COMPUTATIONAL APPROACHES TO IMPROVING THE RECONSTRUCTION OF METABOLIC PATHWAYS

Faizah Aplop

A Thesis
in
the department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Concordia University
Montréal, Québec, Canada

September 2016

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: Mrs. Faizah Aplop

Entitled: Computational Approaches to Improving the Reconstruction of Metabolic Pathways

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr M. Reza Soleymani

_____ External Examiner
Dr Anthony J. Kusalik

_____ Examiner
Dr Nawwaf Kharma

_____ Examiner
Dr Volker Haarslev

_____ Examiner
Dr Lata Narayanan

_____ Supervisor
Dr Gregory Butler

Approved _____
Chair of Department or Graduate Program Director

_____ 20 _____ _____
Dr Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Computational Approaches to Improving the Reconstruction of Metabolic Pathways

Faizah Aplop, Ph.D.
Concordia University, 2016

Metabolic pathway reconstruction is the essence of systems biology where *in silico* modeling and prediction of the cell's function is based on the interaction of the cell's components represented as a network of reactions. The reconstructed model and the associated database of information about the organism's genes and their functional roles facilitate a variety of analysis and simulation techniques that can enrich our understanding. However, there are unresolved issues for genome-scale metabolic network reconstruction, such as our incomplete knowledge of the cell's networks for metabolism, transport, and regulation; the completeness, accuracy, and specificity of the annotation of genomes; and our ability to fully utilise the available information from -omics (genomics, proteomics, metabolomics, etc) for the reconstruction of the networks. These issues result in incomplete metabolic models, which limit our ability to perform analysis of and to make predictions about the cell that are based on the network model.

This dissertation discusses the state-of-the-art of metabolic pathway reconstruction and highlights the outstanding issues. In particular, we consider a number of case studies using genomes of fungi relevant to industrial applications, such as biofuels, to demonstrate the performance of existing techniuqes and illustrate the issues. Our case studies focus on the cell's central metabolism, and the utilisation and transport of sugars as a carbon source, since these are essential concerns for industrial applications.

A significant deficiency in the existing state-of-the-art for the reconstruction of metabolic pathways is the ability to associate genes and proteins to the transport reactions that move specific compounds across the membranes of the cell. The dissertation reviews the state-of-the-art of prediction methods for transmembrane transport proteins by developing a scheme to describe and compare existing methods, and applying the existing techniques to the

fungal genome of *A. niger* CBS 513.88. This reveals the split between those methods that use the Transporter Classification (TC) as their target for prediction, and those that use the type of chemical substrates being transported as their target. Despite this difficulty in comparing approaches, it is clear that the state-of-the-art cannot predict specific substrates being transported, and hence cannot associate genes and proteins to the transport reactions.

The dissertation presents TransATH, which stands for Transporters via ATH (Annotation Transfer by Homology), a system which automates Saier's protocol and includes the computation of subcellular localization and improves the computation of transmembrane segments. The choice of thresholds for the parameters of TransATH is investigated to determine optimal peformance as defined by a gold standard set of transporters and non-transporters from *S. cerevisiae*. The dissertation demonstrates TransATH on the fungal genome of *A. niger* CBS 513.88 and evaluates the correctness of TransATH using the curated information in AspGD (the Aspergillus Database). A website for TransATH is available for use.

Acknowledgments

I wish to express my sincere gratitude to Dr. Gregory Butler for always being there for me. I am extremely grateful and indebted to him for his expertise, time, patience, continuous support, sincere and unwavering guidance, and encouragement extended to me. Also, I would like to place on record my sincere gratitude to my committee members, Dr. Volker Haarslev, Dr. Lata Narayanan and Dr. Nawwaf Kharma for their patience, precious time, constructive advice throughout the processes and challenges in this work. A special thanks to my external examiner, Dr. Anthony J. Kusalik for his support and invaluable advice.

My sincere thanks to my lab mates — Christine Kehyayan, Lin Cheng, Yi Qing, Stuart Thiel, Jun Luo, Jianlong Qi and Stephen Barrett — for being a great source of friendships. I owe a deep sense of gratitude to all the good people at the Centre for Structural and Functional Genomics and the Department of Computer Science and Software Engineering at Concordia University that have provided me with theoretical and technical support.

I thank profusely the good people of Universiti Malaysia Terengganu, my supportive colleagues and friends, Noorasiah Moidu and her team, for being very tolerant with me throughout my study period. Also, I am extremely thankful to my good friends Abu Suffian Abu Bakar, Yoisel Melis Santana and Marie-France Lessard for their continuous support.

It is my privilege to thank my loving husband, Mohd Riduan Abd Rahim for his understanding, sacrifice and support throughout my research period.

To my beloved parents, Aplop Awang and Che' Ramlah Ismail, who have been a source of inspiration to me throughout my life, a very special thank you for your unconditional love, prayer, and nurture. I dedicate this work to my family, many friends and the apple of my eye, Qurratul Aini Aplop.

Above all, my utmost appreciation and grateful to The Almighty God for enabling me to complete this thesis.

Contents

List of Figures

List of Tables

List of Terms and Abbreviations

AAC Amino acid composition: the frequency of each amino acid in a protein

AAindex Database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids

ABC ATP-binding cassette

ADP Adenosine diphosphate

Alignment The process, or its result, of matching sequences to maximize an objective function

Amino acid One of the 20 chemical building blocks that form a polypeptide chain of a protein

AQUA Automated quality improvement for multiple sequence alignment: algorithmm used in construction of eggNOG

AspGD Aspergillus Genome Database `www.aspgd.org`

ATP Adenosine triphosphate

AutoGraph Automatic Transfer by Orthology of Gene Reaction Associations for Pathway Heuristics: semi-automated approach for reconstruction of metabolic pathways with hole-filling using orthology

Base pair Pair of bases held together by hydrogen bonds that form the core of DNA and RNA: A-T, G-C and A-U interactions

BBH Bidirectional best hit: approach to determine orthologs

BiGG Biochemical Genetic and Genomic knowledgebase: repository of systems biology models

BioPAX Biological Pathways Exchange: consortium for standards in pathways

BLAST Basic Local Alignment Search Tool: a heuristic algorithm for pairwise sequence alignment

blastp BLAST program to search a proten sequence as a query against a database of protein sequences

Blast+ Software package from NCBI which is latest version of implementation of BLAST

BP Biological Process domain of the Gene Ontology

BRENDA The Comprehensive Enzyme Information System: database of enzymes and their properties

CC Cellular Component domain of the Gene Ontology

CCM Central carbon metabolism

CDS Coding sequence

ChEBI Chemical Entities of Biological Interest: ontology and related database

Clustal Family of algorithms for multiple sequence alignment

Clustal Omega Latest member of Clustal family

COBRA COnstraints Based Reconstruction and Analysis: toolkit for systems biology

COG Clusters of Orthologous Groups: database for a phylogenetic classification of the proteins

DNA Deoxyribonucleic acid: a basis for genetic material in the cell

EC Enzyme Commission of IUPAC

EC Number Enzyme Commission identifer for an enzyme

eggNOG Orthologous groups and functional annotation database

EM Expectation maximization

EMBL European Molecular Biology Laboratory

EMP/MPW Enzyme and Metabolic Pathways database

Enzyme Class of proteins that are capable of catalyzing chemical reactions by making or breaking chemical bonds

FBA Flux balance analysis

FIG The Fellowship for Interpretation of Genomes

FigFAMS A collection of over 100 000 protein families that are the product of manual curation and close strain comparison

G-BLAST Genome Basic Local Alignment Tool: software from Saier lab for prediction of transporters

G-BLAST(v2) Version 2 of G-BLAST

Gene Unit of inheritance and the region of DNA encoding it

Gene Ontology Set of three controlled vocabularies to describe the role of a gene product

Gene product Protein or RNA that results from expression of a gene

GENRE Genome-scale network reconstruction

GLOBUS Global Biochemical reconstruction Using Sampling: algorithm for hole-filling

GO Gene Ontology

GPR Gene-Protein-Reaction association

HMM Hidden Markov Model

HMMER Software suite for sequence analysis using profile hidden Markov models

HMMTOP Transmembrane topology prediction program

Homology Two or more biological species, systems or molecules that share a common evolutionary ancestor

HSP High scoring pair: region of alignment of two sequences computed by BLAST

IdentiCS Identification of Coding Sequences from Unfinished Genome Sequences

IMP Integral membrane proteins are permanently attached to a membrane

IUBMB International Union of Biochemistry and Molecular Biology

IUPAC International Union of Pure and Applied Chemistry

JDet Software for determining specificity-determining sites given an MSA

JGI Joint Genome Institute

KAAS KEGG Automatic Annotation Server

KEGG Kyoto Encyclopedia of Genes and Genomes

KOBAS KEGG Orthology Based Annotation System

LocTree3 Software for protein subcellular localization prediction

MAFFT MSA program using fast Fourier transforms

MAST Motif Alignment & Search Tool

Mbp Mega base pair: one million base pairs

MCL Markov clustering algorithm and software

MEME Multiple EM for Motif Elicitation

MEMSAT Software for transmembrane helix prediction

Metabolic pathway Series of reactions involved in metabolism

Metabolism The chemical reactions involved in maintaining the living state of the cells and the organism

MetaCyc Highly curated nonredundant reference database of small-molecule metabolism

metaSHARK Metabolic Search And Reconstruction Kit

MF Molecular Function domain of the Gene Ontology

MFS Major Facilitator Superfamily of TCDB

MOD Model organism database

Motif Conserved element of a protein sequence alignment that usually correlates with a particular function

mRNA Messenger RNA

MS Mass spectrometry

MSA Multiple sequence alignment

MSA-AAC Multiple sequence alignment - amino acid composition: vector of frequencies of amino acids in a protein derived from a MSA

MUSCLE MUltiple Sequence Comparison by Log-Expectation: software for MSA

NADPH Nicotinamide Adenine Dinucleotide Phosphate

NGS Next-generation sequencing

NorMD Sum-of-pairs MSA based on Mean Distance used as a measure of quality of an MSA

ORF Open Reading Frame: stretch of DNA that potentially encodes a protein

Ortholog Orthologs are genes in different species that evolved from a common ancestral gene by a speciation event forming two separate species

PAAC Pair amino acid composition: frequency of adjacent pairs of amino acids in a protein

PantoGraph Software for reconstruction of metabolic pathways using orthology

PGDB Pathway genome database created by Pathway Tools

Pfam Collection of protein families represented by multiple sequence alignments and hidden Markov models (HMMs)

Phobius Software for prediction of transmembrane topology and signal peptides

PipeAlign A toolkit for protein family analysis

PPP Pentose Phosphate Pathway

Profile Sequence profile is usually derived from multiple alignments of sequences with a known relationship, and represented as a PSSM or HMM

Protein Macromolecule that consists of a sequence of amino acids

PRIAM PRofils pour l'Identification Automatique du Métabolisme: software to predict EC number of a protein

PseAAC Pseudo amino acid composition

PsePAAC Pseudo pair amino acid composition

PSORTb Protein localization predictor for bacteria

PSSM Position-Specific Scoring Matrix

RASCAL Rapid scanning and correction of MSA: software component of PipeAlign

RAxML Randomized Axelerated Maximum Likelihood: algorithm for construction of a phylogenetic tree

RNA Ribonucleic acid

RNA-Seq Next-generation RNA sequencing

SBML Systems Biology Markup Language

SEED Analysis tool from FIG for annotation of prokaryotes including pathway reconstruction

SGD Saccharomyces Genome Database `www.yeastgenome.org`

SMILES Simplified Molecular Input Line Entry System: text notation for chemical compounds

T-Coffee Algorithm for MSA

TC Transporter classification scheme of IUBMB

TCA Tricarboxylic acid

TCDB Transporter classification database `www.tcdb.org`

TCDB-BLAST Our software for prediction of transporters using blastp search of TCDB

TIP Transport Inference Parser: module in Pathway Tools to predict transporters and transport reactions

TM-Coffee Algorithm for MSA for transmembrane proteins

TMHMM TransMembrane helix prediction using Hidden Markov Models

TMS Transmembrane segment

TransATH Our software for prediction of transporters `transath.umt.edu.my`

Transmembrane protein Protein that spans the membrane

Transmembrane segment The region of a transmembrane protein that actually spans the membrane

Transport The directed movement of a molecule into, out of, or within a cell, or between cells

TransportDB Transporter database primarily for prokaryotes

Transporter Protein carrying out transport

TransportTP A genome-scale membrane transporter prediction and characterization system

TrSSP Transporter Substrate Specificity Prediction Server

Transitivity Clustering Algorithm and software for hierarchical clustering

WHAT Web-based program for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence

WoLF PSORT Protein localization predictor

Chapter 1

Introduction

This thesis deals with computational aspects of the automatic reconstruction of the metabolic pathways of an organism, given an annotated genome of the organism, a body of knowledge and data captured in public web resources, and optionally a collection of other data from modern biotechnological instruments. It is motivated by the critical role of genome-scale network reconstructions (GENREs) of metabolism in systems biology, and the significant impact of systems biology on biology today, especially in industrial applications. It addresses challenges in automating manual steps of the process, and in improving existing algorithms for the steps.

Systems biology has become central to biology after the success of high throughput technology in genome sequencing. It encompasses a holistic approach to the study of biology and the objective is to simultaneously monitor all biological processes operating as an integrated system [Roe12]. According to [Pal08], the complex and dynamic behaviour of living systems drive researchers to innovate from an reductionist approach to an integrative approach in examining how biological components interact to generate whole cell functions.

Biological systems consist of atoms, such as carbon, oxygen, hydrogen, nitrogen, sulphur, and phosphorus, that are the main elements in the building blocks of cell structure and cell function: nucleic acids (DNA and RNA), proteins, carbohydrates and lipids [Wal06]. The genome of the organism encodes the genes for these building blocks.

Systems biology plays an important role in the life science industry specifically in *synthetic biology*. One of its major applications is within the field of *metabolic engineering*, where genetic modifications of cell factories are done [COHA$^+$10]. The goal is to produce strains of

the original organism that can contribute in the manufacturing of bioproducts for industrial use. To achieve such a goal, the functions of genes and gene products, and the relationships between an organism's genome and its phenotypes need to be understood, at least in part. Computer simulation is utilized to perform integrative analysis on genetic characteristics (genotype) in order to predict the physiological properties (phenotype) by reconstructing biochemical reaction networks. An enormous challenge is to integrate the different levels of information pertaining to genes, RNAs, proteins, and pathways that make up a cell or an organism. To study them, qualitative and quantitative measurements of the behaviour of groups of interacting components are taken using genomics, transcriptomics, proteomics and metabolomics, followed by systematic application of bioinformatics tools and technologies. Computational models are used to describe and predict the dynamic behaviour of cellular systems. However, the use of the data obtained from studies with different -omics techniques is not simple; for example, there are situations where genes encode for several different proteins (isozymes) that can complicate data integration [Roe12].

Metabolic pathway reconstruction is a starting point of systems biology where basic biochemical pathways for a specific organism are modelled. One of its main purposes is to understand the function of each gene and the proteins to reveal their roles in that organism [Ray06]. This functional assignment between gene/protein and metabolism can be considered as the first step of the biochemical data integration process [Roe12]. The metabolic model is in the form of a network of interactions of the cell's components. The network is the basis for *in silico* prediction of the cell's mechanisms and behaviour. This metabolic network model becomes the focal point of systems biology and allows the integration of various data types in a form suitable for mathematical analysis [BN05]. The metabolic network can be reconstructed through Gene-Protein-Reaction (GPR) associations and the properties of the reactions enable mathematical constraint-based approaches such as Flux Balance Analysis (FBA) [Pal08]. The key is to transform the metabolic modeling information to mathematical representations such as a stoichiometry matrix in order to facilitate and perform computations [FPG10]. The reconstructed model and the associated database of information about the organism's genes and their functional roles will facilitate a variety of analysis and simulation techniques to help understand the cell system and answer specific biological questions [CBS05].

The integration of -omics data and genome-scale metabolic models through the utilization of computational tools has moved biology from a phenomenological to a predictive science

[COHA+10]. Efforts by researchers in computer science, mathematics, statistics, and biology who are working together in developing the necessary tools to acquire, store, analyze, model, and distribute this information have given rise to the systems biology paradigm of *"components to networks to in silico models to phenotype"* [Pal08].

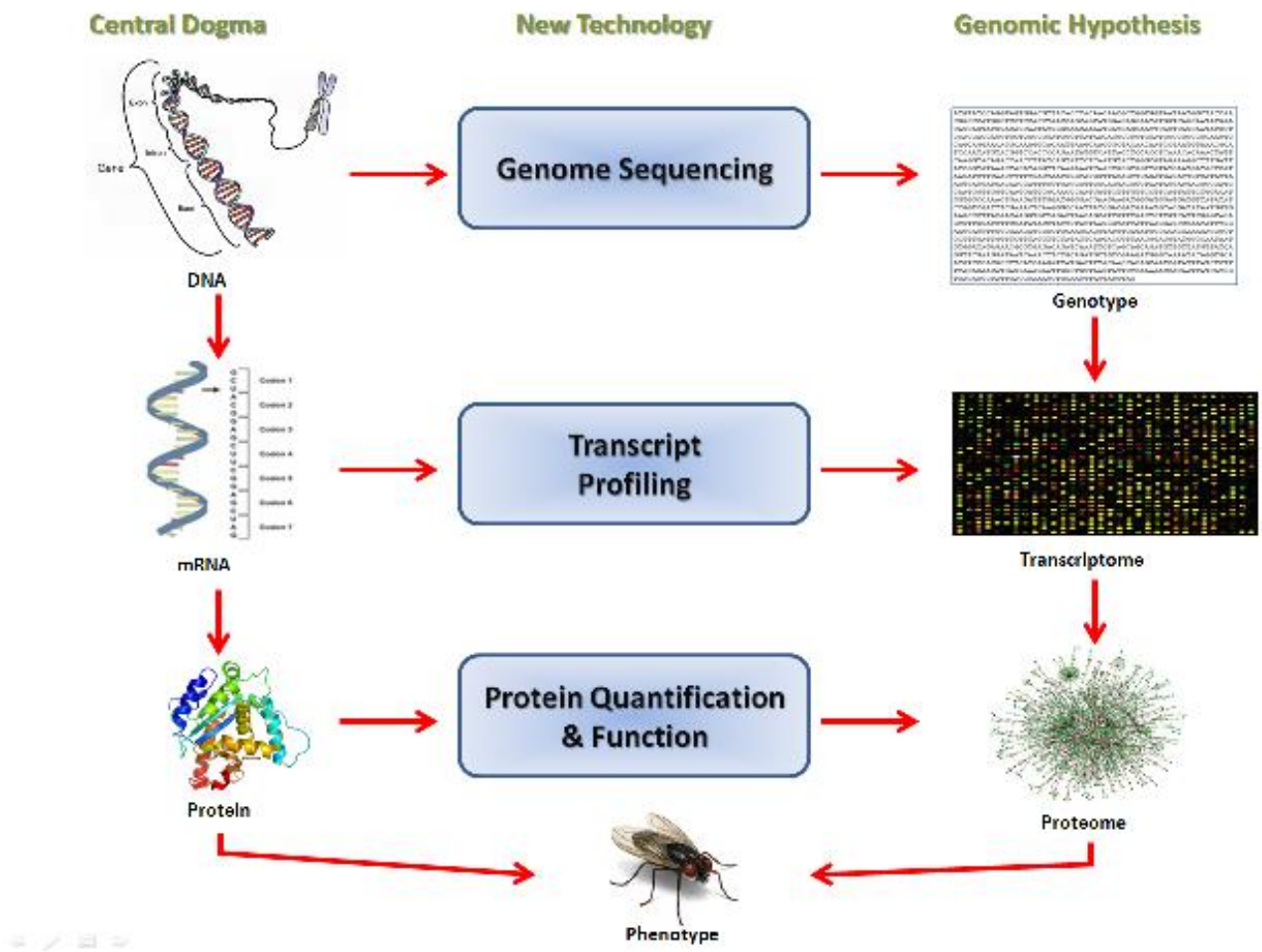


Figure 1: Relating Hypotheses from -Omics to the Central Dogma

In the development of functional genomics technologies, the analysis of genome, transcriptome, proteome and metabolome are critical because understanding interconnections between DNA, gene, RNA, and protein towards function is one of the great biological mysteries. The term *genome* refers to a complete genetic sequence (DNA) of an organism. It contains the entire heredity information of an organism encoded in DNA or RNA. For multicellular organisms, the genome consists of genes and non-coding regions of the organism. The *transcriptome* is the complete set of RNA transcripts produced from the genome at any one time. It includes coding sequences (CDS) that can be translated into proteins for those genes

(potentially) active at the point of time. The *proteome* is the full complement of proteins expressed by the genome at a given time. The *metabolome* consists of all metabolites — that is, small chemical compounds — produced by an organism at a given time. The metabolites are inputs (*substrates*) and outputs (*products*) of reactions catalyzed by enzymes. These reactions form the metabolic pathways. Figure 1 shows how each stage of the central dogma relates to -omics data from the new high-throughput technologies of genome sequencing (next-generation DNA sequencing (NGS)), transcript profiling (RNA-Seq, next-generation sequencing of transcripts) and protein identification and quantification (mass spectrometry (MS)).

1.1 Genome-Scale Network Reconstruction

An organism carries out a range of processes, such as

- reproduction;
- cell growth;
- cell differentiation;
- metabolism;
- response to stimuli; and
- death.

An overview of cell processes can be seen in the Biological Process (BP) aspect of the abbreviated Gene Ontology [The00], the so-called GO Slim.

Thiele and Palsson [TP10] present a comprehensive protocol to develop a GENRE (see Section 2.4) that involves considerable manual curation, iteration, and quality control. In general, the level of curation required limits the application of the protocol to model organisms, or at least those organisms with a well-funded, large research community. Recent advances in biotechnology has improved speed and accuracy, and lowered the cost of sequencing in particular. This has democratized the access to a genome sequence. We aim to democratize the access to a GENRE for those genomes.

