* (2000)). This project also provides a number of tools to access its contents and decrease the time necessary to search and make use of the data provided.

The GO terms define gene products within three separate ontologies: cellular component, biological process and molecular function. Each ontology is a directed acyclic graph structure and within each, a gene product may be assigned to more than one GO term. Because of these two properties a gene annotated to a given node is automatically annotated to all its ancestral nodes.

1.1.2 Structural domain classification

The sequence of amino-acids (also known as primary structure) of a protein determines its 3D conformation. Within the 3D conformation are two types of patterns which constitute the secondary structure. These are the α-helices and the β-sheets. Multiple α-helices and β-sheets can combine into more complex, compact units called *domains*. These structures can be present in different proteins and can be combined with each other in different groups in the same protein (multi-domain proteins) or alone (single-domain), resulting in different enzymatic functions. Domains are also seen as evolutionary units. Within multi-domain proteins, the domains are often structurally and functionally independent of each other.

There are three levels of structural domain classification: fold, superfamily and family (Murzin *et al.* (1995)). Fold is the highest level. It groups together domains that have the same secondary structure elements and the same chain topology. Next to the fold is the superfamily. This level groups together domains that have structure and functional evidence to share a common ancestor. So, in these groups are believed to be the most distant homologous genes. The lowest level is called family. This level groups together domains with clear sequence similarity. Domains from the same family tend to have similar functions.

1.1.3 Metabolic Network

A Metabolic Network is a group of biochemical reactions. The interaction of the reactions is responsible for the metabolic processes that determine the cell functions. Normally, the reactions represent the nodes of this network. For almost all reactions there is an enzyme responsible for its catalysis.

Within the metabolic network are sets of connected chemical reactions that transform a starting molecule into another one (product). These sets of reactions are called *metabolic pathways*. They normally represent the transformation of a main molecule into another. Moreover, the pathways are not independent from each other, having common reactions and molecules.

Because species have different biochemical properties, the same pathway may vary between species or even not exist in others. Some databases collect all these different biochemical properties and have built pathway templates that illustrate the complexity of the metabolic networks and the differences between species. Examples of these pathway templates can be found in KEGG (Kanehisa and Goto (2000); Kanehisa *et al.* (2006, 2008)) and in RAST (Aziz *et al.* (2008)).

1.1.4 Metabolic Reconstruction

Metabolic reconstruction is the process of building a map of an organism's metabolic network using evidence from its genome sequence (Thiele and Palsson (2010)). A motivation to improve these models is, for example, that more accurate metabolic networks of pathogens and parasites will allow the identification of important enzymes or transporters that can be potential targets for new drugs.

The reconstruction process can be described as a sequence of simple steps. After having the whole genome sequenced, the first step is the identification of the coding sequences of possible genes. There are several automatic softwares that can achieve this, such as ERGO or RAST. The methodology used to annotate gene sequences can go from the identification of the start and end codons, to the use of sequence similarity or family profiles.

After having identified the coding sequences of possible genes, the predicted protein sequences are compared against sequences from known, possibly closely related, genomes in order to transfer enzymatic annotations where genes appear to be functionally equivalent. For this step the most common approach relies on sequence comparison methods such as BLAST.

In this way, putative metabolic networks are built. The next steps are more time consuming. They are related not only to the manual curation of the networks and of the assignments that were made but also to the assignments missed. Here, experts try to reconcile the output information with the known biology, in particular with species-specific information (Francke *et al.* (2005)).

Data sources

There are several databases that gather important biological information and are used to support not only the curation of metabolic networks but they are also used by software. The Kyoto Encyclopedia of Genes and Genome (KEGG) database (Kanehisa and Goto (2000); Kanehisa *et al.* (2006, 2008)) gives us a large amount of information about biological systems, ranging from genes and proteins to molecular wiring diagrams of interactions and reaction networks from several species. An example of a more enzyme specific database is BRENDA (BRaunschweig ENzyme DAtabase) (Scheer *et al.* (2011)). This database provides several levels of information regarding enzymes going from nomenclature, relation to reactions and species specificity, etc to connection to other databases and the available literature for each species/enzymatic function.

A number of databases look into the identification and classification of domains. At one level are the Structural Classification of Proteins (SCOP) (Murzin *et al.* (1995)) and CATH (Orengo *et al.* (1997)). These structural databases group the proteins with known structures into several levels of domain similarity. In the case of SCOP, structures are grouped into the fold/superfamily/family levels already mentioned. At another level are the databases that make use of the previous ones and through Hidden Markov Models (HMM) build profiles for each of the structural levels. Examples of these kind of databases are SUPERFAMILY (Gough and Chothia (2002)), Gene3D (Yeats *et al.* (2008)) and Pfam (Bateman *et al.* (2004)).

Databases like FireDB (Lopez *et al.* (2007)) and Catalytic Site Atlas (Porter *et al.* (2004)) look into the protein's functionally important residues. Firedb includes residues that perform binding activities and residues that have catalytic functions. This database has two main sources of information. On the one hand, it uses PDB (Berman *et al.* (2000)) crystal structures to identify the close atomic contacts and, on the other hand, it makes use of the

Catalytic Site Atlas to get reliably annotated catalytic residues.

Some databases offer species specific data. MPMP (Ginsburg (2000)) uses KEGG (Kanehisa and Goto (2000); Kanehisa *et al.* (2006, 2008)) pathways as templates and presents this information together with genomic annotations from other databases like GeneDB (Hertz-Fowler *et al.* (2004)) and PlasmoDB (Bahl *et al.* (2003)). Presently, MPMP has the most current and informed pathways for *Plasmodium falciparum*. It is curated by a world expert on malaria parasite biochemistry.

Available approaches

A possible approach for network reconstruction is based on manual curation. This process just by itself is time consuming. The increasing speed with which we are able to sequence an organism's genome makes bioinformatics an increasingly important source of information. As already mentioned, MPMP (Ginsburg (2000)) is an example of a database that relies a lot on manual reconstruction.

To assist with metabolic reconstruction there are a number of different approaches. Some of them cover most of the steps required to build a metabolic model such as Pathway Tools (Karp *et al.* (2002)), ERGO (Overbeek *et al.* (2003)) and RAST (Aziz *et al.* (2008)). These softwares integrate a set of bioinformatic tools that cover the gene sequence and function annotation, and visualization tools together with integrated databases that help the user to curate the model.

The availability of a fully annotated genome is of crucial importance, because most techniques of network reconstruction start with the assignment of functions to the known and possible enzymes. Some exceptions are metabolic SearcH And Reconstruction Kit (metaSHARK) (Pinney *et al.* (2005); Hyland *et al.* (2006)), ERGO and Rast. These softwares use different approaches to annotated sequences. On one hand, metaSHARK uses its SHARKhunt tool, that uses the PRIAM library (Claudel-Renard *et al.* (2003)) of profile models as the basis of a search tool for finding the DNA sequence regions with significant similarity to known enzymes. On the other hand, a very different approach is used by GLIMMER2 (Delcher *et al.* (1999)) (this software is integrated in Rast). Here, the annotation is made by the use of interpolated Markov models that are trained using curated gene structure data.

The function annotation is still currently mostly based on sequence similarities using, for example, BLAST (Altschul *et al.* (1990)). A common protocol used in this phase is the "reciprocal best hit". Using a sequence comparison software like BLAST, each of the annotated gene sequences is"blasted" against all the gene sequences of a closely related organism. The results for each gene are ranked and only the best one is considered. The same is done in the reverse direction, that is, from the close relative to the organism that is being annotated. The genes that are reciprocal best hits are believed to be orthologs, genes that evolved from a common ancestral gene by speciation, and therefore are likely to have the same function. However, as already stated, the use of sequence similarity is not accurate enough to detect all the enzymatic annotations.

There are some approaches with a more sensitive search protocol. Some, make use of libraries of domain profile models, as in the PRIAM software (Claudel-Renard *et al.* (2003)). PathoLogic, one of the many components of Pathway Tools, uses a text mining approach to predict computationally the metabolic network of any organism whose genome has been sequenced and annotated, creating a pathway/genome database (PGDB). A more recent software tool, EFICAz2 (Enzyme Function Inference by a Combined Approach) (Arakaki *et al.* (2009)), approaches the gene function prediction step by combining the predictions of different methods, including pairwise sequence comparison, support vector machines, hidden Markov models, etc, having as a principal source the Pfam database (Bateman *et al.* (2004)).

1.1.5 Phylogenetic analyses

Since Darwin presented us with his evolutionary theory and sketched the first phylogenetic tree of life, phylogenetic trees have been considered an important way to describe and study evolutionary events. Phylogenetic trees describe evolutionary relatedness through time. At first this relatedness only considered phenotypic traits. After the discovery of DNA and the advent of sequencing technologies, these phylogenetic analyses have progressed to include genotypic information (Felsenstein (2004)).

These analyses are useful to identify and visualise several evolutionary events such as gene duplications/losses or relationships such as orthology and paralogy. Homologous genes are genes that were derived from a common ancestor. If these genes diverged by speciation they are called Orthologous. On the other hand, if they were separated by a gene duplication, they are called paralogous genes. Analogous genes are genes with different structure but able to perform the same function.

1.1.6 Machine learning

As already mentioned, we live in an era where the biggest problem is not lack of data, the main issue now is how to make the most of it. To make use of all the available data sources and to build tools that can improve the accuracy and efficiency of the metabolic reconstruction process, the resources provided by Machine Learning should definitely be applied. Machine Learning is used in the development of methods to understand and help humans with several kinds of problems, often impossible to understand and to resolve by human effort alone.

Nowadays, Machine Learning is applied to a wide variety of tasks including natural language processing, search engines, medical diagnosis, bioinformatics, weather forecasting, parameter estimation, detecting credit card fraud, stock market analysis, speech and handwriting recognition, image processing, game playing and robot locomotion (Langley and Simon (1995)).

Learning paradigms

There are two major learning paradigms, distinguished mainly by the format of the information available: supervised learning (Kotsiantis *et al.* (2007)) and unsupervised learning (Zhang *et al.* (2008)). Supervised learning is performed knowing the desired outputs of a given set of inputs. The aim is to discover the optimal function $f(X)$ that represents a prediction rule so the machine can produce the correct output given a new input. The strategy to achieve the optimal $f(X)$ typically involves the idea of error minimisation. For this, a cost function is used that represents the distance between the function and the learning data set. Normally, there is a set of training data and another set of testing data.

In the case of unsupervised learning, we do not have any learning set linking the input to the desired output. Here we try to detect any emergent collective properties from a given dataset. In a sense, unsupervised learning can be thought of as finding patterns in a given dataset and beyond what would be considered pure unstructured noise. Clustering is an example of unsupervised learning.

The choice of method is difficult and sometimes also a matter of taste. Problems like local maxima and others exist in every algorithm and have been tackled by theoretical work (Chib and Greenberg (1995); Ponce-Ortega *et al.* (2009); Brooks and Morgan (1995)). With the idea that another point of view can open up a whole new world, some effort must also be taken in finding other methodologies and other perspectives. The type of data existent in the problem may have an impact upon the method that can be used. Some methods may not be able to process numerical or categorical data (Alpaydin (2004)).

The objective of the study may also affect the machine learning method used. There are two groups of methods. On one hand we have “white box” models that return insights about the behaviour/predictiveness of the features and enable user to understand the structure of the data and, for example, visualise which properties have a greater predictive value. On the other hand, there are the “black box” models. In contrast to “white box” models, these provide little or no insight into how the data are structured and the importance of each feature within the model.

1.2 Application

All tools are made with a purpose. A bioinformatics tool is not meaningful if it does not show tangible results in a biological context. With this in mind throughout the thesis, every time that is possible, all the methods developed will be applied in a biological context. I have chosen two different species whose metabolic networks are studied for two completely different objectives: *Plasmodium falciparum* and *Geobacillus thermoglucosidasius*.

1.2.1 Malaria

Malaria is one of the most widespread diseases in the world. It is caused by protozoan parasites of the genus *Plasmodium*, passed to humans through mosquito (genus *Anopheles mosquitoes*) bite. Five species of the *Plasmodium* genus cause human malaria. Among these, *Plasmodium falciparum* inflicts the most mortality. In 2010, there were 216 million reported cases, of which 174 million were in Africa. This resulted in 655000 deaths worldwide, a 26% decrease if compared with 2000. Although an improvement, the rates were lower than what had been set has target, a reduction of 50%.

The most affected are young children below 5 years old living mainly in sub-Saharan Africa. Estimates suggest that 40% of the world's population is at risk of malaria (Murray *et al.* (2012)). Much effort has been made to overcome this problem, and many antimalarial drugs have been developed (Breman *et al.* (2007)).

The sequencing of *Plasmodium falciparum* (Gardner *et al.* (2002)) was part of a project started by the Sanger centre, Standford University and the Institute for Genomic Research (TIGR). This ambitious project began in 1996, at a time when large eukaryote whole genome sequencing had not yet been tried. The idea of making all the discoveries and tools freely available gave rise to PlasmoDB, a public web encyclopedia of malaria, providing free access to analysis tools and data.

Later on other *Plasmodium* species and clones were sequenced, such as, *P.yoelli* (funded by US departement of defense) (Carlton *et al.* (2002)), *P. vivax* (TIGR) (Carlton *et al.* (2002)), *P.berghei* (Hall *et al.* (2005)), *P. chabaudi* (Hall *et al.* (2005)), *P. knowlesi* (funded by Welcome Trust) (Pain *et al.* (2008)).

Plasmodium falciparum

Plasmodium falciparum was the first eukaryotic parasite to be sequenced (*Plasmodium falciparum* 3D7). It has 14 chromosomes with 30Mb of DNA in total. The genome is AT-rich (80%) which made its sequencing more challenging because it makes it difficult to clone.

About 60% of the proteins in *P. falciparum* have little or no similarity to proteins in other organisms and most are not functionally annotated. The proportion of these hypothetical proteins is higher in *P. falciparum* than in other organisms. This might be a consequence of the evolutionary distance between *Plasmodium* and other model organisms.

This parasite has a complex life cycle and a intracellular nature. This makes the quest of finding solutions for its eradication more difficult. Its sequencing has offered new means to facilitate and accelerate research towards the development of novel drugs and vaccines.

Moreover, the high mutation rate and difficulties in applying drug policies result in a rapid loss of effectiveness. There is a specific and urgent need for new antimalarial drugs as there is now only one drug family (based on artemisinin (Woodrow *et al.* (2005))) without widespread drug resistance.

1.2.2 Greenhouse gas emission

The increasing rate of carbon dioxide in the atmosphere has been one of the main causes for the global warming problem the world is now facing. If the problem is not solved and the carbon dioxide rate continues to increase it will have terrible consequences such as catastrophic effects on wildlife, large-scale food and water shortages, sea level rise, etc. Dependency on fossil fuels is pointed out to be the main cause (Höök and Tang (2012)).

An increasing effort has been devoted towards research for alternative and renewable sources of energy that have fewer negative environmental consequences than the ones already in use.

Bioengineering is one area of research aiming to find viable alternatives to fossil fuels and to reduce carbon dioxide emissions. Here the main focus is to find biological solutions to solve real-world problems (Endy (2005)).

To solve the above mentioned greenhouse emission problem, *Geobacillus thermoglucosidasius* is a candidate biological solution due to its capability to convert lignocellulose to ethanol (Taylor *et al.* (2009)).

Geobacillus thermoglucosidasius

Geobacillus thermoglucosidasius is a Gram-positive, rod-shaped bacteria able to survive in a variety of environmental conditions. It has around 390000 nucleotides with almost 4000 proteins.

Geobacillus thermoglucosidasius used to be classified as *Bacillus thermoglucosidasius* until 2001 when together with all the thermophilic *Bacillus* strains it was classified into a new genus, *Geobacillus*.

Moreover, *Geobacillus thermoglucosidasius* NCIMB 11955 is a thermophilic bacterium with the potential to convert lignocellulose to ethanol in a highly productive manner. Thermophilic bacteria are especially useful in biofuel production since they can withstand the high temperatures that are unavoidable at certain stages of fermentation.

1.3 Thesis Overview

In this thesis I present several tools targeting the improvement of metabolic reconstruction software and ultimately help to build better models. With the exception of PathwayBooster, all other studies can be coupled with other tools, producing more robust and accurate tools.

Chapter 2 tackles the first handling of a draft metabolic network. The output of an automatic metabolic reconstruction always has a lot of inconsistencies as well as misannotations. This chapter presents PathwayBooster, a software package comprising a unique set of tools and data sources that enables a more rapid detection and correction of possible errors in the model. Here, I present a detailed description of all of these tools together with justifications for their inclusion in this type of analysis. Moreover, in this chapter I also present a practical example of a case study where the advantages of using PathwayBooster are clearly seen. The target organism used is *Geobacillus thermoglucosidasius*.

A gene enzymatic function search tool is presented in Chapter 3. Most software packages currently available are based on sequence based searches. This has been shown to be not sensitive enough to detect all enzyme functions, especially in certain conditions already mentioned. Therefore, this tool was designed with the objective of increasing the domain search sensitivity from a sequence/family level to a structural/superfamily level that groups all the most distantly related domains. This tool was proven to be successful when applied to a known example of evolutionary convergence. This chapter also presents the results of this tool applied to the holes in a manually curated model of *Plasmodium falciparum*.

In Chapter 4, I present a study to tackle the misannotation and error propagation problem, using a network topological perspective. Using a curated set of well annotated and misannotated enzymes, I have built a model that assesses the accuracy of assigned molecular functions. This model is based on simple topological properties and is completely independent from gene sequence analyses. The model is successfully tested using 5-fold and inter-superfamily cross validation analyses. Afterwards it was applied to draft metabolic networks and its results compared with curated metabolic networks of the same organisms. Further sudies were performed showing factors affecting the quality of the current metabolic reconstructions in model organisms.

Finally, the study presented in Chapter 5 gives an evolutionary perspective on enzymatic function change. The protocol described in Chapter 3 was applied to a set of model organisms and several superfamily phylogenetic trees were built. Using different branch length recalculation approaches I have shown that there is a correlation between enzymatic function change and branch length. Two methods were used to reconstruct ancestral functional states. One was based on a parsimony approach and the other on a maximum likelihood approach. Further studies were made to verify if evolutionary correlations were also present for other factors such as the dN/dS ratio, for which no correlation was found.

The final chapter makes a brief summary of the thesis objective followed by a discussion for all the main results and findings of the thesis. Afterwards, I propose possible future steps to the continuation of this research.

Chapter 2

A tool to support the curation of metabolic pathways

2.1 Introduction

As explained in Chapter 1, the manual curation of any organism is a time-consuming and laborious task (Thiele and Palsson (2010)). During the many stages of the metabolic network curation process there are several bioinformatic resources that can reduce the time required for each stage. Moreover, these resources can also have a positive impact on the quality of the model that one is trying to build.

The first stage of a genome-scale metabolic reconstruction is the creation of a draft metabolic model. For this stage, as previously mentioned, some automated resources are available. Good examples mentioned in the introduction chapter are Pathway Tools (Karp *et al.* (2002)), Model SEED (Henry *et al.* (2010)) and ERGO (Overbeek *et al.* (2003)).

However, after this first step there is still a large amount of work to do before reaching an accurate metabolic model. These draft metabolic reconstructions are often found to contain numerous inaccuracies (Kim *et al.* (2011b)), namely in the species specific biochemistry that result from the use of non-species specific databases and reactions.

In the next stages of curation, obvious pathway holes (due to the lack of an assigned enzyme) and false positive reactions (due to enzyme misannotation) need to be found and corrected. To address both of these issues there is a need to collect and analyse evidence for each reaction from the literature and from genomic and metabolic databases, across multiple closely-related species. Without automation this process is tedious and repetitive.

There are already some tools that can tackle this problem, allowing comparative analysis of metabolic pathways, such as Comparative Pathway Analyzer (Oehm *et al.* (2008)), FMM (Chou *et al.* (2009)) and ComPath (Kwangmin and Sun (2008)).

Comparative Pathway Analyzer (CPA) is a web implemented tool with the objective of finding the differences in the metabolic networks between two groups of organisms. The maps and reaction annotation data used are taken from the KEGG database. CPA also contains a pathway-reaction display that enables the easy detection of differences between up to six different genome annotations. Furthermore, it provides cluster analyses that can include any further annotation uploaded by the user.

FMM is a web server with the prime objective of reconstructing metabolic pathways between two metabolites. It is also mainly based on the KEGG database but it integrates other biological databases including UniProtKB/Swiss-Prot (Boutet *et al.* (2007)) and dbPTM (Lee *et al.* (2006)). Moreover, FMM presents the reconstructed pathway by means of a diagram connecting each of the reactions to information such as metabolites and enzymes involved in the pathway as well as comparative analyses from the species chosen by the user.

ComPath is a complex piece of software that integrates several data sources and tools for pathway analyses and gene annotation in multiple genomes. This information is displayed by means of an interactive spreadsheet, enabling access to several data sources simultaneously. Moreover, it provides tools for structural domain analyses as well as sequence comparison and enzyme prediction.

An ideal piece of software for curating a metabolic model would provide a pathway visualiser together with annotation confidence information and existing literature references. None of the packages above contains these features all together.

In this chapter I present PathwayBooster. PathwayBooster is an open-source software tool to support the comparison and curation of metabolic models. Although other tools exist for the comparative analysis of metabolic pathways, PathwayBooster presents a unique combination of features. Amongst other capabilities, PathwayBooster can be used to compare the functional annotations of genes with 'bidirectional best BLAST hits' analyses between the target organism and the relevant related species. It also compiles a list of literature references obtained from BRENDA (Scheer *et al.* (2011)) to support or refute the presence of each enzyme within the selected species. An interactive graphical summary of the evidence found in each organism is produced in the form of a clickable KEGG pathway diagram.

2.2 Colaboration

This software was developed as part of a collaboration with Beata Lisowska (PhD candidate, University of Bath), who used PathwayBooster to support the curation of a genome-scale metabolic model for *Geobacillus thermoglucosidasius*. The sequence and the initial annotation of the *G. thermoglucosidasius* (NCIMB 11955) genome were acquired from the ERGO Integrated Genomics platform. Ms Lisowska has compiled the annotations produced using ERGO (from where the genome was acquired) and RAST with the objective of minimising possible errors in the annotation. The RAST annotation server is a publicly available tool for the annotation of bacterial and archaeal genomes and as such was used for comparative purposes. The RAST server annotation is based on FIGfams (Meyer *et al.* (2009)) (which are protein families that share function and structure) but it also considers the localisation of genes. The decision process used is based on an assumption that proteins in a given family are orthologous to each other and hence share the same molecular function. The sequence of interest is analysed based on the similarity to the other members of a given FIGfam family.

A possible alternative to RAST could have been KAAS (Moriya *et al.* (2007)). KAAS is a web server for automatic annotation that uses bidirectional best hits between the query organism and the KEGG GENE database. The results are divided by KO groups, with a likelihood score assigned to each one. The KO group with the highest score is assigned to the sequence. However, RAST not only uses more sensitive homology methods but it also uses microbe specific pathway templates; the KEGG pathways used by KAAS represent metabolic network knowledge at a more general level.

With Ms Lisowska's collaboration, I have designed several tools to help the next step of manual curation: the identification of falsely annotated or omitted enzymes and reactions and the creation of an SBML model ready for FBA. I have written and put these tools together in one package, PathwayBooster. Using PathwayBooster reports, Ms Lisowska has manually inspected each pathway listed in the KEGG database.

The next step will be to use FBA to make predictions of consumption and production of certain substrates of interest and afterwards compare these *in silico* predictions with *in vivo* experiments. The *in vivo* experiments will use techniques such as gene knockouts, phenotypic analysis and gene cloning and expression.

2.3 PathwayBooster

PathwayBooster is a command-line tool written in Python. The user supplies input in the form of GenBank, EMBL or FASTA files for all the organisms that are to be compared. Output is presented as a browsable set of HTML files, with sections that are described in more detail below (Figure 2.1). Instructions on how to run PathwayBooster, can be found in: http://www.theosysbio.bio.ic.ac.uk/resources/pathwaybooster/

One of the key enabling technologies of PathwayBooster is in the use of KEGG API. This is a web service allows the access to KEGG system in an automated way using the SOAP protocol. This ensures that PathwayBooster always provides up-to-date KEGG data.

Pathway diagram

Using SOAP to retrieve KEGG pathway templates, PathwayBooster returns an interactive image where all the reactions are colour coded according to the presence or absence of a given reaction in each chosen species (Figure 2.1). This is implemented using the SOAP protocol functionality that allows the colouring of each reaction rectangle background with just one colour. PathwayBooster makes use of these functions to pass on to each reaction present in the diagram the information about the species in which the reaction is present. A script using the pathway template in a png format is able to identify the reactions and the colour with the information mentioned before and colour code each reaction as described above.

In the transformed KEGG pathway template, information about each reaction can be accessed via a popup menu which displays the available options for a given enzyme. Information is divided into three groups: annotations, BLAST results and literature. Each choice can be accessed by its own hyperlink, redirecting the user into a new window

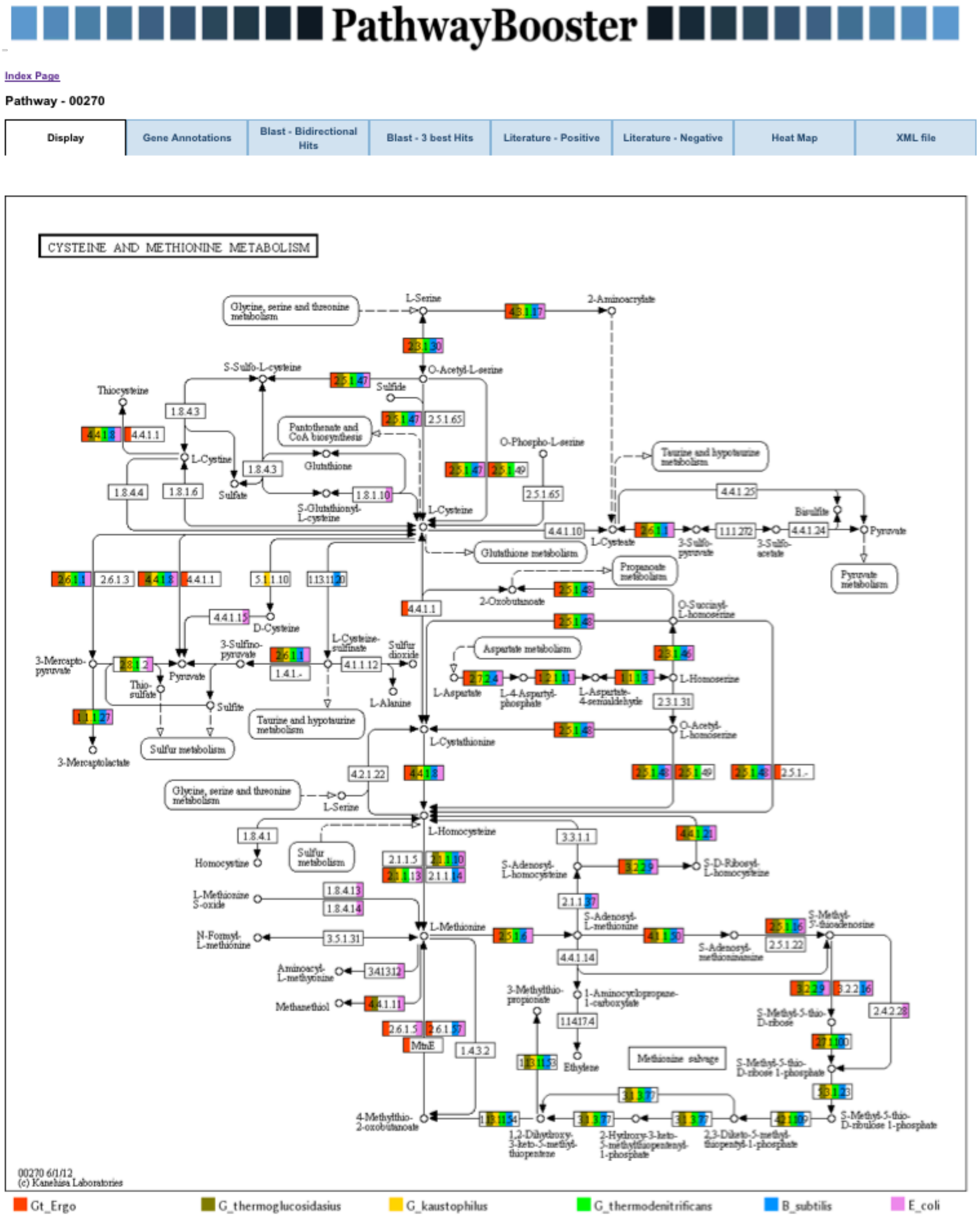


Figure 2.1: Cysteine and methionine metabolism pathway. An example pathway diagram produced by PathwayBooster, showing the cysteine and methionine metabolism. On the top are the tabs directing to different information sources. The coloured blocks show an automated model produced by ERGO$^{\mathrm{TM}}$ (Overbeek *et al.* (2003)) for the thermophilic bacterium *G. thermoglucosidasius* NCIMB 11955 (red) in comparison to selected reference organisms: *G. thermoglucosidasius* C56-YS93 (brown), *G. kaustophilus* (yellow), *G.thermodenitrificans* (green), *B. subtilis* (blue) and *E. coli* (purple).

where the corresponding data can be viewed. All functions can also be accessed through the tabs in the top of the pathway image. An important advantage of using popup menus is that the use of the popup menu will restrict the report data in all the different groups to focus on the enzymatic function specified.

Annotations

The annotation table is divided according to the Enzyme Commission (EC) numbers present in a pathway of interest. Annotated genes are presented by EC number for all specified organisms. Each gene is hyperlinked to the KEGG database, where associated information can be viewed. It also refers to the origin of each annotation. This is relevant when more than one genome annotation source is under consideration. With the exception of KEGG, all annotation sources must be supplied by the user. In the KEGG annotation case, the data is accessed using the SOAP protocol functionalities. For the other enzymatic function input formats (GenBank and EMBL), PathwayBooster parses these input files and stores the gene ID and the enzymatic function. The enzymatic function can be supplied in either GO or EC number format. In the case of GO annotations, the annotations are mapped to EC numbers using the ec2go flat file present in the Gene Ontology website (Ashburner *et al.* (2000)).

Blast results

Two proteins from two different organisms are a best reciprocal hit when each one is the best BLAST hit of the other when a search is performed against the predicted proteome. This is the simplest method used to find pairs of orthologous proteins (Jordan *et al.* (2002)), that is, proteins descending from a common ancestor that have diverged after a speciation event. These proteins tend to have similar sequences and are likely to have similar functions. So, providing this evidence is vital in the curation of a specific reaction. It can be important either to support a given annotation or to find a candidate for a missing one. Based on the genome information provided by the user, BLAST (Altschul *et al.* (1990)) best reciprocal hits are available for proteins from a query organism when compared with the other species. Each hit gene is followed by its annotated function, the respective EC number and the sequence similarity, E-value and BLAST score between the two genes.

The first three BLAST hits can also be viewed for every target gene annotated in a reference species for which a gene sequence annotation was provided, providing possible protein candidates. The report also provides a function annotation and EC number, as well as the sequence similarity, E-value and BLAST score between each candidate gene and the target gene.

To calculate the reciprocal best hits, PathwayBooster makes use of the BLAST $blastp$ function. For a pair of genomes provided in a fasta file format, a script *blasts* each of the protein sequence of one genome against all the proteins sequences of the other genome. For each sequence only the best match is kept. The process is repeated the other way around and again only the best match is kept. Reciprocal best hits are all the pairs of sequences that are simultaneously the best match of each other.

Literature

PathwayBooster makes use of the BRENDA download flat file to provide information about the existing publications available for each organism describing each enzymatic function. Therefore, for the selected pathway, publications taken from BRENDA that assert the presence of each EC number in a specified organism are listed. Publications indicating that a given EC number might be absent in an organism are also available. Each

EC number	Brenda publications	Missed publications	Missed publications %
1.3.1.48	16	9	0.36
1.3.99.18	2	0	0
2.1.1.69	10	1	0.09
2.3.1.103	1	0	0
2.4.1.135	20	16	0.44
2.4.1.187	3	0	0
2.7.11.29	2	0	0
3.1.1.51	3	0	0
3.6.1.24	1	0	0
5.1.3.19	12	6	0.33

Table 2.1: BRENDA literature coverage. This table shows the number of publications provided by BRENDA , the number and the percentage of publications missed by BRENDA for each randomly selected EC number.

publication has a hyperlink to the PubMed website, where its abstract can be viewed. Currently the number of manually annotated references in BRENDA is over 100,000 (Scheer *et al.* (2011)). To have an idea of how comprehensive BRENDA is I have randomly chosen 10 EC numbers and checked against PubMed how many references would have been missed. Overall, BRENDA covers 70% literature of the publications. Table 2.1 shows the number of publications considered by BRENDA for each EC number and the number of publications found in PubMed not present in BRENDA. More than half of the EC numbers are completely covered by BRENDA. However, there are cases where more than 30% of the literature would have been missed. It is important to notice that the comparison against PubMed was made on a string matching bases using enzymatic synonyms. It is likely that some of the literature found are false positives. Therefore, the overall 70% literature coverage by BRENDA may be underestimated.

Heat map

For a given pathway, the Hamming distance between two organisms is the number of enzymatic functions present in one but not in both of those organisms. In the Pathway-

Booster report a heat map is provided to show the Hamming distance between the organisms selected, according to the presence or absence of each function present in the pathway. This simple visualisation of the similarity between pathway structures can be used to support comparative analysis or to summarise the consistency of different annotations. This tool is built by making use of the $matshow$ python function from the Matplotlib library. For each pathway a binary profile of presence and absence of each of the enzymatic functions present in the pathway is built and passed on to the $matshow$ function.

2.3.1 Comparison with existing tools

Platforms such as Model SEED can be used to produce draft metabolic models, but are not designed to support further model curation. PathwayBooster provides a single integrated interface to literature references, BLAST evidence and annotations from alternative sources or related organisms. Most importantly, PathwayBooster provides a logical visual representation of its results, significantly reducing the effort needed to identify enzyme misannotations and pathway holes. The information provided by PathwayBooster can be particularly useful when working with a platform for genome-scale model curation such as MEMOSys (Pabinger *et al.* (2011)) or GEMSiRV (Liao *et al.* (2012)). Although several other tools exist that can support comparative pathway analysis not all were made specifically for this task. Moreover, PathwayBooster provides a unique combination of features that make it particularly suitable for use in model curation.

The most similar published software to PathwayBooster is the Comparative Pathway Analyzer (CPA) (Oehm *et al.* (2008)). Both programs use KEGG pathway templates and are able to display several organisms simultaneously (up to 6 in CPA and up to 7 in PathwayBooster). Moreover, both accept other annotation sources besides KEGG's annotations.

However, unlike CPA, PathwayBooster can accept annotations with GO terms (Ashburner *et al.* (2000)). To visualise the differences between the different organisms in each pathway, CPA uses a hierarchical clustering strategy and PathwayBooster, makes use of a heat map. Unlike CPA, PathwayBooster is able to provide literature information for each enzymatic function and detailed sequence comparison analyses. Therefore, compared with CPA, PathwayBooster is more focused towards pathway curation.

In contrast to ComPath (Kwangmin and Sun (2008)), PathwayBooster uses data provided not only by KEGG, but also publication data from BRENDA and annotation files supplied by the user. In addition to presenting evidence from ‘bidirectional best BLAST hits’, PathwayBooster allows the comparison and compilation of multiple annotations obtained from different sources for a given genome. Another advantage of PathwayBooster is its interactive graphic visual representation. For example, it enables the comparison between different organisms at the same time, making the identification of erroneous annotations easier.

Finally, FMM (Chou *et al.* (2009)) focuses on the possible pathways between two metabolites, whilst PathwayBooster provides a broader view of the complete metabolic network and can be used to curate a metabolic model from the starting point of a genome annotation.

2.4 Case studies

This section presents some examples where the advantages of using PathwayBooster are clearly seen. The case study was developed in collaboration with Beata Lisowska, who used PathwayBooster to support the curation of a genome-scale metabolic model for *Geobacillus thermoglucosidasius*.

Geobacillus thermoglucosidasius

Geobacillus thermoglucosidasius NCIMB 11955 is a thermophilic bacterium with the potential to convert lignocellulose to ethanol in a highly productive manner. Thermophilic bacteria are especially useful in biofuel production since they can withstand the high temperatures that are unavoidable at certain stages of fermentation. Given these interesting properties, we would like to understand the metabolism of this organism in more detail.

As an example, PathwayBooster results for cysteine and methionine metabolism (KEGG id = 00270) are presented. The initial draft metabolic network was built using ERGO (Overbeek *et al.* (2003)). Reference organisms were selected to make full use of the comparative genomics functionalities provided by PathwayBooster. The organisms selected were, *Escherichia coli*, *Bacillus subtilis*, *Geobacillus thermoglucosidasius* C56-YS93, *Geobacillus thermodenitricans* and *Geobacillus kaustophilus*.

Filling pathway holes

The Hamming distance heatmap (Figure 2.2) gives us the first evidence of an unexpected difference between the *Geobacillus thermoglucosidasius* draft metabolic network and the the other organisms. Examining the Pathway diagram, it can easily be seen that the reactions tagged with the EC numbers 4.2.1.109, 3.1.3.77, 1.13.11.53 and 5.3.1.23 are not annotated for the query organism, in contrast to most of the reference organisms. A possible explanation is that the enzymes with these functions were not identified by the ERGO annotation servers.

Making use of the PathwayBooster publication tables for each function present in the pathway, an article can be found relating to the enzyme 4.2.1.109 (5-methylthioribulose-1-phosphate dehydratase) in *Bacillus subtilis*. The article referenced is easily accessed by

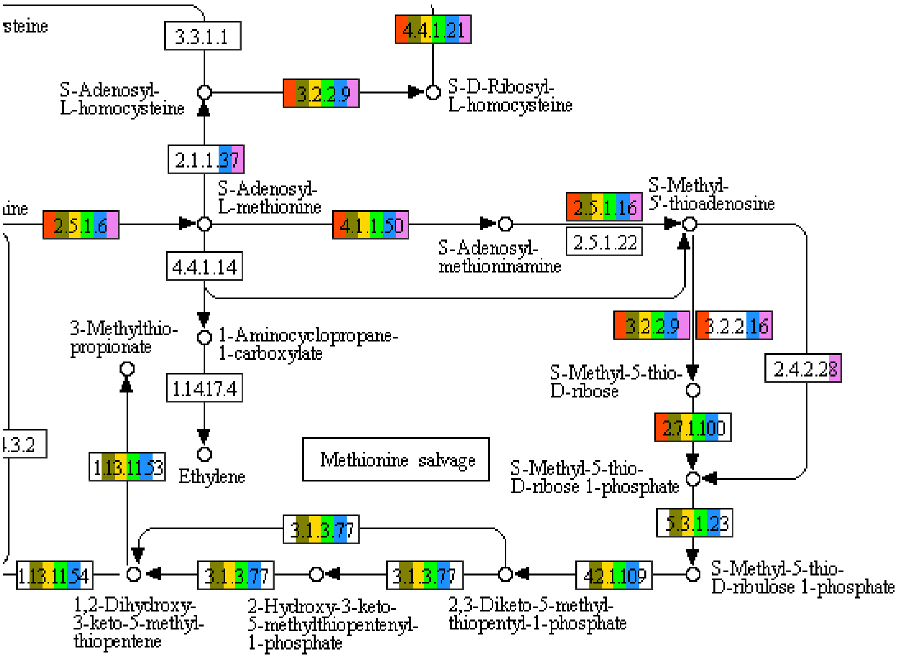

steine
3.3.1.1
4.4.1.21
S-Adenosyl-L-homocysteine
3.2.2.9
S-D-Ribosyl-L-homocysteine
2.1.1.37
S-Adenosyl-L-methionine
ine
2.5.1.6
4.1.1.50
S-Methyl-5'-thioadenosine
2.5.1.16
2.5.1.22
S-Adenosyl-methioninamine
4.4.1.14
3-Methylthio-propionate
1-Aminocyclopropane-1-carboxylate
3.2.2.9
3.2.2.16
2.4.2.28
S-Methyl-5-thio-D-ribose
1.14.17.4
2.7.1.100
.3.2
1.13.11.53
Ethylene
Methionine salvage
S-Methyl-5-thio-D-ribose 1-phosphate
5.3.1.23
3.1.3.77
1.13.11.54
3.1.3.77
3.1.3.77
4.2.1.109
S-Methyl-5-thio-D-ribulose 1-phosphate
1,2-Dihydroxy-3-keto-5-methyl-thiopentene
2-Hydroxy-3-keto-5-methylthiopentenyl-1-phosphate
2,3-Diketo-5-methyl-thiopentyl-1-phosphate

Figure 2.2: Visual representation of the methionine salvage pathway, where *G. thermoglucosidasius* NCIMB 11955 (red) is compared to selected reference organisms: *G. thermoglucosidasius* C56-YS93 (brown), *G. kaustophilus* (yellow), *G.thermodenitrificans* (green), *B. subtilis* (blue) and *E. coli* (purple).

clicking in the hyperlink provided in the table. All genes annotated for all the genomes considered for each function can be found in the Annotations report (Figure 2.4-A). This table provides easy access to further information for each gene via the KEGG database.

To find candidates for filling the enzymatic function 4.2.1.109, PathwayBooster BLAST bidirectional hits report is the indicated resource to use (Figure 2.4-B). Blast searches against *B. subtilis* retrieved a candidate gene within the *G. thermoglucosidasius* NCIMB 11955 genome.

For a less stringent search, PathwayBooster's three BLAST hits report retrieves the three best BLAST hits for each gene against the query genome (Figure 2.4-C). Each hit also reports the sequence similarity information, E-value and overall BLAST score.

The procedure described was also successfully applied to the remainder of the missed

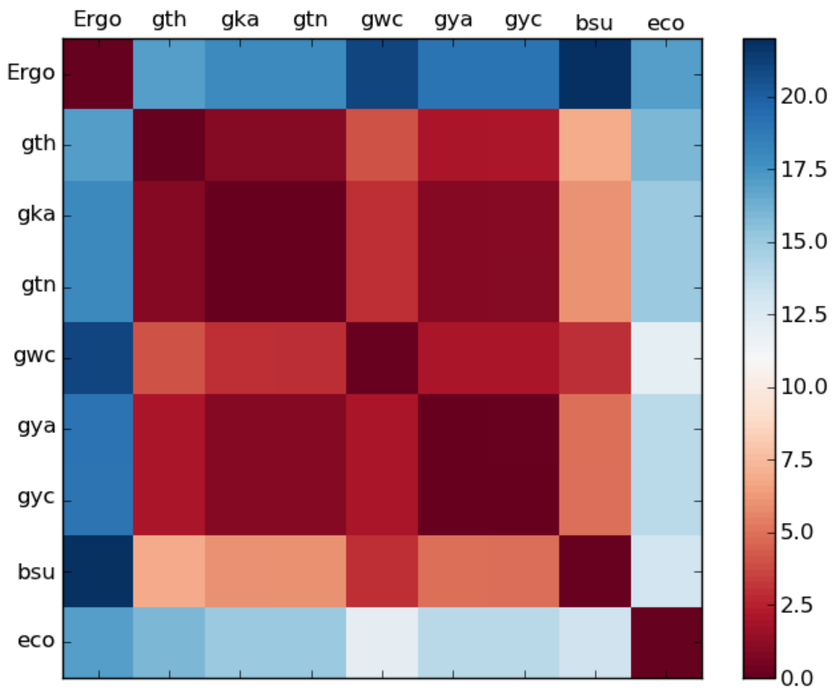
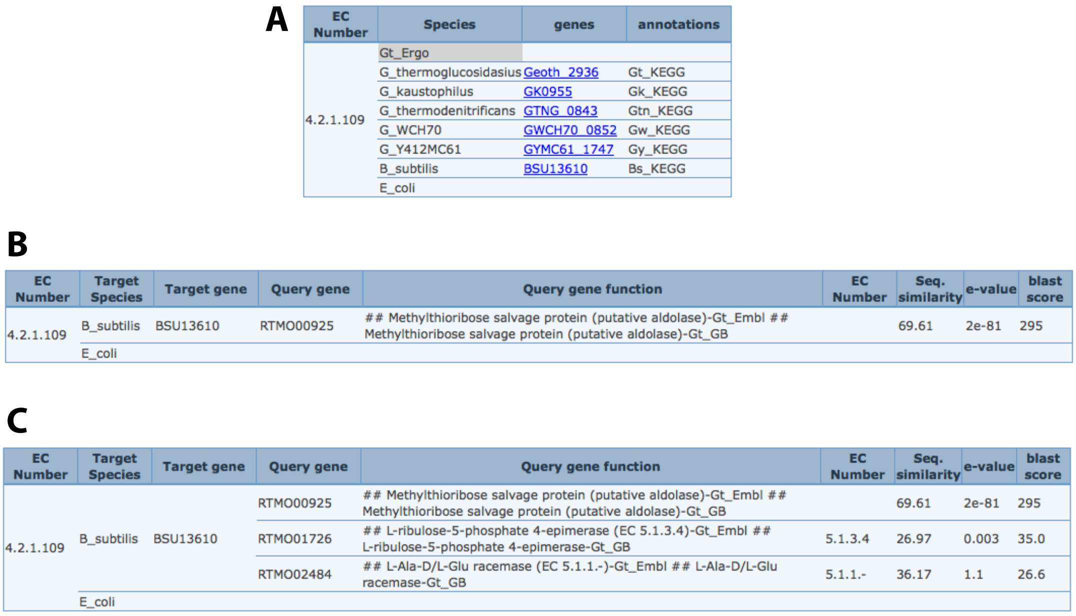


Figure 2.3: Hamming distance heatmap for cysteine and methionine metabolism, showing the similarity between the query species (marked 'Ergo') and reference organisms.



A

EC Number	Species	genes	annotations
4.2.1.109	Gt_Ergo		
	G_thermoglucosidasius	Geoth_2936	Gt_KEGG
	G_kaustophilus	GK0955	Gk_KEGG
	G_thermodenitrificans	GTNG_0843	Gtn_KEGG
	G_WCH70	GWCH70_0852	Gw_KEGG
	G_Y412MC61	GYMC61_1747	Gy_KEGG
	B_subtilis	BSU13610	Bs_KEGG
	E_coli		

B

EC Number	Target Species	Target gene	Query gene	Query gene function	EC Number	Seq. similarity	e-value	blast score
4.2.1.109	B_subtilis	BSU13610	RTMO00925	## Methylthioribose salvage protein (putative aldolase)-Gt_Embl ## Methylthioribose salvage protein (putative aldolase)-Gt_GB		69.61	2e-81	295
	E_coli							

C

EC Number	Target Species	Target gene	Query gene	Query gene function	EC Number	Seq. similarity	e-value	blast score
4.2.1.109	B_subtilis	BSU13610	RTMO00925	## Methylthioribose salvage protein (putative aldolase)-Gt_Embl ## Methylthioribose salvage protein (putative aldolase)-Gt_GB		69.61	2e-81	295
			RTMO01726	## L-ribulose-5-phosphate 4-epimerase (EC 5.1.3.4)-Gt_Embl ## L-ribulose-5-phosphate 4-epimerase-Gt_GB	5.1.3.4	26.97	0.003	35.0
			RTMO02484	## L-Ala-D/L-Glu racemase (EC 5.1.1.-)-Gt_Embl ## L-Ala-D/L-Glu racemase-Gt_GB	5.1.1.-	36.17	1.1	26.6
	E_coli							

Figure 2.4: Information tables for EC 4.2.1.109 (5-methylthioribulose-1-phosphate dehydratase). A - Annotated genes; B - BLAST bidirectional hits; C - Three best BLAST hits

annotations.

Identifying misannotated enzymes

In contrast to the example shown above, the enzyme function 5'-methylthioadenosine nucleosidase (EC 3.2.2.16) was found in the annotation of the query strain and not found in the closely related reference organisms. The two most probable explanations are: the gene assigned to this enzymatic function has been wrongly assigned or *G. thermoglucosidasius* has acquired a new function that is not present in close relatives.

By examining the 'Publications' reports, this function is not found in any relevant literature. Taking a closer look to the assigned gene, RTMO02286, in the 'Annotations' section, it is possible to see that the gene has been assigned with two potential functions: 5-methylthioadenosine nucleosidase (EC 3.2.2.16) and S-adenosylhomocysteine nucleosidase (EC 3.2.2.9). In the EC 3.2.2.9 case, all the reference organisms also had hits to this enzymatic function. This was also supported by the 'BLAST hits' report. Therefore, it was concluded that EC 3.2.2.16 must be a misannotation and that the most probable functional annotation for RTMO02286 is the EC 3.2.2.9 enzymatic function.

2.5 Conclusion

In this chapter has been shown that PathwayBooster contains a unique set of tools that provide relevant and complementary information for the curation of a metabolic network. These information include sequence comparison analyses, literature search, network topology comparisons, enzymatic function annotations as well an intuitive display that enables the visualisation and comparison of several different organisms in each metabolic pathway.

The case studies presented show how these tools complement each other and how PathwayBooster can be used to decrease the time necessary for curating a metabolic network.

Chapter 3

Protocol to identify candidate genes for a missing enzymatic function

3.1 Introduction

As already mentioned in Chapter 1, most software for gene function annotation is still based on sequence similarity searches. These programs use strategies such as “reciprocal best hit” to transfer functional annotations. Such strategies work well if there is a closely related, well annotated organism, otherwise they may not be sufficiently sensitive enough especially for species-specific functions. The overall idea of the work presented in this chapter is to elevate one level in the sensitivity of similarity searches. This work aims to move from BLAST sequence similarity to structural profile similarity, in other words, it aims to move from a family to a superfamily perspective. A superfamily is defined as a group of protein domains that have structural and functional evidence for their descent from a common evolutionary ancestor (Murzin *et al.* (1995)). This level of protein structural classification is in between two other levels: below is the family and above is the fold level. The former groups together those domains that have clear sequence similarities. The latter groups those domains that have the same major secondary structure, with the same chain topology.

Besides the fact that convergent evolution is not a rare phenomenon (Gherardini *et al.* (2007)), there are two important motivations for increasing the sensitivity of sequence similarity searches from family level to superfamily. The first is that the superfamily contains the most distantly related domains and so is the highest level for useful remote homology detection (Dayhoff *et al.* (1976)). Secondly, proteins within the same superfamily often have the same function, and usually but not always have related functions. This study will try to go beyond the immediate orthologous group to examine non-orthologous domains with similar binding sites. If the protein is already able to bind a similar substrate and/or catalyse a similar reaction, it will increase the chances of this protein acquiring the target function during evolution.

There are several examples of convergent evolution within a superfamily. An interesting example of successful hole-filling and of convergent evolution is given by Dittrich and coworkers (Dittrich *et al.* (2008)). This work was based on the idea that an evolving enzyme has more chance to acquire the function of structurally similar enzymes. They manually followed through a bioinformatic protocol to try to detect a functional analog of a missing enzyme (dihydroneopterin aldolase, DHNA) in the *P. falciparum* folate biosynthesis pathway. Afterwards, the most probable candidate was experimentally validated.

They explored how *P. falciparum* is able to cope with the absence of DHNA by coupling bioinformatics and experimental methods. BLAST searches were found to be unsuitable for detecting candidates to fill this missing link either in *P. falciparum* or in any other apicomplexan species. Programs such as 3D-PSSM (Kelley *et al.* (2000)) and GenTHREADER (Jones *et al.* (1999)) were used to search for secondary and tertiary structural similarities. Two significant matches were found: PFF1360w (6-Pyruvoyl tetrahydropterin synthase, putative (PTPS)) and PFL1155w (GTP cyclohydrolase I (GTPCH-I))

Sequence comparison analyses revealed a divergence in residues at the PFF1360w active site. A Cys residue, which is completely conserved in all other known PTPSs, is not only absent in the *P. falciparum* PTPS, but also from other *Plasmodium* species and from the related apicomplexan parasite *T. gondii*.

Structural differences were found between the malarian enzyme and other eukaryotes using crystallographic data for the *P. falciparum* enzyme. The hypothesis that malarial PTPS enzyme may have different catalytic properties compared with other organisms was then tested by cloning the *P. falciparum* PTPS gene into *E. coli*. This showed that the malarian PTPS is able to synthesise two different products, 6-hydroxymethylpterin and pterin, confirming the gene PFF1360w as the missing link. It is important to notice that the missing link was only found by broadening the candidate search to a superfamily level.

If we take a closer look at the folate pathway scheme (Fig. 3.3), it can be seen that the previously assigned reaction for the experimentally validated candidate placed it in a dead end, indicating that this annotation is unlikely to be correct. In an opposite situation, the analysis showed that the other enzyme (annotated as GTP cyclohydrolase I (GTPCH-I)), not only has its reactants produced and its products consumed but also is assigned to four chokepoint reactions. Introduced in Yeh *et al.* (2004), a chokepoint is a compound that is uniquely consumed by a specific reaction and/or is uniquely produced by another one. However the definition of chokepoint used in this thesis is as being a compound that is connected to just two different reactions, distinguishing these case from those that are connect to just one reaction (unpaired compounds) (Figure 3.1). A chokepoint reaction is a reaction where at least one of its products is a chokepoint. So, this kind of analysis together with others, can help to decide an order of priority for a group of candidate proteins.

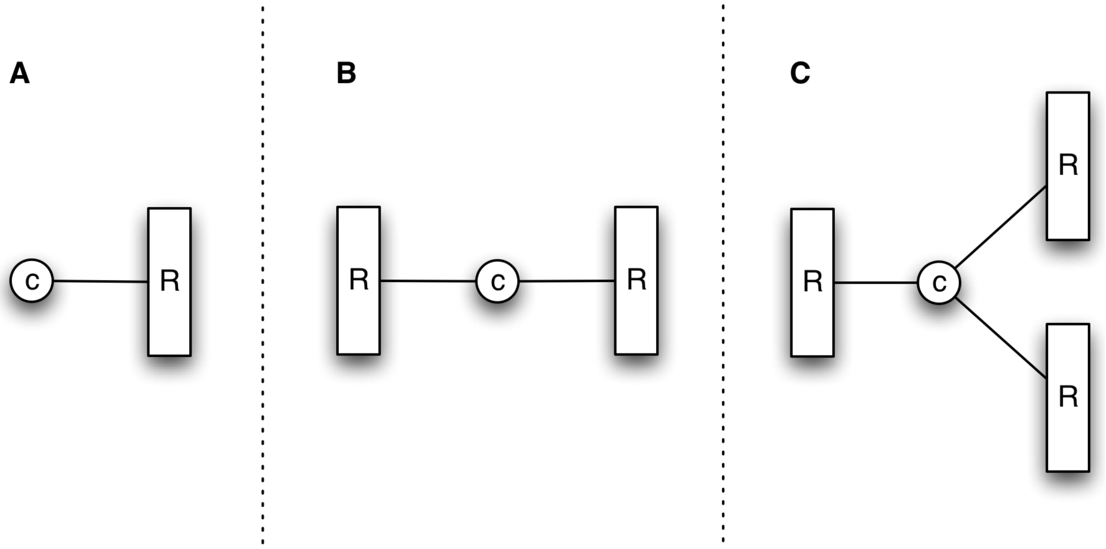


Figure 3.1: An example of an unpaired compound (A), chokepoint (B) and a compound that is neither an unpaired or a chokepoint (C). The circles are compounds and the rectangles are reactions.