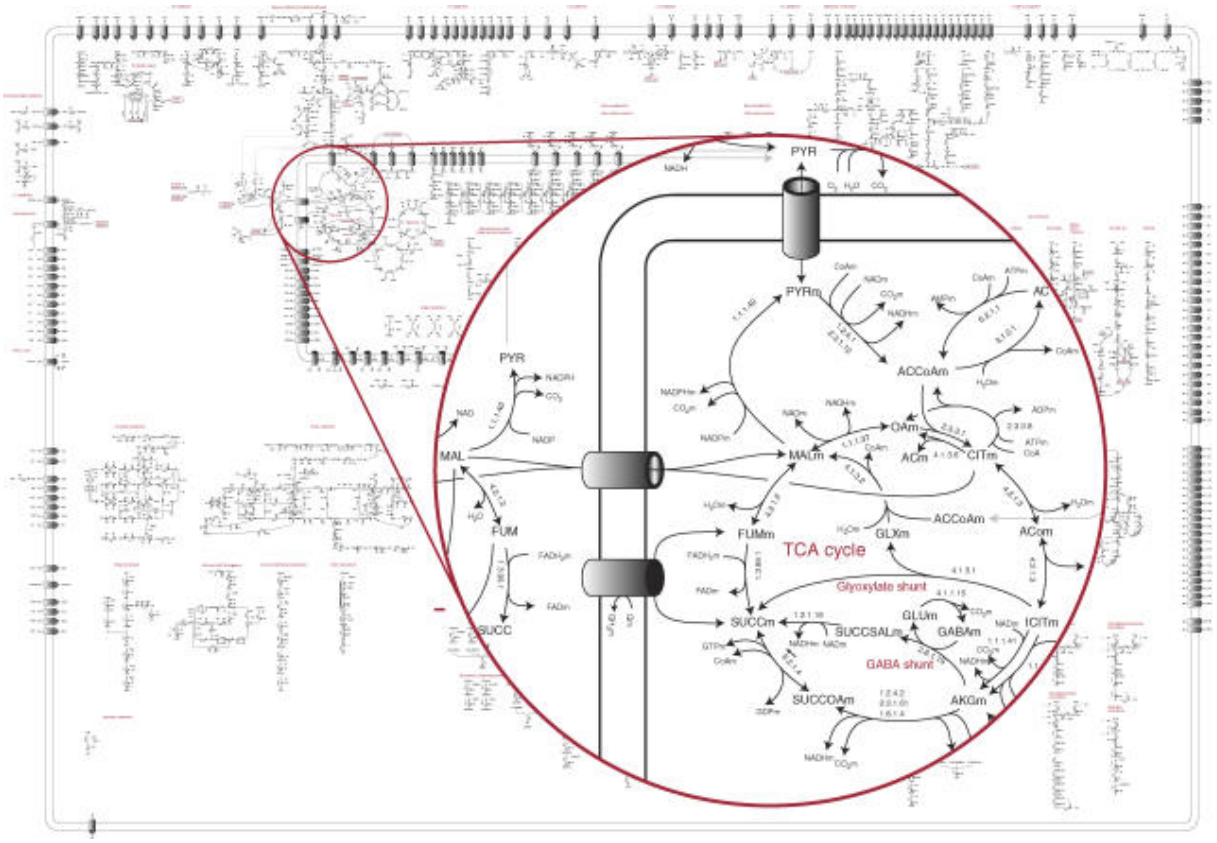


Figure 2: Example of a GENRE

A portion of a GENRE for *Aspergillus niger* CBS 513.88 strain illustrating transport across membrane and metabolic reactions [ANN08]. The highlighted inset shows the mitochondrion where the TCA cycle takes place, its membrane, and three transporters in the membrane.

As an introduction to the concept of a GENRE and the scale and scope of a GENRE, Figure 2 shows a portion of the GENRE for *Aspergillus niger* CBS 513.88 strain developed by Andersen [ANN08]. The highlighted inset shows the mitochondrion where the TCA cycle takes place, its membrane, and three transporters in the membrane.

1.1.1 Some Historical Context

In 1995, the genome of the bacteria *Haemophilus influenzae* was the first full genome to be sequenced [Pal08]. A GENRE was developed 4 years later. It was the first GENRE available and was developed manually. In 1996, the genome of the yeast *Saccharomyces cerevisiae* was the first eukaryotic genome to be completely sequenced. Yeast is one of the best characterized organisms [PL04]. A GENRE of *S. cerevisiae* was developed in 2003 [FFF$^+$03, DHP04]. The

initial reconstruction used the KEGG metabolic pathway database as the reference, and annotated the genes in terms of Enzyme Commission (EC) numbers.

The state of the art in this field obviously is heavily dependent on the history of biology and genomics. What most people regard as the Human Genome Project was actually a larger project to sequence a range of organisms, the so called *model organisms*. The list of model organisms has grown slightly, and is about to grow dramatically with the democratization of genomics. The model organisms were selected due to a number of criteria; mainly how they could throw light on the human genome in terms of cell mechanisms, development, and disease. By default, model organisms had a large scientific community; they had for a long time been organisms of interest to scientists; scientists knew how to perform experiments with them, and how to manipulate their genome. They were generally easy and fast to grow in the lab.

The prokaryotes, bacteria and archea, are simpler organisms with simpler genomes than eukaryotes. In particular, *E. coli* is the basis of recombinant DNA technology. Many prokaryote genomes were sequenced early in the history of genomics, so much of the knowledge and tools for GENREs and its steps are specific to prokaryotes.

GENRE protocols require extensive manual curation of the genome and the model. The commonest approach is to use reconstruction by analogy, that is, the reference template method, that requires a body of knowledge of existing reactions and pathways, and genes that perform those reactions. Hence, most GENREs are developed for model organisms, such as *E. coli*, or for prokaryotes.

Figure 4 shows the history of the *E. coli* GENRE from 1990 to 2007. *E. coli* has approximately 4300 genes, so the latest GENRE is modeling less than 50% of the genes. Note that the y-axis, not only shows the increase in the number of reactions, genes, and metabolites included in the versions of the GENRE, but also shows the knowledge of different cell mechanisms incorporated in the model, as our knowledge, through experimentation, grew:

- biosynthesis of amino acids and nucleotides;
- biosynthesis of cell wall constituents;
- biosynthesis of cofactors;
- fatty acid metabolism;

- alternate carbon utilization;
- quinone; and
- cell wall metabolism.

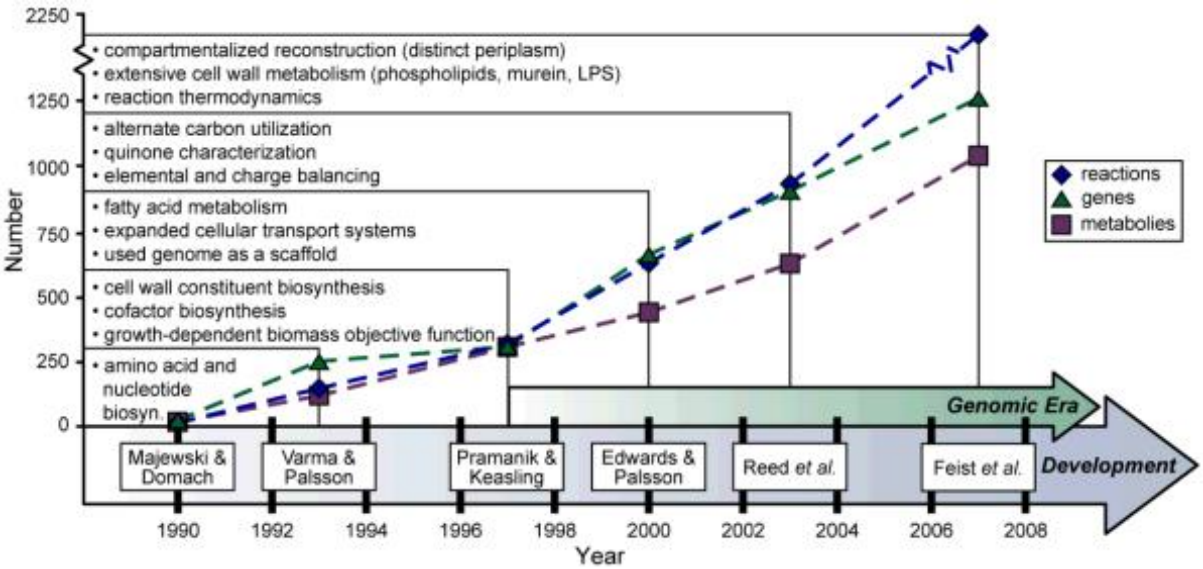


Figure 3: The Ongoing Reconstruction of the *E. coli* Metabolic Network

"History of the *E. coli* metabolic reconstruction. Shown are six milestone efforts contributing to the reconstruction of the *E. coli* metabolic network. For each of the six reconstructions, the number of included reactions (blue diamonds), genes (green triangles) and metabolites (purple squares) are displayed. Also listed are noteworthy properties that each successive reconstruction provided over previous efforts. For example, Varma & Palsson included amino acid and nucleotide biosynthesis pathways in addition to the content that Majewski & Domach characterized. The start of the genomic era (1997) marked a significant increase in included reconstruction components for each successive iteration. The reaction, gene and metabolite values for pre-genomic era reconstructions were estimated from the content outlined in each publication and in some cases, encoding genes for reactions were unclear." [FP08]

Figure 4 shows how the coverage **(c)** of GENREs has expanded to include fungi, plants, and human, though still strongly biased to bacteria **(a)**, and it still does not encompass all the potential reactions as identified in the Enzyme Commission (EC) **(b)**.

1.1.2 Resources

Historically, any work on metabolic pathways would refer back to KEGG [OYH+08] at the Bioinformatics Center, Institute for Chemical Research, Kyoto University and Human

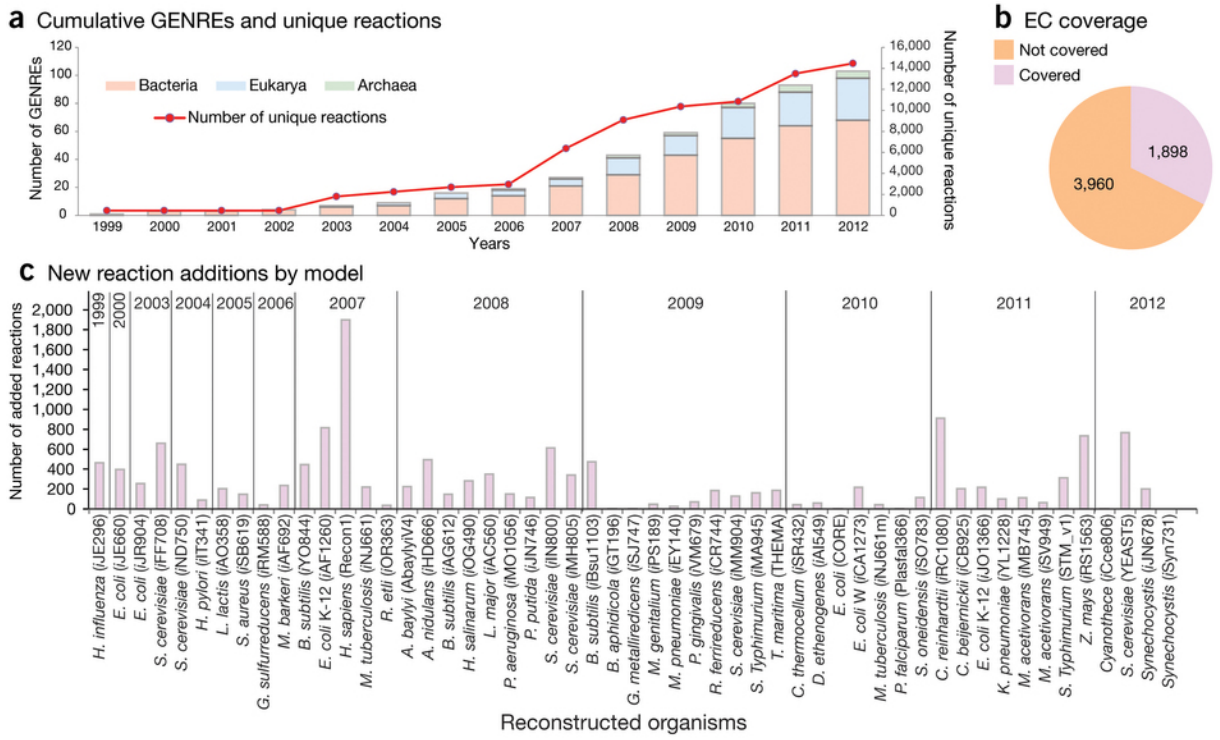


Figure 4: GENREs and their Coverage

"(a) By year, the cumulative number of GENREs published (vertical bars) and unique reactions included in all GENREs (red dots and line). (b) The proportion of Enzyme Commission (EC) numbers included in published GENREs. (c) Contribution to the coverage of metabolic space of each GENRE publication, as determined by the number of unique reactions added by each GENRE at the time of publication. The GENREs are ordered by publication date from *H. influenza* (iJE296) published in 1999, to *Synechocystis* (iSyn731), published in 2012." [MNP14]

Genome Center, Institute of Medical Science, University of Tokyo. KEGG digitized the pathways diagram of the pharmaceutical company Boehringer Ingelheim, and created databases for the pathways and the related enzymes, ligands, and genes. The KEGG information is not curated, so it is not as useful as more recent resources.

MetaCyc [CAD+10] is a curated database from SRI of pathways, reactions, and metabolites, that grew from the modeling and curation efforts of *E. coli*, namely EcoCyc [KCVSZ+11] originally, and now also TransportDB [RKP04] and RegulonDB [SPGGC+13]. It has strong tool support in Pathway Tools [KPK+09] for GENRE.

Today most GENREs can be found at BiGG [SPCP10], *"a Biochemically, Genetically and Genomically structured genome scale metabolic network reconstruction knowledgebase"* at Bernhard Palsson's Systems Biology Lab at UC San Diego. Models are encoded in the systems biology markup language (SBML) [HFS+03]. They develop the COBRA toolkit [SQF+11]

for analysis of GENREs.

Specific to modeling pathways, rather than to systems biology as a whole, is the BioPAX community [DCP+10] for Biological Pathways Exchange in XML. BioPAX is represented in RDF/XML and is defined in OWL.

For annotation of enzymes specifically, there is the Enzyme Commission (EC) classification scheme, which is supported by the BRENDA database [SCP+13] of EC definitions, reactions, metabolites, and enzymes. For annotation of transporters, there is the Transporter Classification (TC) scheme, which is supported by the TC database (TCDB) [STB05]. For annotation in general, one uses the Gene Ontology (GO) [The00]. GO covers enzymes and transporters amongst its collection of terms for annotation. The GOA database [HSMM+15] links gene ontology annotations to the entries in SwissProt and UniProt.

For curated protein sequences and information about the proteins, one consults the SwissProt database [BA00], which is the set of reviewed entries in UniProt [C+14], a resource with both reviewed and unreviewed protein sequences. SwissProt collaborates closely with curators for model organisms, and others, such as the AspGD database [CAI+14] for *Aspergillus* species.

The major software tools for GENREs are reviewed in [HR14] and discussed in Section 2.4.

1.1.3 Issues and Challenges

In modeling the cell, as a step to modeling an organism such as human, there are a number of aspects to consider, namely

- the structure of the cell, such as cell wall, membranes, and organelles;
- the metabolism that transforms metabolites and provides energy to the cell;
- the transport of material into and out of the cell, into and out of the organelles, and about the cell;
- the regulation of the cell processes; and
- the sensing of the environment, and the signaling of that information within the cell and between cells.

Clearly our knowledge is always in a state of flux, and we know more about some aspects above than others. Furthermore, we do not always know how to put that knowledge into practice, often awaiting the development of knowledge representations, reference collections, and algorithms. From electron microscopy we have strong knowledge of the structure of the cell. From our understanding of chemistry and the classification work of the Enzyme Commission, we have a good understanding of metabolism. Our understanding of transport, regulation, and signaling is less well developed.

Many GENREs, however, still do not model cell components fully even though we understand the structure of the cell. For metabolism, the problem arises because there are many EC numbers for which no gene is known, and hence assigning GPR associations by analogy is impossible. Furthermore, reactions may be catalyzed by protein complexes formed from several individual protein molecules. Most GENREs do not model protein complexes, and most functional annotations do not identify protein complexes. Chapter 4 illustrates our limited knowledge of transport.

Curation of the scientific literature in order to create Gold Standard reference sets is time and labour intensive. While one can still obtain funding for the creation of new reference sets it is increasing difficult to obtain funding to maintain existing reference sets.

A result of these two factors, our state of knowledge and the cost of curation, means that many Gold Standard reference sets are small in total size, or have many classes of entity for which the number of examples is small. This hampers machine learning as an approach to develop classifiers. Supervised machine learning requires sufficent data to create a *training set* and a *test set*. The training set should exhibit enough signal to separate the classes from each other, with some redundancy to allow cross-validation. The test set should contain at least one member of each class, but also be large enough to derive meaningful statistical results.

Validation, or evaluation, is a major problem. The quality control steps in GENRE protocols use flux balance analysis to check the self-consistency of the model; this is *internal validation* of the approach. True validation, *external validation* against a ground truth, is established in the wet lab by comparing observed measured behaviour — the *phenotype* — with *in silico* predictions of behaviour based on the nodel. Wet lab work requires collaborators with facilities, expertise, and resources. The experiments take time and effort.

1.2 Contributions

This thesis investigates the reconstruction of metabolic pathways. The goal is to remove obstacles to full automation of the process. To this end, the first contribution of the thesis is to identify those obstacles and identify the issues preventing automation. This is carried out in Chapter 3 through a review of the state of the art and case studies with fungal genomes. The issues identified are as follows.

- The reference template approaches are dependent on the body of existing knowledge, and the effort to manually curate the scientific literature to extract that knowledge and encode it in public databases.

- The evaluation of methods is difficult when applied to new genomes. Internal validation of the model can be measured in terms of numbers of pathways, reactions, and GPR associations to indicate coverage, and by the number of holes to indicate completeness. Further internal validation requires constructing a systems biology model so one can apply flux balance analysis for atoms, charges, energy, etc. External validation requires the scientist to make predictions from the model and then to validate those predictions in the wet lab; this is not expertise usually available to the developer of algorithms.

- The validation of methods for *de novo* discovery of pathways is difficult, even for model organisms. Internal validation shows that the pathways are sound in terms of the chemical transformation of compounds, but external validation of the existence of the pathway in the organism requires extensive wet lab work.

- Even with gap filling, there are typically many holes in the resulting reconstruction. Most approaches to gap-filling do not make use of gene expression data, which today can be readily available even for non-model organisms through RNA-Seq.

- The widely available and widely used tools are biased towards prokaryotes. In particular, they do not model cell compartments such as mitochondrion, Golgi, peroxisome, endoplasmic reticulum (ER), vacuole, or lysosome in their reconstructions.

- Transport reactions are often an afterthought in the modeling of the cell, despite the fact that the reconstruction needs to view the cell as a closed system importing and exporting compounds to its surroundings in order to perform internal validation.

While recognizing the importance of the goal of full automation of the process, there are several of the obstacles above that we could not plausibly attempt to solve. We could not see ourselves resolving the issues of providing a complete reference model of the cell through automation of the discovery of biological knowledge or the extraction of knowledge from the scientific literature. Neither could we resolve the difficulty of evaluation, as at some time, it becomes necessary to perform external validation in the wet lab.

We considered the issue of improving gap filling, especially the incorporation of gene expression data, through the development of new algorithms. However, there has been quite extensive work in the area, mostly with model organisms where the availability of expression data is high. Furthermore, we had no insight into how we might make a breakthrough nor how we could demonstrate through evaluation that we had made an improvement.

In Chapter 4 we investigate the issue of including transport reactions, transporter proteins, and the GPR associations for transport in the reconstruction of metabolic pathways. To clarify the state of the art in that area, we develop a scheme to describe and compare the different approaches. This is necessary so that we can see that the existing work of predicting transport proteins actually is diverse and incomparable. We use a case study to get a deeper understanding of the existing work, and to compare them in a practical setting using a fungal genome of interest. This study reveals several issues:

- the disjointedness of the field with little connection between those that use the Transporter Classification (TC) as their target for prediction, and those that use the chemical substrates being transported as their target for prediction;
- the limited coverage of the predictors, due to the small size of available Gold Standard datasets for transport; and
- the inability of the techniques to predict the specific substrate, or specific collection of substrates, that is transported across the membrane by the transport protein, even though they could identify the type of substrate in some cases.

In Section 4.4 we automate a protocol for determining the transporters in a genome that is used in the lab of Milton Saier, who develops the Transporter Classification and maintains the TCDB. In Section 4.6 we explore how to predict specific substrates of transporters. This is a very difficult problem, so we do not find a solution. Based on our experience, in Section 4.7 we propose a framework for the overall problem of predicting transporters, which includes the problem of determining specific substrates.

1.3 Organization of the Thesis

The thesis is organized as follows:

Chapter 2 contains the background material that is important to the understanding of this dissertation. Key are the Gene-Protein-Reaction (GPR) associations that are the units of the metabolic pathway reconstructions. They relate the central dogma of biology that genes through the processes of transcription and translation produce proteins, and these proteins in turn carry out the functional roles of the cell, including the enzymatic reactions of metabolism and the transport reactions across membranes. Section 2.1 introduces the concepts of genomics and the central dogma of molecular biology; Section 2.2 introduces metabolism, metabolic pathways, enzymes and reactions, illustrated by central carbon metabolism; Section 2.3 introduces transport of molecules and ions across cell membranes by transmembrane transport proteins; Section 2.4 provides an overview of techniques for genome-scale network reconstruction; and Section 2.5 briefly introduces the important aspects of machine learning and bioinformatics for this thesis.

Chapter 3 focuses on one aspect in the automation of systems biology, namely the reconstruction of the metabolic pathways. This step begins with an annotated genome of an organism, and perhaps with other data such as RNA-Seq expression data, and produces a model of the metabolism of the organism's cell. Section 3.1 reviews the state of the art for this step in the overall process; Section 3.2 looks at those fungal genomes that are well curated in order to see the completeness (or non-completeness) of their functional annotations; Section 3.3 presents our case studies in reconstructing metabolic pathway models for fungi; and Section 3.4 presents the lessons learned about the strengths and weaknesses of metabolic pathway reconstruction.

Chapter 4 investigates how to include transport reactions, transporter proteins, and the GPR associations for transport in the reconstruction of metabolic pathways. For prokaryotes, it is sufficient to model the transport across the cell membrane. However, eukaryotes have internal organelles, therefore the reconstruction requires modeling of the cell internal components and the intracellular transport across their membranes. The transport reaction should represent the transport of one or more specific substrates across a specific membrane. The GPR association should identify the transmembrane protein that performs the movement of those substrates across that membrane. Section 4.2 presents the scheme for describing and comparing existing methods, and presents the state of the art; Section 4.3 presents the case

study of the existing methods when applied to a fungal genome; Section 4.4 presents the automation of Saier's protocol and demonstrates how the implementation works on the fungal genome of the case study; Section 4.6 explores approaches to predicting specific substrates given a transport protein; Section 4.7 proposes a framework for the transport prediction problem; and Section 4.8 presents the lessons learned.

Chapter 5 concludes the thesis. It recaps the thesis work, and presents a summary of challenges addressed, the progress made, and the current state of the art. Section 5.1 presents the contributions of our work; Section 5.2 discusses the limitations of our work; and Section 5.3 offers some directions for future work.

The appendices contain details that support the thesis argument but are not vital to the understanding of the main body of the work.

Chapter 2

Background

This chapter contains the background material that is important to the understanding of this dissertation.

Key are the Gene-Protein-Reaction (GPR) associations that are the units of the metabolic pathway reconstructions. They relate the central dogma of biology that genes through the processes of transcription and translation produce proteins, and these proteins in turn carry out the functional roles of the cell, including the enzymatic reactions of metabolism and the transport reactions across membranes.

Our knowledge of genes and the roles of their proteins are captured in public web resources, such as SwissProt. The data about roles is represented as terms in ontologies or classification schemes. For metabolic reactions, the important classifications are the Enzyme Commission (EC) numbers, and the Gene Ontology (GO). Protein domain classification provided by the Pfam and InterPro resources is an important means of automatic annotation, so maps between the various schemes and GO have been created and are widely used. For transport reactions, the important classifications are the Transporter Classification (TC) scheme, and the Gene Ontology; however, the classification of transport is more recent, more in development, and less harmonized than metabolism. Again, protein domains play important roles in annotation, but maps between TC and the other schemes have not been developed yet.

Important techniques for this work from bioinformatics and machine learning are introduced. Many good references are available for this material, so we are brief. The key techniques are sequence similarity, the BLAST tool, and its results for *e-values*, *percent identity*, and

sequence coverage; amino acid composition and its variations that provide features for machine learning; profile Hidden Markov Models (HMM) representing sequence families, and the related use of multiple sequence alignment (MSA) and phylogenetic trees.

Draft reconstructions are based on analogy with knowledge available about the organism of interest, and related organisms. Public web resources act as reference templates for forming Gene-Protein-Reaction (GPR) associations. The Gold Standard resources are based on experimental results in the scientific literature that are manually curated. These include SwissProt, for proteins and their properties; MetaCyc, for pathways and reactions; TCDB, for transport proteins; and model organism databases, especially those of *E. coli* (bacteria), *S. cerevisiae* (fungus), and *A. thaliana* (plant). The KEGG pathway database was the first pathway resource and is still widely used even though its pathway templates are not all based on manual curation of experimental results.

The chapter organization is as follows: Section 2.1 introduces the concepts of genomics and the central dogma of molecular biology; Section 2.2 introduces metabolism, metabolic pathways, enzymes and reactions, illustrated by central carbon metabolism; Section 2.3 introduces transport of molecules and ions across cell membranes by transmembrane transport proteins; Section 2.4 provides an overview of techniques for genome-scale network reconstruction; and Section 2.5 briefly introduces the important aspects of machine learning and bioinformatics for this thesis.

2.1 Basic Concepts from Biology

The cell is the unit of life and knowing the cell components and how they work is the fundamental quest of biological science. Cell biology is the scientific discipline that studies the cell including its life cycle, physiological properties, structure, components, their behaviour, and how the cell interacts with environment. Today this is done at a molecular level. Understanding the molecular mechanisms and processes in living cells has been critical in understanding the basis for many cell process, and how they go wrong in diseases. The genome is the "program" that determines how a cell develops, its structure, and its functions. Figure 5 shows the components of a eukaryotic cell. Each cellular compartment plays specific roles in the cell processes.

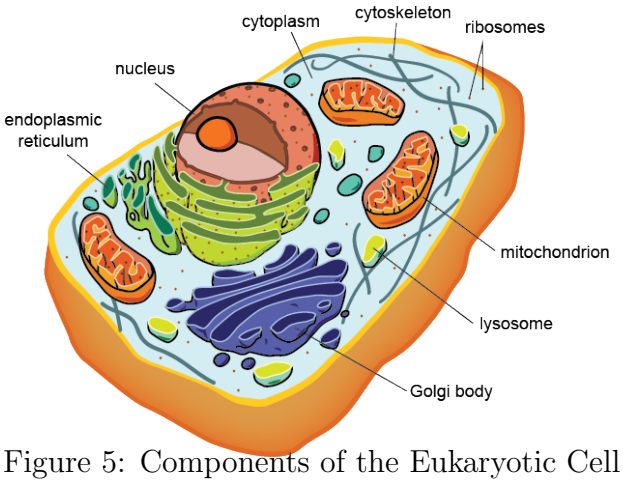


Figure 5: Components of the Eukaryotic Cell [`http://www.shmoop.com/biology-cells/all-eukaryotic-cells.html`]

2.1.1 Nucleic Acids

Nucleic acids are long biological molecules formed from smaller molecules called *nucleotides*. They carry the genetic information of an organism. There are two types of nucleic acids: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). The genetic information in DNA is coded with four *bases*: adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of bases are arranged in two strands that form a spiral called a double helix. Each type of base on one strand is paired up with a specific type of base on the other strand to form a unit called *base pair*. A is paired with T and C with G. DNA is found in the nucleus of eukaryotic cells and in the cytoplasm of prokaryotic cells. RNAs are usually single stranded and are assembled as a sequence of A, G, C, and uracil (U) bases. RNA molecules are synthesized on DNA templates and are used in protein synthesis in the cytoplasm.

2.1.2 Central Dogma of Molecular Biology

The genetic information on DNA sequence — or *genes* — of a biological system is used to synthesize messenger RNA (mRNA) molecules through a process called *transcription*. The information present in mRNA molecules is subsequently used to synthesize proteins through a process called *translation*. This flow of genetic information through transcription and translation is referred to as the central dogma of molecular biology and was first stated in 1958 by Francis Crick.

There is a difference in the transcription process of eukaryotic and prokaryotic cells. In eukaryotic cells transcription occurs in the nucleus and mRNA molecules are then transported to the cytoplasm to be translated. Transcription in prokaryotic cells occurs in the cytoplasm. Another major difference is that a eukaryotic gene has interleaved coding and non-coding segments, called *exons* and *introns*, respectively. Transcription in eukaryotic cells produces pre-mRNA strands that are subsequently converted into mRNA by removing introns and splicing exons.

The translation process synthesizes *proteins* from the mRNA molecules produced during transcription. Translation happens in the cytoplasm where an rRNA molecule called a *ribosome* attaches itself to mRNA and moves along it to produce a specific amino acid sequence based on codon to amino acid mapping. A *codon* is a triplet of bases coding for a specific amino acid. There are 20 standard amino acids. The mapping of codons to amino acids was determined experimentally and is called the *genetic code* [CBBWT61]. There are 64 possible codons, therefore an amino acid can be coded by more than one codon.

2.1.3 Proteins

The *primary structure* of a protein is the sequence of its amino acid molecules. Each amino acid is represented by a letter from the English alphabet. A protein sequence is represented as a string of letters from a set of English alphabet of size 20. See the one-letter code in Table 1. An important aspect of proteins is their function. The function of a protein is the role that the protein plays in a cell; it can be inferred from the three-dimensional structure of the protein, which in turn can be obtained from its primary structure [ARC+54, Anf73, WP99]. A corollary to the central dogma is that proteins that share sequence similarity are expected to have similar functions. Therefore, it is important to quantify sequence similarity to determine whether proteins perform similar function or not.

Two protein sequences are said to be *homologous* if they share a common evolutionary origin. Homology is a qualitative inference, i.e., there is no degree of homology, proteins are either homologous or not. Sequence similarity, however, is a quantitative inference measured by sequence alignment algorithms. Homologous proteins are derived from two evolutionary events, gene duplication and gene speciation. Gene duplication occurs when regions of DNA containing genes are duplicated giving rise to duplicates in an organism [Ohn70]. Duplicates are free to evolve new functions.

Amino Acid	3-letter code	1-letter code	Properties		
			Hydrophobic	Functional	Structural
Alanine	Ala	A	Hydrophobic	Non-polar	Ambivalent
Isoleucine	Ile	I		Non-polar	Internal
Leucine	Leu	L		Non-polar	Internal
Methionine	Met	M		Non-polar	Internal
Phenylalanine	Phe	F		Non-polar	Internal
Proline	Pro	P		Non-polar	Ambivalent
Tryptophan	Trp	W		Non-polar	Ambivalent
Valine	Val	V		Non-polar	Internal
Arginine	Arg	R	Hydrophilic	Polar; Basic	External
Asparagine	Asn	N		Polar; Uncharged	External
Aspartate	Asp	D		Polar; Acidic	External
Cysteine	Cys	C		Polar; Uncharged	Ambivalent
Glutamate	Glu	E		Polar; Acidic	External
Glutamine	Gln	Q		Polar; Uncharged	External
Glycine	Gly	G		Polar; Uncharged	Ambivalent
Histidine	His	H		Polar; Basic	External
Lysine	Lys	K		Polar; Basic	External
Serine	Ser	S		Polar; Uncharged	Ambivalent
Threonine	Thr	T		Polar; Uncharged	Ambivalent
Tyrosine	Tyr	Y		Polar; Uncharged	Ambivalent

Table 1: Amino Acids

The amino acids are grouped by their hydrophobic properties together with their functional and structural alphabets.

2.1.4 Domains

A *protein domain* is a substring of a protein sequence that can fold into a three-dimensional structure independent from the rest of the protein sequence. As such, it can have a function of its own. A protein sequence can have more than one domain, and if each performs different function, the result is a multi-functional protein sequence. For this reason, considering protein domains on their own is important in protein functional annotation. Protein domain databases exist that organize protein sequences into protein families based on their domains. Examples of commonly used domain databases are Pfam [PCE+12] and Conserved Domain Database (CDD) [MBZC+13].

2.1.5 Classification Schemes for Enzymes

2.1.5.1 EC Numbers

Enzymes are proteins that act as catalysts for biochemical reactions that occur in the cells of living organisms. A reaction is a chemical transformation in which chemical bonds are formed, broken or both. As stated in [Bai00], there are approximately 4000 known biochemical reactions being catalyzed by enzymes, which are classified into six classes (see Table 2) by the types of chemical reactions they catalyze. Many of these reactions are reversible.

	Enzymes Group Name	**Catalyzed Reaction**
EC 1	Oxidoreductases	Oxidation-reduction reactions
EC 2	Transferases	Transfer of functional groups
EC 3	Hydrolases	Hydrolysis reactions
EC 4	Lyases	Addition to double bonds or single bonds
EC 5	Isomerases	Isomerization reactions
EC 6	Ligases	Formation of bonds with ATP cleavage

Table 2: Enzymes classification

The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) is an organization responsible for the standardized numerical scheme, the Enzyme Commission number (EC number), to specify enzyme-catalyzed reactions [IUB]. This scheme has six major EC number classification groups (EC 1 to EC 6).

2.1.5.2 Gene Ontology

The Gene Ontology (GO) [The00] defines terms to describe the roles of the gene products of an organism. The terms are organized hierarchically as a directed acyclic graph, and categorized in three aspects: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Molecular Function includes function at a molecular level and describes the essential activities of a gene or gene product. Biological Process includes the processes that occur in living system that are mediated by gene products. Cellular Component describes the site of the activities.

The modeling of enzymes in the Gene Ontology MF mirrors closely the organization of EC. There is a standard mapping EC2GO translating between EC numbers and GO terms.

2.2 Metabolic Pathways

Metabolism is the essential part of cell maintenance to allow organisms to grow, reproduce, maintain structures and respond to environments. It takes place within each cell of a living organism where food is converted into energy through a series of chemical reactions that are catalyzed by enzymes. The energy then can be used for other important processes such as synthesizing organic materials, facilitating messages between cells, and the replication of DNA.

The products of metabolism are small molecules known as *metabolites*. They can be the final end products or intermediates (substrates) to other enzymatic reactions. These chemical reactions are organized into *metabolic pathways* where several enzymes and cofactors are responsible for transforming one molecule into another molecule. The pathways form a *metabolic network*. The speed and efficiency of the transformation of molecules relies on the enzymes. *Enzymes* are the proteins that act as the catalysts for biochemical reactions that occur in the cell. The set of enzymes determine which metabolic pathways occur in a cell. A *reaction* can be defined as a chemical transformation in which chemical bonds are formed, broken or both [KR93]. All this information on cell metabolism can be organized through the reconstruction of a metabolic model and development of a specific organism database.

The relationships of biochemical compounds that form a metabolic network M can be defined as

$$M = \langle C, \Re, E, P \rangle$$

where C is the set of compounds c, $\Re$ is the set of reactions r, E is the set of enzymes e, and P is the set of pathways p. A pathway p is a set of connected reactions r, and a reaction r is a tuple $\langle I, O, e \rangle$, where $I \subseteq C$, $O \subseteq C$, and $e \subseteq E$. I is the set of input compounds, O is the set of output compounds and e represents the enzyme catalyst(s).

Conventionally, to perform *in silico* computations and analysis, the transformation and relationship of biochemical compounds in a metabolic network are represented using graph theory [PSM+11, SYC09, DGHW03, HCL+07, CJ10, AS06, HWGW02].

Most cellular processes such as metabolism, gene expression, transferring molecules across cell membranes and cell communication require energy. In other words, energy allows cells to work, grow, move, maintain their structure, and perform specific functions. Eukaryotes, other than plants, obtain energy from foods, which contain nutrients such as sugar, fatty

acids and amino acids. The cells turn these nutrients into chemical bond energy through a series of chemical reactions known as *cellular respiration*.

Cellular respiration is the catabolic metabolism responsible for breaking down large molecules to produce energy in the form of adenosine triphosphate (ATP) [SHHB09]. ATP is the molecule that supplies energy to the whole cellular system, which includes powering metabolism, constructing new cell structures, synthesizing macromolecules (DNA, RNA, and proteins), and for enzymes to catalyze chemical reactions. *Aerobic respiration* and *anaerobic respiration* are the two types of cellular respiration. The former requires oxygen as one of its reactants to generate ATP and the later does not require oxygen.

Carbohydrates or sugars are the main nutrients that provide energy to the cell system via both aerobic and anaerobic respiration. A good source of energy are the simple sugars known as *monosaccharides*, such as glucose, fructose and lactose. These monosaccharides are the building blocks of dissacharides (e.g. sucrose). The other types of sugars are oligossaccharides (e.g. oligofructose) and polysaccharides (e.g. starch). For eukaryotes, the cellular respiration occurs in both the cytosol and the mitochondria. Respiration involves central carbon metabolism and the transport of molecules across cell membranes [SHHB09].

2.2.1 Central Carbon Metabolism

One example of the interaction of genes, proteins and metabolites in a cellular system is its central carbon metabolism (CCM). This pathway is crucial for examining biochemical yields in pathway engineering as the primary metabolites involved can determine the nutritional and growth status [RB09]. The essential pathways of central carbon metabolism are: Glycolysis (Figure 6); the Pentose Phosphate Pathway (PPP) (Figure 7); and the Tricarboxylic Acid (TCA) cycle (Figure 8).

2.2.1.1 Glycolysis

Glucose is the simplest sugar that fuels cellular respiration. It is the precursor metabolite for glycolysis in cell central carbon metabolism. Glycolysis, which occurs in the cytosol is the enzymatic breakdown of one glucose molecule to form two pyruvic acid molecules [SHHB09]. In other words, it degrades 6-carbon compounds (glucose) to form 3-carbon compounds (pyruvate) as end products. Then, pyruvic acid becomes the precursor molecule

for the TCA cycle. Two essential functions of glycolysis are [SRIa]: 1) to oxidize hexoses to generate ATP, reductants and pyruvate, and 2) being a pathway that can perform catabolic metabolism. Figure 6 shows the model of glycolytic system inferred in YeastCyc. There are 23 compounds altogether, with 14 enzymes, 21 genes, and 9 chemical reactions involved in YeastCyc glycolysis metabolism. Known variations of the glycolysis pathway are shown in Table 3.

2.2.1.2 Pentose Phosphate Pathway

The pentose phosphate pathway (PPP) is a linear pathway that has two distinct phases: the oxidative (irreversible reactions) and non-oxidative synthesis (reversible reactions). This pathway occurs in the cytosol and starts from glucose 6-phosphate (G6P) in glycolysis [Pal11]. The PPP is responsible for producing precursor substrates, known as pentose phosphates, for pentose sugars (ribose and deoxyribose) required for nucleic acids and Nicotinamide Adenine Dinucleotide Phosphate (NADPH), a reducing agent in redox reactions. The PPP also provides a precursor for aromatic amino acids [RP]. MetaCyc shows that the evidence code for both phases is EV-EXP, which means they were inferred from wet-lab experiments. Figure 7 shows chemical compounds involved in PPP as inferred in YeastCyc.

2.2.1.3 Tricarboxylic Acid Cycle

The tricarboxylic acid cycle (TCA cycle), once called the Krebs cycle, is a cyclic pathway that occurs in mitochondria of a cell. The mitochondrion is known as the cell's power house. The TCA cycle is the heart of aerobic metabolism and it produces most of the ATP for cellular activities. In MetaCyc, there are 6 models for TCA cycles as shown in Table 4. Figure 8 is the model inferred by YeastCyc.

2.2.1.4 Sugar Transport in Central Carbon Metabolism

Transmembrane transport proteins are proteins in cell membranes responsible for moving molecules and ions across the membrane [SHHB09]. They play important roles in cellular metabolism and signaling. The transport of small molecules occurs from mitochondria into the cytosol or vice versa, and across the cell membrane. In central carbon metabolism of a eukaryotic cell, both glycolysis and PPP occur in the cytosol while the TCA cycle

Instances	No. of Reactions	Evidence
Glycolysis I (**from glucose-6P**)	11	**EV-EXP-TAS**: EcoSal "Escherichia coli and Salmonella: Cellular and Molecular Biology." Online edition.
Glycolysis II (**from glucose-6P**)	10	**EV:EXP:TAS**: EcoSal "Escherichia coli and Salmonella: Cellular and Molecular Biology." Online edition.
Glycolysis III (**from glucose**)	10	**EV-EXP-TAS**: Dang CV (2012). "Links between metabolism and cancer." Genes Dev 26(9);877-90. PMID: 22549953 **EV-EXP-IDA**: (1) Hansen T, Schonheit P (2003). "ATP-dependent glucokinase from the hyperthermophilic bacterium Thermotoga maritima represents an extremely thermophilic ROK glucokinase with high substrate specificity." FEMS Microbiol Lett 226(2);405-11. PMID: 14553940; (2) Schroder C, Selig M, Schonheit P "Glucose fermentation to acetate, CO2 and H2 in the anaerobic hyperthermophilic eubacterium Thermotoga maritima: involvement of the Embden-Meyerhof pathway." Archives of Microbiology 161:460-470 (1994); (3) Selig M, Xavier KB, Santos H, Schonheit P (1997). "Comparative analysis of Embden-Meyerhof and Entner-Doudoroff glycolytic pathways in hyperthermophilic archaea and the bacterium Thermotoga." Arch Microbiol 1997;167(4);217-32. PMID: 9075622.
Glycolysis IV(**Plant cytosol**)	10	**EV-EXP-TAS**: (1) William C. Plaxton "The organization and regulation of plant glycolysis." Annu. Rev. Plant Physiol. Plant Mol. Biol. 1996. 47:185-214; (2) Fernie AR, Carrari F, Sweetlove LJ (2004). "Respiratory metabolism: glycolysis, the TCA cycle and mitochondrial electron transport." Curr Opin Plant Biol 7(3);254-61. PMID: 15134745; (3) Dey, PM, Harborne, JB "Plant Biochemistry." Academic Press 1997 **EV-EXP**: Giege P, Heazlewood JL, Roessner-Tunali U, Millar AH, Fernie AR, Leaver CJ, Sweetlove LJ (2003). "Enzymes of glycolysis are functionally associated with the mitochondrion in Arabidopsis cells." Plant Cell 15(9);2140-51. PMID: 12953116.
Glycolysis V (**Pyrococcus**)	9	**EV-EXP-TAS**: (1) Sakuraba H, Ohshima T (2002). "Novel energy metabolism in anaerobic hyperthermophilic archaea: a modified Embden-Meyerhof pathway." J Biosci Bioeng 93(5);441-8. PMID: 16233230; (2) Verhees CH, Kengen SW, Tuininga JE, Schut GJ, Adams MW, De Vos WM, Van Der Oost J (2003). "The unique features of glycolytic pathways in Archaea." Biochem J 375(Pt 2);231-46. PMID: 12921536, **EV-EXP-IDA**: Kengen SW, de Bok FA, van Loo ND, Dijkema C, Stams AJ, de Vos WM (1994). "Evidence for the operation of a novel Embden-Meyerhof pathway that involves ADP-dependent kinases during sugar fermentation by Pyrococcus furiosus." J Biol Chem 269(26);17537-41. PMID: 8021261.
Glycolysis V (**Metazoan**)	10	**EV-EXP-TAS**: Dang CV (2012). "Links between metabolism and cancer." Genes Dev 26(9);877-90. PMID: 22549953.

Table 3: Variations of Glycolysis Pathway in MetaCyc

Known variations from the literature curated in MetaCyc as of March, 2014 [SRIb].

Instances	No. of Reactions	Evidence
TCA cycle I (**Prokaryotic**)	11	**EV:EXP**: Baldwin JE, Krebs H (1981). "The evolution of metabolic cycles." Nature 291(5814);381-2. PMID: 7242661
TCA Cycle II (**Plant & Fungi**)	9	**EV-EXP-IDA**:(1) Krebs HA, Johnson WA (1937). "Acetopyruvic acid ($\alpha\gamma$-diketovaleric acid) as an intermediate metabolite in animal tissues." Biochem J 31(5);772-9. PMID: 16746397; (2) Krebs HA, Salvin E, Johnson WA (1938). "The formation of citric and α-ketoglutaric acids in the mammalian body." Biochem J 32(1);113-7. PMID: 16746585; (3) Krebs HA, Eggleston LV (1945). "Metabolism of acetoacetate in animal tissues. 1." Biochem J 39(5);408-19. PMID: 16747930.
TCA Cycle III (**Helicobacter**)	9	**EV-EXP-IDA**: Hughes NJ, Clayton CL, Chalk PA, Kelly DJ (1998). "Helicobacter pylori porCDAB and oorDABC genes encode distinct pyruvate:flavodoxin and 2-oxoglutarate:acceptor oxidoreductases which mediate electron transport to NADP." J Bacteriol 1998;180(5);1119-28. PMID: 9495749.
TCA Cycle IV (**2-oxoglutarate decarboxylase**)	11	**EV-EXP-IDA**: Tian J, Bryk R, Itoh M, Suematsu M, Nathan C (2005). "Variant tricarboxylic acid cycle in Mycobacterium tuberculosis: identification of alpha-ketoglutarate decarboxylase." Proc Natl Acad Sci U S A 102(30);10670-5. PMID: 16027371.
TCA Cycle V (**2-oxoglutarate: ferredoxin oxidoreductase**)	12	**EV-EXP-IDA**: Tian J, Bryk R, Itoh M, Suematsu M, Nathan C (2005). "Variant tricarboxylic acid cycle in Mycobacterium tuberculosis: identification of alpha-ketoglutarate decarboxylase." Proc Natl Acad Sci U S A 102(30);10670-5. PMID: 16027371.
TCA Cycle VI (**Obligate autotrophs**)	11	**EV-EXP**: Smith AJ, London J, Stanier RY (1967). "Biochemical basis of obligate autotrophy in blue-green algae and thiobacilli." J Bacteriol 94(4);972-83. PMID: 4963789.
TCA Cycle VII (**Acetate-producers**)	9	**EV-EXP-IDA**: Mullins EA, Francois JA, Kappock TJ (2008). "A specialized citric acid cycle requiring succinyl-coenzyme A (CoA):acetate CoA-transferase (AarC) confers acetic acid resistance on the acidophile Acetobacter aceti." J Bacteriol 190(14);4933-40. PMID: 18502856.
TCA Cycle VII (**Metazoan**)	10	**EV-EXP-IDA**: (1) Krebs HA, Salvin E, Johnson WA (1938). "The formation of citric and α-ketoglutaric acids in the mammalian body." Biochem J 32(1);113-7. PMID: 16746585; (2) Krebs HA, Eggleston LV (1945). "Metabolism of acetoacetate in animal tissues. 1." Biochem J 39(5);408-19. PMID: 16747930.