Another case of convergent evolution is presented by Christopher M. Bruns and coworkers (Bruns *et al.* (1997)). In this case two proteins (*Haemophilus influenza* hFBP and the *Homo sapiens* transferrin) from two different families of the same superfamily have developed a Fe+3 binding site independently. In fact, their common ancestor is though to have been an anion-binding protein. A similar case is presented by Kuriyan and coworkers (Kuriyan *et al.* (1991)). They have discovered that the textitE. coli thoredoxin reductase, although it catalyses similar reactions as the human glutathione reductase, uses a different mechanism with different active sites. They also show that these two proteins have evolved from a common ancestor. Makaroca and coworkers have studied the Zn-peptidase superfamily (Makarova and Grishin (1999)). They have shown that two distantly related families (ZnCAP and ZnCP) have acquired in parallel the ability to catalyse the same reactions after diverging from a common ancestor. In a more recent study, Gaskell and coworkers (Gaskell *et al.* (2009)) have found two almost identical proteins with the same bifunctional properties (tyrosine and and phenylalanine hydroxylases) in *Toxoplasma gondii*. These two functions are performed by two different enzymes in *H.*

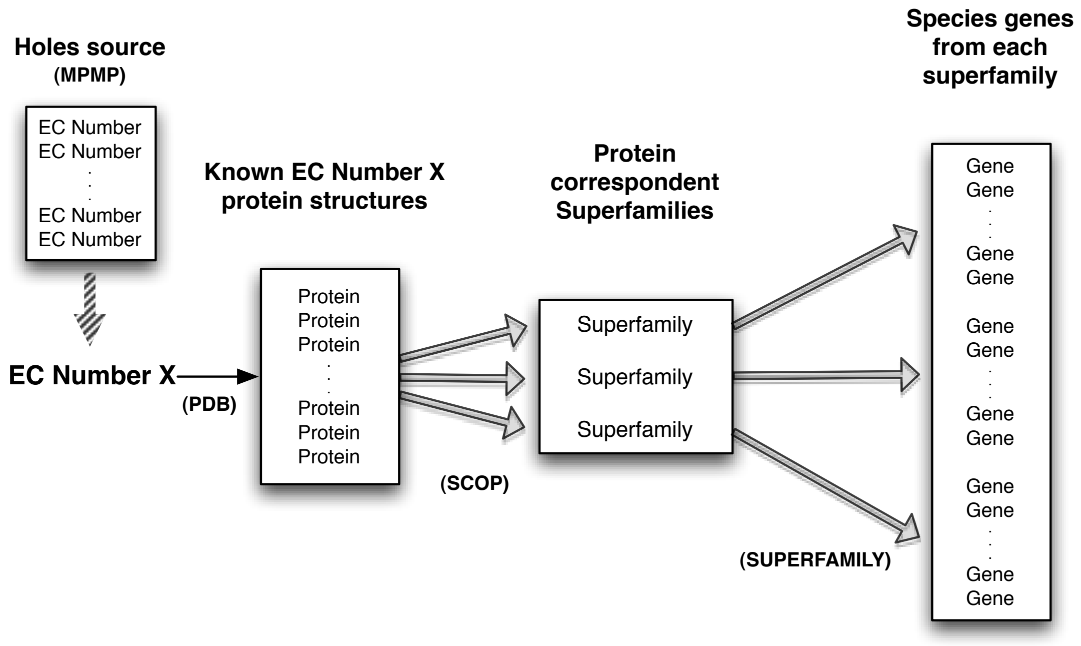


Figure 3.2: Hole candidate search protocol.

sapiens.

3.2 Methods

The protocol built makes use of three different databases: PDB, SCOP and SUPERFAMILY. The overall scheme is represented in Figure 3.2.

3.2.1 Protocol

Given an EC Number, the protocol starts by searching in the PDB for all known proteins known to perform that function. This search is made in a flat file from the PDB database with all proteins in the database mapped to an EC Number.

SCOP groups the proteins with known structure into three different levels of evolutionary relationships: Fold, superfamily and family (see Introduction). This information is accessible through a flat file that this protocol uses to identify the superfamily of each of the identified proteins in the PDB.

The SUPERFAMILY database provides HMM profile models built from the SCOP database structural assignments. These profiles are used to scan all the annotated genes in each species. All of this information is accessible in the form of a database. The protocol, given a species and a superfamily list, queries the SUPERFAMILY database for all genes assigned to this superfamily.

3.2.2 Known case

Dittrich and coworkers found a protein to fulfil a pathway hole (Dittrich *et al.* (2008)). This pathway hole is labelled with the enzymatic function EC4.1.2.25. Therefore, the protocol started by searching the PDB database for all known proteins with this enzymatic function assigned and following the steps presented in the previous section (Table 3.1.).

3.2.3 *Plasmodium falciparum* pathway holes

The input EC numbers used for the protocol were the pathway holes suggested by the *Plasmodium falciparum* manually curated metabolic network from the MPMP database (Ginsburg (2000)).

3.3 Results and Discussion

As a first step to accomplish the proposed objective, this study started to focus on the task of filling pathway holes. To find a candidate enzyme for a given hole, we need to use some nomenclature that enables us to identify the function that is missing and relate it to possible enzymes. For this reason, to tag the missing links (holes), the nomenclature used is the EC numbers (Webb *et al.* (1992)). As described in more detail in the introduction chapter, the Enzyme Commission number (EC number) is a hierarchical numerical classification scheme for enzymes, based on the chemical reactions they catalyse. Every enzyme code consists of the letters EC followed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzymatic function. Unlike for example GO terms, EC number is an enzyme-specific nomenclature.

Inspired by the Dittrich and coworkers example described above, I have constructed a computational protocol (Fig. 3.2) that, given an EC number assigned to a reaction node of the metabolic network of a given species, returns not only possible candidates using the philosophy mentioned above, but also phylogenetic trees of genes from the target species and evolutionarily closely related ones. This last part is explained in more detailed in the Chapter 5.

3.3.1 Candidate search protocol

For each required EC number, the protocol searches the PDB for all the proteins known to perform that function. Using the SCOP database (Murzin *et al.* (1995)), for each protein gathered, it returns the superfamily into which the protein is classified. The Structural Classification of Proteins (SCOP) database uses the nomenclature mentioned above (family, superfamily and fold). This database contains all PDB structures released before January 2005 plus a part of the PDB releases between January 2005 and October 2008, making it a good tool for the study objective.

The next step is, for each superfamily, to see which genes from the target species have domains classified to the given superfamily. These genes are the set of candidates to fill the hole. For this the protocol makes use of the SUPERFAMILY database (Gough and Chothia (2002)). This is a database of structural and functional protein annotations for all completely sequenced organisms using a set of profile hidden Markov models. Like SCOP, there are other databases that also group protein structures in to different evolutionary levels. A good example of that is the CATH database (Orengo *et al.* (1997)). CATH, compared to SCOP, makes more use of automated procedures in the classification of protein domains. Although there are some differences between the methodologies, the classifications are quite similar (Hadley and Jones (1999)).

However, SUPERFAMILY using SCOP provides a natural connection between structural classification and HMM superfamily profiles. The SUPERFAMILY database integrates these two components into a single relational database, making it a powerful tool for this chapter's objective.

Moreover, it is not possible to query directly SUPERFAMILY database with EC codes. Besides the curated superfamily assignments, SCOP works as a link between PDB EC number assignments and SUPERFAMILY HMM's models.

After having obtained a set of candidates for a pathway hole, any further evidence can be useful for discriminating between the candidates to see which may have a higher probability of being the correct one. Intuitively, any obvious properties in a network, for example dead ends or disconnected components, could be an indicator of how likely an annotation already assigned to a candidate gene is to be correct. In Figure 3.1 there are examples of both unpaired compounds and chokepoints. Considering all reactions as reversible, in Figure 3.1-A, the compound shown is only produced/consumed by one reaction. In Figure 3.1-B, the compound is an example of a chokepoint. It is only linked to two reactions. Therefore, these two reactions connected with this compound are chokepoint reactions. According to the definition presented in Yeh *et al.* (2004), the case shown in Figure 3.1-A would also be considered to be a chokepoint.

3.3.2 Known case

As a way to test if the gene candidate procedure was working well, it was applied to a former pathway hole investigated by Dittrich and coworkers (Dittrich *et al.* (2008)). The KEGG database has not yet updated this pathway. This pathway hole is labelled with the EC number 4.1.2.25. The protocol started by identifying all the known proteins in the PDB database with this function assigned. There were 56 proteins in these conditions. All these proteins were within the same superfamily identified by the id 55620. Afterwards, the protocol search for all the genes with domains assigned with this superfamily. The final results are presented in Table 3.1.

The protocol was able to successfully obtain two candidates that corresponded to the two top hits obtained by Dittrich and co-workers, including the experimentally validated enzyme (PFF1360), annotated as 6-Pyruvoyl tetrahydropterin synthase (PTPS). Analysing the unpaired and chokepoint compounds (see Table 3.1), it could also be seen that some compounds of PTPS were not produced/consumed in the current KEGG annotation. So, this kind of analysis together with others, can help to decide the most likely enzyme within a group of candidates.

3.3.3 *P. falciparum* case study

The same analysis was applied to the curated holes given by the MPMP database for the *Plasmodium falciparum* metabolic network. The Malaria Parasite Metabolic Pathways (MPMP) database is a set of the most likely metabolic enzymes in *P. falciparum*, curated by a world expert (Ginsburg (2000)). This database identifies a filtered list of reactions that are strongly believed to be present, making it much more valuable than information collected from the more general sites.

Hole	Hole function	Superf.	Seq. id	EC #	Assigned function	KEGG	
						UC	CP
4.1.2.25	dihydroneopterin aldolase (DHNA)	55620	PFF1360	4.2.3.12	6-pyruvoyltetrahydropterin synthase (PTPS)	1	0
			PFL1155w	3.5.4.16	GTP cyclohydrolase I (GTPCHI)	1	4

Table 3.1: Results from applying the search protocol to the pathway hole found by Dittrich *et al.* (2008). The first and second columns show the EC number and function assigned to the hole, respectively. The third column shows the superfamilies of the candidate genes found. The fourth, fifth and sixth columns show the candidate gene names, their currently assigned EC number and its function, respectively. The two last columns show if each candidate, in the KEEGs metabolic network, is assigned to a reaction with unpaired compounds (UC) or a chokepoint (CP). For each case, it gives the number of reactions.

Hole	Hole function	Superf.	Seq. id	EC #	Assigned function	UC		CP		evidence
						KEGG	MPMP	KEGG	MPMP	
1.4.4.2	Glycine dehydrogenase	53383	PFL0255c	2.9.1.2	OphosphoseryltRNA:selenium transferase	0	0	0	0	ISS
			PFD0285c	4.1.1.18	Lysine decarboxylase	0	0	0	0	ISS
			PF14_0155		None	-	-	-	-	
			PFL2210w	2.3.1.37	5aminolevulinate synthase	1	1	1	1	ISS
			PF07_0068		None	-	-	-	-	
			PF14_0534		None	-	-	-	-	
			PFB0200c	2.6.1.1	aspartate aminotransferase	1	1	2	2	ND
			MAL7P1.150	2.8.1.7	cysteine desulfurase	0	-	1	-	ISS
			PFF0435w	2.6.1.13	ornithineoxo-acid transaminase	0	1	1	1	TAS
			PFL1720w	2.1.2.1	glycine hydroxymethyltransferase	1	1	1	1	TAS
2.2.1.2	Transaldolase	51569	PF10_0210	4.1.2.4	deoxyribose-phosphate aldolase	0	0	0	0	ISS
2.5.1.54	3deoxy7phosphoheptulonate synthase		PF14_0381	4.2.1.24	porphobilinogen synthase	1	1	1	1	IDA
4.2.1.10	3dehydroquinate dehydratase		PF14_0425	4.1.2.13	fructosebisphosphate aldolase	1	1	0	1	TAS
2.7.1.23	NAD+ kinase	111331	PFI0650c		None	-	-	-	-	
4.2.1.75	UroporphyrinogenIII synthase	69618	PFL2285c		None	-	-	-	-	
1.1.1.25	shikimate dehydrogenase	53223	PF08_0132	1.4.1.2	glutamate dehydrogenase	1	1	0	0	ISS
			PFF1490w		None	-	-	-	-	
			PF14_0164	1.4.1.4	glutamate dehydrogenase (NADP+)	1	1	0	0	IDA
			PF14_0286	1.4.1.4	glutamate dehydrogenase (NADP+)	1	1	0	0	ND

Table 3.2: Results from applying the search protocol to the pathway holes proposed by Ginsburg (2000). The first and second columns show the EC number and its function assigned to hole, respectively. The third column shows the superfamilies of the candidate genes found. The fourth, fifth and sixth columns show the candidate gene names, their currently assigned EC number and its function, respectively. The seventh and eighth columns show if each candidate, in the KEEGs and MPMP metabolic networks, is assigned to a reaction with unpaired compounds (UC) and chokepoint (CP). For both, it says the number of reactions for each case. The last column indicates the GO evidence code for each gene enzymatic function assigned. IDA - Inferred from Direct Assay; ISS - Inferred from Sequence or Structural Similarity; TAS - Traceable Author Statement; ND - No biological Data available

The MPMP database currently contains 18 holes across the whole metabolic network (Ginsburg (2000)). Of these, the method was applied to those that had a complete EC number assigned. The candidates for each hole for which at least one candidate was found are shown in Table 3.2. Some of them already have an enzymatic function assigned. This function might have been assigned from similarity methods or validated experimentally. Ultimately, It will be important to be able to discriminate between assignments that have been experimentally verified and those that are assigned based on sequence similarity.

Analysing the results, the method was able to find candidates for 7 out of the 18 holes submitted. Among these 7 holes, 3 were within the same superfamily, so there were 5 different sets of candidates. For each hole, only one superfamily candidate set was found. These sets range from single genes to 10 superfamily members, where most already have annotated functions. The fact that for most of the pathway holes, there were not any candidates found may be a reflection of two main factors: The first one is that there is no closely related, well annotated model organism which means that the specific biochemical properties of *P. falciparum* may be very different from those previously observed. A second reason is the interaction between *P. falciparum* and its host: the parasite may be able to take some molecules from its host which may lead to the mistaken inference of missing enzymes.

Most of the candidates have an enzymatic function assigned. However, examining the last column of the tables, it is possible to verify that only two of them have experimental evidence of their existence and of their enzymatic function. These two proteins, PF14_0381 and PF14_0164 are therefore considered bad candidates for filling a missing link because they are already known to perform a different enzymatic function. Although there are examples of promiscuous enzymes capable of catalysing multiple reactions with a single active site, this type of enzymes is considered to be rare. Most of the other candidate's enzymatic function has been inferred by sequence similarity methodologies. As already discussed, this type of enzymatic assignment is prone to error, making them less reliable and at the same time increasing their chances of being the right protein for filling the missing link.

The seventh and eighth columns show the results of the unpaired and chokepoint analysis applied to the assigned function of each candidate. Comparing the topological position of the candidate annotation between KEGG and MPMP models there are not many differences. The most relevant one is the MAL7P1.150 gene that has been annotated only by KEGG with EC2.8.1.7. Nine of the enzymes catalyse reactions with compounds that are only consumed or produced by the reaction (unpaired compounds). This might be evidence that these reactions are in a dead end or belong to pathways that are not biologically meaningful to the *P. falciparum* metabolic network. At the same time, there are 11 enzymes responsible for reactions with chokepoint compounds. These reactions are essential for the pathway, as without them the pathway would have a hole.

These types of evidence, together with others, will help to rank all the candidates found for each of the missing links.

3.4 Conclusion

In this chapter I have shown many cases of convergent evolution and the importance of increasing the sensitivity search from a family level to a superfamily level. I have presented a protocol that makes use of this idea using well known public sources of different levels of enzyme information, such as: PDB for protein annotations, SCOP for protein structure classification and SUPERFAMILY for HMM superfamily profiles. The protocol presented in this chapter has not only shown to be successful in finding a missing link using a known test case but has also been able to provide candidates to a set of pathway holes suggested by a world expert in malaria. I have also shown that information about the different levels of enzymatic function assignments, together with topological properties and other types of information will have an important role in the ranking of the candidates found.

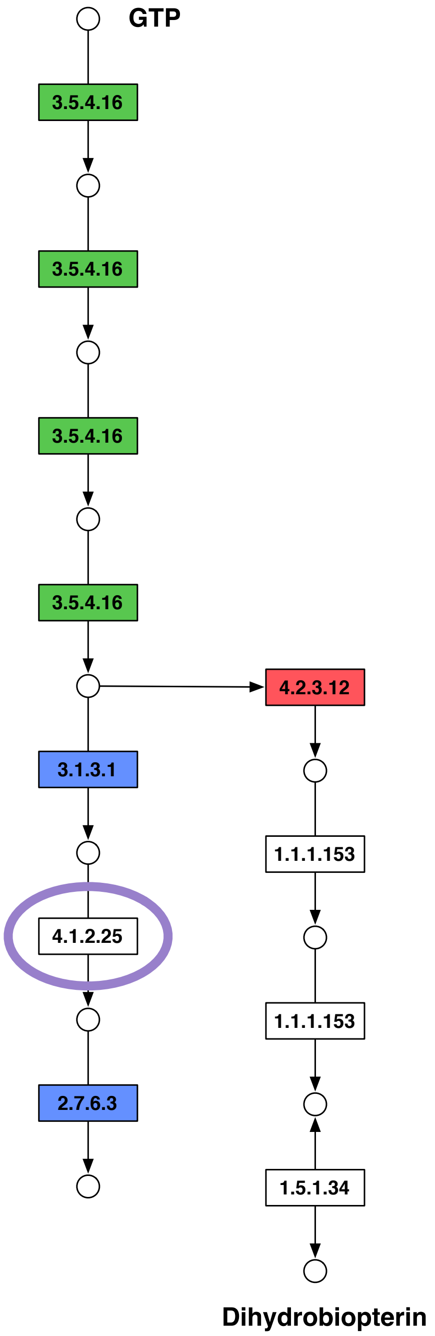


Figure 3.3: Part of the *Plamodium falciparum* Folate Biosynthesis pathway. Highlighted by a purple ellipse is the missing link referred by the paper Dittrich *et al.* (2008). All the reactions that do not have an enzyme assigned to them have a white background. The two main candidates found by the Dittrich study were originally assigned to the functions EC3.5.4.16 (PFL1155w) and EC4.2.3.12 (PFF1360w) that also belong to the Folate Biosynthesis pathway (with green and red background, respectively). The missing link experimentally validated is the gene PFF1360w. As can be seen, the reaction catalysed by PTPS (red background) is not supported by the rest of the pathway, being in a dead end.

Chapter 4

Topological analyses predict misannotations in a metabolic network

4.1 Introduction

Misannotation in sequence databases has been a recognised problem for more than a decade. Early studies reported the emergence of this issue (Galperin (1998); Brenner *et al.* (1999)) and estimated that up to 30% of proteins were misannotated in public databases Devos and Valencia (2001). More recent studies have confirmed that this problem is still a reality (Jones *et al.* (2007)) and some even suggest that it has been getting worse over time (Schnoes *et al.* (2009)), identifying over-prediction and error propagation as the main sources of error. Since experimental verification of gene function is expected to remain a highly time consuming process, it is unlikely that it will be able to keep pace with the increasing amount of genome sequence data being deposited in public databases. More accurate computational methods for functional annotation and assessment of confidence in gene annotations are therefore increasingly necessary.

In the area of automated functional annotation, several approaches moving beyond basic sequence similarity are now available (Jones *et al.* (2007)). Some recent annotation software will classify proteins based on locally conserved sequence patterns that are normally related with function (Forslund and Sonnhammer (2008)). Other approaches take into account the evolutionary relationships between proteins by integrating evidence across phylogenetic trees (Engelhardt *et al.* (2009)) or use additional information such as protein-protein interaction data (Ta and Holm (2009)) or genomic correlations (Hsiao *et al.* (2010)).

However, functional annotation is still mainly based on sequence similarity. Given this fact, the accuracy of existing annotations has a crucial impact on that of future annotations (Jones *et al.* (2007)). This dependency can lead to error propagation and a consequent increase in the number of annotation errors (Gilks *et al.* (2002)). Moreover, as information on the origin of annotation is often scarce, this error propagation does not have an easy solution. The problem becomes even clearer when we note that the proportion of manually annotated proteins is less than 5% and continues to decrease (Frishman (2007)).

Any evidence that is independent of sequence may therefore be useful for discriminating between true and false functional annotations. The concept of gene function implies interaction with some part of the cell or the environment, and almost all functions of interest are the result of interactions among several components (Hartwell *et al.* (1999)). Modeling these interactions by means of networks and studying their topological properties is therefore one way to understand the context of these molecular functions. Intuitively, any obvious problems in such a network, for example dead ends or disconnected components, could therefore be an indication of misannotation.

One easily accessible example of a well-defined molecular network derived from a set of gene annotations is a draft metabolic network, such as those available in the KEGG database (Kanehisa *et al.* (2008, 2006); Kanehisa and Goto (2000)). The topological properties of these networks have been studied previously in the contexts of network evolution (Wagner and Fell (2001)) and drug target discovery (Yeh *et al.* (2004)). For example, the metabolic networks of parasitic species are known to be distinguishable from non-parasitic species on the basis of their topology (Nerima *et al.* (2010); Borenstein and Feldman (2009)). In this Chapter, I propose a supervised machine learning methodology to assess the accuracy of assigned molecular functions, based on simple topological properties of an organism's draft metabolic network. I show that this approach is able to separate correct annotations from incorrect ones with an accuracy of up to 86%. Being entirely independent of sequence properties, it can be used to complement existing approaches and hence contribute to the detection and correction of errors in functional annotation.

No other studies that have tried to detect misannotations based only on metabolic network topological properties were found. However, there are studies that consider topological properties of protein-protein interaction networks. For example, Natasa Przulj's group (Milenkoviæ and Pržulj (2008)) demonstrates that local node topological structures and enzymatic biological function are correlated. They have used a clustering methodology together with 2 to 5 node graphlet vectors to group topologically similar proteins.

4.2 Methods

4.2.1 Metabolic networks

Bipartite (reaction and compound) graphs were used to represent metabolic networks, generated using the KEGG LIGAND database (Kanehisa *et al.* (2008)). To reconstruct the metabolic network for each species, all gene functions annotated for that species were collected. The reactions mapped to each function were then retrieved. Finally, the compounds attached to each reaction were added to produce a bipartite metabolic network for each species. All reactions were considered as being reversible. For each reaction were calculated 24 different features from which 22 were network topological features. Network topological properties comprising eccentricity and betweeness concepts were calculated using the NetworkX library in Python. The other features used elementary mathematical functions in Python.

4.2.2 Training data

Schnoes and coworkers previously examined the annotation errors in four large public protein databases (KEGG, GenBank NR, UniprotKB/TrEMBL and UniProtKB/SwissProt) (Schnoes *et al.* (2009)). Schnoes and coworkers work provides gold standard sets of correct and incorrect EC number assignments within 331 species in KEGG, across six enzyme superfamilies. In addition to sequence similarity approaches at the superfamily and family levels, the authors used information on functionally important residues to infer misannotations, making this one of the most reliable data sources suitable for our purposes. From their correct and incorrect annotation data, only the annotations with EC number identified were considered. In total there were 834 correct and 477 incorrect annotations considered. Each annotated function was mapped to a reaction according to KEGG. Where an EC function was mapped to more than one reaction, one of these was chosen at random.

4.2.3 Machine Learning

As with any supervised machine learning task, it is necessary to choose a machine learning method and a set of features from which to learn. Random forests (Breiman (2001)) were found to be a suitable machine learning approach for our aims. The advantages of using Random Forests in this work are their ability to process both numerical and categorical data and the interpretability of their output (a so-called ‘white box’ model). In contrast to other machine learning methods such as neural networks or support vector machines, random forests can provide insights into the signals that are useful for classification.

The approach used to separate correct from incorrect predictions was Random Forests, implemented in the randomForest R package (Liaw and Wiener (2002)). The random forests algorithm implemented is the one described in Breiman (2001). The parameters used in both the $randomForest$ and $predict$ functions were the default ones. The Random Forest classifier is a set of decision trees. Each testing entry is classified using all the trees present in the model. The probability of an entry being a misannotation is equal to the proportion of trees that have classified the entry as a misannotation. For building the ROC curves, the $type = "prob"$ option in the $predict$ function was used.

4.2.4 Features

In total, 22 different topological features were considered in training the classifier. These features can be placed into three broad groups: local, semi-local and global features (Table 4.1). Two other features were considered: domain (Archaea, Bacteria and Eukaryota) and whether or not the species is related to disease (Table 4.1-4E). The reason for considering these two features was to cope with potential topological differences between the domain and disease related categories. It has been shown that the network topology can be affected by the selection pressures applied by the environment during the evolutionary process. These pressures can sometime result in detectable topological properties (Parter *et al.* (2007); Borenstein and Feldman (2009); Kreimer *et al.* (2008)).

Local topological features capture the properties of the immediate neighbourhood of each reaction. Several of these features are related to the compounds involved in the reaction,

Group		Feature	Definition
1	A	m	Number of compounds connected to more than 2 reactions.
		u	Number of unpaired compounds.
		t	Reaction type: 1 - unpaired compounds on both sides of the reaction, 2 - unpaired compounds on only one side, 3 - no unpaired compounds.
		h	Number of chokepoint compounds.
		c	Number of compounds.
		$c_{<10}$	Number of compounds connected to more than 2 and less than 10 reactions.
		c_{10-50}	Number of compounds connected to 10 to 50 reactions
		$c_{>50}$	Number of compounds connected to more than 50 reactions.
		R	Number of other reactions sharing a compound with this reaction.
		$\bar{r}$	Mean number of other reactions connected to each compound.
	B	r_1	Number of connections of the least connected compound.
		r_2	Number of connections of the second least connected compound.
		r_3	Number of connections of the third least connected compound.
		r_4	Number of connections of the fourth least connected compound.
2	C	e	Eccentricity using unweighted edges,
		$\hat{e}$	Normalized eccentricity using unweighted edges.
		e_w	Eccentricity using weighted edges
		$\hat{e}_w$	Normalized eccentricity using weighted edges
		b	Betweeness using unweighted edges
		b_w	Betweeness using weighted edges
		N	Number of reactions in the connected component.
3	D	$t_{1,2}$	Fraction of reactions of type 1 or 2 in the network.
4	E	G	Domain: 1 - Bacteria, 2 - Eukaryota, 3 - Archaea.
		D	1 - species is related to disease, 0 - species is not related to disease.