Table 4: Variations of TCA Cycle Pathway in MetaCyc

Known variations from the literature curated in MetaCyc as of March, 2014 [SRIb].

occurs in mitochondria. Therefore, compounds such as pyruvate, ATP and ADP need to be transported across the mitochondrial membrane for energy metabolism.

In yeast, the uptake of sugar compounds requires transporters as these compounds do not freely permeate biological membranes [Lag93]. The most widely studied carbon sources in yeast are glucose, fructose, galactose and mannose (hexoses), and maltose and sucrose (dissacharides) [RLL06].

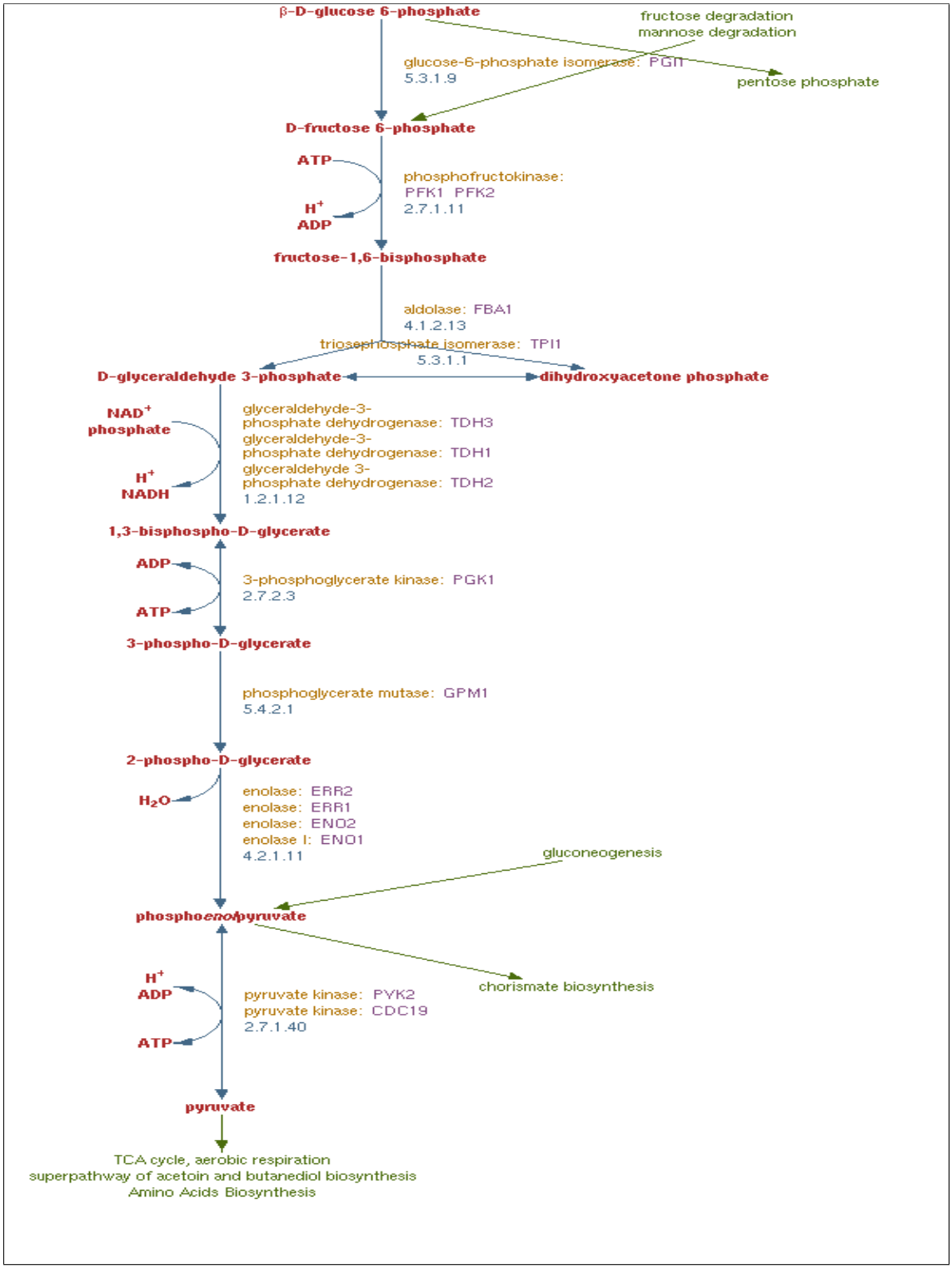


Figure 6: Computationally Inferred Glycolysis I Pathway of *S. cerevisiae* in YeastCyc

Compounds are represented in red, enzymes in orange, genes in purple, and pathways in green. Numbers separated by dots and in blue color are EC numbers designating the chemical reactions. From YeastCyc [SUb].

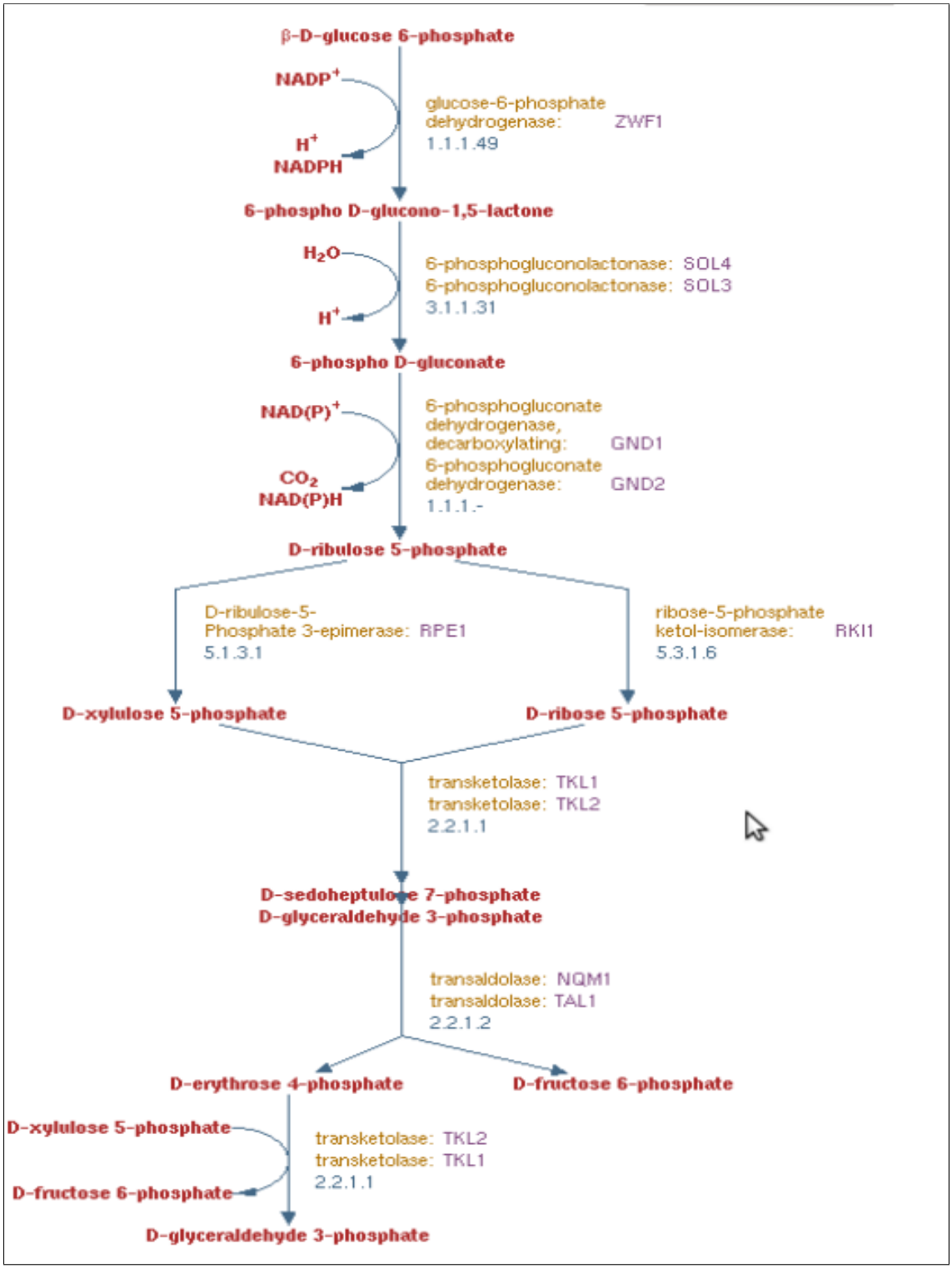


Figure 7: Computationally Inferred PPP Pathway of *S. cerevisiae* in YeastCyc

Compounds are represented in red, enzymes in orange, genes in purple, and pathways in green. Numbers separated by dots and in blue color are EC numbers designating the chemical reactions. From YeastCyc [SUb].

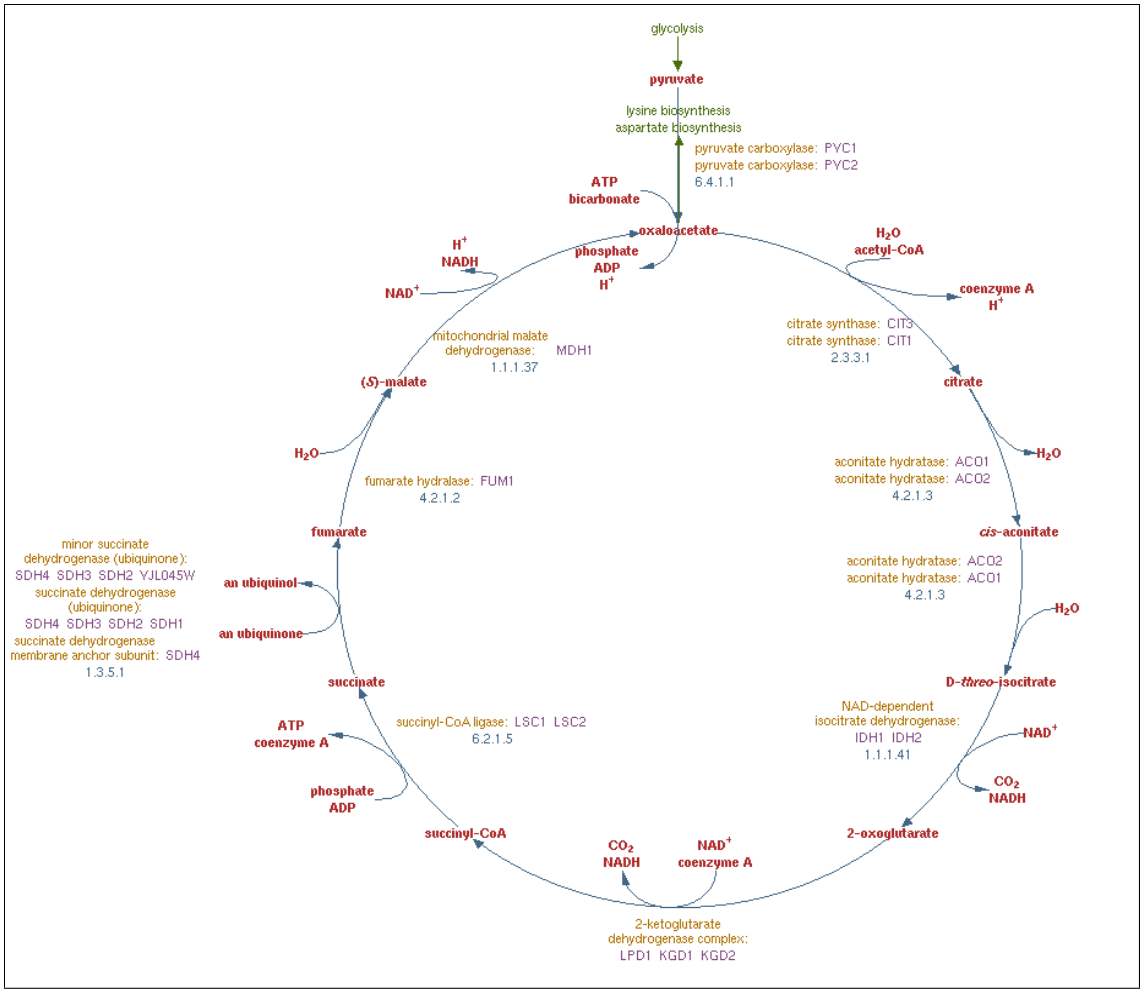


Figure 8: Computationally Inferred TCA Cycle II of *S. cerevisiae* in YeastCyc

Compounds are represented in red, enzymes in orange, genes in purple, and pathways in green. Numbers separated by dots and in blue color are EC numbers designating the chemical reactions. From YeastCyc [SUb].

2.3 Transport

A eukaryotic cell is surrounded by a plasma membrane and contains cell organelles, that are themselves defined by membranes and perform their own specific functions [Kuy08]. The membrane is a phospholipid bilayer as shown in Figure 9. There are two major classes of membrane proteins defined by their position relative to the membrane: the *peripheral membrane proteins* and the *integral membrane proteins* (IMP). The IMP are further classified into two groups: the *integral polytopic proteins*, which span the entire membrane, and the *integral monotopic proteins*, which do not. The polytopic proteins are also called *transmembrane proteins*.

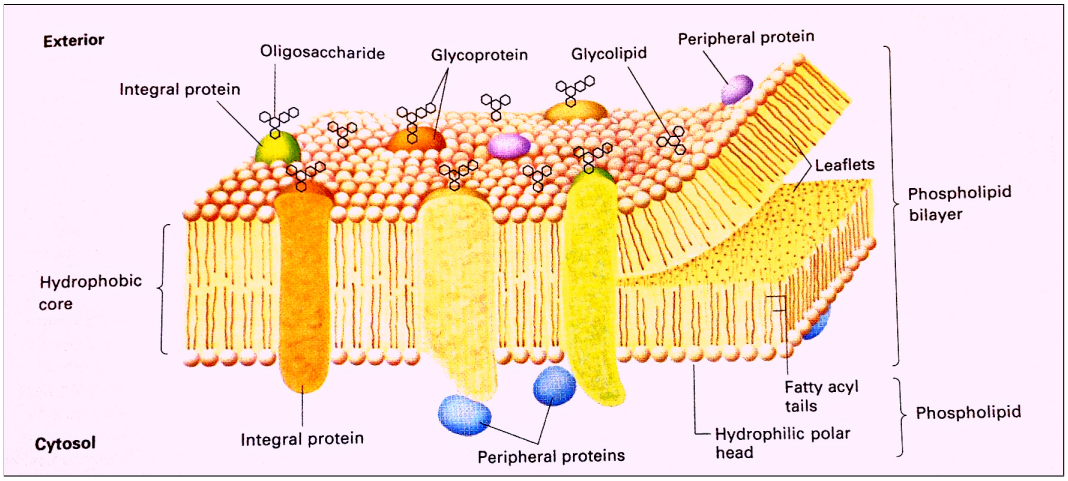


Figure 9: Typical Membrane Proteins in a Biological Membrane
From [LBZ+00].

Structurally, the eukaryote transmembrane proteins have α-helices that span the membrane [WW99]. In gram negative bacteria, there are transmembrane strand proteins that span the membrane with β-strands [Sch03]. These are called transmembrane segments (TMS). Figure 10 shows α-helices spanning a membrane.

Functionally, membrane proteins are classified as

- transporters, which transport ions or molecules across the membrane;
- ion channels, which provide a hydrophilic pathway across the membrane for ions; and
- receptors, which are proteins in the membrane that attach to molecules such as hormones and neurotransmitters and trigger cell changes.

Transporters move molecules and ions across the membrane [SHHB09]. Transporters constitute up to 30% of all cellular proteins [SSM10], and they play important roles in cellular metabolism [RP05]. Transporters have a high degree of substrate specificity and bind to one or a few substrate molecules [LBZ+00]. The different forms of molecule transport are [Kuy08]:

(I) Diffusion of small hydrophilic or hydrophobic particles driven by a concentration gradient;

(II) Diffusion of hydrophilic or charged particles driven by a voltage gradient;

(III) Osmosis, diffusion of solute driven by a concentration gradient of a non-permeable compound;

(IV) Facilitated diffusion; and

(V) Active transport against a concentration gradient.

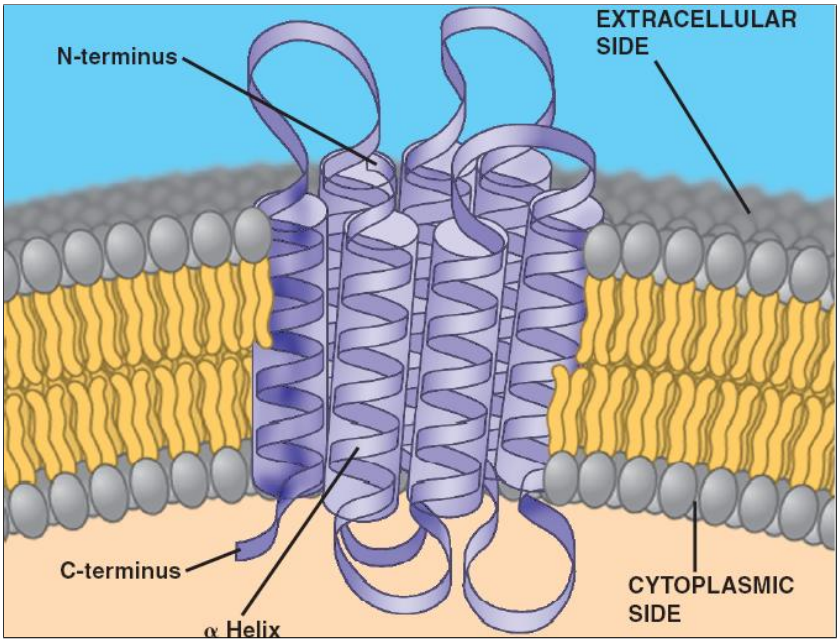


Figure 10: Transmembrane Segments: Helices cross a Membrane
[`http://bio1151b.nicerweb.net/Locked/media/ch07/`]

The transport of sugar across membranes is an example of active transport, which requires energy. Figure 11 illustrates the mechanism of active transport of glucose. It shows the transmembrane transport protein forming a V in order to accept the glucose molecule from the outside of the cell, and then inverting the V in order to release the glucose molecule into the cytosol. GLUT1 is the glucose transporter in mammals. Figure 12 shows a representation of part of the 3D structure of GAL2, the yeast galactose transporter, with a glucose molecule *in situ*. The figure highlights the few important sites where amino acids in the middle of certain TMS — TM5, TM8, and TM10 — of the transporter interact with the glucose molecule.

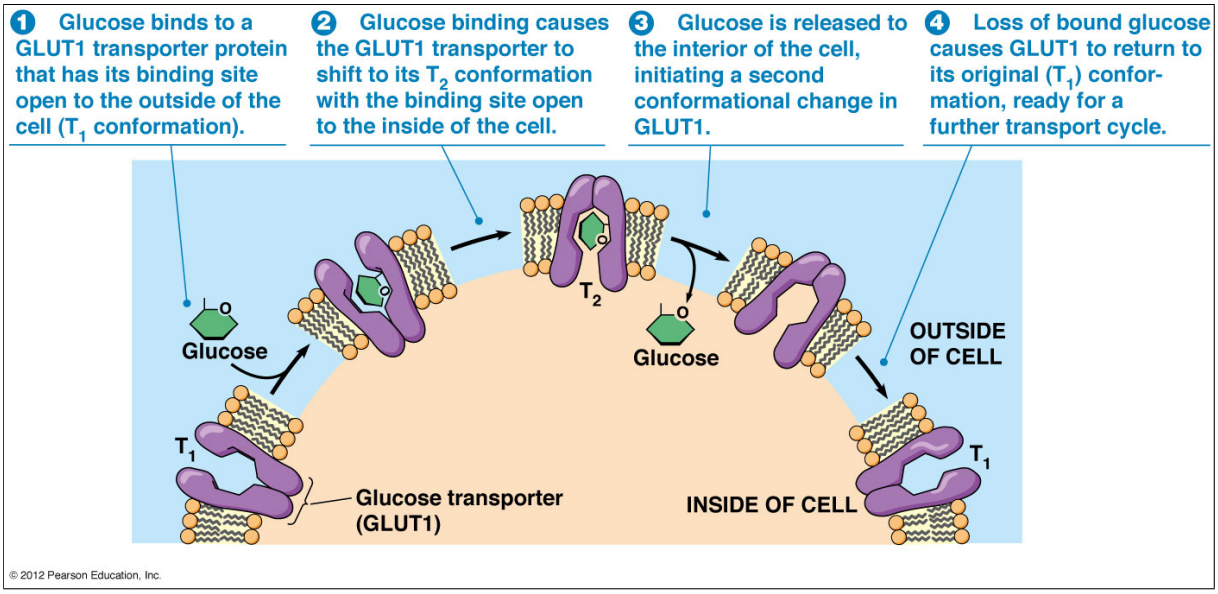


Figure 11: Mechanism of Transport for an Active Transport

Active transport of glucose by the GLUT1 transporter in mammals. It shows the transmembrane transport protein forming a V in order to accept the glucose molecule from the outside of the cell, and then inverting the V in order to release the glucose molecule into the cytosol. ©Pearson Education, Inc.

2.3.1 Classification Schemes

Transporters are classified according to different criteria, such as mechanism, substrate, and family. While functional annotation in general targets the Gene Ontology as the description or annotation, predictors for transport proteins target either the Transporter Classification scheme, or the substrate category. It would be useful if these three approaches were cross-referenced with each other, and with the protein domains [CVP+15], so that the correspondence between classifications were clear.. Here we briefly overview the three schemes.

2.3.1.1 Transporter Classification System

The International Union of Biochemistry and Molecular Biology (IUBMB) introduced the Transporter Classification System (TC) [BS04] in June 2001 for classifying membrane transport proteins. The TC system is analogous to EC numbers for classifying enzymes. A TC identifier such as TC 2.A.1.1.35 has five components representing

1. the transporter class (TC-class), eg 2;

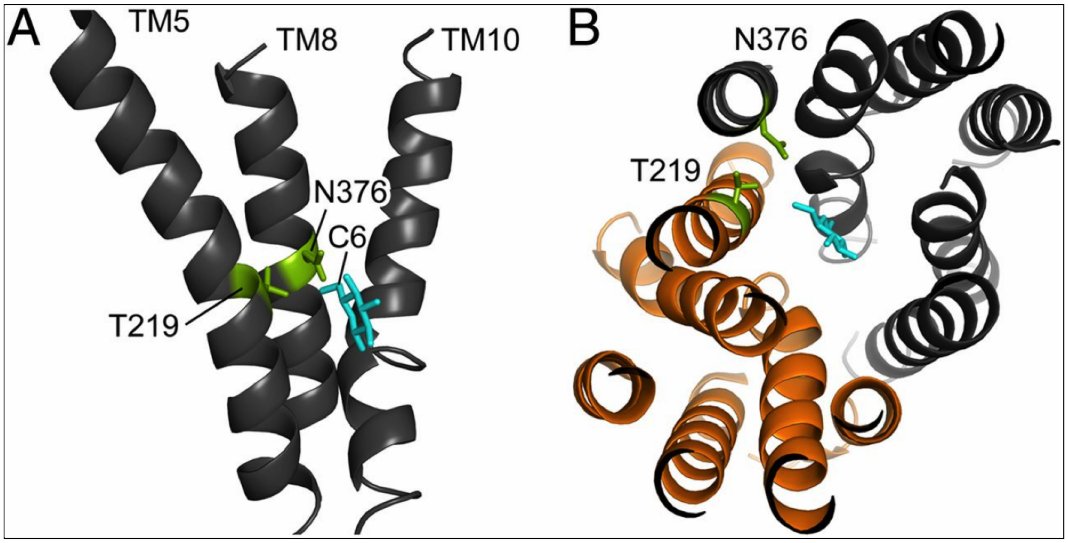


Figure 12: Important Residues for Glucose Transport

"Homology model of the Gal2 structure. The model is based on the outward-facing partly occluded structure of E. coli *XylE with bound glucose (PDB ID code 4GBZ). (A) Side view of Gal2; for reasons of clarity, only TMs 5, 8, and 10 are shown. The two amino acid residues T219 and N376 (green) are located at the center of their respective helix, with their side chains protruding toward the C6 of glucose (cyan). (B) Top view of Gal2 from the extracellular side, with a cross-sectional plane for better view; glucose (cyan) is found in between subdomains N (orange) and C (dark gray). The 3D images were created with PyMOL."* [FBS+14]

2. the transporter subclass (TC-Subclass), eg 2.A;

3. the transporter family (TC-Family), eg 2.A.1, which in some cases is a superfamily;

4. the transporter subfamily, eg 2.A.1.1; and

5. the specific transporter (TC-ID), eg 2.A.1.1.35.

A superfamily is a large divergent family, in which the distant clades are considered families within the larger superfamily. The categorization and classification of transporters is described in Table 5. The grouping of transport proteins is determined by sequence homology and phylogenetic analysis into the various classes and families and stored in the TC Database (TCDB) [SYN+08]. As of May 28, 2014, the TCDB contained more than 10,000 published references with 11,574 unique protein sequences, classified into more than 800 transporter families and 53 transporter superfamilies [SJRTV14].

Name of TC Class	TC Subclass	Description of TC Subclass
Channels/pores	1.A	α-type channels
	1.B	β-Barrel porins
	1.C	Pore-forming toxins (proteins and peptides)
	1.D	Non-ribosomally synthesized channels
	1.E	Holins
	1.F	Vesicle fusion pores
	1.G	Viral Fusion Pores
	1.H	Paracellular channels
	1.I	Membrane-bounded channels
Electrochemical potential-driven transporters	2.A	Porters (uniporters, symporters, antiporters)
	2.B	Nonribosomally synthesized porters
	2.C	Ion-gradient-driven energizers
Primary active transporters	3.A	P-P-bond-hydrolysis-driven transporters
	3.B	Decarboxylation-driven transporters
	3.C	Methyltransfer-driven transporters
	3.D	Oxidoreduction-driven transporters
	3.E	Light absorption-driven transporters
Group translocator	4.A	Phosphotransfer-driven group translocator
	4.B	Nicotinamide ribonucleoside uptake transporters
	4.C	Acyl CoA ligase-coupled transporters
Transport electron carriers	5.A	Transmembrane 2-electron transfer carriers
	5.B	Transmembrane 1-electron transfer carriers
Accessory factors involved in transport	8.A	Auxiliary transport proteins
	8.B	Ribosomally synthesized protein/peptide toxins that target channels and carriers
	8.C	Non-ribosomally synthesized toxins that target channels and carriers
Incompletely characterized transport systems	9.A	Recognized transporters of known biochemical mechanism
	9.B	Putative transport proteins
	9.C	Functionally characterized transporters lacking identified sequences

Table 5: Transporter Classification System in TCDB

As of September 2014.

2.3.1.2 Substrates

The molecule transported by a transporter is essential information in the annotation or description of the transport protein. Chemical molecules have a systematic name as determined by IUPAC (International Union of Pure and Applied Chemistry). The company Daylight Chemical Information Systems has a linear textual notation SMILES (Simplified Molecular Input Line Entry System) for representing chemicals and reactions (`http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`). SMILES aids computation as it support a canonical form which determines equality or identity of different chemicals, though it is not truly canonical. SMILES is widely used in cheminformatics.

In bioinformatics, specific substrates are documented using the Chemical Entities of Biological Interest (ChEBI) ontology [HdMD+13], but the organization of ChEBI has not influenced the substrate grouping in the prediction of transport. There prediction occurs at the level of substrate category or class — amino acid, anion, cation, electron, protein/mRNA (oligopeptide), sugar, and other — but the notation is not standardized.

Milton Saier, who leads the Transporter Classification effort, uses the following groupings in his work [PVL+14]. A high-level grouping is shown [PVL+14, Figure 2(A)]:

1. Inorganic compounds;
2. Carbon sources;
3. Amino acids and their derivatives;
4. Drugs, dyes, sterols, and toxics;
5. Bases and derivatives; and
6. Macromolecules.

This is broken down into *Substrate Groups* [PVL+14, Figure 2(B)]:

- Nonselective ions;
- Cations;
- Anions;
- Electrons;

- H2O;
- Sugar and polyols;
- Monocarboxylates;
- Di- and tri-carboxylates;
- Organoions;
- Aromatic compounds;
- Amino acids and conjugates;
- Amines, amides, polyamines, and organocations;
- Peptides;
- Siderophores, siderophores-Fe complexes;
- Substrate cofactors;
- Multiple drugs;
- Specific drugs;
- Other hydrophobic substrates;
- Nucleobases;
- Nucleosides;
- Polysaccharides;
- Proteins;
- Lipids;
- Nucleic acids; and
- Unknown.

Milton Saier [PVL+14, Table 1] additionally includes Substrate Groups for *Cofactor* and *Dicarbonate*, and includes a column for the Specific Substrate; though the entry is often identical to the Substrate Group.

2.3.1.3 Gene Ontology

The Gene Ontology (GO) [The00] defines terms to describe gene products of an organism. The terms are organized hierarchically as direct acyclic graph, and categorized in three aspects: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC).

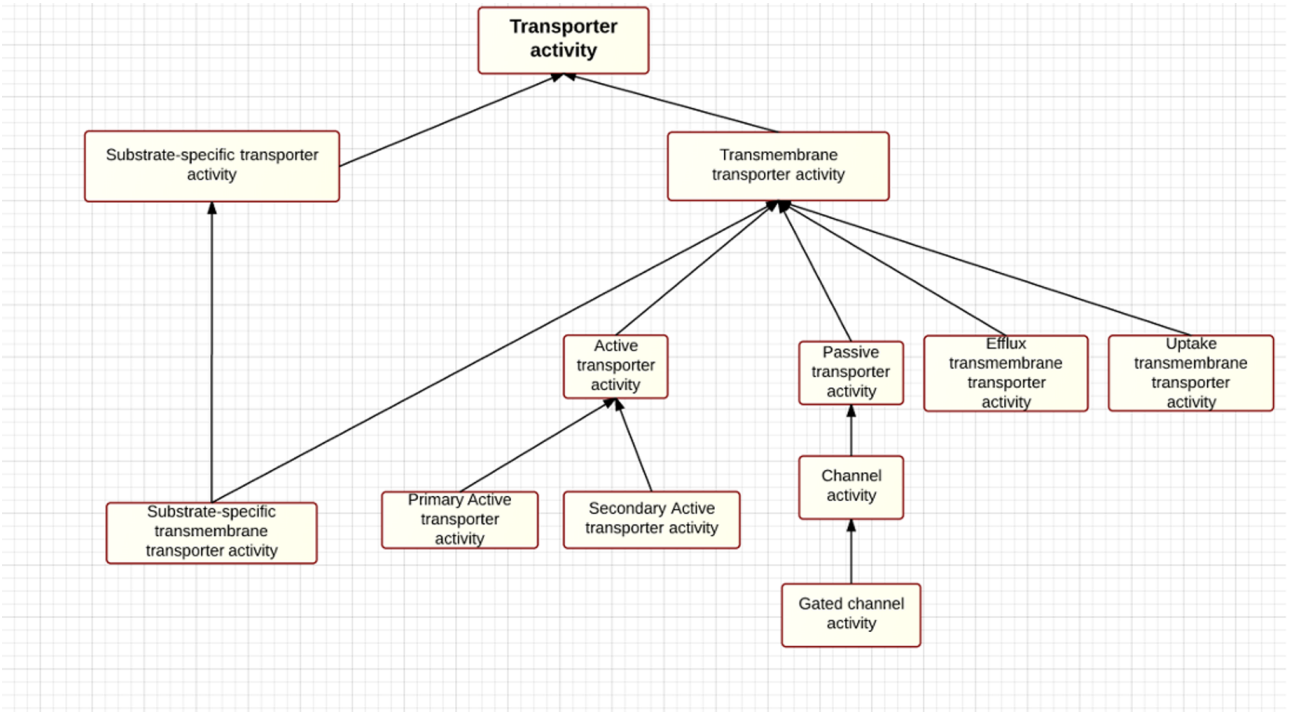


Figure 13: GO Molecular Function Hierarchy for Transport

The transporter activity is the general term representing the molecular function of transporters. Note that the children for primary and secondary active transporters activities, and gated channel activities were excluded.

The guidelines for transporters (http://geneontology.org/page/transport-and-transporters) relates terms across the three aspects and considers localization, substrate, transport mechanism, affinity to the substrate, constitutive versus inducible activity, and the D- and L-forms of substrates (see Figure 13).

The hierarchical nature of GO allows a term to capture the level of precision of the substrate, eg, in Biological Process, see Figure 14.

```
GO:0006810 transport
   GO:0008643 carbohydrate transport
      GO:0015749 monosaccharide transport
         GO:0008645 hexose transport
            GO:0015762 rhamnose transport
```

Figure 14: GO Transport Subtree for Biological Process

A selection of terms in the GO BP subtree rooted at the term for transport.

2.4 Genome-Scale Network Reconstruction

A genome-scale network reconstruction (GENRE) for an organism models the working of the genes, proteins, and metabolites within the organism. This ideally covers metabolism, transport, regulation, and signaling. Ideally a GENRE should be quantitative, and not just qualitative. Aa typical GENRE models metabolism quite well, and can assign Gene-Protein-Reaction (GPR) associations of genes to reactions, based on Enzyme Commission (EC) classification. The GENRE may include transport reactions in the model, but not be able to assign GPR associations of genes to transport reactions.

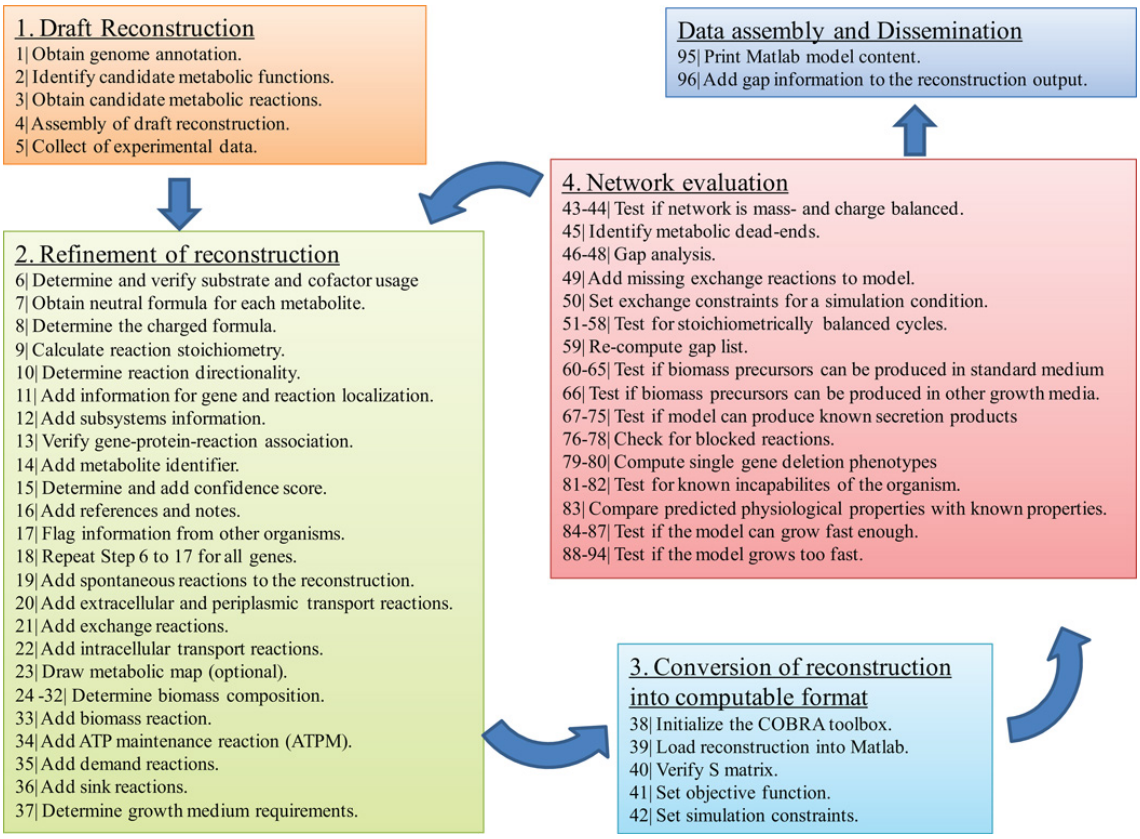


Figure 15: Thiele and Palsson 2010 Protocol for GENRE

An overview of a detailed protocol [TP10] for the constructtion of a GENRE.

A major reference is the detailed protocol of Thiele and Palsson [TP10] summarized in Figure 15. The techniques for reconstructing the draft metabolic network can be categorized [KTY+13] as:

- reference methods, that build a model by analogy to existing pathways; and

- *de novo* methods, that discover novel pathways. These can be categorized as
  - compound-filling methods, where the input and output compounds of the network are known, and the method uses both compounds and reactions to reconstruct the network; and
  - reaction-filling methods, where all the compounds involved in the network are known, and the method uses reactions to reconstruct the network.

The existing reviews [FST05, PRU10, FS11, OP10, RGM+12, SCM14, HR14] can be summarized in Table 16 [HR14] for the major software tools. Note that none of them fully automate the process, and that steps 20 and 22 for transport are poorly handled by existing tools.

2.5 Machine Learning in Bioinformatics

This section highlights key aspects of bioinformatics and machine learning relevant to this thesis: the classification of machine learning problems into binary, multi-class, and multi-label; BLAST for sequence similarity; amino acid composition; and Hidden Markov Models.

2.5.1 Binary, Multi-Class and Multi-Label Classifiers

In supervised learning, the examples are described by a set of *features* and known to be assigned to specific *classes*, $C_1, C_2, ..., C_k$. The aim is to build a *classifier* that can look at a new example and determine its classification. The simplest case is a *binary* classifier for a class C, which is simply required to determine whether the new example is a member of C, or is not a member of C. A *multi-class* classifier is required to determine to which class C_i the new example belongs. There is an implicit assumption that the classes are disjoint. For *multi-label* classifiers, this assumption is dropped, and the classifier is required to determine whether or not the new example belongs to each class C_i; that is, what subset of classes does the new example belong to. This is important in Chapter 4, where different tools adopt differing requirements for their classifiers.

Automatic
Assistance
No Support
*** Manual inspection recommended

	Step	Activity	SuBliMinaL	Model SEED	RAVEN	Pathway Tools
Stage 1: Draft Reconstruction	1	Obtain genome annotation				
	2	Identify candidate metabolic functions				
	3	Obtain candidate metabolic reactions				
	4	Assemble draft reconstruction			***	***
Stage 2: Refinement / Curation	6	Determine substrate and cofactor usage		***		
	7,8	Obtain charged formula for each metabolite	***	***		***
	9, 43-44	Mass- and charge-balance reactions	***	***		***
	10	Determine reaction directionality	***	***	***	***
	11	Reaction localization	***		***	
	12	Add subsystems information				
	13	Verify gene−protein-reaction association	***	***	***	***
	14	Add metabolite identifiers				
	15	Determine and add confidence score				***
	16	Add references and notes				
	17	Flag information from other organisms				
	19	Add spontaneous reactions				
	20	Add extracellular transport reactions	***	***		***
	22	Add intracellular transport reactions	***			
	23	Draw metabolic map				
	24−33	Determine biomass composition	***	***		
	34	Add ATP-maintenance reaction				
	35, 36	Add demand and sink reactions				
	37	Determine growth requirements		***		
Stage 4: Network Evaluation	45	Identify metabolic dead-ends				
	46−48	Perform gap analysis				
	51−58	Test for Stoichiometrically Balanced Cycles				
	60−66	Test production of biomass precursors		***		
	67−75	Test production of secretion products				
	76−78	Check for blocked reactions				
	79−80	Compute single gene deletion phenotypes				
	81−83	Test other physiological properties				
	84−94	Test for model growth rate				
Steps Omitted	5, 18, 21, 38-42, 49-50, 59, 95-96					

Figure 16: Review of Software for GENRE

Comparison of the systems SuBliMinal [SSM+11], Model SEED [ADD+12], RAVEN [ALS+13], and Pathway Tools [KPK+09] from the paper [HR14] according to the steps in the protocol of Thiele and Palsson [TP10]. The colour green indicates automatic execution of the step; yellow indicates that the software provides assistance; and red indicates that the software provides no support. The asterisks "***" indicate the need to manually inspect the results of the software.

2.5.2 Basic Local Alignment Search Tool

Sequence alignment algorithms are typically used to align a query sequence against all sequences in a sequence database to find similar sequences or matches. Sequence databases can contain millions of sequences making optimal alignments computationally expensive. As such, fast alignment algorithms were developed. A popular one is Basic Local Alignment Search Tool (BLAST) [AGM+90, AMS+97]. BLAST uses a heuristic algorithm to compute local alignments. The idea is that similar proteins must have short matches.
blast generates all possible short words or substrings of the query sequence. The default length of a word for protein sequences is 3 and for nucleic acid sequences is 11. The algorithm scans a sequence database for sequences that match the words with some threshold. Such matches are called *seeds*. The original BLAST then extends the seeds to the right and left using ungapped alignments [AGM+90]. In following releases, BLAST uses gapped alignments [AMS+97]. The algorithm terminates when the score of the extended alignment falls below some threshold. BLAST reports the extended alignments or hits that have a score at or above the threshold with their statistical significance. Such hits are called High Scoring Pairs (HSPs).

BLAST uses a substitution matrix to compute the scores of each HSP. Statistical analysis of BLAST alignment scores have been performed in the literature [ABGW94, AG96, PJ01]. The statistical significance of a BLAST score S is given by the expected number, *e-value*, of alignments with a score equivalent to or better than S that one would expect with a random sequence. The lower the e-value, the more significant the score and the alignment are.

For a pair of query and subject sequences, BLAST reports all HSPs and their associated measurements. The measurements of interest for the purpose of this document are query coverage, subject coverage, percent identity, e-value, and score. *Query coverage* is the ratio of the length of the HSP in the query sequence to the full length of the query sequence. *Subject coverage* is the ratio of the length of the hit in the subject sequence to the full length of the subject sequence. For protein sequences *percent identity* is the percentage of identical amino acids at the same positions in the alignment with respect to the alignment length. *Score* is the bit score, which is the raw score calculated from the substitution matrix normalized to parameters including the database size [AMS+97].

2.5.3 Amino Acid Composition

The composition of a protein in terms of its amino acids and their physicochemical properties can be crucial in determining the protein structure and function. For example, the helical TMS of a transporter consist of hydrophobic amino acids to be compatible with the hydrophobic bilipid membrane. Table 6 shows properties of the amino acids.

Amino Acids	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Hydrophobic	-	+	+	+	-	+	+	-	+	-	-	+	-	+	+	+	+	-	-	+
Structural	a	a	x	x	i	a	x	i	x	i	i	x	a	x	x	a	a	i	a	a
Chemical	al	s	ac	ac	ar	al	b	al	b	al	s	am	i	am	b	h	h	al	ar	ar
Functional	n	p	ac	ac	n	p	b	n	b	n	n	p	n	p	b	p	p	n	n	p
Charge	n	n	ac	ac	n	n	b	n	b	n	n	n	n	n	b	n	n	n	n	n
Volume	t	s	s	m	x	t	m	l	l	l	l	s	s	m	l	t	s	m	x	x

Table 6: Amino Acid Alphabets

From [BB01, p.117], Hydrophobic: hydrophobic (-), hydrophilic (+); Structural: ambivalent (a), external (x), internal (i); Chemical: acidic (ac), aliphatic (al), amide (am), aromatic (ar), basic (b), hydroxyl (h), imino (i), sulphur (s); Functional: acidic (ac), basic (b), hydrophobic nonpolar (n), polar uncharged (p); Charge: acidic (ac), basic (b), neutral (n). From [PLS+04], Volume: tiny (t), small (s), medium (m), large (l), very large (x).

There are variations [SCH10] of amino acid composition of a protein that are used as features for machine learning.

AAC: The frequency of each amino acid in the protein is the standard amino acid composition (AAC) of a protein, which is a vector of length 20.

PAAC: The frequency of dipeptides of a protein is recorded by the pair amino acid composition PAAC [PK03] which is a vector of length 400.

PseAAC: The pseudo amino acid composition PseAAC [Cho00] of a protein is an extended version of AAC that has λ additional entries which incorporate correlation within a neighbourhood of amino acid physicochemical properties, such as mass, hydrophobicity, or isoelectric point (pI). PseAAC is parameterised by the choice of properties, the choice of λ, and a set of weights.

PsePAAC: A combination of PAAC with the λ last entries of PseAAC is termed PsePAAC. PsePAAC consists of $400 + \lambda$ entries, where the first 400 correspond to PAAC, the frequencies of all amino acid pairs, and the other λ to the neighbourhood correlations of PseAAC.

MSA-AAC: There is a profile-based version called MSA-AAC [PHH07]. The MSA-AAC uses a multiple sequence alignment (MSA) of the protein. For example, the MSA may be built by ClustalW from a maximum of 1000 homologous sequences found using BLAST against the nr non-redundant database. Often sequences with an identity below 25% are removed. The MSA-AAC vector of length 20 records the frequency of each amino acid in all sequences of the MSA.