Table 4.1: Feature analysis. The features were divided into 3 groups as shown in the first column: 1 - local topological features, 2 - semi-local topological features and 3 - global topological features.

each of which can be classified according to their connectivity (degree) as an unpaired, chokepoint or 'normal' metabolite. Based on this classification, several integer attributes were defined for each reaction (Table 4.1-1A). It was noticed that the connectivity of compounds involved in a reaction tends to vary depending on enzyme class, so four additional features were defined to capture this variation. These features correspond to the ranked connectivities of the reaction's four least-connected compounds (Table 4.1-1B)

The semi-local topological features describe the position of each reaction within the network. These features are based on the graph theoretical concepts of betweenness centrality and eccentricity. The betweenness of a node is the fraction of shortest paths (geodesics) between all pairs of nodes in the network that include that node, whilst the eccentricity of a node is the length of the longest geodesic between the node and all other nodes in the network. In both cases these values were also calculated including weights on the edges of the networks (Table 4.1-2C). Weighted metabolic networks have previously proved useful in the automatic identification of biologically meaningful pathways within a metabolic network (Croes *et al.* (2006)). This is a simple way to exclude spurious links via very highly connected compounds such as water or ATP. Here, we place a weight on each compound equal to its connectivity. To take variations in network size into account, a variant of eccentricity is also considered which is normalised by dividing by the diameter of the connected component to which the reaction belongs.

In addition to these reaction-based features, some global topological features of the network may be relevant if the amount of human curation varies between species. We use the proportion of reactions that have a dead-end compound on one or both sides as a proxy for the overall reliability of the network (Table 4.1-3D).

4.2.5 Validation of classifier

Fivefold cross validation

The cross validation process used was to start with the original data (D) and divide it into 5 equal sets. Each of the sets was used as an independent test set (D_{test}) . Random Forests considering all the features was applied to the remaining 4 sets (D_{train}). The Random Forests predictor built was then tested on D_{test}.

Inter-superfamily cross validation

The training data were grouped by enzyme superfamily. Owing to the paucity of data in most superfamilies, only the four most populated superfamilies were taken forward to cross-validation. Each superfamily in turn was removed from the balanced dataset SF to form the test set SF_{test}. Random Forests was applied to the remainder (SF_{train}). The model built was then tested on SF_{test}.

Final classifier

Random Forests were trained on the whole of the original data using all features. The $importance$ function from the randomForest R package was used to assess each feature's individual performance after training the model with the full learning set.

Comparison against curated models

To further validate the classifier, it was applied to 24 KEGG metabolic networks and the results compared with curated genome-scale metabolic models for these species (Table 4.4). The species used were the all genome models listed in Feist *et al.* (2009) for which the functions were labeled with EC numbers. For each KEGG model considered each annotated function was mapped to a reaction according to KEGG. Where a EC function was mapped to more than one reaction, one of these was chosen at random. The classifier was applied to this data. Afterwards, the results were compared with the curated models verifying the presence or absence in the curated models of the functions assigned in the KEGG models.

Tree of life analysis

Ciccarelli and coworkers Ciccarelli *et al.* (2006) reconstructed a highly resolved tree of life. Their species tree was built from a concatenation of 31 unambiguous orthologs present in 191 species. This tree and the multiple alignment used to build it were download from iTOL Letunic and Bork (2007, 2011). iTOL also provides other types of data related to these species, including genome sizes, domains per genome and publication dates. The multiple alignment was used to calculate the distances between the species using protdist from PHYLIP Felsenstein (1993), a package of programs for inferring phylogenies. The classifier was applied to the metabolic networks present in KEGG for each species included in the iTOL phylogeny.

4.3 Results/Discussion

To gain intuition of which topological features may have a greater influence in the results, the performance of each individual feature was evaluated independently. Histograms of the correct and incorrect annotation data provide a visual summary (Figure 4.3). A quantitative evaluation of each features performance was also obtained using the $importance$ function from the randomForest package Liaw and Wiener (2002). This function evaluates the accuracy increase and the entropy decrease for each feature (Figure 4.4). The column on the left shows the average increase of the accuracy after using each feature. The column on the right shows the average decrease of the entropy after using each feature.

All metrics show a similar ranking between the features, with those based on the concepts of betweenness and eccentricity seen to be the most highly predictive. The weighted network factor seems to improve the performance of both eccentricity and betweenness

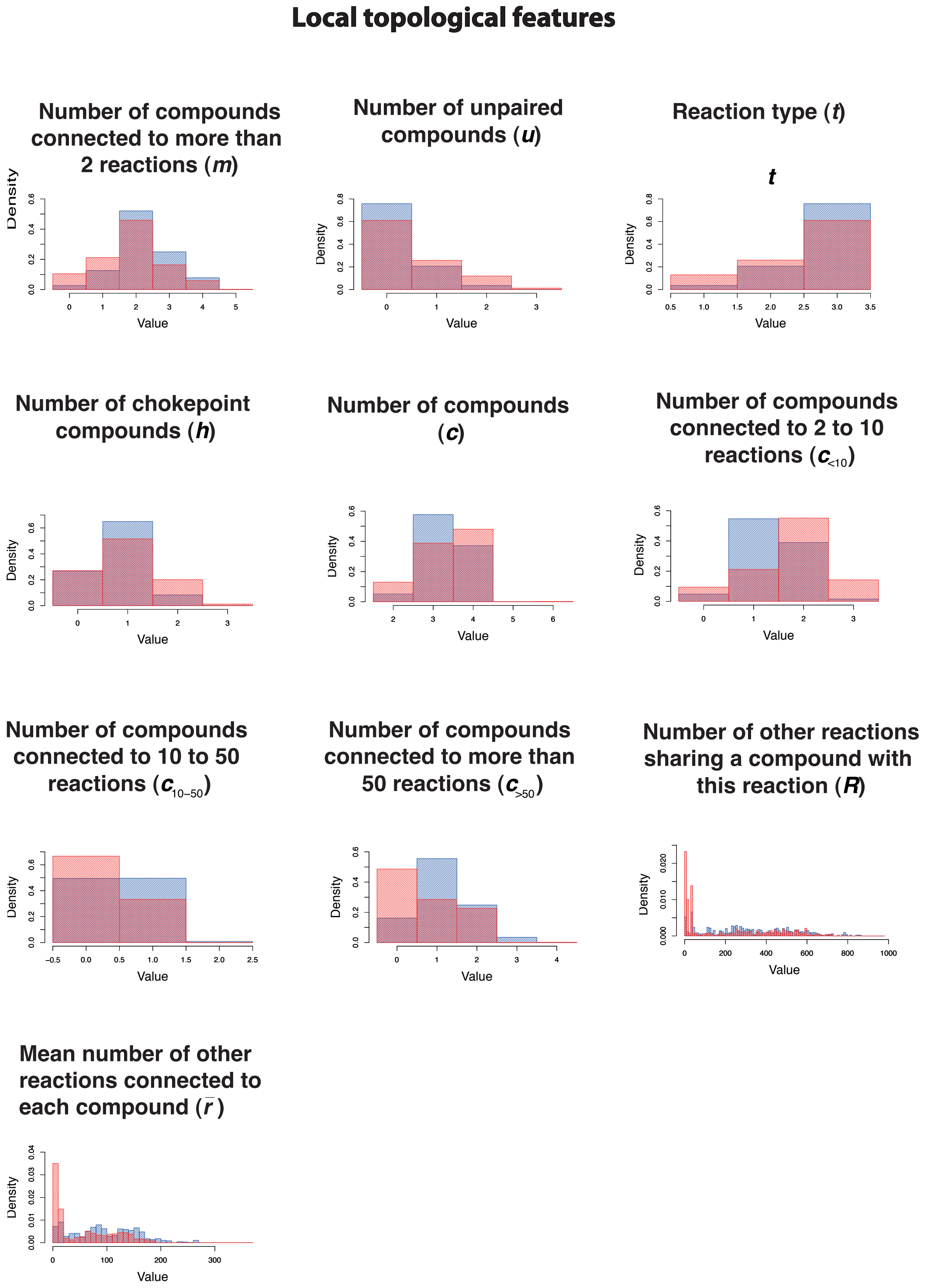


Figure 4.1: Feature histograms - Local topological features. Visualisation of the potential value of each attribute in distinguishing the correct functional assignments from the incorrect ones (red - incorrect annotations; blue - correct annotations).

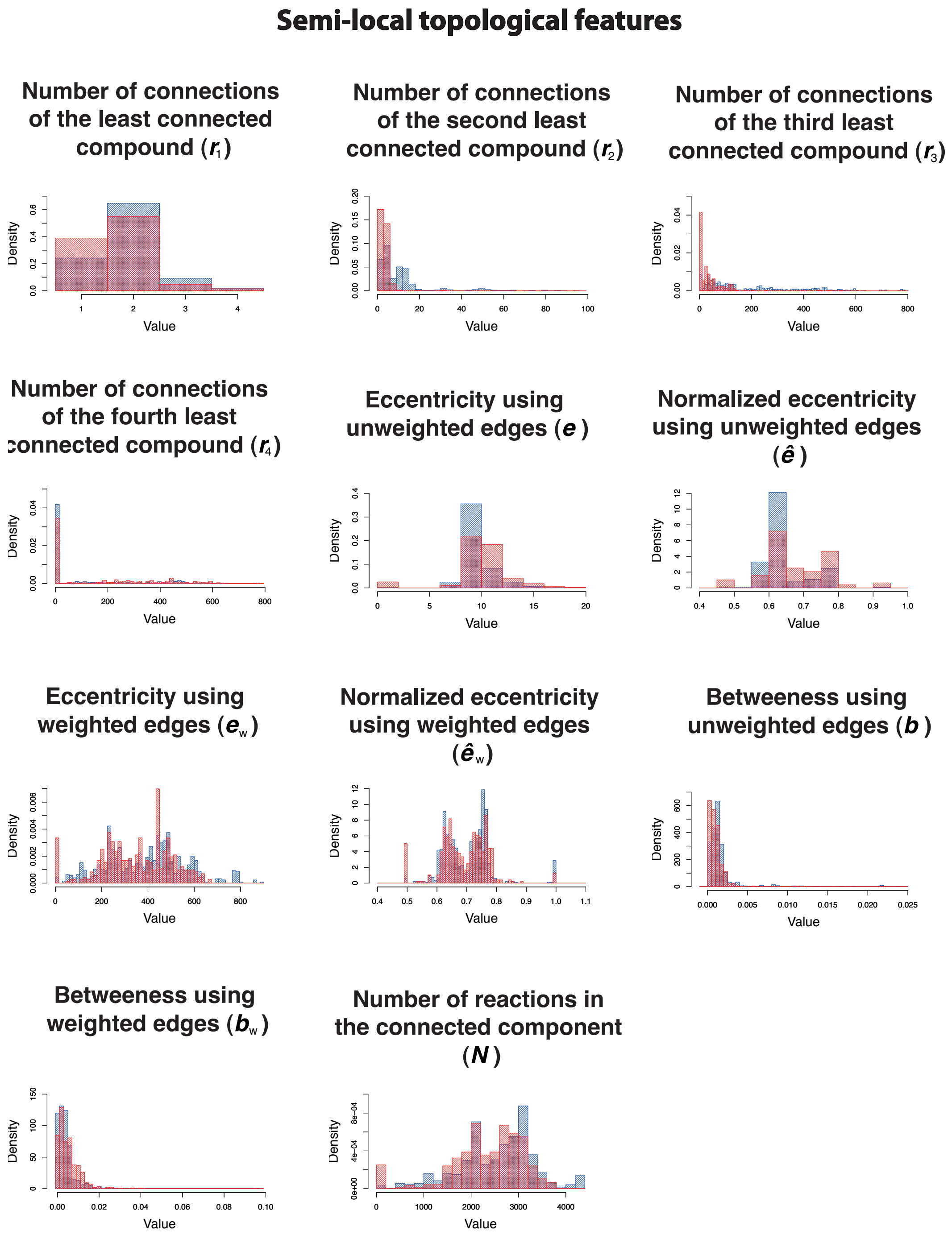


Figure 4.2: Feature histograms - Semi-local topological features. Visualisation of the potential value of each attribute in distinguishing the correct functional assignments from the incorrect ones (red - incorrect annotations; blue - correct annotations).

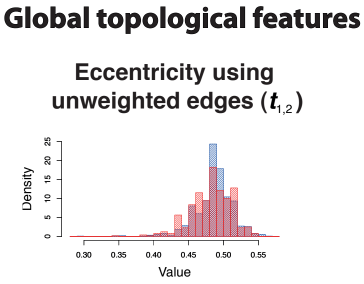


Figure 4.3: Feature histograms - Global topological features. Visualisation of the potential value of each attribute in distinguishing the correct functional assignments from the incorrect ones (red - incorrect annotations; blue - correct annotations).

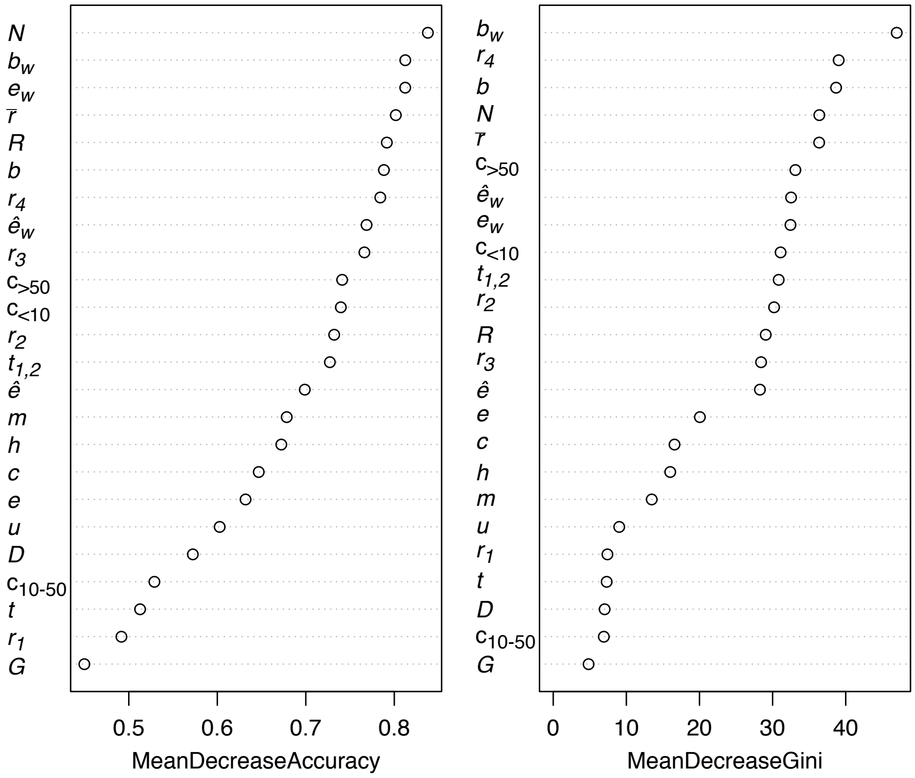


Figure 4.4: Feature predictiveness. These scores, obtained from the `importance` function of the randomForest R package, are used to assess the relative contribution of each feature to the performance of the predictor. left: average accuracy decrease when each feature is removed. right: average entropy decrease for each feature.

features, although it is more clearly seen in the case of eccentricity.

The taxonomic domain is the least informative feature. This may imply that the features already considered, such as the connected component size, may already be capturing any differences between species from different domains. The same might be happening with the disease-related feature. For example, parasitic species may be expected to have a larger number of unpaired compounds and smaller connected components, making this feature less informative. However, both features still show some predictive power.

In binary classifiers, such as the one presented here, accuracy values close to 50% on a balanced input data set show that the classifier is close to randomness and does not contain any information. The closer to 100% the more powerful the classifier is (Sonego *et al.* (2008)).The results for this classifier have been consistently around 60% showing that it returns some valuable information.

This classifier would not be suitable as a sole means to detect misannotations in a curated database such as Swiss-Prot, given the high rate of false positives that it would return (Baldi *et al.* (2000)). However, this classifier would be useful in discriminating between a set of possible candidates as shown with the Dittrich study example. Another possibility to consider is to use the classifier in an ensemble of methods. Given that it does not use sequence similarity searches, as most of the current methods do, it is likely to have a positive impact in performance.

4.3.1 Cross validation

The performance of the classifier on unseen data was assessed using two types of cross-validation. In fivefold cross-validation experiments (Table 4.2), the model obtained has an accuracy of almost 86%. Figure 4.5-A shows the receiver-operator characteristic (ROC) curves obtained for each of the cross validation folds. The mean area under the ROC curve

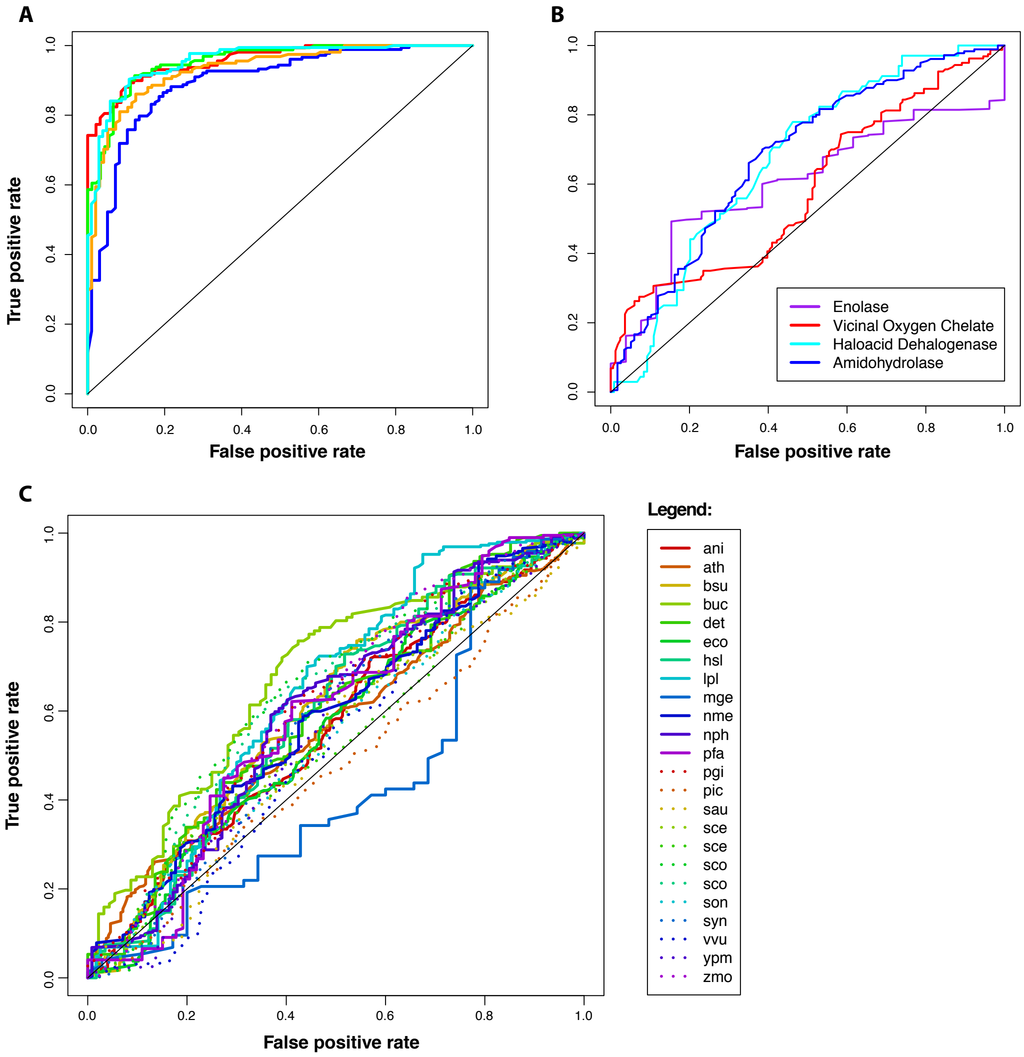


Figure 4.5: Cross validation ROC curves. A- fivefold cross-validation ROC curves; B- superfamily cross-validation ROC curves; C- classifier performance in curated models ROC curves

Accuracy	Precision	Recall	AUC
0.86	0.91	0.88	0.92

Table 4.2: 5 fold cross validation results. The predictive model performance was assessed by a 5 fold cross validation. The table shows the accuracy, precision, recall and AUC of this analysis.

Superfamily	Accuracy	AUC
Enolase	0.60	0.60
Vicinal Oxygen Chelate	0.52	0.59
Haloacid Dehalogenase	0.60	0.67
Amidohydrolase	0.66	0.68

Table 4.3: Superfamily cross validation results. To test performance on unseen enzyme classes, the classifier was assessed in a leave-one-out cross validation at the superfamily level. The table shows the accuracy and the AUC of each analysis, where each superfamily in turn was used as the test data set.

(AUROC) was 0.92%. Another important aspect of performance is how well the predictor would be expected to perform on enzymes from unseen superfamilies. To this end, a second cross validation was performed, using as a training set the enzymes for three out of the four superfamilies with data and testing on the enzymes from the fourth (Table 4.3). In this experiment, with the exception of the Vicinal Oxygen Chelate superfamily, the accuracy of the predictor was consistently above 60%. Figure 4.5-B shows the ROC curves for each superfamily. The area under the curve varied between 0.59 and 0.68. These results suggest that the functional classes covered in the training data do have an effect on the rules obtained. For example, enzyme classes may occupy topologically distinct positions in the network, and/or be subject to particular types of misannotation. However, these results indicate that the classifier trained on the entire available data set should still be informative when applied more generally.

4.3.2 Comparison to a manually curated network

In order to assess the performance of the model, the classifier was applied to 24 KEGG genome annotations. These results were compared with recent manually curated genome-scale metabolic models as gold standards (Table 4.4 and Figure 4.5-C). The species used were the full genome models listed in Feist *et al.* (2009) for which the functions were labeled with EC numbers.

The AUC results were almost entirely above 0.5, showing a performance better than random. In fact, in almost half of the species tested the classifier produced an AUC of 0.6 or above. There were only two cases where AUC was found to be below 0.5. The worst result was found with *Mycoplasma genitalium*, perhaps related to the fact that this is the smallest prokaryote genome sequenced.

4.3.3 An atypical orthologue case

An interesting example of the successful identification of an unexpected enzyme function is given by Dittrich and co-workers (Dittrich *et al.* (2008)). This work was based on the idea that an evolving enzyme has more chance to acquire the function of structurally similar enzymes. A bioinformatic protocol was followed to draw up a shortlist of candidate functional analogs of a missing enzyme (dihydroneopterin aldolase, DHNA) in the *Plasmodium falciparum* folate biosynthesis pathway.

During the process, the authors found two candidates for filling the role of the missing enzyme. Both enzymes already had an assigned function in KEGG: PFF1360w is annotated as a putative 6-Pyruvoyl tetrahydropterin synthase (PTPS, EC4.2.3.12) and PFL1155w as GTP cyclohydrolase I (GTPCH-I, EC3.5.4.16). Although PFF1360w was subsequently experimentally validated as performing the missing DHNA function, KEGG has not yet

KEGG ID	**Species name**	**AUC**	citation
ani	*Aspergillus nidulans*	0.56	David *et al.* (2008)
ath	*Arabidopsis thaliana*	0.57	de Oliveira Dal'Molin *et al.* (2010)
bsu	*Bacillus subtilis*	0.61	Oh *et al.* (2007)
buc	*buchnera aphidicola*	0.68	Thomas *et al.* (2009)
det	*Dehalococcoides ethenogenes*	0.60	Islam *et al.* (2010)
eco	*Escherichia coli* K-12	0.55	Reed *et al.* (2003)
hsl	*Halobacterium salinarum*	0.60	Gonzalez *et al.* (2008)
lpl	*Lactobacillus plantarum*	0.64	Teusink *et al.* (2006)
mge	*Mycoplasma genitalium*	0.43	Suthers *et al.* (2009)
nme	*Neisseria meningitidis*	0.58	Baart *et al.* (2007)
nph	*Natronomonas pharaonis*	0.60	Gonzalez *et al.* (2010)
pfa	*Plasmodium falciparum*	0.59	Plata *et al.* (2010)
pgi	*Porphyromonas gingivalis*	0.60	Mazumdar *et al.* (2009)
pic	*Pichia stipitis*	0.48	Caspeta *et al.* (2012)
sau	*Staphylococcus aureus*	0.52	Lee *et al.* (2009)
sce	*Saccharomyces cerevisiae*	0.56	Herrgård *et al.* (2008)
sce	*Saccharomyces cerevisiae*	0.53	Förster *et al.* (2003)
sco	*Streptomyces coelicolor*	0.64	Borodina *et al.* (2005)
sco	*Streptomyces coelicolor*	0.63	Alam *et al.* (2010)
son	*Shewanella oneidensis*	0.55	Pinchuk *et al.* (2010)
syn	*Synechocystis* PCC6803	0.57	Nogales *et al.* (2012)
vvu	*Vibrio vulnificus*	0.52	Kim *et al.* (2011a)
ypm	*Yersinia pestis*	0.55	Navid and Almaas (2009)
zmo	*Zymomonas mobilis*	0.61	Widiastuti *et al.* (2011)

Table 4.4: Genome-scale model validation results. The final classifier was applied to KEGG metabolic models and the results compared with curated genome-scale metabolic models for these species.

updated this annotation. This enables us to apply the classifier to the KEGG *Plasmodium falciparum* metabolic network to study this case.

Taking a closer look at the two annotated reactions in their network context (Figure 3.3), it can be seen that the PTPS reaction appears to be a dead end, indicating that this annotation is unlikely to be correct. In contrast, the GTPCH-I enzyme not only has its reactants produced and its products consumed, but is also assigned to four chokepoint reactions.

Applying our classifier to these two enzymatic functions, it returned a probability of 0.94 for the GTPCH-I reaction, indicating that this function seems to make biological sense within its network context. On the other hand, the PTPS reaction scores only a probability of 0.21 to be a correct annotation. This simple case study shows that the classifier has successfully captured the same network topological features that provided evidence for an incorrect annotation in the published manual analysis of this enzyme.

4.3.4 Holes revisited

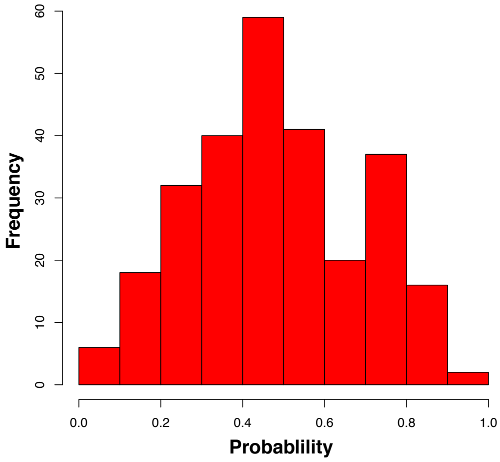


Figure 4.6: Classifier probability distribution for KEGG's enzymatic annotations for *Plasmodium falciparum*

The protocol presented in the previous chapter had beed applied to a set of suggested holes from a manually curated metabolic model of *Plasmodium falciparum*. To decide which of the candidates for each hole is most likely to be the best candidate we need to gather as much additional information as possible.