2.5.4 Hidden Markov Models for Protein Sequences

Hidden Markov models (HMMs) were first described in the late 1960's and subsequently employed in speech processing. In the area of speech processing, an HMM models sounds forming a word or phoneme and generates an output distribution with a high probability for the sounds of the word or phoneme it models. A satisfactory model is that which assigns high probability to the sounds of the word it models and low probability to the sounds of any other word. It was not until the late 1980's that HMMs were employed in several applications in computational biology including modeling homologous nucleotide or protein sequences [KBM+94].

Given the multiple sequence alignment of protein sequences of a protein family, the functional sites of the proteins are projected on the multiple sequence alignment as sites with conserved amino acids. Other sites with no particular features are less conserved. Therefore, each site has a distinct probability distribution over the 20 amino acids that measures the likelihood of each amino acid occurring at that site of the protein family, as well as the probability of no amino acid occurring. A multiple sequence alignment can then be modeled by a probabilistic model that captures the consensus nature of a multiple sequence alignment [KBM+94].

One widely used HMM tool is the HMMER package [Edd98]. It has a number of HMM related programs including *hmmbuild* to train HMMs and *hmmscan* to scan protein sequences against trained HMMs. We use *hmmbuild* to train HMMs and subsequently use *hmmscan* to scan protein sequences against trained HMMs.

2.6 Genomics Resources

Biochemical reference databases contain information related to genome, transcriptome, proteome, and metabolome of organisms. In metabolic pathway reconstruction, this information

acts as a source of genome, gene annotations, and functional annotations, as well as providing reference templates of pathways and reactions. Some of the most widely used databases are shown in Table 7.

Name	**Web Address**	Type
ENZYME	http://www.expasy.ch/enzyme	Enzyme
BRENDA	http://www.brenda.uni-koeln.de	Enzyme
GO	http://www.geneontology.org	Protein classification & annotation
UniProtKB	http://www.uniprot.org	Protein sequence & annotation
GenBank	http://www.ncbi.nlm.nih.gov	Genome sequence
InterPro	https://www.ebi.ac.uk/interpro	Protein families, domain & functional sites
PFAM	http://pfam.sanger.ac.uk	Protein Family, domain, & functional sites
PROSITE	http://prosite.expasy.org/prosite.html	Protein Family, domain, & functional sites
SMART	http://smart.embl-heidelberg.de	Protein domain & annotation
Broad Institute*	http://www.broad.mit.edu/annotation/fgi	Fungal genomes information
MetaCyc	http://MetaCyc.org	Genome & Pathway
KEGG (Pathway)	http://www.genome.jp/kegg	Genome & Pathway
Joint genome Institute (JGI)*	http://genome.jgi.doe.gov/	Genomes
PathGuide	http://www.pathguide.org	Pathway
PUMA2	http://compbio.mcs.anl.gov/puma2	Genome & pathway
Reactome	http://www.reactome.org	Pathway (human)
GOLD*	http://genomesonline.org/	Genome & Metagenomes sequencing projects
TCDB	http://www.tcdb.org	Transporter classification
MIPS	http://www.helmholtz-muenchen.de/en/ibis	Genome & protein sequences
AspGD*	http://www.aspgd.org/	Aspergillus biological information
SGD*	http://www.http://www.yeastgenome.org	*S. cerevisiae* comprehensive information
PubMED	http://www.ncbi.nlm.nih.gov/pubmed	Scientific literature (MEDLINE references)

Table 7: Reference Databases

Genome, enzyme, protein sequence, protein classification, pathway and transporter classification databases are the reference databases that contain biological information crucial for metabolic pathway reconstruction. Those databases marked with "*" provide additional and more specific biochemical information for implicated fungal genomes.

Historically, any work on metabolic pathways would refer back to KEGG [OYH+08] at Kyoto University. KEGG digitized the pathway charts of Boehringer Ingelheim, and created databases for the pathways and the related enzymes, ligands, and genes. See Table 8. The KEGG information is not curated, so it is not as useful as more recent resources.

The KEGG PATHWAY database contains a collection of manually drawn pathway maps to represent molecular interactions, reactions, and pathways. The KEGG pathway maps have been used as the template for developing metabolic models by several software tools, for instance, the RAVEN toolbox of BioMet.

MetaCyc [CAD+10] is a curated database from SRI of pathways, reactions, and metabolites, that grew from the modeling and curation efforts of *E. coli*, namely EcoCyc [KCVSZ+11] originally, and now also TransportDB [RKP04] and RegulonDB [SPGGC+13]. MetaCyc has strong tool support in Pathway Tools [KPK+09] for GENRE.

KEGG Database			
Category	**Entry point**	**Description**	**Instances**
Info. Systems	KEGG PATHWAY	Pathway maps	Metabolism Genetic Information Processing Environmental Information Processing Cellular Processes Organismal Systems Human Diseases Drug Development
	KEGG BRITE	BRITE Functional hierarchies	Pathways & Ontologies Genes & Proteins Organisms & Cells Compounds & Reactions
	KEGG MODULE	modules	Pathway module Structural Complex Functional Set Signature Module
	KEGG MAPPER	Analysis Tools	Mapping tool for PATHWAY, BRITE, MODULES & TAXONOMY
	KEGG ATLAS	Analysis Tools	Navigation tool to explore KEGG global maps
Genomic Info	KEGG GENOME	Collection of genomes	**Prokaryotes (2750)**: Bacteria (2585) Archaea (165) **Eukaryotes (228)**: Animals (81) Plants (35) Fungi (71) Protists (41)
	KEGG GENE	Collection of gene catalogs	GENES: Complete genomes DGENES: Draft genomes EGENES: EST datasets MGENES: Metagenomes VGENES: Viruses
	KEGG Orthology	Ortholog groups	Metabolism Genetic Information Processing Environmental Information Processing Cellular Processes Organismal Systems Human Diseases Drug Development
Chemical Info	KEGG LIGAND	Contains information on chemical substances and reactions	Databases in LIGAND: COMPOUND GLYCAN REACTION RPAIR RCLASS ENZYME

Table 8: KEGG Database

Information systems, genomic and chemical information contained in KEGG [KL] used to reconstruct metabolic networks. Note that this table does not represent all the biological information and analysis tools provided by KEGG.

MetaCyc is the reference template used to reconstruct a metabolic pathway model and the associated database, called a Pathway/Genome Database (PGDB) using Pathway Tools. BioCyc is the collection of PGDBs, which numbers over 2000 genomes.

Today most GENREs can be found at BiGG [SPCP10], *"a Biochemically, Genetically and Genomically structured genome scale metabolic network reconstruction knowledgebase"* at Bernhard Palsson's Systems Biology Lab at UC San Diego. Models are encoded in the systems biology markup language (SBML) [HFS+03]. They develop the COBRA toolkit [SQF+11] for analysis of GENREs.

Specific to modeling pathways, rather than to systems biology as a whole, is the BioPAX community [DCP+10] for Biological Pathways Exchange in XML. BioPAX is represented in RDF/XML and is defined in OWL.

For annotation of enzymes specifically, there is the Enzyme Commission (EC) classification scheme, which is supported by the BRENDA database [SCP+13] of EC definitions, reactions, metabolites, and enzymes. For annotation of transporters, there is the Transporter Classification (TC) scheme, which is supported by the TC database (TCDB) [STB05]. For annotation in general, one uses the Gene Ontology (GO) [The00]. GO covers enzymes and transporters amongst its collection of terms for annotation. The GOA database [HSMM+15] links gene ontology annotations to the entries in SwissProt and UniProt.

ENZYME and BRENDA are two widely used enzyme databases for genome annotation. ENZYME is maintained by the Swiss Institute of Bioinformatics. BRENDA is developed and maintained by Department of Bioinformatics and Biochemistry, Technische Universität Braunschweig, Germany. Both support the Enzyme Commission (EC) number official classification of enzymes based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). BRENDA incorporates over 1,000,000 enzymes, of which more than 65,000 are manually curated.

For curated protein sequences and information about the proteins, one consults the SwissProt database [BA00], which is the set of reviewed entries in UniProt [C+14], a resource with both reviewed and unreviewed protein sequences. SwissProt collaborates closely with curators for model organisms, and others, such as the AspGD database [CAI+14] for *Aspergillus* species.

UniprotKB is a protein knowledgebase comprised of two different sections: (1) SwissProt for manually annotated and reviewed proteins; and (2) TrEMBL for protein sequences that are automatically annotated but not reviewed.

The Transporter Classification database (TCDB) [Gro] contains information on characterized transporters based on the Transporter Classification (TC) system of IUBMB. It is a curated database of more than 10,000 proteins and more than 10,000 literature references for more than 800 transporter families.

The Saccharomyces Genome Database (SGD) [SUa] is a manually curated database about the yeast model organism *Saccharomyces cerevisiae*. The Aspergillus Genome Database (AspGD) [MGD$^+$] is a database of filamentous fungi of the genus *Aspergillus*. It also acts as a multispecies comparative genomics browser tool.

Chapter 3

Metabolic Pathway Reconstruction

This chapter focuses on one aspect in the automation of systems biology, namely the reconstruction of the metabolic pathways. This step begins with an annotated genome of an organism, and perhaps with other data such as RNA-Seq expression data, and produces a model of the metabolism of the organism's cell. The model includes metabolic reactions organised into pathways that transform metabolites, and may include information on transport and regulation.

Automation of the reconstruction of metabolic pathways is necessary if we wish to study non-model organisms. Any manual aspect in the process of constructing models and quality control of models is time-consuming. Experience indicates that manual reconstruction takes upwards of six months to two years [TP10, p. 2]. Our experience in this chapter shows that Pathway Tools takes less than one hour on a workstation to construct a metabolic pathway model of a fungal genome.

While there are many toolkits that automate some steps of the process of reconstruction, there are only two software systems that one would consider as automating the full reconstruction process; they are SEED [ADD+12] and Pathway Tools [KPK+09]. Both provide semi-automation and not full automation. Both work best on genomes of prokaryotes, and Pathway Tools is the only one that can claim to work with eukaryotes. We work with fungi, which are eukaryotes, so this chapter uses Pathway Tools in case studies in order to better understand the strengths and weaknesses of the state of the art.

The chapter organization is as follows: Section 3.1 reviews the state of the art for this step in the overall process; Section 3.2 looks at those fungal genomes that are well-curated in

order to see the completeness (or non-completeness) of their functional annotations; Section 3.3 presents our case studies in reconstructing metabolic pathway models for fungi; and Section 3.4 presents the lessons learned about the strengths and weaknesses of metabolic pathway reconstruction.

3.1 The State of the Art

This section reviews the state of the art for the reconstruction of metabolic networks, which is the starting point for systems biology. This section complements existing reviews [FST05, PRU10, FS11, OP10, RGM+12, SCM14, HR14]. A common approach is to construct a draft network model based on a reference of known pathways and reactions, as typified by Pathway Tools [KPK+09], which uses the MetaCyc knowledgebase of pathways curated from the literature to provide a template of metabolic, transport, and regulatory pathways against which to match the roles of proteins in a genome. The primary input to the process is an annotated genome. The steps in reconstruction are:

1. Establish the Gene-Protein-Reaction (GPR) associations, based on the functional annotation of the genes. The reaction types may include one or more of metabolism, transport, and regulation.

   Techniques may use the annotation of each gene in terms of a text description, GO terms, and EC numbers [KLC11]; homology and orthology [NEF+06]; or HMMs for protein families, eg FigFAMs [ADD+12].

2. Determine the pathways present in the organism, based on the reactions present in the GPR associations [KLC11, DPK10].

3. Perform *hole filling*, also called *gap filling*, by considering each reaction in the pathways present in the organism that are not associated with a Gene-Protein [OP10].

   The pathways present may have holes; that is, there are orphan reactions in the pathway that are not assigned to a gene. The hole-filling algorithm [GK04, GK07] uses a Bayesian approach to rank the genes in the genome with each hole, and the software allows curators to accept or reject a match. There are alternative hole-filling approaches that use orthology (AutoGraph [NEF+06]) and expression data (GLOBUS [PFH+12]).

In systems biology, these steps are followed by quantitative modeling and quality control that balances flux, charge, energy, etc.

There are also *de novo* approaches to predicting previously undiscovered pathways. They may use comparative genomics [FK11], expression profiles [SUS07], or gene clusters in prokaryotes [ADD+12]. These may use a knowledge of chemistry and the reactions in the organism to predict a set of pathways that connect the given reactions and suggest other required reactions. Alternatively, our knowledge of chemistry and data on the metabolites present in the organism can be used to predict the reactions and pathways that match the set of given metabolites. These approaches are called *compound-filling* and *reaction-filling* in [KTY+13] as compared to the *reference-based* approach above.

3.1.1 Pathway Tools

The Pathway Tools software is an integrated system that employs the metabolic pathway ontology to develop a specific organism pathway database. This software tool was developed by Peter Karp and his co-workers at Stanford Research Institute (SRI) and its development has been continuously ongoing since 1990s following the successful construction of the *E.coli* pathway database (EcoCyc) [KPK+09]. The EcoCyc database is the first model-organism database (MOD) developed within SRI. The MOD created by Pathway Tools is called a Pathway/Genome Database (PGDB). EcoCyc is the only PGDB based on information derived from the biomedical literature [KPR02], prior to the construction of YeastCyc for *S. cerevisiae*.

Pathway Tools [KPK+09] uses the MetaCyc knowledgebase of pathways that have been curated from the literature to provide a template of metabolic, transport, and regulatory reactions and pathways. Using an existing functional annotation, the tools first match genes to reactions, then determine whether each pathway is present or not in the organism [KLC11, DPK10]. The pathways present may have holes; that is, there are orphan reactions in the pathway that are not assigned to a gene. The hole-filling algorithm [GK04, GK07] uses a Bayesian approach to rank the unassigned genes in the genome with each hole, and the software allows curators to accept or reject a match.

MetaCyc [SRIb] is the reference database for all the PGDBs constructed using Pathway Tools. It is a freely available comprehensive knowledgebase that contains biological information on metabolic pathways and enzymes from all domains of life, which are extracted

from the scientific literature [CAD+10]. PGDBs constructed using Pathway Tools integrate information of the genome of an organism such as genome sequence, biochemical data such as metabolites, substrates, pathways, metabolic network, and the genetic network of an organism [KPR02].

Pathway Tools uses the Metabolic Pathway Ontology (MPO) to encode high fidelity biological information. The output is a Pathway/Genome Database (PGDB) [KPK+09]. There are three ontologies within Pathway Tools: the evidence ontology, the cell component ontology, and the protein feature ontology [Kar]. They capture genomic datatypes by a rich set of classes, attributes and relationships for biological data modeling [KPR02]. According to [GK06], the performance of Pathway Tools depends on these ontologies.

The main component of the Pathway Tools software is known as PathoLogic, which infers probable metabolic pathways based on genome annotation, infers transport reactions using the Transport Inference Parser (TIP), and assists users to perform refinement on the created PGDB, such as filling pathway holes. Pathway Tools also provides a user-friendly navigation interface that allows user to perform large-scale data analysis, querying, and visualization; curation tools to edit or update existing information; and MetaFlux for flux-balance analysis. Pathway Tools can be installed locally and used from the desktop or a web browser.

Tier	Databases	Description
1	EcoCyc MetaCyc HumanCyc AraCyc YeastCyc LeishCyc	- EcoCyc and MetaCyc were created through intensive manual efforts based on experiment information elucidated from scientific literatures. - The rest of the PGDBs in this tier were created using Pathway Tools Software - All these PGDBs received literature-based curation by scientist continuously (at least once a year).
2	36 databases; 16 eukaryotes 20 prokaryotes.	- PGDBs were generated computationally using PathoLogic. - Undergone moderate amounts of manual reviews (e.g. removing false-positive pathway predictions), updates, and polishing steps (e.g. defining protein complexes). - Undergone short period of literature based curation. Most PGDBs undergo 1–4 months of curation - Only 1 database for fungi but it is unavailable (*Penicillium chrysogenum Wisconsin 54-1255)*
3	2950 databases	- PGDBs were created using PathoLogic but without any manual review nor subsequent literature-based curation. - Do not even run pathway hole filler for predicting missing enzymes

Table 9: Tiers in BioCyc

Tiers in BioCyc as of March 2014 [SRIa]

Nowadays, many researchers use Pathway Tools software to reconstruct metabolic networks. As reported by [KLC11], the SRI BioCyc database collection [SRIa] contains PGDBs for more than 1000 genomes. Table 9 shows the different categories, or tiers, of PGDBs in

BioCyc. The popularity is believed due to the state-of-the-art algorithm of PathoLogic that can automatically infer metabolic pathways and quickly create a new PGDB. Pathway Tools also provides manual, semi-automated and automated database refinement tools for curation purposes. In FungiCyc [BI], there are more than 20 genome-scale metabolic networks of fungi constructed using Pathway Tools, including YeastCyc for *S. cerevisiae* and AspCyc for the Aspergillus genomes.

3.1.2 SEED

The metabolic network models in Model SEED [TFfIoG] were constructed computationally using a custom pipeline of automated and manual steps. The aim was to reconstruct metabolic network with consistent, high quality and rapid genome annotations from a newly sequenced genome based on the subsystems approach [ABB+08, ADD+12]. The term *subsystem* is the general concept of a pathway. A subsystem is represented as a graph consisting of proteins such as enzymes and transporters, and compounds as nodes, while edges link the nodes. However, compounds like cofactors are omitted in these linkages. The variants of the subsystem are produced as a subgraph. The variant detection is performed using integer programming and visualized using Graphviz [YOOG05]. The majority of the model SEED are for bacteria; one good example is the gram-positive bacteria *Bacillus subtilis* [HZCS09]. SEED cannot produce models for eukaryotes.

To quote from [HZCS09], *"The Model SEED integrates existing methods and introduces techniques to automate nearly every step of this process, taking approximately 48 hours to reconstruct a metabolic model from an assembled genome sequence. We apply this resource to generate 130 genome-scale metabolic models representing a taxonomically diverse set of bacteria. Twenty-two of the models were validated against available gene essentiality and biological data, with the average model accuracy determined to be 66% before optimization and 87% after optimization."*

3.1.2.1 What Is There (WIT)

A precursor to SEED was What Is There (WIT) [OLP+00]. WIT performed comparative genome analysis and reconstruction of metabolic pathways based on the Enzyme and Metabolic Pathways (EMP/MPW) family of databases. WIT processed genomes of prokaryotes, performing gene finding, gene annotation, finding gene clusters on chromosomes, and

clustering orthologs as bidirectional best hits across related genomes. The metabolic model could be viewed in both textual and graphical form. Model refinement by curators was supported, allowing evaluation of the model against biochemical data and phenotypes known from the literature.

3.1.3 Pathway Analyst

Pathway Analyst [PPS+05, PSLG06] is a freely accessible web server that can be used to predict metabolic pathways from the protein sequences of an organism. Pathway Analyst uses each of Support Vector Machines (SVM), BLAST and Hidden Markov Models (HMM) predict matches between sequences in the set of model organism pathways and the sequences in the target organism to predict metabolic pathways.

3.1.4 AUTOGRAPH

The key steps of AUTOGRAPH (Automatic Transfer by Orthology of Gene Reaction Associations for Pathway Heuristics) [NEF+06] apply comparative genomics using orthology as determined by InParanoid [ÖSF+10] rather than sequence similarity. Models from comparative organisms act as reference templates that supply the reactions and the pathways. These models also have GPR associations. The protein sequences are used by InParanoid to find matches between the target organism and the comparative organisms. Once a match is found, the reaction can be assigned to the target Gene-Protein. AUTOGRAPH is compared to PathoLogic from Pathway Tools on a bacterial genome, *L. lactis* in [NEF+06]. AUTOGRAPH assigned reactions to 186 more genes than PathoLogic, of which 43% were transport reactions. The AUTOGRAPH method should be considered a protocol as it is not implemented as software, but rather executed by hand.

3.1.5 Pantograph

Pantograph is the first system for metabolic pathway reconstruction that was designed from the bottom-up for fungal genomes. It includes cellular components, including the peroxisome, and specifically modeled transport across the membrane of the peroxisome. Pantograph was designed and implemented in the PhD thesis [Loi12] of Nicolas Loira at Bordeaux, and

applied to reconstruct the metabolic pathways of the yeast *Yarrowia lipolytica* which accumulates lipids in the peroxisome component of the cell. The Pantograph method [LZS15] relies on a database of profile HMMs for fungal protein families and their annotations that is maintained at Génolevures in Bordeaux. The protein families are designed to be orthologous proteins. It also relies on a reference template, which Pantograph calls the *scaffold model*, which also models the cell compartments. The Pantograph algorithm first assigns GPR associations, and then must decide what to include in the draft model based on these associations. Like PathoLogic, this includes selecting which pathways are present in the organism. Unlike PathoLogic, Pantograph also selects which compartments should be included in the model for the organism.

The scaffold model, which is the reference template for Pantograph, was manually curated to include 421 transport reactions. The associated transport protein families of orthologs were manually identified in the Génolevures collection.

The Pantograph software, written in Python, is available for download at `http://pathtastic.gforge.inria.fr/`. The distribution includes the scaffold model in SBML (Systems Biology Markup Language). The scaffold is intended to cover yeasts, while our lab work deals with another kind of fungi, the filamentous fungi.

3.1.6 Other Tools

Two systems that take a genome sequence as input, and combine the steps of identification of genes (that is, coding sequences in prokaryotes), functional annotation of genes using EC numbers, and reconstruction of metabolic pathways are IdentiCS [SZ04] and metaSHARK [PSMW05]. IdentiCS (Identification of Coding Sequences from Unfinished Genome Sequences) uses BLAST to search the genome for matches to genes and proteins in the public databases KEGG and SwissProt that have EC number annotations. Having identified the coding sequences for proteins that are enzymes, it constructs the pathways from the templates in KEGG. The metaSHARK (metabolic SearcH And Reconstruction Kit) system uses HMM profiles to search the genome to identify such coding sequences and proteins. It also uses the KEGG templates to reconstruct the metabolic pathways. The HMM profiles are based on the PRIAM [CRCFK03] profiles and sequences. Once a coding region is identified, the Wise2 [BCD04] gene predictor is applied to identify the gene.

KOBAS (KEGG Orthology Based Annotation System) [WMC+06, XMH+11] annotates

genes and proteins against the KEGG databases. It identifies the pathway and reaction associated with the sequence. However, KOBAS does not reconstruct metabolic pathways. KAAS (KEGG Automatic Annotation Server) [MIO+07, OYH+08] is similar annotation tool that is designed to process genomes and reconstruct metabolic pathways..

ComPath [CK08] is an interactive tool that integrates various databases and computational analysis tools in the interactive spreadsheet to reconstruct pathway and annotation of an organism. Information from sequence, structure and domain databases, and KEGG, are integrated with computational tools into an interactive spreadsheet. Its main aim is to identify GPR assocations, and perform pathway analysis.