The classifier presented in this chapter can be one of those sources of information. Possibly the best candidates are the genes that do not have any function assigned. However, if an enzymatic function assigned to a gene is classified as unlikely to exist, this gene also becomes a strong candidate. Table 4.5 shows the quality annotation score for each of the annotated functions assigned from KEGG for each of the candidates. The genes without any enzymatic function assigned were not included in this table.

According to KEGG, the gene MAL7P1.150 has the enzymatic function cysteine desulfurase with EC number EC2.8.1.7, although this gene was not annotated in the MPMP database. Given that MPMP is a curated database by an expert in malaria, this annotation is likely to be incorrect. The classifier presented in this chapter has given a score of 0.17 to this gene being in accordance with its KEGG annotation. For the hole with the enzymatic function EC1.4.4.2, there are candidates with relatively high probabilities for their annotated functions. The two with highest scores are PFD0285c and PFL2210w with 0.716 and 0.604, respectively. Therefore, from a topological perspective, these two genes would be the two worse candidates, because their assigned functions seem to be biological meaningful. On the other hand, MAL7P1.150, as already mentioned, has a probability of 0.17. Figure 4.6 shows this value is placed in the extreme left part of the distribution, which makes its annotation likely to be incorrect, making this gene a good candidate to perform the missing enzymatic function.

For the holes EC2.2.1.2, EC2.5.1.54 and EC4.2.1.10 a similar situation is present. On one side there is PF14_0381 with a probability of 0.754 and on the other side is PF10_0210 with a probability of 0.228. Therefore, for the same reason presented above, from a topological perspective, PF10_0210 would the best candidate from the three genes presented in the table.

4.3.5 Comparison of predicted annotation quality across multiple species

To investigate how annotation quality varies between species, the classifier was applied to the KEGG metabolic networks of the species present in the tree of life provided by iTOL (Letunic and Bork (2007, 2011)). The proportion of enzymatic functions predicted

Hole	**Superfamily**	**Seq. id**	**EC Numb**	**UC**		**CP**		Score
				KEGG	MPMP	KEGG	MPMP	
1.4.4.2	53383	PFL0255c	2.9.1.2	0	0	0	0	0.208
		PFD0285c	4.1.1.18	0	0	1	0	0.716
		PFL2210w	2.3.1.37	1	1	1	1	0.604
		PFB0200c	2.6.1.1	1	1	2	2	0.282
		MAL7P1.150	2.8.1.7	0	-	1	-	0.17
		PFF0435w	2.6.1.13	0	1	1	1	0.578
		PFL1720w	2.1.2.1	1	1	1	1	0.41
2.2.1.2	51569	PF10_0210	4.1.2.4	0	0	0	0	0.228
2.5.1.54		PF14_0381	4.2.1.24	1	1	1	1	0.754
4.2.1.10		PF14_0425	4.1.2.13	1	1	0	0	0.594
1.1.1.25	53223	PF08_0132	1.4.1.2	1	1	0	0	0.504
		PF14_0164	1.4.1.4	1	1	0	0	0.71
		PF14_0286	1.4.1.4	1	1	0	0	0.71

Table 4.5: The result for MPMP holes. The final classifier was applied to the pathway holes suggested by MPMP. The first column represents the hole's EC number. The second column indicates the superfamily to which the candidates belong. The $Seq.id$ column are the gene IDs of each candidate. The fourth column shows the enzymatic function with which each candidate was annotated, according to KEGG. The fifth and sixth columns show the number of reactions with unpair compounds each enzyme catalyses according to KEGG and MPMP, respectively. The seventh and eighth columns show the number of reactions with chokepoint compounds each enzyme catalyses according to KEGG and MPMP, respectively. In the last column are the classifier probability results for each of the candidate.

to be correctly annotated in the network of each species (i.e. the predicted precision of the set of enzymatic functions reported by KEGG for that organism) was taken as a measure of annotation quality. Figure 4.7 shows the prokaryote phylogenetic tree and quality scores for each of the species. The *E. coli* strains and the most closely related species produce the highest scores, indicating their higher levels of curation. With the exception of Chlamydiae/Verrucomicrobia and the Cyanobacteria, all phyla show a wide variety of accuracy scores.

The number of eukaryotic species provided by iTOL is much smaller than the number of prokaryotes. Figure 4.8 shows the eukaryote phylogenetic tree and the quality scores of the KEGG metabolic networks for each of the species. The vertebrates and plants produce higher scores than the other species. A possibly unexpected result is the relatively low score reported for the yeast *Saccharomyces cerevisiae* and the fruit fly *Drosophila*

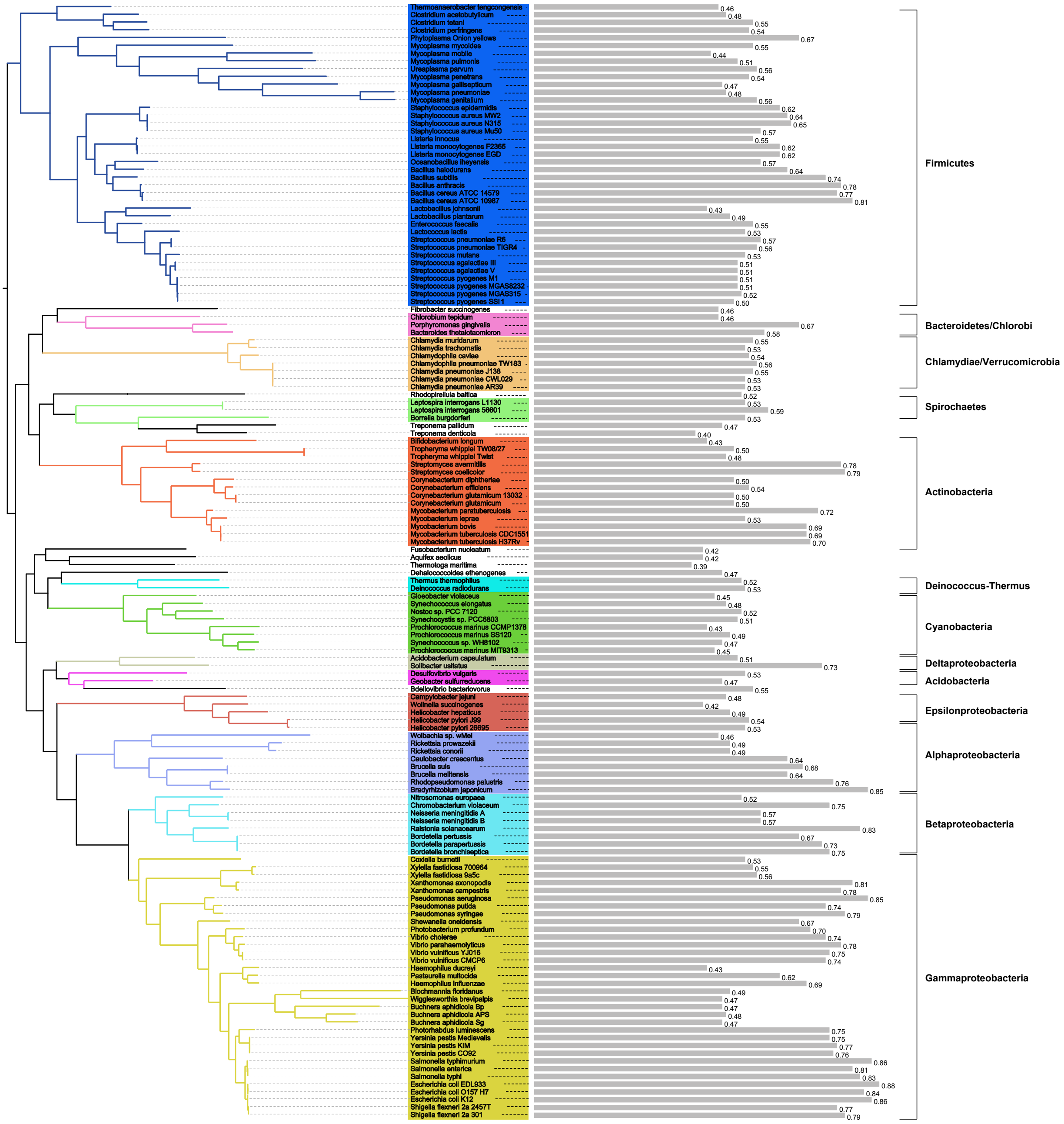


Figure 4.7: Predicted quality of draft metabolic networks across a prokaryote phylogeny. The classifier was applied to all prokaryote species present in the iTOL phylogeny (Letunic and Bork (2007, 2011)). Coloured clades represent the different phyla present (only phyla with more than 1 species were coloured). The names of the phyla are shown to the right. Predicted annotation quality values are represented by grey bars next to the species name.

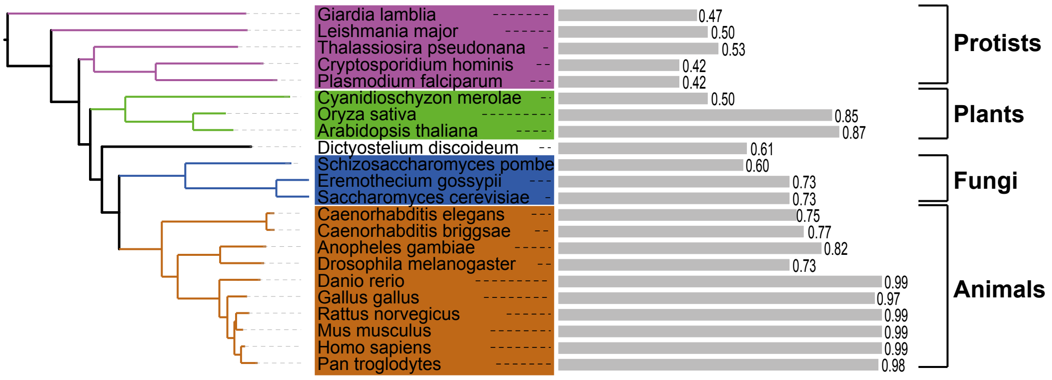


Figure 4.8: Predicted quality of draft metabolic networks across a eukaryote phylogeny. The classifier was applied to all eukaryote species present in iTOL. To the left is the eukaryote phylogenetic tree. The quality values are represented by bars next to the species names.

melanogaster (both 0.73), especially when compared with those achieved by the vertebrates. However, this most probably reflects the massive amount of study that human biochemistry has received relative to any other eukaryote, including these two important model organisms.

It is reasonable to expect that the quality of a draft metabolic network should be better for species that are closely related to organisms with well characterised biochemistry. Figure 4.9 shows that this is indeed the case: there is a clear negative correlation ($R^2 = 0.393$) between the predicted annotation quality in prokaryotes and the phylogenetic distance to *E. coli* and an even stronger negative correlation ($R^2 = 0.779$) between the predicted annotation quality in eukaryotes and the phylogenetic distance to *H. sapiens*.

To check for any dependency between annotation quality and genome size, a similar scatter plot was drawn (Figure 4.10). Although a positive correlation appears to be present, this may be explainable by other factors. In particular, the intracellular obligate species (highlighted in green in Figure 4.10) and the well curated species (highlighted in orange), constituted by the *E. coli* strains and very closely related species (*Salmonella* and *Yersinia*), have distinctly low and high quality scores, respectively. Since intracellular obligate species will tend to have lost many genes that are necessary for free-living or-

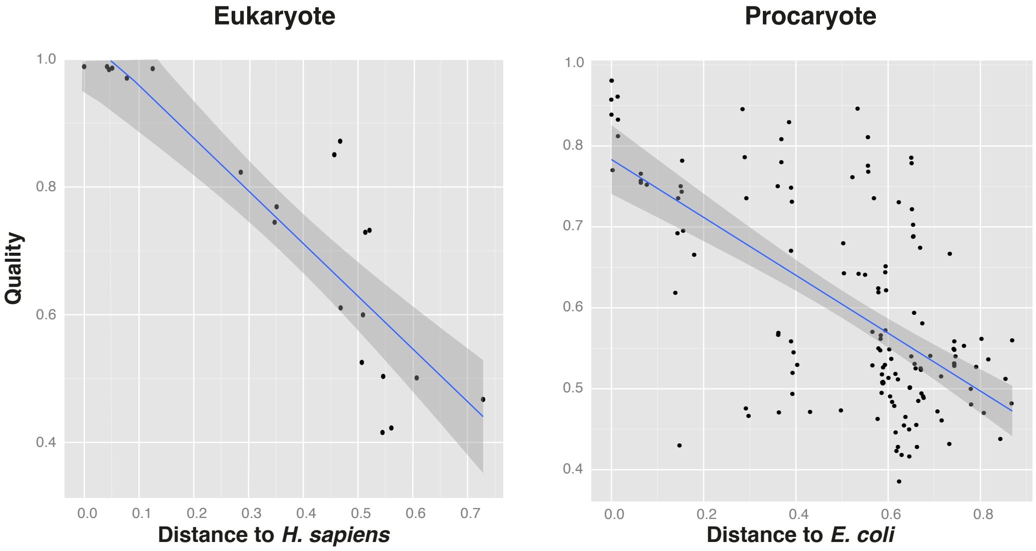


Figure 4.9: Variation of predicted quality of draft metabolic networks. left: Scatter-plot showing predicted annotation quality (precision of annotated reactions according to the classifier) for eukaryotes against phylogenetic distance to *H. sapiens*. right: Scatter-plot showing predicted annotation quality (precision of annotated reactions according to the classifier) for prokaryotes against phylogenetic distance to *E. coli* (Ciccarelli *et al.* (2006)).The shaded region shows the 95% confidence interval for the regression line.

ganisms Ochman and Moran (2001), their genomes are smaller than average: intracellular obligates are almost exclusively at the bottom left of the plot. The low quality scores for this group of species (Figure 4.10) may indicate either an increased difficulty in reconstructing their metabolic networks by automatic methods or simply the known general topological differences between their metabolic networks and those of the other prokaryotes (Ochman and Moran (2001)). These two groups of species tend to enhance the correlation between predicted annotation quality and genome size. Without these species the correlation becomes slightly weaker (changing from $R^2 = 0.51$ to $R^2 = 0.48$).

In addition to the intracellular obligates and well-studied bacteria, the box plots in Figure 4.11 show the predicted annotation quality for two further sets of species: those with available manually curated genome-scale reconstructions (GENREs, Price *et al.* (2004)) and those that are facultatively intracellular. We can clearly see the low quality scores in

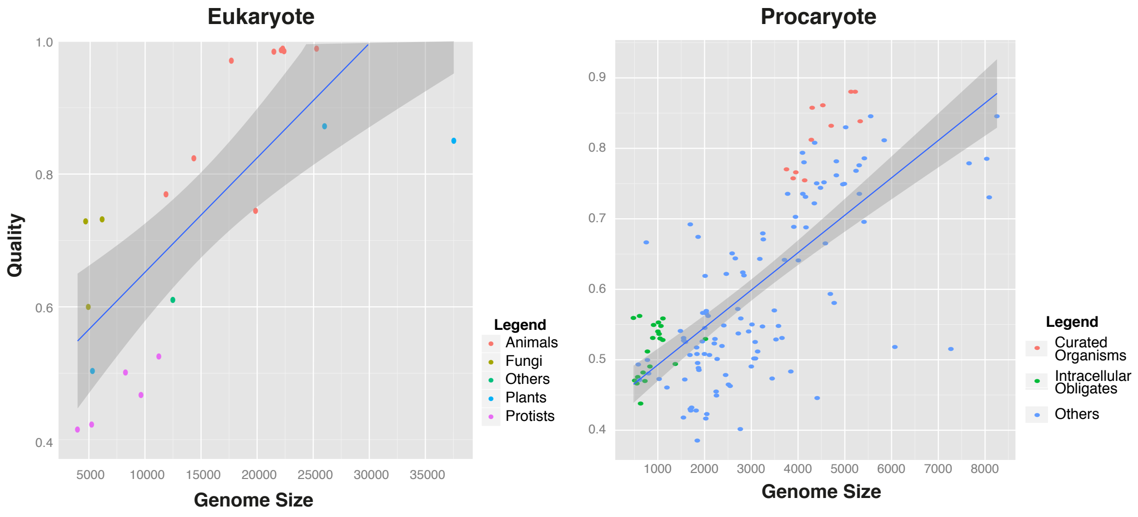


Figure 4.10: Variation of predicted quality of draft metabolic networks. left: Scatter-plot showing predicted annotation quality against genome size in eukaryotes: The species are divided in Animals, Fungi, Plants, Protists and others. The shaded region shows the 95% confidence interval for the regression line. right: Scatter-plot showing predicted annotation quality against genome size: orange - well studied species (*E. coli* strains and the very closely related species (*Salmonella* and *Yersinia*)); green intracellular obligate species. The shaded region shows the 95% confidence interval for the regression line.

the obligate (though not the facultative) intracellular species (p-value=1.158e-08) and the high accuracy scores in the well-studied species set (p-value=3.055e-06). However, the extra curation possibly provided by the existence of a GENRE is not seen to be reflected in the semi-automated annotations within KEGG.

For prokaryotes, possible dependencies on other species attributes were also considered: motility, phylum, pathogenicity, oxygen requirement and habitat (Figure 4.11). The quality scores do not appear to depend on these attributes, with the exception of habitat: the species living in specialised habitats have lower accuracy scores compared to all other species (p-value=4.334e-08). As stated above, specialised environments may be responsible for differences in selective pressures that could result in detectable differences in metabolic network topologies.

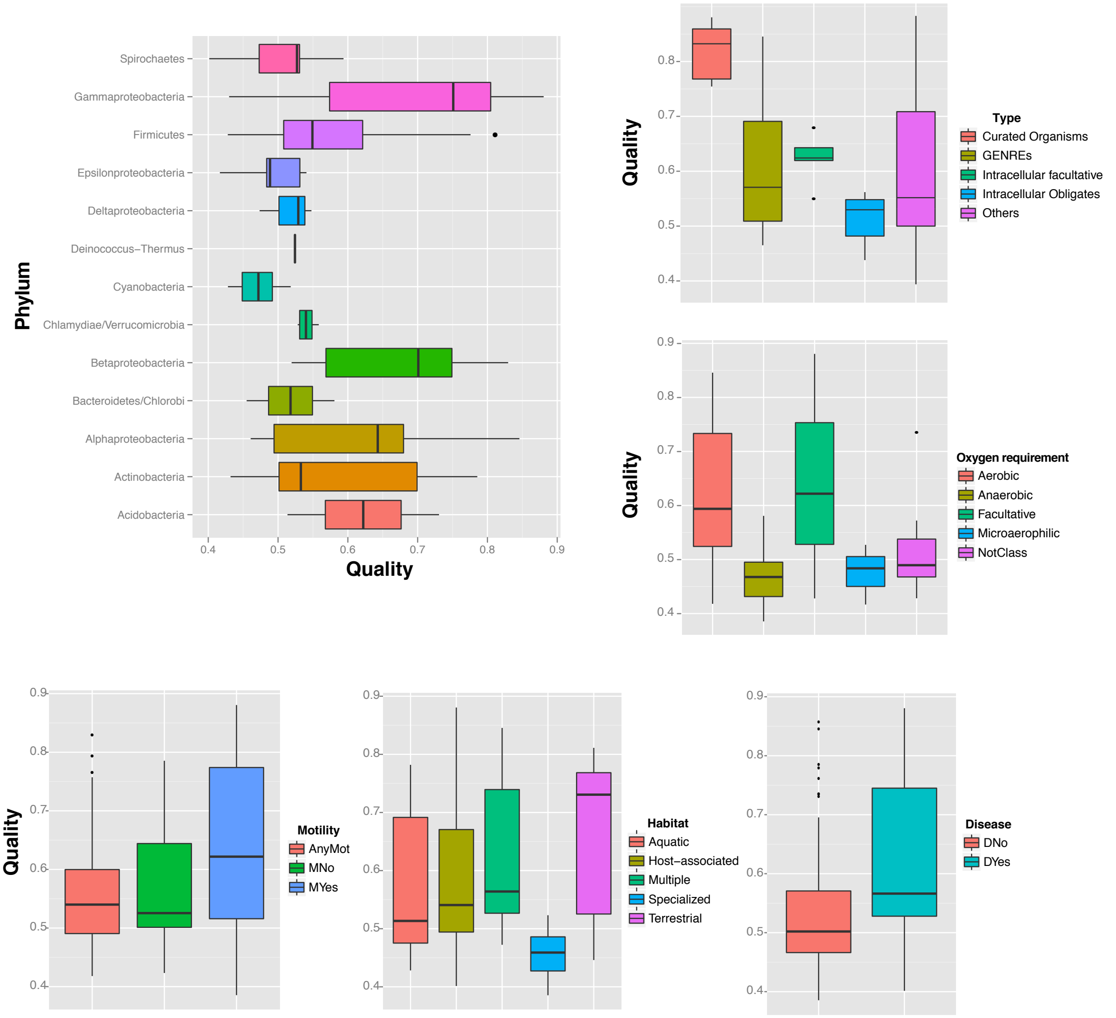


Figure 4.11: Variation of predicted quality of draft metabolic networks. Box-plot of the distribution of the quality scores in different sets of prokaryote species: orange and light green - same as in B; purple species for which there is a GENRE (Price *et al.* (2004)) available; dark green facultative intracellular species; blue all the species.

The possible link between annotation quality and genome size was also checked in eukaryotes. According to Figure 4.10 a positive correlation could be present. However, closer inspection shows that there are two well defined groups that contribute to this correlation. Towards the bottom left (small genomes, low annotation quality) are the protists and the fungal species and at the top right are a group of animals (mostly vertebrates) and plants. Taken together with the fact that the number of species present is small, there does not appear to be any strong evidence for a direct link between genome size and annotation quality.

For both eukaryotes and prokaryotes other possible dependencies were also studied. Examples of these were the number of publications existent for each species and the year that the models considered were published. However there were no significant correlations between the quality of the model and these factors (data not shown).

4.4 Conclusion

The study presented in this chapter has demonstrated that simple topological features can be used to predict incorrect functional annotations within metabolic networks. The Random Forest classifier has not only achieved high overall cross validation accuracy but has also been shown to be informative when applied to enzymes belonging to superfamilies that were not used in training. This approach is entirely independent of sequence properties, so could be used to support automated metabolic reconstruction pipelines as well as helping to identify incorrectly annotated enzymes within public databases.

For both prokaryotes and eukaryotes, it appears that the quality of automated metabolic reconstruction decreases with phylogenetic distance to the major model organism for biochemistry, *E. coli* and human, respectively. However, differences in network topology between free-living organisms and obligate intracellular species may make the classifier less accurate when applied to the latter group of species. Given a larger amount of training data, it should be possible to produce separate classifiers for each of these two groups.

Chapter 5

Phylogenetic analyses predict functional changes in an enzyme superfamily

5.1 Introduction

Genome evolution is the result of complex events such as gene duplication, loss and speciation. These events are essential for species to adapt to the environment and the identification of these events is fundamental in order to understand the emergence and loss of cellular functions. For example, a gene is constantly suffering mutations and when a gene gets duplicated, the genome then has two identical genes whose encoded proteins have the same function. This may result in the accumulation of mutations in one or both sequences without affecting the fitness of the organism. These mutations can cause changes in the 3D structure of the protein which may result in the acquisition of a new function or the subfunctionalisation of existing functions (Tokuriki and Tawfik (2009); Nowak (1997)).

Moreover, the identification of conserved residues may be a decisive resource to assign functions to currently unannotated proteins. This conservation is a result of evolutionary constraints suggesting that changes on these residues may cause severe changes in the protein. Residues that directly interact with the ligand or that are related with the protein structure tend to be more conserved during evolution. Hence, the presence of similar conserved residues may indicate that two proteins preform similar functions. Therefore, to make more accurate functional transfer it is important to understand how new enzymatic functions evolve from existing ones.

There has been already some research regarding the understanding of the evolutionary mechanisms. Good examples are the SCOP (Murzin *et al.* (1995)) and CATH (Orengo *et al.* (1997)) databases that attempt to organise the several known domains into several levels using evolutionary relations. These databases are used as the basis for studies of homolog gene searches, as for example shown in Chapter 3. At a more complex level there is FunTree. FunTree brings together several data sources important for the study of function evolution. Publicly accessible, it uses the evolutionary levels groups built by CATH, together with functional information from reliable sources such as UniProtKB (Boutet *et al.* (2007)) and enzyme reaction mechanisms from MACiE (Holliday *et al.* (2005)), taking also into consideration catalytic residues from the Catalytic Site Atlas (Porter *et al.* (2004)). It groups the proteins by superfamily domains and structurally similar groups and presents them in a phylogenetic perspective. These types of resources allow the study of divergence and convergence of enzymatic functions, functional diversification within and across superfamilies, and the relationship between structure and catalytic function.

This chapter provides an evolutionary perspective on gene function annotation. Phylogenetic trees will be the basis for our comparative genomic analyses and allow us to analyse the evolution of homologous domains between closely related species. Moreover, phylogenetic trees can help us to understand how and when a group of genes have acquired a different function from their orthologues and, for example, detect cases of convergent evolution like the one identified in Dittrich *et al.* (2008).