Rahnuma [MPH09] is a hypergraph tool used to perform metabolic pathways predictions and analysis. It is written in JAVA and uses a MySQL database to store data from KEGG. Rahnuma has three main modules: network analysis module that builds a metabolic network over a phylogeny of related organisms; pathway analysis module to perform pathway predictions; and comparative analysis module that allows the user to compare two metabolic networks. However, there is no available information on the metabolic pathway predictions of an organism using this tool.

3.2 Well-Curated Fungal Genomes

Our research (https://www.fungalgenomics.ca) searches fungal genomes for secreted enzymes that have potential industrial applications such as biofuel, textiles, pulp bleaching, paper deinking, food processing, and feed processing for livestock. So functional annotation focuses on fungal genomes. There are 8 well studied fungal genomes where significant effort on manual curation has been done (see Table 10 and Table 11). Table 12 shows the number of annotations with GO terms for the three aspects — biological process (BP), molecular function (MF), and cellular component (CC) — for both automatic and manual annotations. Table 13 shows the number of proteins with manually annotated GO terms across different combinations of the three aspects: biological process (BP), molecular function (MF), and cellular component (CC). Together the tables show the level of incompleteness of our knowledge of the role of proteins. This incompleteness is the status in general, as seen in Section 1.1, and not particular to only fungal genomes.

Genome	**Size (Mbp)**	**Source**	**Genetic Elements**	**No.**
S. cerevisiae S288C	12	http://www.yeastgenome.org/	Chromosomes	16
S. pombe ASM294	13	http://www.pombase.org/	Chromosomes	3
C. albicans SC5314	29	http://www.candidagenome.org/	Contigs	22
A. fumigatus Af293	29	http://www.aspergillusgenome.org/	Chromosomes	8
A. nidulans FGSCA4	30	http://www.aspergillusgenome.org/	Chromosomes	8
A. niger CBS513.88	34	http://www.aspergillusgenome.org/	Contigs	19
A. oryzae RIB40	38	http://www.aspergillusgenome.org/	Chromosomes Contigs	8 3
N. crassa OR74A	40	https://www.broadinstitute.org/	Supercontig Chromosomes	7 1

Table 10: Sources of Well-Curated Fungal Genomes

The summary of 8 well-curated fungal genomes. Column 1 contains the name of the strain, followed by the *size* column that indicates the size for each strain in megabase pair *(Mbp)*, the *source* websites, the type of *Genetic Elements*, and the last column (*No.*) displays the number of genetic elements.

Organism	ORFs Total	ORFs Verified	ORFs GO
S. cerivisiae S288C	6607	5061	5910
S. pombe ASM294	5123	NA	5456
N. crassa OR74A	9730	NA	NA
C. albicans SC5314	6214	1504	6045
A. nidulans FGSCA4	10678	1113	10750
A. niger CBS513_88	14056	214	14386
A. fumigatus Af293	9783	449	10070
A. oryzae RIB40	11902	157	12173
Total	74093		

Table 11: Well-Curated Fungal Genomes

The table indicates the number of proteins (actually open reading frames (ORFs)) based on the gene models of the genome. The total number of ORFs is given, as well as those ORFs verified by the existence of some other experimental data such as transcripts or proteins. Finally, the number of ORFs for which there is at least one Gene Ontology (GO) term, regardless of whether the term is electronically annotated or manually assigned. Note that for *N. crassa* where the genome comes from the Broad Institute, the downloaded files contain only those ORFs that have at least one manual annotation. So in *N. crassa* all proteins are verified and have at least one manually annotated GO term. (As of August 2013)

Organism	ORFs Total	GO BP	GO MF	GO CC	GO Total
S. cerivisiae S288C	6607	31192	25980	34597	91769
S. pombe ASM294	5123	13323	9383	14636	37342
N. crassa OR74A	9730	3261	1222	1996	6479
C. albicans SC5314	6214	7291	6870	6085	20246
A. nidulans FGSCA4	10678	6160	5973	4886	17019
A. niger CBS513.88	14056	6543	6445	4980	17968
A. fumigatus Af293	9783	5569	5460	4607	15636
A. oryzae RIB40	11902	6561	6412	4913	17886

Table 12: GO Annotation of Well-Curated Fungal Genomes

The table indicates the number of GO annotations of proteins (actually open reading frames (ORFs)) based on the gene models of the genome. The columns list the number of annotations in the three aspects biological process (BP), molecular function (MF), and cellular component (CC), and the total number of GO annotations. Note that a protein may have more than one GO annotation in an aspect. Note that for *N. crassa* where the genome comes from the Broad Institute, the downloaded files contain only those ORFs that have at least one manual annotation.

Organism	ORFs BP	ORFs MF	ORFs CC	ORFs BPMF	ORFs BPCC	ORFs MFCC	ORFs None	ORFs ≥ 1	ORFs BPMFCC
S. cerivisiae _S288C	4771	3996	5193	3857	4581	3814	480	5430	3722
S. pombe ASM294	4423	3558	5094	3420	4338	3481	264	5192	3356
N. crassa OR74A	1530	891	1082	812	870	754	0	1786	719
C. albicans SC5314	1475	983	984	863	674	535	4173	1872	502
A. nidulans FGSCA4	1338	954	554	920	321	212	9156	1594	201
A. niger CBS513.88	494	399	197	371	90	85	13761	625	81
A. fumigatus Af293	537	362	170	304	59	33	9376	694	21
A. oryzae RIB40	380	346	52	320	24	22	11743	430	18
Total								17633	8620

Table 13: Number of Proteins with Manual GO Annotations by Aspect

The table presents the number of proteins with manually assigned Gene Ontology (GO) terms for different combinations of the three aspects: biological process (BP), molecular function (MF), and cellular component (CC).

3.3 Case Studies

The case study investigated the application of Pathway Tools, a widely used tool for the reconstruction of metabolic pathways, to a range of fungal genomes. Five of them are from the list of well-curated fungal genomes in Section 3.2, while the other is a genome of interest, *Phanerochaete chrysosporium RP78*. One aim was to see how variable the results were, and whether there was a link between the functional annotation and the result, both in terms of quality and quantity of the annotation. For this reason, we include *P. chrysosporium RP78* that was automatically annotated at the Joint Genome Institute (JGI).

This section describes the Datasets, the Methods, the Results, and then presents the case of *P. chrysosporium RP78* in detail. This is followed by Discussion.

3.3.1 Datasets

The protein sequences and annotation information of these fungi are gathered from three different resources. The *Aspergillus* genomes from AspGD, the *N. crassa* genome from the Broad Insittute, and *P. chrysosporium* from JGI. Table 14 shows the summary of the genomes involved in this study.

Genome	**Size (Mbp)**	**Source**	**Genetic Elements**	**No.**
A. fumigatus Af293	29	http://www.aspergillusgenome.org/	Chromosomes	8
A. nidulans FGSCA4	30	http://www.aspergillusgenome.org/	Chromosomes	8
A. niger CBS513.88	34	http://www.aspergillusgenome.org/	Contigs	19
A. oryzae RIB40	38	http://www.aspergillusgenome.org/	Chromosomes Contigs	8 3
N. crassa OR74A	40	https://www.broadinstitute.org/	Supercontig Chromosomes	7 1
P. chrysosporium RP78	35	http://jgi.doe.gov/	Scaffolds	178

Table 14: Sources of Fungal Genomes for Case Study

The summary of six fungal genomes: five well-curated fungal genomes and one automatically annotated fungal genome *P. chrysosporium RP78*. Column 1 contains the name of the strains, followed by the *size* column that indicates the size for each strain in mega base pair *(Mbp)*, the *source* websites, the type of *Genetic Elements*, and the last column (*No.*) displays the number of genetic elements.

The datasets for *A. fumigatus Af293*, *A. nidulans FGSCA4* and *A. niger CBS513.88* are as of March 2014, and the datasets for *A. oryzae RIB40* are as of June 2014. The annotations

for the genes are GO terms from either manual curation, or orthology to a gene in another *Aspergillus* species that is manually curated.

The download site at AspGD provides protein sequences as .fasta files; genome information with gene definitions in .gff files; and GO annotations for all genomes in one file in standard GAF format in the `sequences`, `gff` and `go` directories respectively.

`A_nidulans_FGSC_A4_version_current_orf_trans_all.fasta`

`A_nidulans_FGSC_A4_version_current_features.gff`

`gene_association.aspgd`

The Broad Institute information for *N. crass OR74A* is available at `http://www.broadinstitute.org/annotation/genome/neurospora/MultiDownloads.html`. It is equivalent information, though formatted differently: the gff files use the suffix `gtf`, and the `tsv` file for the GO terms is not in GAF format. The only annotations in the files from Broad are manually curated annotations.

`neurospora_crassa_or74a_12_transcripts.gtf`

`neurospora_crassa_or74a_12_proteins.fasta`

`http://www.broadinstitute.org/annotation/genome/neurospora/assets/go_for_nc12.tsv`

The dataset of *P. chrysosporium RP78* v2.1 is downloaded from JGI. Table 15 shows the annotation files that contain information used to create the input file for Pathway Tools. The annotations are automatically computed by the pipeline at JGI.

Table 16 shows the number of curated pathways in MetaCyc from the different kingdoms of life. Note the predominance of pathways from bacteria. Statistics on the reference pathways in MetaCyc used in reconstructions is shown in Table 17.

3.3.2 Methods

We develop metabolic models and pathway genome databases (PGDBs) using PathoLogic of Pathway Tools Software v17.5. The annotation input files for each genome are formatted according to the PathoLogic (.pf) format. This format accepts information on the roles of

Description	Filename	Total
Scaffolds	Pchrysosporium_BestModelsv2.1.gff.gz	178
Transcripts in FASTA	BestModels2.1.transcripts.gz	10048
Transcripts (KEGG)	Pchrysosporium_ecpathwayinfo_BestModels2.1.tab.gz	4012
EC (KEGG)	Pchrysosporium_ecpathwayinfo_BestModels2.1.tab.gz	2155
Pathways (KEGG)	Pchrysosporium_ecpathwayinfo_BestModels2.1.tab.gz	107

Table 15: Annotation for *P. chrysosporium* RP78

This table displays genome annotation for *P. chrysosporium* (version 2.0) downloaded from JGI. There are 412 scaffolds being annotated in gff but only 178 can be used for PathoLogic annotation. The FASTA file represents DNA sequences where 10048 genes were annotated. But only 4012 genes were annotated by KEGG, together with 2155 EC numbers assigned and a total of 107 pathways. All this information is used for the PathoLogic input file annotation.

Source	Total
Bacteria	883
Plants	607
Fungi	199
Mammals	159
Archaea	112

Table 16: Source of Curated Pathways in MetaCyc

Indicative source of curated pathways in MetaCyc, as of v13.1 As of v17.5 MetaCyc has about 35% more pathways.

Description	Total
Pathways	2089
Polypeptides	10885
Protein Complexes	3356
Enzymes	9146
Enzymatic reactions	11410
Compounds	10965
Transporters	101
Transport reactions	154

Table 17: Biological Entities in MetaCyc

Biological entities in MetaCyc version 17.5 used in case studies.

genes in terms of text descriptions, EC numbers, GO terms, and KEGG pathways. For the *Aspergillus* genomes, the information on genes and sequence assemblies are extracted from the gff, sequence, and GO directories on the AspGD download site. The EC numbers for enzymes are retrieved from Uniprot. For *N. crassa OR74A* from the Broad Institute the information on genes, GO annotations, and EC numbers are in the downloads. For *P. chrysosporium RP78* v2.0 from JGI there are 412 scaffolds in the gff file, but only 178 with genes used for PathoLogic. From the downloads, there are EC numbers assigned to 2155 proteins that is used together with the GO annotations.

MetaCyc contains a list of reactions and a list of pathways defined in terms of the reactions. PathoLogic identifies a list of potential reactions for an organism from the gene annotations, primarily the EC numbers assigned to genes. From the list of reactions PathoLogic selects the pathways most likely to occur in the organism using an algorithm based on random forests [KLC11, DPK10]. In addition PathoLogic runs the Transport Inference Parser (TIP) [LPK08] to predict transport reactions based on keywords in the gene descriptions and annotations.

3.3.3 Results

The reconstruction of each metabolic model took between 25 to 35 minutes on a workstation with 3.4GHz processor and 16GB memory running Linux. Table 18 shows a summary of the statistics for each model.

3.3.4 Details for *P.chrysosporium RP78*

Phanerochaete chrysosporium is the model organism for white-rot fungi which have extraordinary capabilities to degrade lignin and a wide range of toxic chemical pollutants. Its genome was the first genome from a basidomycote fungus to be sequenced [MLP$^+$04]. It is the most extensively studied white rot organism due to its unique ability to degrade dioxins, polychlorinated biphenyls (PCBs) and other chloroorganics. This makes it a spearhead fungi in bioremediation research. Its gene complement of glycosyl hydrolases, cytochrome P450 peroxidases, and oxidases is impressive. Therefore, *in silico* metabolic network reconstruction of *P. chrysosporium* is anticipated to help in understanding its metabolic capabilities and in predicting the functions of genes and proteins.

Description	Afu	And	Ang	Aor	Ncr	Pch
Pathway	287	312	319	302	299	227
Enzymatic Reaction	1871	1868	1963	1875	1900	1480
Transport Reaction	12	11	10	12	13	10
Genes	10073	10983	14296	12176	10812	9624
Polypeptides	10074	10923	14970	12176	10815	10007
Enzymes	1615	1580	1997	1782	1327	1742
Transporters	37	41	38	38	44	41
Compounds	1326	1350	1434	1311	1461	1212
Pathway Holes (%)	315 (32)	335 (31)	343 (32)	340 (33)	248 (25)	311 (37)

Table 18: Statistics on PGDBs for Six Fungal Genomes

Afu: *A. fumigatus Af293*, And: *A. nidulans FGSC A4*, Abg: *A. niger CBS513.88*, Aor: *A. oryzae RIB40*, Ncr: *N. crassa OR74A*, and Pch: *P. chrysosporium RP78*. The *pathway* indicates the number of base pathways, *enzymatic reactions* are reactions that are catalyzed only by enzymes, *transport reactions* are reactions occured in cellular compartments where the involved substrates reside, *genes* and *polypeptides* are to represent the number of predicted genes and proteins respectively, *enzymes* are the proteins that catalyze reactions, *transporters* represent total number of membrane transport proteins in each fungal genome, *compounds* are small molecules used in the reactions, while the *pathway holes* represents the number of missing enzymes or gaps that exist in base pathways, together with the percentage displayed in the brackets.

Based on the dataset, a PGDB for *P. chrysosporium* is developed using PathoLogic and is given the name PHACHCyc. The predicted number of pathways and other biological entities are shown in Table 18.

An analysis of the completeness of the model looked at both the overall number of pathways (Table 18) the percentage of holes (Table 18), the pathways with and without holes (Table 19), and the distribution of holes across pathways (Figure 17).

Description	Total	Percentage
Pathway reactions that are holes	311	36.5
Pathway reactions that are not holes	540	63.5
Total no. of pathway with holes	125	55.1
Total no of pathway without holes	102	44.9

Table 19: Pathway Holes Predicted by Pathway Hole Filler

The table shows the pathway holes predicted by Pathway Hole Filler. These are the total number of pathway holes occur in PHACHCyc, the number of pathways affected and the percentage for each case.

The single pathway with the highest number of holes, 24, in Figure 17 is the Palmitate

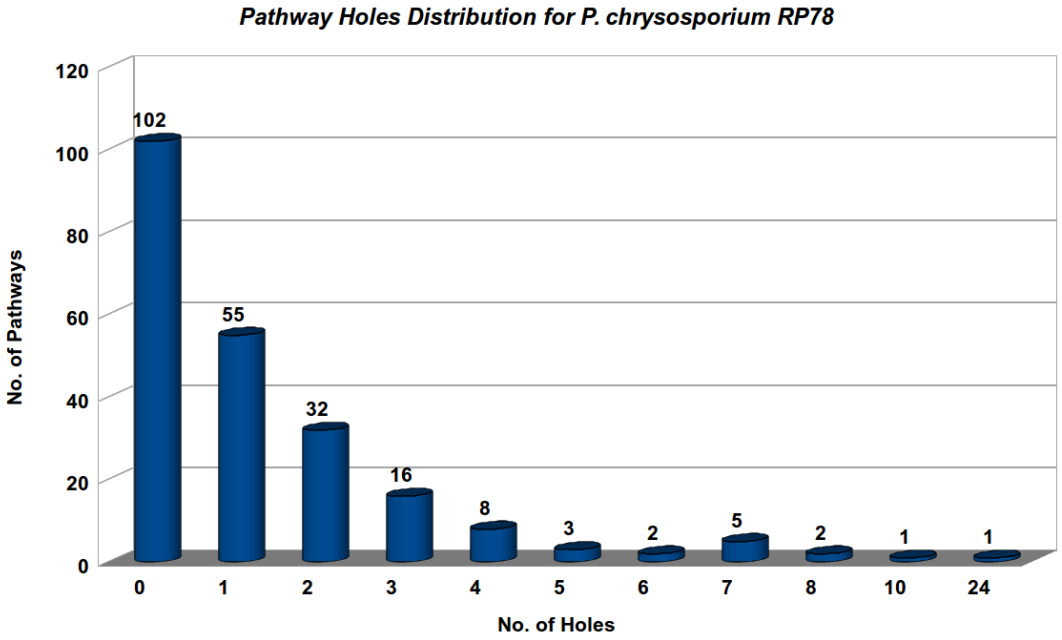


Figure 17: Histogram of the Pathway Hole Distribution for PHACHCyc
The pathway hole distribution for PHACHCyc. The number on the top of each bar represents the number of pathways corresponding to the number of holes.

Biosynthesis I (animal & fungi) pathway which is essential for fatty acid biosynthesis. The pathway has 32 reactions in total. One of the pathways that is missing only a single GPR is the TCA cycle II (plants and fungi) in Figure 18. The figure has an arrow pointing to the hole.

The impact of hole-filling is investigated to see the effectiveness of the Pathway Hole Filler program [GK04] of Pathway Tools. Hole-filling is a semi-automated process that returns a ranked list of candidate genes for the hole, together with a score given as a percentage. The default threshold for accepting a gene to fill a hole is 90%. Figure 19 considers a range of thresholds from 90% down to 50% and shows the number of holes filled, and the number of holes remaining. At a threshold of 90% about 45% of the holes are filled. Without further experimental results, we do not know whether a correct Gene-Protein-Reaction association has been made to fill a hole.

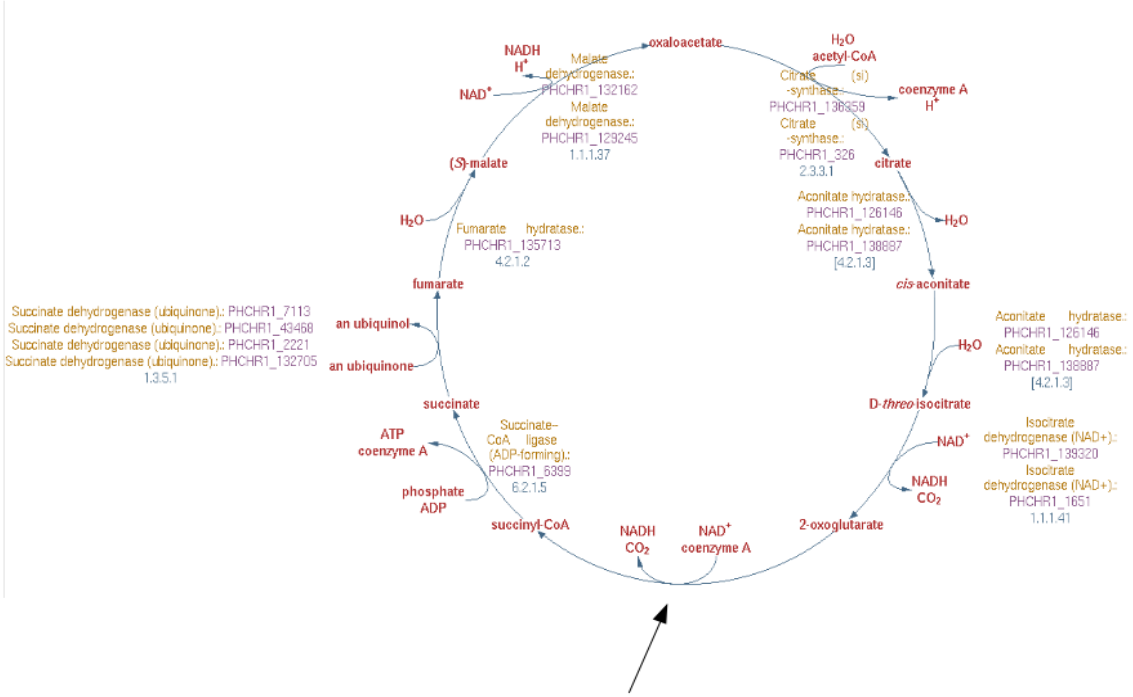


Figure 18: TCA Cycle Model for PHACHCyc

The TCA cycle predicted for *P. chrysosporium RP78* is TCA cycle II (plants and fungi). The black arrow is pointing to a missing reaction in this pathway.

3.3.5 Discussion

This section discusses all aspects of the case studies including *P. chrysosporium*. Evaluation of the methods for GENRE is very problematic because new organisms do not have a ground truth available, and similarly novel pathways in model organisms do not have a ground truth available. Validation of predictions requires wet lab experiments. Therefore the arguments in this section are internal validations based on statistics of the reconstructed metabolic pathway models.

Quality of Genome Assembly and Annotation Affects GENRE

The selected fungal genomes exhibit a range of completeness for their genome assemblies, as shown in Table 14. Several have assemblies that contain complete chromosomes: *A. fumigatus Af293*, *A. nidulans FGSC A4*, *A. oryzae RIB40*, and *N. crassa OR74A*. *A. niger CBS 513.88* has an assembly with approximately two contigs per chromosome, while *P. chrysosporium RP78* has many hundreds of contigs, which would be considered a moderate quality assembly.