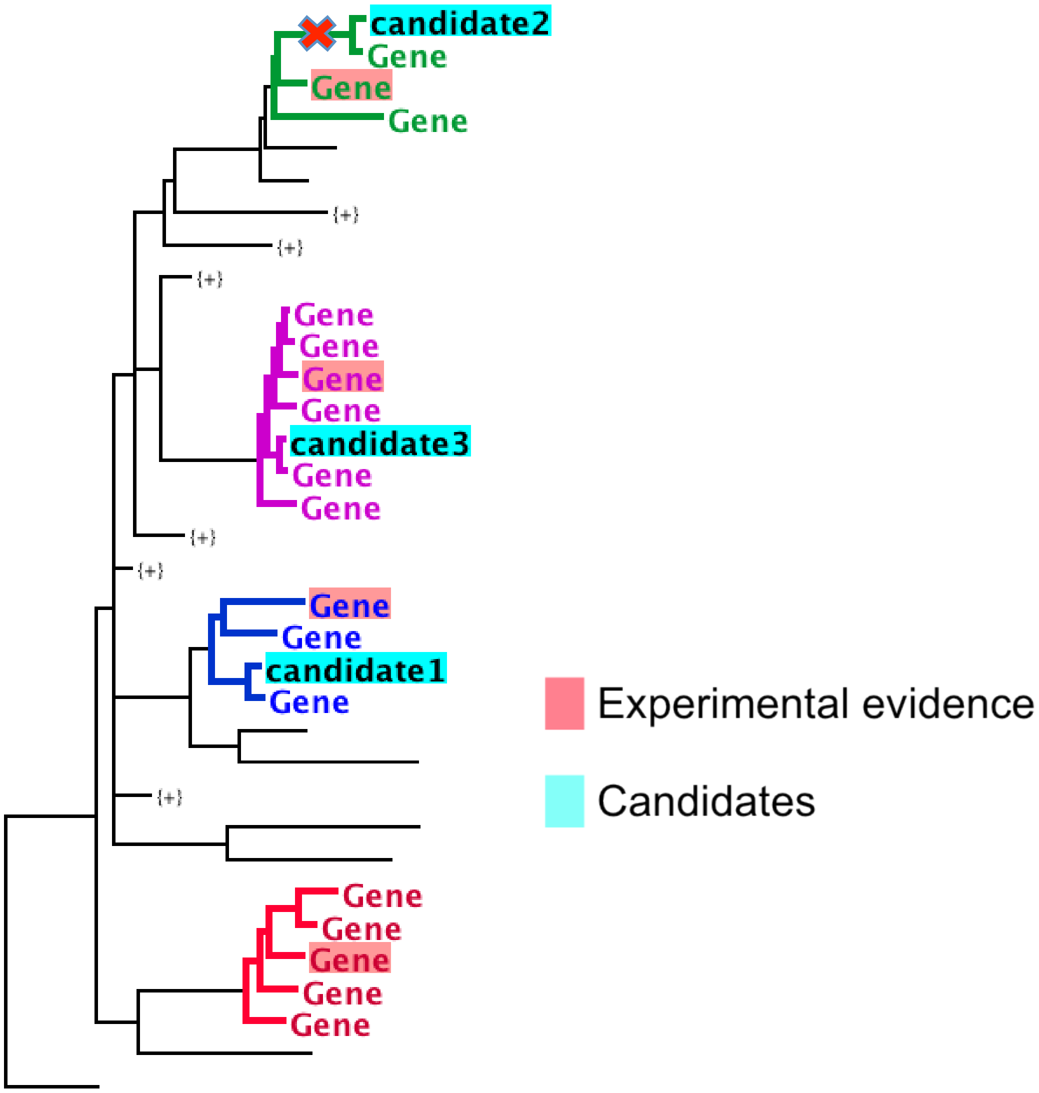


Figure 5.1: Phylogenetic tree analyses objective. The objective of this chapter is to build a phylogenetic tree for each superfamily where its branch lengths describe the enzymatic function evolution. Therefore, longer branches would be more correlated with enzymatic function change. In a normal tree we expect to have several clusters with genes with similar enzymatic functions (represented with green, pink, blue and red). Some of the genes annotations are based on experimental evidence (light red background) and some of the cluster would include candidate genes for the species in focus (Cyan background). Although candidate2 gene may be expected to have the same enzymatic function as the other genes in the green cluster, the long branch that precedes it (marked with a red cross) might be an indication of a enzymatic functional change making it a good candidate for the hole we are trying to fill.

Briefly, I have built a predictor able to find evidence that might lead us to identify the misannotation of a gene or, possibly, evidence that might help us assign a new function to a certain domain. So, the aim is to detect possible function changes, by correlating functional change with tree branch lengths and with evolutionary events such as gene duplication and selective pressure scores (Figure 5.1).

5.2 Methods

Figure 5.2 provides with an overview of the different steps in the analyses made in this chapter.

5.2.1 Homolog search

The protocol used to build the phylogenic trees integrates the protocol presented in Chapter 3. Briefly, the input of the protocol from Chapter 3 is an enzymatic function EC number. Using the PDB database, the protocol searches for all the PDB protein entries with the EC number input assigned. Afterwards, using the SCOP database, the protocol retrieves all the different superfamilies IDs to which the proteins gathered from the PDB database belong. The last step queries the SUPERFAMILY database for all the genes of the target species that have domains classified to one of the superfamily ids returned from SCOP (see Chapter 3 for more details).

5.2.2 Phylogenetic trees

To build the phylogenetic trees, the protocol described above was applied to a set of model organisms covering the tree domains: Eukaryota, Archaea and Bacteria (Table 5.1).

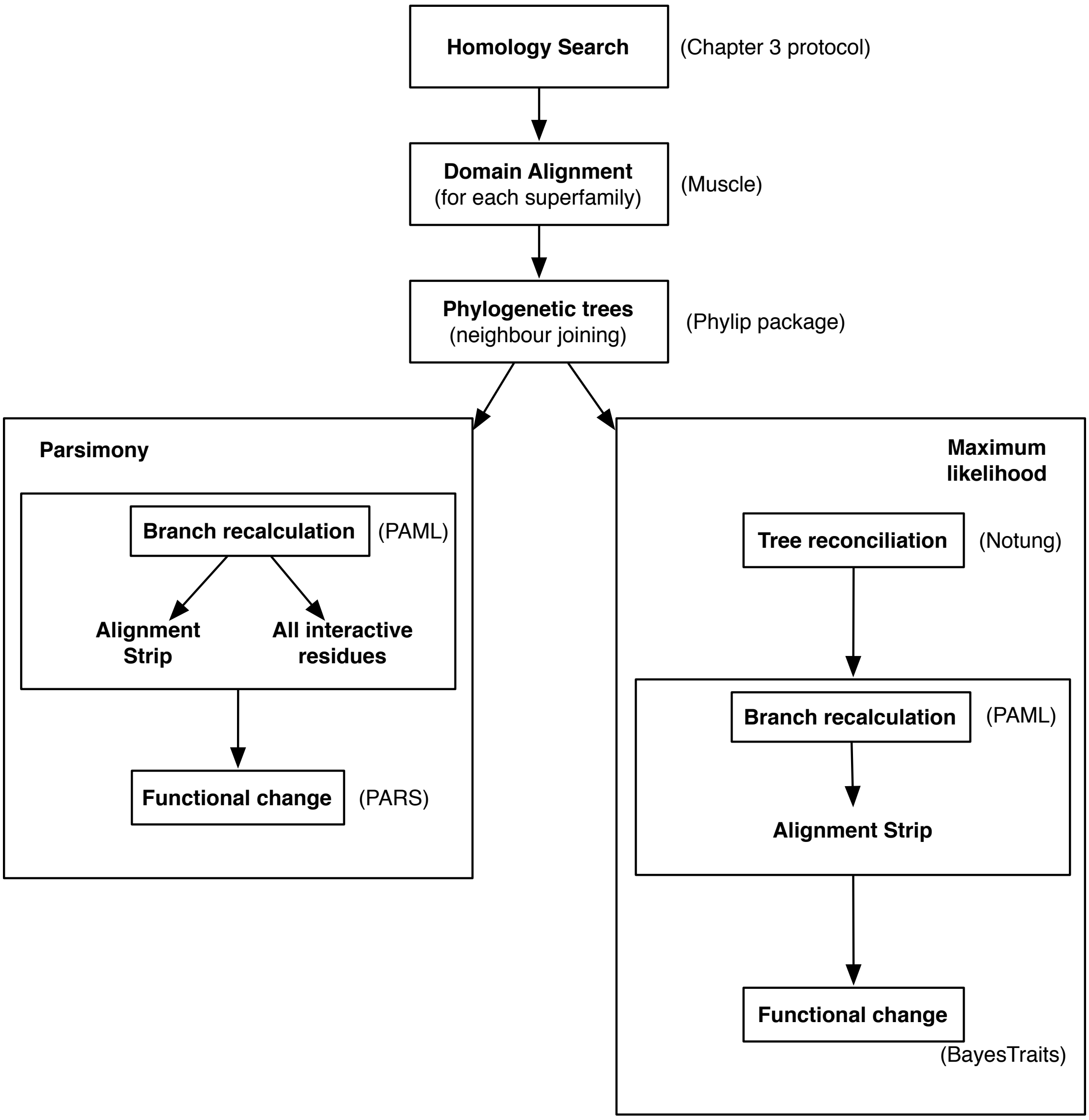


Figure 5.2: Overview of the different steps and analyses made in this chapter. The analysis pipeline starts with the protocol presented in Chapter 3. Afterwards, for each superfamily the domains found are aligned using Muscle and then the phylogenetic trees are built using a neighbour joining approach. To infer functional change two different approaches were used: parsimony (left) and maximum likelihood (right). For the parsimony method, two different branch recalculation methods were considered: alignment strip and all interactive residues. For each one the functional change inference was done using PARS. For the Maximum likelihood evolutionary traits were also considered, for which Notung was used to make the tree reconciliation. The branch recalculation was done using only the Alignment strip approach. Finally, the functional change inference was done using BayesTraits.

Each of the superfamily groups obtained using the protocol described above was aligned using MUSCLE (Edgar (2004)). After aligning, to build the phylogenetic tree I used a neighbor joining approach (Saitou and Nei (1987)). I first calculated the matrix distance with *protdist* and then with *bionj* built the tree. The matrix distance calculates the pairwise distance between all the sequences. This matrix is then used by *bionj* that tries to build the tree that better explains those distances. To add more confidence to the constructed tree, the tree is bootstrapped using *seqboot*. Both *protdist* and *seqboot* belong to *phylip* (Felsenstein (1993)), a package of programs for inferring phylogenies. To build a tree that better describes the evolutionary history of a superfamily group, the branches with low bootstrap score must be collapsed because there is little support that such branches do really exist. Therefore, the branches that had a bootstrap score of less that 50% were collapsed (Górniak *et al.* (2010)).

5.2.3 Branch length recalculation

The branch recalculation was done by applying a maximum likelihood strategy using *codeml* from PAML [37] (Figure 5.4). PAML (Yang (1997)) is a package of programs for phylogenetic analyses using maximum likelihood. Two different approaches were used: *alignment strip* (Figure 5.3-A) and *all interacting residues* (Figure 5.3-B) approaches.

The *alignment strip* approach is the simplest one. It does not require any additional information besides the multiple alignment. It simply extracts the columns with less than 30% of gaps. Several cutoffs values were tested. However, little difference was found between the resulted mutiple alignments above 30%. For this reason I decided to consider 30%.This is done to remove the residue columns that express little or no evolutionary restrictions, so that the distance between the sequences expresses as much as possible these residues that define the protein structure and function. The filtered alignment is then used

Domain	**Species**
Eukaryota	
	Trypanosoma brucei gambiense
	Aspergillus nidulans
	Neurospora crassa
	Schizosaccharomyces pombe
	Arabidopsis thaliana
	Caenorhabditis elegans
	Drosophila melanogaster
	Anopheles gambiae
	Mus musculus
	Homo sapiens
	Xenopus laevis
	Gallus gallus
	Saccharomyces cerevisiae
	Dictyostelium discoideum
Archaea	
	Halobacterium salinarum
	Sulfolobus tokodaii
Bacteria	
	Escherichia coli
	Neisseria meningitidis
	Helicobacter pylori
	Bacillus subtilis
	Mycoplasma genitalium
	Mycobacterium tuberculosis
	Chlamydia trachomatis
	Treponema pallidum
	Aquifex aeolicus

Table 5.1: Model species selected. The protocol was applied to all the species present in this table covering the Eukaryota, Archaea and Bacteria domains

to feed the PAML software (Figure 5.4). To do the branch length recalculation, a maximum likelihood strategy was chosen because, unlike neighbour joining, it allows branch length recalculation without changing a given tree topology. So, the branch recalculation was done by applying a maximum likelihood strategy using *codeml* from PAML.

The other method, in more detail, performs further transformations on the initial multiple alignment before feeding the PAML software. It starts by taking from FireDB the proteins that were assigned by SCOP to the superfamily in focus. The FireDB database (Lopez

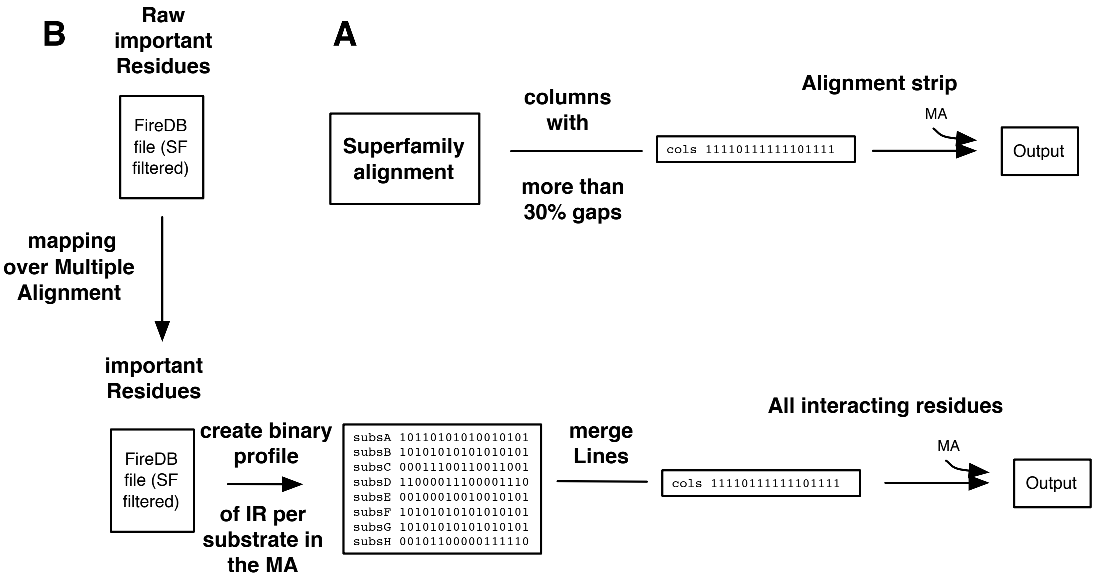
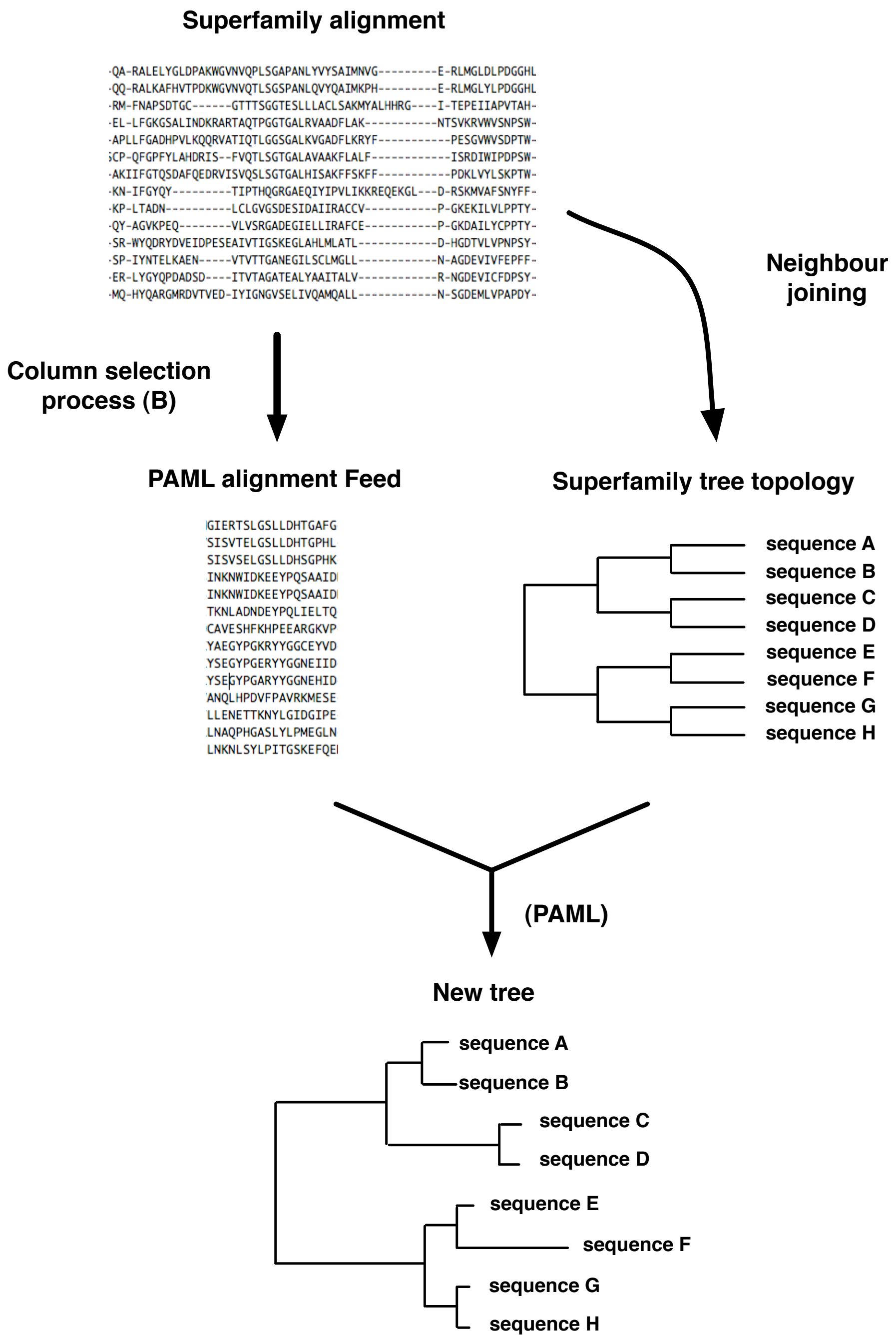


Figure 5.3: Branch recalculation protocol - column extraction approaches. A- *alignment strip* approach; B- *all interacting residues* approach.

et al. (2007)) gathers information about known functionally important residues. This includes residues that perform binding activities and residues that have catalytic functions. This database has two main sources of information. On the one hand it uses PDB crystal structures to identify the close atomic contacts and, on the other hand, it makes use of the Catalytic Site Atlas (Porter *et al.* (2004)) to get reliably annotated catalytic residues.

Using MUSCLE, every protein sequence is aligned to the superfamily multiple alignment. Knowing, from the FireDB file, the functionally important residue positions of the proteins, these positions are mapped onto the multiple alignment, identifying in this way the columns where each important residue was aligned. To collect all columns that contain important residues, a binary profile of important residues per substrate in the multiple alignment is made, identifying the different columns where each substrate has an interacting residue. this has a 1 if the column corresponds to an important residue and 0, otherwise. Afterwards, the profile lines are merged to identify all the potential interacting columns. This information is then used to strip the superfamily alignment where every



```
-QA-RALELYGLDPAKWGVNVQPLSGAPANLYVYSAIMNVG---------E-RLMGLDLPDGGHL
-QQ-RALKAFHVTPDKWGVNVQTLSGSPANLQVYQAIMKPH---------E-RLMGLYLPDGGHL
-RM-FNAPSDTGC------GTTTSGGTESLLLACLSAKMYALHHRG----I-TEPEIIAPVTAH-
-EL-LFGKGSALINDKRARTAQTPGGTGALRVAADFLAK-----------NTSVKRVWVSNPSW-
-APLLFGADHPVLKQQRVATIQTLGGSGALKVGADFLKRYF-----------PESGVWVSDPTW-
SCP-QFGPFYLAHDRIS--FVQTLSGTGALAVAAKFLALF------------ISRDIWIPDPSW-
-AKIIFGTQSDAFQEDRVISVQSLSGTGALHISAKFFSKFF-----------PDKLVYLSKPTW-
-KN-IFGYQY---------TIPTHQGRGAEQIYIPVLIKKREQEKGL---D-RSKMVAFSNYFF-
-KP-LTADN----------LCLGVGSDESIDAIIRACCV-----------P-GKEKILVLPPTY-
-QY-AGVKPEQ--------VLVSRGADEGIELLIRAFCE-----------P-GKDAILYCPPTY-
-SR-WYQDRYDVEIDPESEAIVTIGSKEGLAHLMLATL------------D-HGDTVLVPNPSY-
-SP-IYNTELKAEN-----VTVTTGANEGILSCLMGLL------------N-AGDEVIVFEPFF-
-ER-LYGYQPDADSD----ITVTAGATEALYAAITALV------------R-NGDEVICFDPSY-
-MQ-HYQARGMRDVTVED-IYIGNGVSELIVQAMQALL------------N-SGDEMLVPAPDY-
```

```
GIERTSLGSLLDHTGAFG
SISVTELGSLLDHTGPHL
SISVSELGSLLDHSGPHK
INKNWIDKEEYPQSAAID
INKNWIDKEEYPQSAAID
TKNLADNDEYPQLIELTQ
CAVESHFKHPEEARGKVP
YAEGYPGKRYYGGCEYVD
YSEGYPGERYYGGNEIID
YSEGYPGARYYGGNEHID
ANQLHPDVFPAVRKMESE
LLENETTKNYLGIDGIPE
LNAQPHGASLYLPMEGLN
LNKNLSYLPITGSKEFQE
```

Figure 5.4: **Branch recalculation protocol.**

column without any important residue is removed.

5.2.4 Functional change

The gene functional annotations used were taken from the KEGG database. To have the cleanest trees possible, only the genes with complete EC numbers were considered. To infer functional change events on tree branches I used two different methods. The first method applied *pars* from the *phylip* phylogenetic programs package (Felsenstein (1993)) to the collapsed trees. *Pars* applies a parsimony approach to infer the functional change history of a phylogenetic tree. The parsimony approach considers the tree with the least number of functional changes as the best evolutionary explanation. Ordinarily, *pars* does not handle more than 8 different states. However, in most of our phylogentic trees there are more than 8 different functions. So, to overcome this limitation we had to change the part of the code that restricted the number of states.

The parsimony approach only accounts for the optimal trees. However, this can be a problem if there are several equally optimal trees that differ in their ancestor states. To take this uncertainty into account, the second method used was *BayesTraits* (Pagel and Meade (2007)). *BayesTraits* can only be applied to binary trees. The methodology used to build the binary trees is described below.

5.2.5 Tree reconciliation

One possible way to account for evolutionary traits such as duplication and speciation events when studying gene evolution is through the use of gene tree reconciliation. This aims to describe the gene tree evolution within a species tree. Using an accurate species tree and a gene tree, tree reconciliation expresses the evolution of the gene tree using gene duplications and losses as explanations for incongruences between the species tree and the gene tree.

I used *Notung* (Chen *et al.* (2000)) to perform tree reconciliation in the superfamily trees. *Notung* is a framework that incorporates several phylogenetic tasks such as tree reconciliation, identification of gene duplications events or tree rooting. As gene duplication and gene loss are rare events, *Notung* uses the parsimony approach to infer these evolutionary events. Tree reconciliation requires a species phylogenetic tree and a gene phylogenetic tree. The phylogenetic tree was taken from the SUPERFAMILY database. SUPERFAMILY uses a maximum-likelihood phylogenetic estimation with RAxML Stamatakis (2006)) to construct the species tree for all completely sequenced genomes. They consider in all the genomes the presence/absence of molecular characters like domain architectures, superfamilies and families.

The gene tree used previously had been collapsed for the branches with less than 50% bootstrap score. To make analyses on evolutionary events it is better to use a binary tree otherwise for example, we might get the situation where a node has 10 children and 5 duplications. In this cases there would not be possible to define a branch/duplication event association. This was done using *Notung*. Afterwards, using *Notung* the trees were rerooted so as to describe the evolutionary history of the gene tree that had the least duplication/loss events. This was done for the reason already mentioned above, that duplication and loss events are believed to be rare. Finally, the tree's branch lengths were recalculated using PAML and tested using the *alignment strip* approach.

5.2.6 Selective pressures

To search for adaptive evolution events I used the *HyPhy package* (Pond and Muse (2005)), a platform that provides likelihood-based tools for sequence evolution analyses. I ran the branch-site model (Yang and Nielsen (2002)) to detect instances of episodic diversifying selection, as it might be expected for a functional change event. Between the output measures returned by *HyPhy*, I considered two. One of the measures represents the average dN/dS between all the ancestor states and the other represents an uncorrected p-value derived from a likelihood ratio test for the hypothesis that dN/dS > 1 for some of the sites.

5.2.7 Machine learning

The approach used to identify function change events was Random Forests, implemented in the randomForest R package (Liaw and Wiener (2002)). The random forests algorithm implemented is the one described in (Breiman (2001)). The parameters used in both $randomForest$ and $predict$ functions were the default ones. For building the ROC curves, the $type = “prob''$ option in the $predict$ function was used.

5.3 Results and Discussion

To find evolutionary features that correlate with changes in function, we need to first build the tree that represents the evolutionary relationships of the given superfamily as accurately as possible.

We have constructed a computational protocol that, given an EC number returns gene domain sequences grouped by superfamilies from the target species (See methods). Afterwards, the domains for each superfamily are aligned using *MUSCLE*. The protocol was applied to 21 EC Numbers, resulting in 14 unique superfamilies. Only the superfamilies with at least two different enzymatic functions and 10 domains with a complete EC number annotation were considered, reducing the number to 9 unique superfamilies (Table 5.2).

EC number	Superfamily	Selected	Bayestraits
1.1.1.25	53223	X	
1.3.99.2	56645	X	X
1.3.99.3	56645	X	X
1.4.4.2	53383	X	-
1.13.11.27	54593	X	X
2.1.1.13	52242		-
	56507		-
2.2.1.2	51569	X	-
2.5.1.18	52833	X	-
	54593	X	X
2.5.1.54	51569	X	-
2.7.1.23	111331	-	-
2.7.4.16	-	-	-
3.1.3.25	56655	X	X
3.1.3.27	-	-	-
3.1.3.57	56655	X	X
3.5.1.63	-	-	-
4.1.2.25	55620	X	X
4.2.1.10	51569	X	-
	52304	-	-
4.2.1.75	69618	-	-
4.2.3.4	-	-	-
5.3.3.8	52096	X	-
6.1.1.24	-	-	-

Table 5.2: EC numbers used as input to the protocol. The protocol was applied to all the EC numbers present in the first column. In the second column are all the superfamilies to which at least one structure with the respective EC number was found in the PDB. In the *Selected* column are indicated the superfamilies with more than one different enzymatic function and more than 10 domains with a complete EC number. The fourth column indicates which superfamilies was possible to apply the *Bayestraits* approach to infer functional change events.

5.3.1 Branch recalculation

Tree branch recalculation using more restricted and meaningful information, like the interacting residues, might improve the description of the function history of the sequences of a given superfamily (Figure 5.4). This may also increase the sensitivity of function change prediction from the branch lengths. To investigate this hypothesis, two methods were tested.

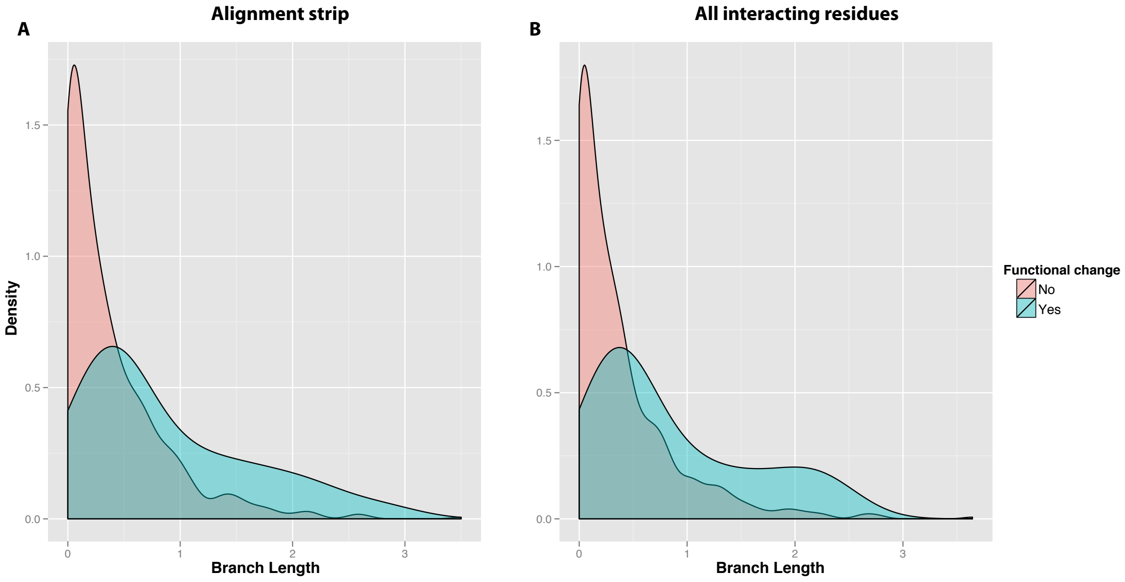


Figure 5.5: Branch length recalculation distributions Visualisation of potential value of each of the approaches (*Alignment strip* and *All interacting residues*) in distinguishing between functional change and no functional change events (red - no functional change; green - functional change).

We used two different approaches in order to identify the best way to describe the superfamily gene history (Figure 5.3). What differs between these approaches is the specific set of multiple alignment columns extracted from the overall superfamily alignment. Each alignment is fed to PAML together with the superfamily phylogenetic tree, previously built using a neighbour joining approach. As described in the methods sections, *alignment strip* is focused on the most conserved columns in the alignment and the *all interacting residues* approach uses information about functionally important residues.

5.3.2 Function change - parsimony approach

As described in the methods, to infer the functional change for the parsimony approach I used *pars*. For each branch in the tree, *pars* returns one of three possible states: ‘yes’ if there is a function change; ‘no’ if there is not a function change; ‘maybe’ if is not conclusive. I did not consider the branches assigned with a 'maybe'. This has decreased the number of branches analysed. Using the *Kolmogorov-Smirnov* test, the predictive potential of the *Alignment strip* (p-value = 0.0004774) and *All interacting residues* (p-value = 0.0003775) approaches using a parsimony approach is shown in Figure 5.5.

Overall, no matter what approach we used, the branches that correspond to no functional change tend to have smaller branch lengths. For a 4 EC digit change the AUC is around 73% for both approaches. As expected, the AUC is greater for the changes in the third EC number digit, being on both cases around 80%. Therefore, this shows that there is a strong correlation between branch length and functional change. Regarding the two main methods for choosing more meaningful residues, the results show little difference between the two. The fact that the *important residues* do not strongly increase the performance compared with the strip approach might be a result of the large functional diversity that a superfamily can include.

5.3.3 Function change - Bayesian approach

Using now a Bayesian approach, the branch length and functional change was again analysed. Although the parsimony approach has achieved promising results, as stated before, it did not return a conclusive result for every branch. Figure 5.6 shows the branch length distributions resulted from using BayesTraits. Branches with function change tend to be longer. This difference between the two cases is reinforced by the Kolmogorov-Smirnov test (p-value = 1.408e^{-11}). This corroborates the results achieved with the parsimony approach.

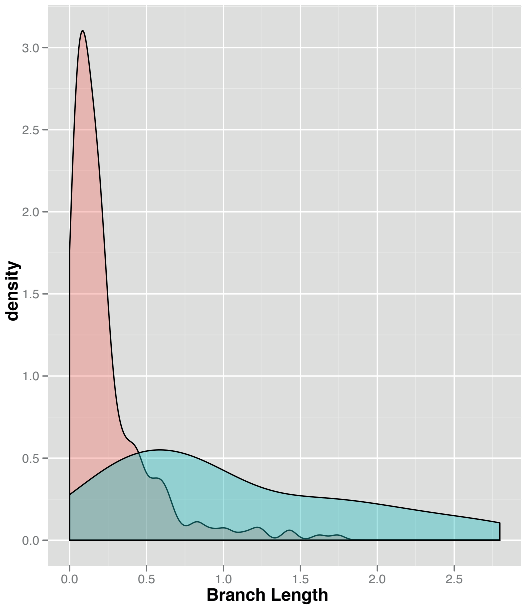


Figure 5.6: **BayesTraits histogram result.**

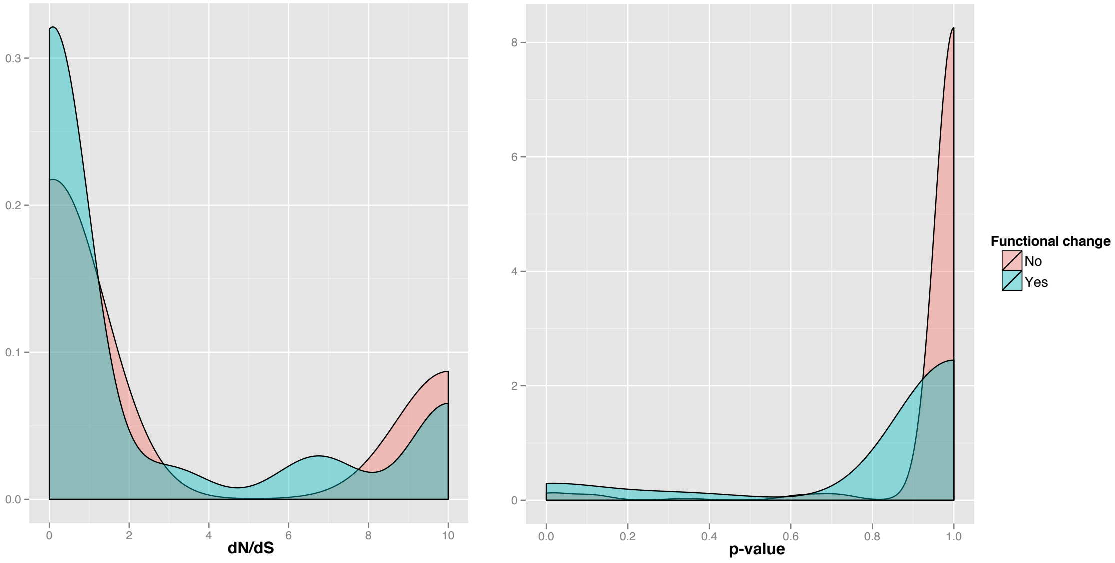


Figure 5.7: **Hyphy histogram results.**

		Duplication	
		0	1
Function change	0	100	25
	1	18	20

Table 5.3: **Gene duplication results.**

Accuracy	Specificity	Sensitivity
0.736196	0.52631	0.8

Table 5.4: **Simple predictor results.**

5.3.4 Gene duplication and selective pressure

Besides the branch length I also tested the correlation between function change following duplication events and function change with two selective pressure scores (dN/dS and p-value). In the case of the duplication events the results show a correlation. Using Fisher's exact test we can see that there is a significant difference (p-value = 0.0001671) between the proportion of function change events in relation to duplication events. The proportion of function change events after duplication is around 44%, otherwise is 15%. To study in more detail this correlation we defined a simple predictor as follows: if there is a duplication event there is a function change; if not, there is not a function change. Table 5.4 shows the accuracy, specificity and sensitivity of the predictor. The accuracy is above 73% and it has a low specificity and a high sensitivity. This means that, although a duplication event does not imply a function change, from these results we can conclude that gene duplications favour functional change.

For the two selective pressure scores, the results were not as promising as the correlation presented above for duplication events. The distribution in the two cases did not show any significant difference between the existence or not of functional changes (dN/dS: p-value= 0.3154 ; p-value: p-value=0.6953).

5.3.5 Predictor testing

To build a predictor to detect function change I have selected two of the several features tested in this study: branch length and duplication events. By their distributions, these two features were the only ones that were demonstrated to have predictive qualities. As

AUC	S. d.
0.891	0.046

Table 5.5: 4 fold cross validation results. The predictive model performance was assessed by a 4 fold cross validation. The table shows the AUC and the standard deviation of this analysis.

55620	56645	56655	54593
0.753	0.988	0.925	0.969

Table 5.6: Superfamily cross validation results. To test performance on unseen enzyme classes, the classifier was assessed in a leave-one-out cross validation at the superfamily level. The table shows the AUC of each analysis, where each superfamily in turn was used as the test data set.

mentioned in methods, the Machine Learning method used was Random Forests with the same R package as used in Chapter 4.

To test the predictor I performed several cross-validation tests (Figure 5.8). Figure 5.8-A shows the 4-fold cross validation between function change and no change. The predictor showed a very high performance (Table 5.5) of almost 0.9 AUC. To study if these results could be extrapolated to other phylogenetic trees I performed a cross validation test where the model was trained with all trees leaving just one out in turn that would be used for testing (Figure 5.8-B). Once again the accuracy results were very high (Table 5.6). Changes on the 4th EC number digit correspond to a function change at the level of the substrate specificity. Therefore, the difference between the sequence of two proteins whose enzymatic functions only differ in the last EC number digit may be very subtle and more challenging to detect than differences at a higher EC number level. To see how well the predictor could detect these differences, the function changes were divided into three different groups. Besides the group with no changes, in one group were included the changes at the 4th EC number digit level, and in the third group joined all the other types of enzymatic function changes that will be referred as 3 digit change group. Using the three groups, the predictor was tested to see if it could distinguish between the three enzymatic function levels. Figure 5.8-C/D/E and (Table 5.7) show the results

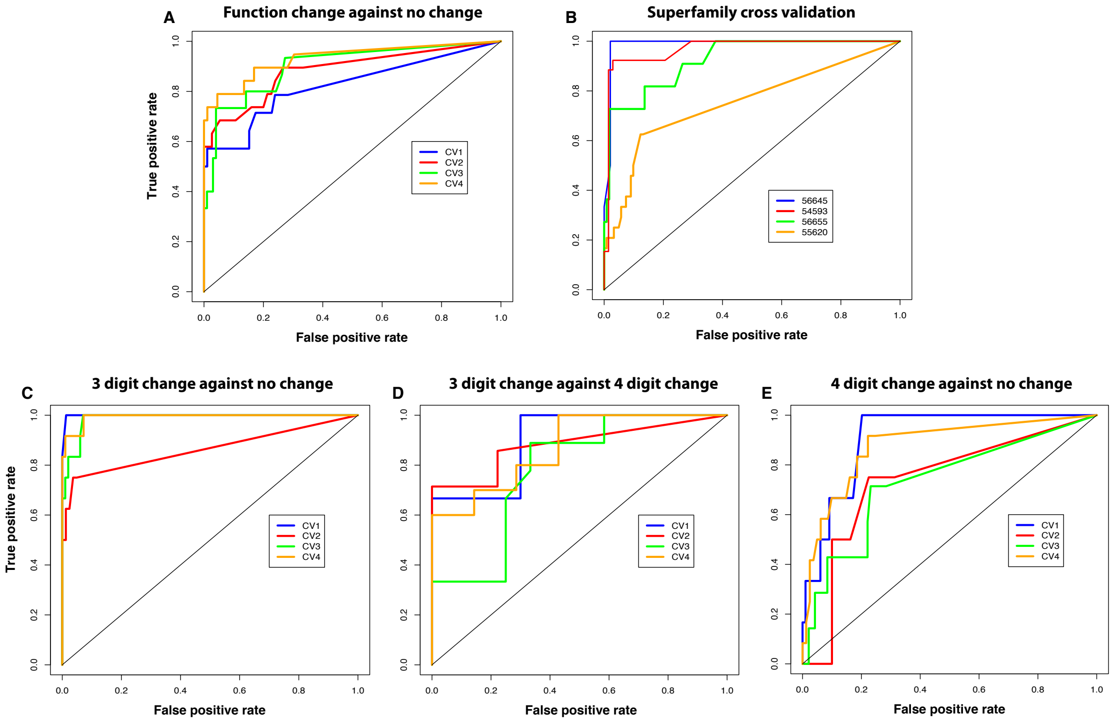


Figure 5.8: **Predictor ROC curves results.**

Functional change	**AUC**	**S. d.**
3EC - No change	0.99	0.0649
3EC - 4EC	0.786	0.112
4EC - No change	0.801	0.093

Table 5.7: Predictor results. The predictor test performance distinguishing between different levels of functional change. The table shows the AUC and the standard deviation of each analysis.

when testing: 3 digit change against no change; 4 digit change against 3 digit change and 4 digit change against no change. In all the cases the predictor had a significant high performance. As expected, it was revealed to be more difficult to distinguish between a change at the substrate level and no change than between a change at a higher EC number level and no functional change. However, the AUC was still around 80%. The predictor showed similar results when distinguishing the changes at the substrate level from the other ones.

5.4 Conclusion

The results of this study have shown that evolutionary events such as branch length and gene duplication have a high correlation with function change. Moreover, when joined in a predictive model high accuracy results were achieved. The classifier has not only been able to identify function change events but also to successfully distinguish between different levels of enzymatic functional change.

Chapter 6

Discussion

6.1 Conclusions

Fully automated metabolic reconstructions are still far behind manually curated models in terms of accuracy and completeness. This thesis has not only contributed to diminishing the gap between automatic and manual approaches, but has also shown that they can be complementary.

This thesis has presented novel approaches for the improvement of automated models, which exploit a variety of different techniques and concepts.

PathwayBooster provides a bridge between automated and manual modelling by making relevant information more easily accessible to modellers and hence decreasing the time required for curation. It combines information from well known metabolic and enzymatic databases such as KEGG and BRENDA. It also provides sequence based information using reciprocal best hits from BLAST. Moreover, based on KEGG's pathways, PathwayBooster provides a means to visualise information relating to several species at the same time, which is essential for the detection of inconsistencies and allows differences between different species to be spotted. Chapter 2 also presented examples where PathwayBooster has been used and highlighted some of its properties.

Compared with the available software, PathwayBooster is the most focused in providing easily accessible information essential for pathway curation. It is the only available software that combines intuitive pathway visualisation, literature information, sequence comparisons information and species pathway similarity information in a rapid, accessible way.

Recently, BRENDA has created a new database that provides literature information based on data mining search (FRENDA). In the future, PathwayBooster can also query this database providing more complete literature information. Currently, PathwayBooster uses BLAST to provide sequence similarity information on gene annotations and possible candidates for each EC number. Other, more sensitive methods may also be considered such as the protocol to identify candidate genes for a missing enzymatic function presented in Chapter 3. Another way would be to consider tools such as SHARKhunt that uses the PRIAM library to scan for sequences that are similar to genes that have a given function. This way we would likely reach to more candidates. Furthermore, these candidates could be supported by other sorts of information such as topological properties and functional evolution information like the ones described in chapter 4 and 5.

Moving beyond the information already available to curators, my objective was to build a protocol that increased the sensitivity of searches for candidate proteins that could potentially fulfil a missing enzymatic function known as “pathway hole”. The main idea behind the protocol is that if a protein shares a common domain structure with other proteins that are known to have a specific enzymatic function, it is more likely to be able to acquire the same function than a randomly selected protein. Therefore, instead of conducting the sequence searches at the protein family level, the protocol described in chapter 3 searches for proteins that have similar 3D structure. This corresponds to searching at the superfamily level where the most distant homologues are believed to reside. The protocol is based on several existing tools. It starts with PDB, passing through SCOP and SUPERFAMILY. This protocol was successfully tested on an example where a similar strategy had been applied to find a missing enzyme for a hole in the *P. falciparum* folate pathway. It is important to notice that KEGG has not yet corrected this annotation. Afterwards, the protocol was applied to other putative holes from the same species, and provided a set of candidates for many of the holes. One reason why it may be difficult to identify an enzyme to fill a pathway hole is that the protein responsible may be incorrectly annotated with a different function.

Chapter 4 tackles misannotation in public databases. The use of automated methods to annotate gene functions is the major contributor to this important problem. Existing methods are still mainly based on sequence similarity searches. As well as being less accurate than experimental validation, these can lead to error propagation. Therefore, the main objective of this chapter was to build a sequence-independent classifier to detect misannotations based on metabolic network topological properties. Our results demonstrate that topological properties can be used to detect misannotations, showing that using this information could be part of the solution to the misannotation problem in public databases.

The classifier was carefully tested using not only traditional 5 fold cross validation but also using inter-superfamilies cross validation. In the latter case, the results suggest that different enzymatic functions may occupy different topological positions in the metabolic network. However, they also indicate that the trained model should still be informative when applied more generally. Furthermore, the Random Forest classifier was applied to several KEGG genomic metabolic networks and the results were compared with curated models. The accuracy of the model was around 60%, confirming its effectiveness. After validating the model, Chapter 4 also demonstrated that there is a clear negative correlation between the quality of automated metabolic networks and their phylogenetic distance to the major model organism for biochemistry (*Escherichia coli* for prokaryotes and *Homo sapiens* for eukaryotes). This last result highlights the misannotation problem in public databases and the fact that sequence similarity approaches may be less accurate when transferring enzymatic function from distantly related species.

Chapter 5 provides an evolutionary perspective on gene function annotation. Evolutionary events such as gene duplication, loss and speciation have been proven to influence gene function changes (Nowak (1997)). The objective of this chapter was to build a tool to detect enzymatic function change based on the evolutionary events just mentioned. This chapter also took into account that within a protein the amino acids suffer different evolutionary pressures. Residues that are essential for function tend to be more conserved during evolution. Therefore, when building the phylogenetic trees, two different amino acid selection procedure were used.

In this chapter I have shown that in the superfamily phylogenetic tree, branch length is correlated with function change. This was supported by two different methods to infer functional change. Moreover, I have shown that duplication events favour function change. I have also checked if there was any correlation between selective pressure scores such as dN/dS, for which the results indicate that there is not. Using branch length and duplication features, I have built a predictor that is able to predict function change. To access the accuracy, I have used 4 fold cross-validation and inter-superfamily trees cross validation. The area under the curve was around 0.9 in the 4 fold cross validation and varied between 0.75 and 0.99 in the latter. The predictor was also shown to be able to distinguish function change at the substrate level, which corresponds to a change in the 4th EC number digit. These analyses can be also used ranking candidates for a missing pathway link. The classifier can be employed on existent enzymatic functional annotations and access the likelihood of a possible functional change.

With the objective of basing possible candidate genes on the most curated annotations possible, the starting point of the chapter 3 protocol was the PDB database. One drawback is the restricted number of superfamilies that result from the protocol and therefore the number of candidate genes for filling a missing link. To increase the number of superfamilies we may try to adopt a less stringent search when searching for known enzymes that perform a certain enzymatic function.

Instead of searching for only enzymes present in PDB, we could make use of databases such as UniProtKB/Swiss-Prot, that contain manually annotated data. To connect these sequences to the SUPERFAMILY database, SCOP would no longer be suitable. To overcome this, we could make use of the SUPERFAMILY HMM's models.

Another way would be to consider a hierarchical protocol which would start by using tools such as SHARKhunt to scan for sequences that are similar to genes having a given function. Afterwards, we could cluster the genes found using Pfam or SUPERFAMILY HMMs. This approach would be likely to yield more candidates.

The objectives of the phylogenetic method and the topological properties method were different and therefore their comparison may lead to misleading conclusions. As already mentioned, the purpose of the phylogenetic method is to identify enzymatic function change and that of the topological properties method is to predict misannotations. Part of the explanation of the very good results that the phylogenetic method achieved may be by the fact that this method relies on gene sequence, which has been the basis for almost all automatic function predictions in the past. Although the results obtained using the topological properties method may not seem as impressive as those obtained with the phylogenetic method, the topological properties method has presented a novel approach to detecting misannotations. The biggest difference to current approaches is the fact that this method is sequence-independent. It is important to note that the training data set was restricted to only 4 well-represented superfamilies. It is likely that with a training dataset that contains information from a larger number of superfamilies, the results would greatly improve. Nevertheless, the results were significantly positive.

6.2 Future Work

The tools presented here can be used for filling pathway holes where each tool provides a different aspect of information for hole identification, candidate search and candidate evaluation. However, these tools can also be used in a wider context. They can be applied at the general node analysis level in the metabolic network. Using the same methodology, we can aim to detect and study exaptations and promiscuous proteins and the robustness of the network, with the ultimate aim of producing improved, evolution-aware software for automated metabolic network reconstruction.

So, future work would be to combine the built tools with other tools and other sources of information, gathering as much information as possible for each functional assignment. At the level of enzyme information, an example would be the confidence of its annotation. This includes the way in which the annotation was performed and wether there is experimental data to confirm the annotation.

In terms of additional tools, a potentially useful resource is DETECT (Hung *et al.* (2010)). DETECT returns probability values for enzyme annotations that take into account the existing variation within each enzyme family. It uses enzymatic family profiles built using the proteins and annotations from the Swiss-Prot database. These profiles are used together with a Bayesian statistical framework to assign a probability value to an annotation.

Afterwards, all these data for each annotation would be used to rank the candidate enzymes for a particular enzymatic function. Each model instance including different combinations of enzymes sampled from the ranking mentioned before, would be compared

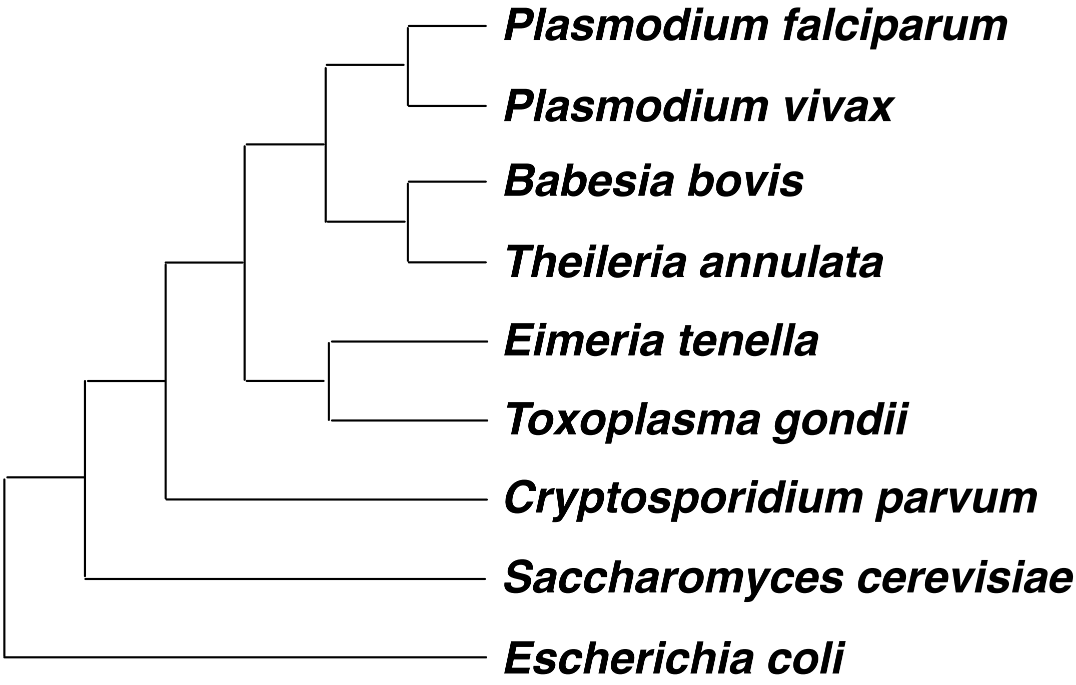


Figure 6.1: Apicomplexa phylogeny. This phylogeny is a result of genome-scale phylogenetic analysis in the phylum (Kuo *et al.* (2008)). Three different phylogenetic methods were used to infer the phylogenetic tree: maximum likelihood, maximum parsimony and Neighbor-Joining. To this tree I have added to model organisms: *Saccharomyces cerevisiae* and *Escherichia coli*

against experimental knowledge using Flux Balance analyses.

Although we have focused in the metabolic reconstruction of a specific species, such as *Plasmodium falciparum* or *Geobacillus thermoglucosidasius*, if we consider the evolutionary history of all the species in the phylogenetic tree exemplified in the Figure 6.1 when assigning a proposed function to an enzyme, we can see that this assignment will have repercussions for their ancestral states and consequently also for other present-day species. So, the best way to rebuild the metabolic network for a species of interest is to also consider the consistency and robustness of the network in all the other species. The final step in this process will therefore be to consider all the different evolutionary histories originated by the possible assignments and use a likelihood framework to decide which of them is most plausible, given the available data.

Appendix A

PathwayBooster

A.1 Introduction

PathwayBooster is an open-source software tool to support the comparison and curation of metabolic models. It combines gene annotations from GenBank files and other sources with information retrieved from the metabolic databases BRENDA and KEGG to produce a set of pathway diagrams and reports summarising the evidence for the presence of a reaction in a given organism's metabolic network. By comparing multiple sources of evidence within a common framework, PathwayBooster assists the curator in the identification of likely false positive (misannotated enzyme) and false negative (pathway hole) reactions. Reaction evidence may be taken from alternative annotations of the same genome and/or a set of closely related organisms.

This document provides information on how to install and run PathwayBooster. The software has been built and tested with Python 2.6 and newer, on Windows, Mac OS X, and Linux platforms.

PathwayBooster may be downloaded from

`http://www.theosysbio.bio.ic.ac.uk/resources/pathwaybooster/`.

For support and other queries, send e-mail to j.pinney@imperial.ac.uk.

A.2 Setup instructions

A.2.1 Prerequisites

- Python[1]. Tested with versions 2.6+.
- BLAST[2]. Tested with version 2.2.24.
- SOAPpy[3] and PIL[4] python modules. If you encounter problems installing PIL or when running PathwayBooster, with an error like '`ImportError: The _imagingft C module is not installed`', you should try installing a PIL version with precompiled libraries[5]. Using Enthought python, PIL should already be installed.
- BRENDA flatfile[6]. Once extracted from the zipfile, move the file `brenda_download.txt` into the `PathwayBooster/files` directory.

[1] e.g. `http://enthought.com/repo/free/`

[2] `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/`

[3] `http://pypi.python.org/pypi/SOAPpy/`

[4] `http://www.pythonware.com/products/pil/`

[5] available from `http://www.lfd.uci.edu/~gohlke/pythonlibs/`

[6] `http://www.brenda-enzymes.org/brenda_download/`

A.2.2 Running PathwayBooster

To run PathwayBooster you need a setup file as explained below. Then from the PathwayBooster directory, type:

```
python PathwayBooster.py [setupFilename.xml]
```

Since BLAST runs may take a while, there is the option of pre-compiling the BLAST files without running the other analyses. To do so, type:

```
python PathwayBooster.py -blast [setupFilename.xml]
```

By default, PathwayBooster saves the results in a directory named `PathwayBoosterReports`. To change the output directory, type:

```
python PathwayBooster.py [setupFilename.xml] -outDir [newOutPutDirector
```

PathwayBooster does not need to run from the PathwayBooster directory. To run from a different directory, just type:

```
python path/to/PathwayBooster.py [setupFilename.xml]
```

and the results will be saved in the current working directory.

A.2.3 Setup File

The setup file is constructed in XML format and is divided into three parts that reflect groups of information to be provided by the user: `<pathwayList>`, `<genomeList>` and, optionally, `<blockList>`. An example setup file is shown in Fig. A.1.

<pathwayList>

In this section, the user specifies the set of KEGG pathways to be processed as a series of `<pathway>` elements. There are two ways to specify pathways:

- using KEGG metabolic function groups[7], e.g. carbohydrate metabolism has id 1.1. The declaration `<pathway id=1.1>` means that all pathways in this group are processed.
- using the global KEGG id of an individual pathway, e.g. `<pathway id = 00010>` corresponds to Glycolysis/Gluconeogenesis.

<genomeList>

This section requires the user to specify genome information for the species of interest and other reference organisms. For each organism to be included, the user must define a `<genome>` element. The attribute `name` refers to a species identifier, which will be used by the software for display. For each `<genome>`, the user may provide multiple `<annotation>` sources. These can be of three different kinds, defined by the attribute `type`: `kegg`, `genbank` or `embl`. For `genbank` and `embl`, the user must provide a `filename` for a genome annotation in the respective file format. For the `kegg` annotations, the user provides the `keggId` for the given genome. For example, in the case of *Bacillus subtilis*, set `keggId=bsu`. The user can provide more than one annotation of each type, however all the annotations must have a unique `id` attribute.

For each `<genome>` there are multiple options available, specified by the following attributes:

[7] http://www.kegg.jp/kegg/pathway.html

- `filename`

  The user may supply a FASTA-format file containing amino acid sequences for the predicted proteome.

- `query`

  When set to `true`, this signifies that this genome is the one of main interest. If none of the genomes is set with `query=true`, the first genome with a genome annotation sequence file provided will be considered as the query genome.

- `brenda`

  The full taxonomic name of the organism, which will be used by PathwayBooster to search the BRENDA database in order to retrieve publication information.

- `color`

  The color that should be used to identify the genome in the PathwayBooster display. The accepted format is the RGB color model. This format is constituted by 3 numbers between 0 and 255 separated by commas. An example would be `color="30,40,200"`. If the user does not specify a color, PathwayBooster will attribute one automatically.

- `pathway`

  This controls whether the genome is included in the pathway visualisation. (Default is `true`). There can be a maximum of 7 genomes displayed. A reaction is considered as present if it has either an annotated gene (from any of the annotations provided) or literature evidence (if `brenda` is provided).

- `hamming`

  This controls whether the genome is included in the Hamming distance matrix. (Default is `true`).

<blockList>

This optional section can be used to specify more complex display preferences, for example if the user wants to compare the annotations that were obtained from two different sources for the same organism, these can be separated into different `<block>` elements. All options available for a `<genome>` are also available for a `<block>`, with the exception of `filename`. When the `<blockList>` section is present, the `<genome>` attributes will be overridden for the pathway map, Hamming distance and literature evidence displays.

Within each `<block>` element, the user specifies one or more `<annotationReference>` elements, with an `id` matching that of an `<annotation>` specified previously. By choosing annotations from multiple organisms, it is possible to compare groups of genomes against the query organism.

```
<xml>

<pathwayList>
        <pathway id=”00270”/>
</pathwayList>

<genomeList>
<genome name=”Gt_Ergo” filename=”ERGO/TMO_protein.txt”>
        <annotation type=”embl” id=”Gt_Embl” filename=”ERGO/TMO_embl.txt”/>
        <annotation type=”genbank” id=”Gt_GB” filename=”ERGO/TMO_genbank.txt”/>
</genome>
<genome name=”G_thermoglucosidasius” brenda=”Geobacillus␣thermoglucosidasius”>
        <annotation type=”kegg” id=”Gt_KEGG” keggId=”gth”/>
</genome>
<genome name=”G_kaustophilus” brenda=”Geobacillus␣kaustophilus”>
    <annotation type=”kegg” id=”Gk_KEGG” keggId=”gka”/>
</genome>
<genome name=”G_thermodenitrificans” brenda=”Geobacillus␣thermodenitrificans”>
        <annotation type=”kegg” id=”Gtn_KEGG” keggId=”gtn”/>
</genome>
<genome name=”G_WCH70” brenda=”Geobacillus␣sp.␣WCH70” pathway=”false”>
        <annotation type=”kegg” id=”Gw_KEGG” keggId=”gwc”/>
</genome>
<genome name=”G_Y412MC61” brenda=”Geobacillus␣sp.␣Y412MC61” pathway=”false”>
        <annotation type=”kegg” id=”Gy_KEGG” keggId=”gyc”/>
</genome>
<genome name=”B_subtilis” brenda=”Bacillus␣subtilis” filename=”b.subtilis.pep”>
        <annotation type=”kegg” id=”Bs_KEGG” keggId=”bsu”/>
</genome>
<genome name=”E_coli” brenda=”Escherichia␣coli” filename=”e.coli.pep”>
        <annotation type=”kegg” id=”Ec_KEGG” keggId=”eco”/>
</genome>
</genomeList>


<blocklist>
<block name=”Gt_Ergo_Embl” query=”true” color=”255,0,0” pathway=”true” filename=”ERGO/TMO_protein.txt”>
        <annotationReference id=”Gt_Embl”/>
</block>
<block name=”Gt_Ergo_GB” color=”0,255,0” pathway=”true”>
    <annotationReference id=”Gt_GB”/>
</block>
<block name=”Gt_Kegg” color=”0,0,255” pathway=”true”>
        <annotationReference id=”Gt_KEGG”/>
</block>
<block name=”Gt_Kegg_pub” color=”100,0,100” pathway=”true” brenda=”Geobacillus␣thermoglucosidasius”>
</block>
<block name=”G_WCH70_Y412MC61” pathway=”false”>
        <annotationReference id=”Gk_KEGG”/>
        <annotationReference id=”Gtn_KEGG”/>
</block>
<block name=”G_kaust” pathway=”true” brenda=”Geobacillus␣kaustophilus”>
        <annotationReference id=”Gk_KEGG”/>
</block>
<block name=”G_thermo” pathway=”true” brenda=”Geobacillus␣thermodenitrificans”>
        <annotationReference id=”Gtn_KEGG”/>
</block>
<block name=”Bs_Ec” pathway=”true”>
        <annotationReference id=”Bs_KEGG”/>
        <annotationReference id=”Ec_KEGG”/>
</block>
<block name=”B_sub” pathway=”false” hamming=”false” brenda=”Bacillus␣subtilis” filename=”b.subtilis.pep”>
                <annotationReference id=”Bs_KEGG”/>
</block>
<block name=”E_co”  pathway=”false” hamming=”false” brenda=”Escherichia␣coli” filename=”e.coli.pep”>
                <annotationReference id=”Ec_KEGG”/>
</block>
</blocklist>

</xml>
}
```

Figure A.1: Setup file example for PathwayBooster.

References

Alam, M., Merlo, M., *et al.* (2010). Metabolic modeling and analysis of the metabolic switch in streptomyces coelicolor. *BMC genomics*, **11**(1), 202.

Alpaydin, E. (2004). *Introduction to machine learning*. MIT press.

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., *et al.* (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

Arakaki, A., Huang, Y., and Skolnick, J. (2009). Eficaz2: enzyme function inference by a combined approach enhanced by machine learning. *BMC bioinformatics*, **10**(1), 107.

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. *Nature genetics*, **25**(1), 25.

Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., *et al.* (2008). The rast server: rapid annotations using subsystems technology. *BMC genomics*, **9**(1), 75.

Baart, G., Zomer, B., De Haan, A., Van Der Pol, L., Beuvery, E., Tramper, J., Martens, D., *et al.* (2007). Modeling neisseria meningitidis metabolism: from genome to metabolic fluxes. *Genome Biol*, **8**(7), R136.

Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M., Gajria, B., Grant, G., Ginsburg, H., Gupta, D., Kissinger, J., Labo, P., *et al.* (2003). Plasmodb: the plasmodium genome resource. a database integrating experimental and computational data. *Nucleic acids research*, **31**(1), 212–215.

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

Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., *et al.* (2004). The pfam protein families database. *Nucleic acids research*, **32**(suppl 1), D138–D141.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic acids research*, **28**(1), 235–242.

Borenstein, E. and Feldman, M. W. (2009). Topological signatures of species interactions in metabolic networks. *J Comput Biol*, **16**(2), 191–200.

Borodina, I., Krabben, P., and Nielsen, J. (2005). Genome-scale analysis of streptomyces coelicolor a3 (2) metabolism. *Genome research*, **15**(6), 820–829.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). Uniprotkb/swiss-prot. *database*, **2**, 3.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32. 10.1023/A:1010933404324.

Breman, J., Alilio, M., White, N., Teklehaimanot, A., McCord, G., and Sachs, J. (2007). Scaling up malaria control in africa: An economic and epidemiological assessment.

Brenner, S. *et al.* (1999). Errors in genome annotation. *Trends Genet*, **15**(4), 132–3.

Brooks, S. and Morgan, B. (1995). Optimization using simulated annealing. *The Statistician*, pages 241–257.

Bruns, C., Nowalk, A., Arvai, A., McTigue, M., Vaughan, K., Mietzner, T., and McRee, D. (1997). Structure of haemophilus influenzae fe&plus; 3-binding protein reveals convergent evolution within a superfamily. *Nature Structural & Molecular Biology*, **4**(11), 919–924.

Carlton, J., Angiuoli, S., Suh, B., Kooij, T., Pertea, M., Silva, J., Ermolaeva, M., Allen, J., Selengut, J., Koo, H., *et al.* (2002). Genome sequence and comparative analysis of the model rodent malaria parasite plasmodium yoelii yoelii. *Nature*, **419**(6906), 512–519.

Caspeta, L., Shoaie, S., Agren, R., Nookaew, I., and Nielsen, J. (2012). Genome-scale metabolic reconstructions of pichia stipitis and pichia pastoris and in silico evaluation of their potentials. *BMC Systems Biology*, **6**(1), 24.

Chen, K., Durand, D., and Farach-Colton, M. (2000). Notung: a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology*, **7**(3-4), 429–447.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, **49**(4), 327–335.

Chou, C., Chang, W., Chiu, C., Huang, C., and Huang, H. (2009). Fmm: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic acids research*, **37**(suppl 2), W129–W134.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**(5765), 1283–7.

Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. (2003). Enzyme-specific profiles for genome annotation: Priam. *Nucleic acids research*, **31**(22), 6633–6639.

Croes, D., Couche, F., Wodak, S. J., and van Helden, J. (2006). Inferring meaningful pathways in weighted metabolic networks. *Journal of molecular biology*, **356**(1), 222–36.

David, H., Özçelik, İ., Hofmann, G., and Nielsen, J. (2008). Analysis of aspergillus nidulans metabolism at the genome-scale. *BMC genomics*, **9**(1), 163.

Dayhoff, M. *et al.* (1976). The origin and evolution of protein superfamilies. In *Federation Proceedings*, volume 35, page 2132.

de Oliveira Dal'Molin, C., Quek, L., Palfreyman, R., Brumbley, S., and Nielsen, L. (2010). Aragem, a genome-scale reconstruction of the primary metabolic network in arabidopsis. *Plant Physiology*, **152**(2), 579–589.

Delcher, A., Harmon, D., Kasif, S., White, O., and Salzberg, S. (1999). Improved microbial gene identification with glimmer. *Nucleic acids research*, **27**(23), 4636–4641.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet*, **17**(8), 429–31.

Dittrich, S., Mitchell, S., Blagborough, A., Wang, Q., Wang, P., Sims, P., and Hyde, J. (2008). An atypical orthologue of 6-pyruvoyltetrahydropterin synthase can provide the missing link in the folate biosynthesis pathway of malaria parasites. *Molecular microbiology*, **67**(3), 609–618.

Edgar, R. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**(5), 1792–1797.

Endy, D. (2005). Foundations for engineering biology. *Nature*, **438**(7067), 449–453.

Engelhardt, B. E., Jordan, M. I., Repo, S. T., and Brenner, S. E. (2009). Phylogenetic molecular function annotation. *J Phys*, **180**(1), 12024.

Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L., and Palsson, B. O. (2009). Reconstruction of biochemical networks in microorganisms. *Nat Rev Micro*, **7**, 129–143.

Felsenstein, J. (1993). PHYLIP - Phylogeny Inference Package (Version 3.5). *Distributed by the author. Department of Genetics, University of Washington, Seattle.*

Felsenstein, J. (2004). Inferring phytogenies. *Sunderland, Massachusetts: Sinauer Associates.*

Forslund, K. and Sonnhammer, E. L. L. (2008). Predicting protein function from domain content. *Bioinformatics*, **24**(15), 1681–7.

Förster, J., Famili, I., Fu, P., Palsson, B., and Nielsen, J. (2003). Genome-scale reconstruction of the saccharomyces cerevisiae metabolic network. *Genome research*, **13**(2), 244–253.

Francke, C., Siezen, R., and Teusink, B. (2005). Reconstructing the metabolic network of a bacterium from its genome. *TRENDS in Microbiology*, **13**(11), 550–558.

Frishman, D. (2007). Protein annotation at genomic scale: the current status. *Chem Rev*, **107**(8), 3448–66.

Galperin, M. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biology*.

Gardner, M., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R., Carlton, J., Pain, A., Nelson, K., Bowman, S., *et al.* (2002). Genome sequence of the human malaria parasite plasmodium falciparum. *Nature*, **419**(6906), 498–511.

Gaskell, E., Smith, J., Pinney, J., Westhead, D., and McConkey, G. (2009). A unique dual activity amino acid hydroxylase in toxoplasma gondii. *PLoS One*, **4**(3), e4801.

Gherardini, P., Wass, M., Helmer-Citterich, M., and Sternberg, M. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of molecular biology*, **372**(3), 817–845.

Gilks, W. R., Audit, B., Angelis, D. D., Tsoka, S., and Ouzounis, C. A. (2002). Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**(12), 1641–9.

Ginsburg, H. (2000). Malaria parasite metabolic pathways.

Gonzalez, O., Gronau, S., Falb, M., Pfeiffer, F., Mendoza, E., Zimmer, R., and Oesterhelt, D. (2008). Reconstruction, modeling & analysis of halobacterium salinarum r-1 metabolism. *Mol. BioSyst.*, **4**(2), 148–159.

Gonzalez, O., Oberwinkler, T., Mansueto, L., Pfeiffer, F., Mendoza, E., Zimmer, R., and Oesterhelt, D. (2010). Characterization of growth and metabolism of the haloalkaliphile natronomonas pharaonis. *PLoS computational biology*, **6**(6), e1000799.

Górniak, M., Paun, O., and Chase, M. W. (2010). Phylogenetic relationships within orchidaceae based on a low-copy nuclear coding gene,¡ i¿ xdh¡/i¿: Congruence with organellar and nuclear ribosomal dna results. *Molecular phylogenetics and evolution*, **56**(2), 784–795.

Gough, J. and Chothia, C. (2002). Superfamily: Hmms representing all proteins of known structure. scop sequence searches, alignments and genome assignments. *Nucleic Acids Research*, **30**(1), 268–272.

Green, M. and Karp, P. (2004). A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC bioinformatics*, **5**(1), 76.

Hadley, C. and Jones, D. (1999). A systematic comparison of protein structure classifications: Scop, cath and fssp. *Structure*, **7**(9), 1099–1112.

Hall, N., Karras, M., Raine, J., Carlton, J., Kooij, T., Berriman, M., Florens, L., Janssen, C., Pain, A., Christophides, G., *et al.* (2005). A comprehensive survey of the plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, **307**(5706), 82–86.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, **402**(6761 Suppl), C47–52.

Henry, C., DeJongh, M., Best, A., Frybarger, P., Linsay, B., and Stevens, R. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature biotechnology*, **28**(9), 977–982.

Herrgård, M., Swainston, N., Dobson, P., Dunn, W., Arga, K., Arvas, M., Büthgen, N., Borger, S., Costenoble, R., Heinemann, M., *et al.* (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature biotechnology*, **26**(10), 1155–1160.

Hertz-Fowler, C., Peacock, C., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., *et al.* (2004). Genedb: a resource for prokaryotic and eukaryotic organisms. *Nucleic acids research*, **32**(suppl 1), D339–D343.

Holliday, G., Bartlett, G., Almonacid, D., O'Boyle, N., Murray-Rust, P., Thornton, J., and Mitchell, J. (2005). Macie: a database of enzyme reaction mechanisms. *Bioinformatics*, **21**(23), 4315–4316.

Höök, M. and Tang, X. (2012). Depletion of fossil fuels and anthropogenic climate changea review. *Energy Policy*.

Hsiao, T.-L., Revelles, O., Chen, L., Sauer, U., and Vitkup, D. (2010). Automatic policing of biochemical annotations using genomic correlations. *Nature Chemical Biology*, **6**(1), 34–40.

Hung, S., Wasmuth, J., Sanford, C., and Parkinson, J. (2010). Detecta density estimation tool for enzyme classification and its application to plasmodium falciparum. *Bioinformatics*, **26**(14), 1690–1698.

Hyland, C., Pinney, J., McConkey, G., and Westhead, D. (2006). metashark: a www platform for interactive exploration of metabolic networks. *Nucleic acids research*, **34**(suppl 2), W725–W728.

Islam, M., Edwards, E., and Mahadevan, R. (2010). Characterizing the metabolism of dehalococcoides with a constraint-based model. *PLoS computational biology*, **6**(8), e1000887.

Jones, C. E., Brown, A. L., and Baumann, U. (2007). Estimating the annotation error rate of curated go database sequence annotations. *BMC Bioinformatics*, **8**, 170.

Jones, D. T. *et al.* (1999). Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology*, **287**(4), 797–815.

Jordan, I., Rogozin, I., Wolf, Y., and Koonin, E. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome research*, **12**(6), 962–968.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, **34**(Database issue), D354–7.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic acids research*, **36**(Database issue), D480–4.

Kapranov, P. and Laurent, G. (2012). Genomic dark matter: Implications for understanding human disease mechanisms, diagnostics, and cures. *Frontiers in Genetics*, **3**.

Karp, P., Paley, S., and Romero, P. (2002). The pathway tools software. *Bioinformatics*, **18**(suppl 1), S225–S232.

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., *et al.* (2010). Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, **11**(1), 40–79.

Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3d-pssm. *Journal of molecular biology*, **299**(2), 501–522.

Kim, H., Kim, S., Jeong, H., Kim, T., Kim, J., Choy, H., Yi, K., Rhee, J., and Lee, S. (2011a). Integrative genome-scale metabolic analysis of vibrio vulnificus for drug targeting and discovery. *Molecular systems biology*, **7**(1).

Kim, T., Sohn, S., Kim, Y., Kim, W., and Lee, S. (2011b). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current opinion in biotechnology*.

Kotsiantis, S., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Frontiers in Artificial Intelligence and Applications*, **160**, 3.

Kreimer, A., Borenstein, E., Gophna, U., and Ruppin, E. (2008). The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, **105**(19), 6976.

Kuo, C., Wares, J., and Kissinger, J. (2008). The apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular biology and evolution*, **25**(12), 2689–2698.

Kuriyan, J., Krishna, T., Wong, L., Guenther, B., Pahler, A., Williams, C., and Model, P. (1991). Convergent evolution of similar function in 2 structurally divergent enzymes.

Kwangmin, C. and Sun, K. (2008). Compath: comparative enzyme analysis and annotation in pathway/subsystem contexts. *BMC Bioinformatics*, **9**.

Langley, P. and Simon, H. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, **38**(11), 54–64.

Lee, D., Burd, H., Liu, J., Almaas, E., Wiest, O., Barabási, A., Oltvai, Z., and Kapatral, V. (2009). Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple staphylococcus aureus genomes identify novel antimicrobial drug targets. *Journal of bacteriology*, **191**(12), 4015–4024.

Lee, T., Huang, H., Hung, J., Huang, H., Yang, Y., and Wang, T. (2006). dbptm: an information repository of protein post-translational modification. *Nucleic acids research*, **34**(suppl 1), D622–D627.

Letunic, I. and Bork, P. (2007). Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**(1), 127–8.

Letunic, I. and Bork, P. (2011). Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic acids research*, **39**(Web Server issue), W475–8.

Liao, Y., Tsai, M., Chen, F., and Hsiung, C. (2012). Gemsirv: a software platform for genome-scale metabolic model simulation, reconstruction and visualization. *Bioinformatics*, **28**(13), 1752–1758.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, **2**(3), 18–22.

Lopez, G., Valencia, A., and Tress, M. (2007). Firedba database of functionally important residues from proteins of known structure. *Nucleic acids research*, **35**(suppl 1), D219–D223.

Makarova, K. and Grishin, N. (1999). The zn-peptidase superfamily: functional convergence after evolutionary divergence. *Journal of molecular biology*, **292**(1), 11–17.

Mazumdar, V., Snitkin, E., Amar, S., and Segrè, D. (2009). Metabolic network model of a human oral pathogen. *Journal of bacteriology*, **191**(1), 74–90.

Meyer, F., Overbeek, R., and Rodriguez, A. (2009). Figfams: yet another set of protein families. *Nucleic acids research*, **37**(20), 6643–6654.

Milenkoviæ, T. and Pržulj, N. (2008). Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, **6**, 257.

Morett, E., Korbel, J., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B., and Bork, P. (2003). Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature biotechnology*, **21**(7), 790–795.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*, **35**(suppl 2), W182–W185.

Murray, C., Rosenfeld, L., Lim, S., Andrews, K., Foreman, K., Haring, D., Fullman, N., Naghavi, M., Lozano, R., and Lopez, A. (2012). Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet*, **379**(9814), 413–431.

Murzin, A., Brenner, S., Hubbard, T., Chothia, C., *et al.* (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, **247**(4), 536–540.

Navid, A. and Almaas, E. (2009). Genome-scale reconstruction of the metabolic network in yersinia pestis, strain 91001. *Mol. BioSyst.*, **5**(4), 368–375.

Nerima, B., Nilsson, D., and Mäser, P. (2010). Comparative genomics of metabolic networks of free-living and parasitic eukaryotes. *BMC genomics*, **11**, 217.

Nogales, J., Gudmundsson, S., Knight, E., Palsson, B., and Thiele, I. (2012). Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proceedings of the National Academy of Sciences*, **109**(7), 2678–2683.

Nowak, M. A. (1997). Evolution of genetic redundancy. *Nature*, **388**, 167.

Ochman, H. and Moran, N. A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*, **292**(5519), 1096–9.

Oehm, S., Gilbert, D., Tauch, A., Stoye, J., and Goesmann, A. (2008). Comparative pathway analyzer: a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms. *Nucleic acids research*, **36**(suppl 2), W433–W437.

Oh, Y., Palsson, B., Park, S., Schilling, C., and Mahadevan, R. (2007). Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *Journal of Biological Chemistry*, **282**(39), 28791–28799.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., and Thornton, J. (1997). Cath–a hierarchic classification of protein domain structures. *Structure*, **5**(8), 1093–1109.

Osterman, A. and Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. *Current opinion in chemical biology*, **7**(2), 238–251.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov Jr, E., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., *et al.* (2003). The ergotm genome analysis and discovery system. *Nucleic acids research*, **31**(1), 164–171.

Pabinger, S., Rader, R., Agren, R., Nielsen, J., and Trajanoski, Z. (2011). Memosys: Bioinformatics platform for genome-scale metabolic models. *BMC systems biology*, **5**(1), 20.

Pagel, M. and Meade, A. (2007). Bayestraits. *Univ. Reading, URL:[http://www. evolution. rdg. ac. uk/BayesTraits. html].*

Pain, A., Böhme, U., Berry, A., Mungall, K., Finn, R., Jackson, A., Mourier, T., Mistry, J., Pasini, E., Aslett, M., *et al.* (2008). The genome of the simian and human malaria parasite plasmodium knowlesi. *Nature*, **455**(7214), 799–803.

Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, **7**(1), 169.

Pinchuk, G., Hill, E., Geydebrekht, O., De Ingeniis, J., Zhang, X., Osterman, A., Scott, J., Reed, S., Romine, M., Konopka, A., *et al.* (2010). Constraint-based model of shewanella oneidensis mr-1 metabolism: a tool for data analysis and hypothesis generation. *PLoS computational biology*, **6**(6), e1000822.

Pinney, J., Shirley, M., McConkey, G., and Westhead, D. (2005). metashark: software for automated metabolic network prediction from dna sequence and its application to the genomes of plasmodium falciparum and eimeria tenella. *Nucleic acids research*, **33**(4), 1399–1409.

Pinney, J., Papp, B., Hyland, C., Wambua, L., Westhead, D., and McConkey, G. (2007). Metabolic reconstruction and analysis for parasite genomes. *TRENDS in Parasitology*, **23**(11), 548–554.

Plata, G., Hsiao, T.-L., Olszewski, K. L., Llinás, M., and Vitkup, D. (2010). Reconstruction and flux-balance analysis of the plasmodium falciparum metabolic network. *Mol Syst Biol*, **6**, 408.

Ponce-Ortega, J., Serna-González, M., and Jiménez-Gutiérrez, A. (2009). Use of genetic algorithms for the optimal design of shell-and-tube heat exchangers. *Applied Thermal Engineering*, **29**(2), 203–209.

Pond, S. and Muse, S. (2005). Hyphy: hypothesis testing using phylogenies. *Statistical methods in molecular evolution*, pages 125–181.

Porter, C., Bartlett, G., and Thornton, J. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids research*, **32**(suppl 1), D129–D133.

Price, N. D., Reed, J. L., and Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, **2**(11), 886–97.

Reed, J., Vo, T., Schilling, C., Palsson, B., *et al.* (2003). An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, **4**(9), R54.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2004). Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, **6**(1), R2.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4**(4), 406–425.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J., and Schomburg, D. (2011). Brenda, the enzyme information system in 2011. *Nucleic acids research*, **39**(suppl 1), D670–D676.

Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, **5**(12), e1000605.

Sonego, P., Kocsor, A., and Pongor, S. (2008). Roc analysis: applications to the classification of biological sequences and 3d structures. *Briefings in bioinformatics*, **9**(3), 198–209.

Stamatakis, A. (2006). Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**(21), 2688–2690.

Suthers, P., Dasika, M., Kumar, V., Denisov, G., Glass, J., and Maranas, C. (2009). A genome-scale metabolic reconstruction of mycoplasma genitalium, ips189. *PLoS computational biology*, **5**(2), e1000285.

Ta, H. X. and Holm, L. (2009). Evaluation of different domain-based methods in protein interaction prediction. *Biochemical and biophysical research communications*, **390**(3), 357–62.

Taylor, M., Eley, K., Martin, S., Tuffin, M., Burton, S., and Cowan, D. (2009). Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. *Trends in biotechnology*, **27**(7), 398–405.

Teusink, B., Wiersma, A., Molenaar, D., Francke, C., De Vos, W., Siezen, R., and Smid, E. (2006). Analysis of growth of lactobacillus plantarum wcfs1 on a complex medium using a genome-scale metabolic model. *Journal of Biological Chemistry*, **281**(52), 40041–40048.

Thiele, I. and Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, **5**(1), 93–121.

Thomas, G., Zucker, J., Macdonald, S., Sorokin, A., Goryanin, I., and Douglas, A. (2009). A fragile metabolic network adapted for cooperation in the symbiotic bacterium buchnera aphidicola. *BMC Systems Biology*, **3**(1), 24.

Tokuriki, N. and Tawfik, D. (2009). Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, **19**(5), 596–604.

Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc Biol Sci*, **268**(1478), 1803–10.

Webb, E. *et al.* (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press.

Widiastuti, H., Kim, J., Selvarasu, S., Karimi, I., Kim, H., Seo, J., and Lee, D. (2011). Genome-scale modeling and in silico analysis of ethanologenic bacteria zymomonas mobilis. *Biotechnology and Bioengineering*, **108**(3), 655–665.

Woodrow, C., Haynes, R., and Krishna, S. (2005). Artemisinins. *Postgraduate medical journal*, **81**(952), 71–78.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS*, **13**(5), 555–556.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, **19**(6), 908–917.

Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X., and Orengo, C. (2008).

Gene3d: comprehensive structural and functional annotation of genomes. *Nucleic acids research*, **36**(suppl 1), D414–D418.

Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D., and Altman, R. B. (2004). Computational analysis of plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome research*, **14**(5), 917–24.

Zhang, H., Fritts, J., and Goldman, S. (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, **110**(2), 260–280.