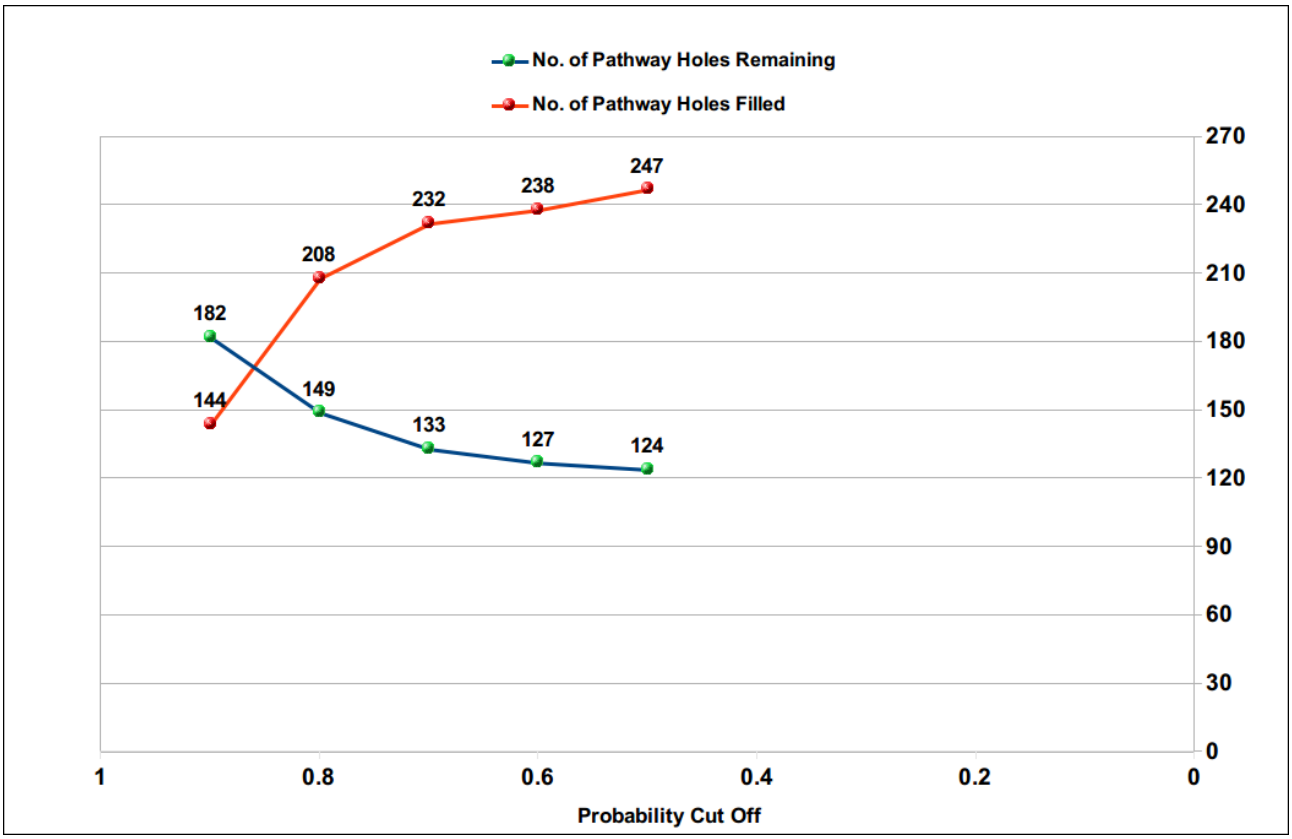


Figure 19: Number of Hole Candidates versus Cutoff

The number of hole candidates based on the probability cut off. The number at each point of the blue line represents the total number of candidate genes that filled the holes. The number at each point of the green line represents the total number of holes remaining after hole filling. Due to some double counting of holes, the sum of holes filled and holes unfilled is more than 311, the total number of holes.

The selected fungi display some phylogenetic diversity. *N. crassa OR74A* is a yeast, while the others are filamentous fungi. *P. chrysosporium RP78* is a basidomycote, while the others are ascomycote.

The genome annotations range from fully automatic (*P. chrysosporium RP78*) to fully manual (*N. crassa OR74A*) with the *Aspergillus* genomes combining manual curation and annotation transfer by orthology from other *Aspergillus* species. One would deem manual annotations to be high quality and automatic annotation to be low quality, as a rule of thumb. This is supported by Table 11 and Table 13. The number of verified ORFs indicates the number of gene predictions that are supported by experimental evidence. This is substantially higher for *A. nidulans FGSC A4* at 1113, than for the other *Aspergillus* genomes; it is not available for *N. crassa OR74A*, and is presumably zero for *P. chrysosporium RP78*.

The number of manually curated GO terms by aspect (BP, MF, CC) in Table 13 shows that cellular components in *N. crassa OR74A*, a yeast, are better known than for the filamentous fungi, and that the cellular components for *A. nidulans FGSC A4* are much better studied than the other *Aspergillus* genomes. This relationship holds true for the number of ORFs for which there is a curated GO term for each of the three aspects. Note that *P. chrysosporium RP78* has no manually curated GO terms at all.

However, the results in Table 18 do not show substantial differences between the models of the *Aspergillus* genomes themselves, nor with the model of *N. crassa OR74A* in terms of number of pathways. On the other hand, the model of *P. chrysosporium RP78* has 227 pathways compared to the approximately 300 pathways of the other models.

So the evidence shows a clear distinction between automatic annotation for a moderate quality assembly and manual annotation for a high quality assembly. However, the evidence does not show a difference between manual curation alone, as in *N. crassa OR74A*, and manual curation plus limited automatic annotation in the *Aspergillus* genomes.

Automated GENREs are incomplete

Table 18 shows that the number of holes in pathways accounts for over 30% of all reactions. For the worst assembly and annotation of a genome, namely *P. chrysosporium RP78*, the percentage reaches 37%. Table 19 for *P. chrysosporium RP78* shows that only 102 or 45% of pathways in the model have no holes at all. Therefore, the PGDB has incomplete GPR associations.

Automated GENREs after hole-filling are incomplete

The recommended threshold in Pathway Tools for accepting a putative GPR for a hole is 90%. At this level, Figure 19 shows that 45% of the holes in the model for *P. chrysosporium RP78* are filled. However, this leaves 182 holes, which is 20% of all reactions. Therefore, even after hole-filling, the PGDB has incomplete GPR associations.

Transporters are very incomplete

Table 18 shows that Pathway Tools identifies 10–13 transport reactions, and associates 37–44 genes with the transport reactions for the fungal genomes. This is from a repertoire of 154 transport reactions and 101 transporter proteins in MetaCyc (Table 17). In Chapter 4 we see that there are 205 transport reactions for *A. niger CBS 513.88* in a manual GENRE [ANN08], and that a fungal genome has about 500 transporters predicted. Therefore the information in a PGDB about transport is very incomplete.

3.4 Conclusion

This chapter presents the experiments and evaluation of the metabolic pathways reconstruction of six fungal genomes using Pathway Tools. MetaCyc is a well-established curated reference template for GENREs; however, MetaCyc is by no means complete. Furthermore, most of the information is for prokaryotes. In order to be able to move to the next steps in GENRE and perform flux balance analysis, the metabolic pathway model needs to be connected and to include the core metabolism of the organism.

Our evaluation relied on internal validation of the model through counts of entities, in particular, using the number of pathways to gauge the extent of the model, and the number of holes before and after hole-filling to measure the completeness of the model. We had no ground truth, nor access to wet lab validation, in order to perform external validation. This is true for known pathways, and more so for *de novo* pathways.

The Transport Inference Parser (TIP) [LPK08] is very limited in the prediction of transporters. Furthermore, Pathway Tools models only the extracellular space, the periplasmic space, the cytosol, and the mitochondrion.

The heavy dependence on genome annotation for GENRE, in our experience, only had an impact for automated annotations for which there was no review, as in the case of *P. chrysosporium*. The PGDBs for the other genomes were roughly equivalent. It did not matter whether there were only manual annotations, as in *N. crassa*, or those manual annotations were augmented by additional trusted annotations from orthologs, as in the *Aspergillus* genomes. Presumably they each covered the core known metabolism in each of them, as this would be the first step in manual annotation.

The issues identified for eukaryotes in particular are the need to model a cell's internal organelles, predict localization of proteins, and predict transport proteins with their specific substrate and membrane localization. In summary:

- The reference template approaches are dependent on the body of existing knowledge, and the effort to manually curate the scientific literature to extract that knowledge and encode it in public databases.

- The evaluation of methods is difficult when applied to new genomes. Internal validation of the model can be measured in terms of numbers of pathways, reactions, and GPR associations to indicate coverage, and by the number of holes to indicate completeness.

Further internal validation requires constructing a systems biology model so one can apply flux balance analysis for atoms, charges, energy, etc. External validation requires the scientist to make predictions from the model and then to validate those predictions in the wet lab; this is not expertise available usually to the developer of algorithms.

- The validation of methods for *de novo* discovery of pathways is difficult, even for model organisms. Internal validation shows that the pathways are sound in terms of the chemical transformation of compounds, but external validation of the existence of the pathway in the organism requires extensive wet lab work.

- Even with gap filling, there are typically many holes in the resulting reconstruction. Most approaches to gap-filling do not make use of gene expression data, which today can be readily available even for non-model organisms through RNA-Seq.

- The widely available and widely used tools are biased towards prokaryotes. In particular, they do not model cell compartments such as mitochondrion, Golgi, peroxisome, ER, vacuole, or lysosome in their reconstructions.

- Transport reactions are often an afterthought in the modeling of the cell, despite the fact that the reconstruction needs to view the cell as a closed system importing and exporting compounds to its surroundings in order to perform internal validation.

Chapter 4

Prediction of Transport Proteins

This chapter investigates how to include transport reactions, transporter proteins, and the GPR associations for transport in the reconstruction of metabolic pathways. For prokaryotes, it is sufficient to model the transport across the cell membrane. However, eukaryotes have internal organelles, therefore the reconstruction requires modeling of the cell internal components and the intracellular transport across their membranes. The transport reaction should represent the transport of one or more specific substrates across a specific membrane. The GPR association should identify the transmembrane protein that performs the movement of those substrates across that membrane.

The official home of transporters is the Transporter Classification (TC) scheme and its associated collection of transporters, the Transporter Classification Database (TCDB). Some predictors of transport proteins target the TC as the goal of the predictor. However, the TCDB does not explicitly identify a transport reaction, the specific substrate, or the membrane for each of its entries. Therefore, other predictors target the prediction of substrates directly. However, they are able to predict the type of substrate being transported, but not the specific substrate. Unfortunately, the actual problem addressed by each predictor of transport proteins is so diverse that meaningful comparison of their performance is impossible. We develop a scheme to describe and compare the existing work, and carry out a case study on a fungal genome to get a deeper understanding of the existing work, and to compare them in a practical setting.

The most useful approach seemed a direct application of sequence similarity as used in the protocol of Milton Saier's lab. So we automate the protocol and include localization to

identify which organelle membrane is involved in the transport reaction.

The prediction of which specific substrate is transported by the transport protein is beyond the current state of the art. We explore various approaches that offer potential solutions, but we do not solve the problem. The lack of characterized examples is a major factor in our failure; there are often sufficient examples within a type of substrate to effectively train a predictor, while for each specific substrate the number of examples is insufficient for this task.

In order to make effective use of available examples, and to prepare for the day when sufficient examples are available for specific substrates, we propose a framework for the transport prediction problem that draws on our experience. This is a proposal, not a worked solution, that relates the exsiting sources of knowledge and identifies two key aspects: first, that the problem is hierarchical in nature, corresponding to the subset grouping of the substrates based on their chemical and physical properties; and second, that it is a multi-label machine learning problem.

The chapter is organized as follows: Section 4.1 presents the scheme for describing and comparing existing methods; Section 4.2 presents the state of the art; Section 4.3 presents the case study of the existing methods when applied to a fungal genome; Section 4.4 presents the TransATH system which automates Saier's protocol and demonstrates TransATH on the fungal genome of the case study; Section 4.5 presents an evaluation of the thresholds to use for blastp and the correctness of TransATH; Section 4.6 explores approaches to predicting specific substrates given a transport protein; Section 4.7 proposes a framework for the transport prediction problem; and Section 4.8 presents the lessons learned.

4.1 A Scheme to Compare Transport Predictors

The existing work on predicting transporters is quite diverse, and lacks any clear comparisons between the different schools of work. Therefore, to make the similarities and differences between the approaches clear, we required a scheme for structuring the descriptions. Table 23 presents an overview of the work on the transport protein prediction problem using the scheme.

For the purposes of GENRE and assigning a gene to a transport reaction, the prediction must target the specific substrate(s) transported, and the membrane across which the transport

takes place. Predicting the specific substrate is difficult because the specificity depends on a small number of residues at specific sites in the protein, and the number of characterized transporters is small.

Existing work on the prediction of whether a protein is a transporter adopts one of two approaches, either

TC: classifying the protein according to the Transporter Classification (TC) of the International Union of Biochemistry and Molecular Biology (IUBMB), or

Substrate: classifying according to the type of substrate transported: amino acid, anion, cation, electron, protein/mRNA, sugar, and other.

There are many interpretations in the literature of the prediction problem for transporters. This makes comparison of the existing approaches difficult to compare. In describing the work on the problem of transporter prediction we introduce the following dimensions with their values:

Scale: **P** (protein), **G** (genome);

Classifier: **B** (binary), **MC** (multi-class), **ML** (multi-label);

Target: **Transporter**, **TC-Superfamily**, **TC-Family**, **TC-Subfamily**, **TC-ID**, **SubstrateType**, **Substrate**;

Scope: **All**, **B** (Bacteria/Prokaryote), **F** (Fungi), **P** (Plant), **H** (Human);

Localization: **NoLoc**, **Loc**;

Two important variations are whether the problem is to classify a particular protein (**P**), or to classify all proteins in an organism's genome (**G**). More importantly, to take advantage of the fact that all transporters in the genome are the goal, and use techniques such as gap-filling. No existing predictor works at the genome scale.

A second distinction is the type of classifier, be it a binary classifier (**B**), multi-label classifier (**ML**), or multi-class classifier (**MC**). For example, on sugars, for a protein p, (**B**) does p transport the substrate *glucose*? (**MC**) which single substrate in the set {*glucose*, *maltose*, *xylose*} does p transport? and (**ML**) which subset of substrates in the set {*glucose*, *maltose*, *xylose*} does p transport?

The basic classification task is also interpreted depending on whether the target of the prediction is to classify transporters, their TC family, or their substrate. We identify the following specific prediction tasks and their targets:

Transporter: Given a protein p, is p a transport protein?

TC-Superfamily: Given a protein p, and a superfamily X, is p a transport protein in X?

TC-Family: Given a protein p, and a family X, is p a transport protein in X?

TC-Subfamily: Given a protein p, and a subfamily X, is p a transport protein in X?

TC-ID: Given a protein p, and a TCDB protein with identifier X, is p a transport protein with X as its nearest neighbour in TCDB?

SubstrateType: Given a protein p, and a category of substrates S, does p transport a substrate in S?

Substrate: Given a protein p, and a substrate s, does p transport the substrate s?

The scope of the classifier is also important. Most approaches present themselves as generic, that is, covering all kingdoms of life, even though they are trained, evaluated and tested on a few specific organisms, or the TCDB which is biased towards the model organisms.

Finally, the issue of predicting the localization of transport is important in eukaryote cells. Most existing approaches treat this as a separate problem.

4.2 The State of the Art

For most of the work done on the prediction of transport proteins [GO14], there is no available software, so it is difficult to reproduce the work and to compare the results of different articles. The two schools of predicting substrate category or TC family further complicate any comparisons. A summary of the work using our dimensions of the transport protein prediction problem is given in Table 23.

Research on prediction of transporters has three main sources of gold standard datasets:

1. the model organism databases for *E. coli*, *S. cerevisiae*, and *A. thaliana*;
2. the UniProt/SwissProt database of reviewed protein annotations that includes the data from (1); and

3. the Transporter Classification Database (TCDB) [SLBG].

The number of experimentally characterized transport proteins is quite small. So one is either restricted to small datasets and a restricted range of target classes for prediction, or one includes proteins with electronic annotations.

In the TCDB, there is great imbalance between the size of families, which impacts the evaluation of the predictors, or restricts the range of target classes. Of the 835 families, 137 have only a single member, and 734 have size from 1 to 20; therefore there are 101 families of size greater than 20. The largest families are 3.A.1, The ATP-binding cassette (ABC) superfamily, of size 1569, and 2.A.1, The major facilitator superfamily (MFS), of size 720. Further details are given in Table 20.

Size	Number
1	137
2–20	597
21–50	77
51–100	12
101–200	6
201–300	4
700+	2

Table 20: Number of TC Families of Given Sizes

The table shows the number of TC families of size within the specified range. The number 20 is taken as an indication that the family is large enough to support the training of a predictor using machine learning. As of May 2014.

4.2.1 TransAAP

The TransAAP [RKP04] is a semi-automated analysis pipeline to input data into TransportDB. TransAAP targets only prokaryotes. A new genome is matched against the curated set of TransportDB proteins with assigned family using BLAST with e-value cut-off of 1e-3. Information from these BLAST searches against TransportDB are collected, as is information from searches against non-transporters in the nr protein database, and classification by COG. A web-based interface displays the information to help a human annotator decide and assign possible substrates or functions.

4.2.2 Transport Inference Parser

Pathway Tools includes the Transport Inference Parser (TIP) [LPK08] which analyses keywords in a gene annotation to assign GPR associations to transport reactions in MetaCyc.

4.2.3 Saier Lab

G-Blast [RS12] screens proteins against all entries in TCDB using BLAST to retrieve the top hit, and HMMTOP to determine information about TMS for the query and the hit sequence. It is an integral part of a manual protocol to predict the transport proteins for a genome [PVL+14] developed by Saier's lab.

> [G-Blast] Run blastp against TCDB with e-value 1e-3 and no low complexity filter.
> [G-Blast] Run HMMTOP to determine TMS.
> [TMS check] Use WHAT [ZSJ01] with window size of 19 and angle of 100 degrees to create hydropathy plot.
> [TMS check (Manual)] Check plot and TMS prediction.
> [TMS?] Reject any protein with zero TMS in target or query.
>
> [G-Blast] Run blastp with e-value 0.1.
> [Putative transporters (Manual)] A new hit (query) may be member of new transporter family.
>
> [Beta-barrel proteins] Run BOMP (Beta-barrel integral Outer Membrane Proteins) program (http://services.cbu.uib.no/tools/bomp/handleForm).
> [Manual review] Hits may be putative transporters.

Figure 20: Protocol of Saier Lab

On the basis of sequence similarity, and on the basis of the number and location of TMS, with entries of known function in TCDB, the transport proteins are classified into families and subfamilies which often allows the *"prediction of substrate type with confidence."* [PVL+14].

4.2.4 Zhao Lab

The Zhao Lab has developed three methods: a nearest neighbour approach [LDZ08]; TransportTP [LBUZ09]; and TrSSP [MCZ14]. The nearest neighbour approach achieved a balanced accuracy of 67.0%.

TransportTP [LBUZ09] is a two-phase algorithm that combines homology and machine learning to predict TC family of one or more proteins. For training and cross-validation testing,

TransportTP used the yeast proteome. For testing, it used 10 genomes from the TransportDB database [RCP07] of annotated prokaryote transporters. As an independent test, TransportTP is trained on the proteome of the plant *A. thaliana* and then used to predict the transporters in 4 other plant proteomes.

The overall process consists of a pre-processor phase, a phase to construct an initial classifier, and a phase to refine the classifier.

The preprocessing phase uses (1) the TCDB database of transporters classified into TC superfamilies and TC families; (2) the Pfam database of protein domains; and (3) the Gene Ontology subgraph rooted at term `GO:0022857 transmembrane transporter` and the associated sequences. The preprocessing phase constructs (A) a HMM for each TC superfamily and for each TC family that had sufficient members using the SAM program; and (B) a mapping of Pfam domains to TC families or superfamilies using an all-vs-all HMM search of Pfam against the TCDB.

The initial classifier integrated the results of BLAST search and HMM search. The BLAST search is performed against the TCDB, while the HMM search is performed against the collection (A) of HMMs from the preprocessing phase.

The final classifier is constructed as an ensemble of balanced SVMs from a large feature space of the transporters identified by the initial classifier. The intent of the ensemble is to refine the classification and remove false positives. The feature space has seven parts, each derived separately:

1. The first category of features is the e-values of the protein against each entry in the TCDB generated during the BLAST search and the HMM search during the initial classification phase;

2. The second category are binary features (whether or not the classification falls into channels, carriers, or primary active transporter) and the sizes of the initially classified families;

3. The third category is the number of transmembrane segments for the protein and for the TC families;

4. The fourth category is the consistency of TC family amongst the top k-homologs from the initial search;

5. The fifth category is the occurrence of Pfam domains in the protein from those domains that map to TC families or superfamilies;

6. The sixth category is the occurrence of a GO term by BLAST search of the protein against the associated sequences; and

7. The seventh category is an indication of non-transport function as measured by keywords associated with the top BLAST neighbours in SwissProt.

TransportTP achieved a balanced accuracy of 81.8%. They compared TransportTP with their earlier work [LDZ08] using nearest neighbours with balanced accuracy of 67.0%, the initial classifier, and the individual components BLAST search and HMM search of the initial classifier. Compared with the SVM-Prot classifier [LHC+06], on the five TC superfamilies and three families used by SVM-Prot, TransportTP achieved better performance in recall and precision: SVM-Prot achieved an average recall of 81.0% and an average precision of 26.1%.

The Transporter Substrate Specificity Prediction Server (TrSSP) [MCZ14] is a web server to predict membrane transport proteins and their substrate category. The substrate categories are: (1) oligopeptides (amino acid); (2) anion; (3) cation; (4) electron; (5) protein/mRNA; (6) sugar; and (7) other. TrSSP makes a top-level prediction of whether the protein is a transporter, or not. A SVM is applied with highest accuracy being reported using amino acid index (AAindex) and Position-Specific Scoring Matrix (PSSM).

4.2.5 Gromiha Lab

Gromiha and Yabuki [GY08] reported that a k-nearest neighbour method using the amino acid composition could discriminate non-transporters and transporters with accuracy about 80%. The use of PSSM profiles and 49 amino acid physicochemical properties showed an increase of 5–10% in discrimination accuracy [OCG10].

Gromiha and Yabuki [GY08] used amino acid composition for discriminating channels/pores, electrochemical and active transporters, with an accuracy of 64%. Again, using PSSM profiles and amino acid properties, they obtained an average accuracy of 78% [OCG10].

Ou et al. [OCG10] also considered six major families in TCDB. Their method based on PSSM profiles and amino acid properties showed an average accuracy of 69%, with an improvement of 8% over amino acid composition.

Chen et al. [COLG11] considered four major classes of substrates: (i) electron, (ii) protein/mRNA, (iii) ion and (iv) others. They analyzed the characteristic features of transporters associated with these targets using amino acid properties. They used various features, amino acid composition, residue pair preference, amino acid properties and PSSM profiles and developed an algorithm based on radial basis function (RBF) networks to discriminate transporters with different substrates with an AUC of 0.90, 0.86, 0.77 and 0.86, respectively.

4.2.6 Helms Lab

Schaadt et al. [SCH10] used amino acid composition, characteristics of amino acid residues and conservation to detect transporters based on different substrates, amino acids, oligopeptides, phosphates and hexoses and showed an accuracy of 75% to 90%. They classified to four substrate categories: amino acid, oligopeptide, phosphate, and hexose. The number of characterized transporters in *A. thaliana* for the four substrates numbered from 13 to 17. They constructed a vector for each protein using various types of amino acid composition, AAC, PAAC, PseAAC, PsePAAC, MSA-AAC, and used Euclidean distance from the query protein's vector to the known vectors to rank the substrate categories. They found that AAC did not yield accurate results. However, PAAC performed as well as the more complicated PsePAAC and MSA-AAC, yielding accuracy over 90%.

Schaadt and Helms [SH12] compared the similarity of transporters in TCDB and annotated transporters in *A. thaliana* using amino acid composition and classified the proteins into three families. By distinguishing the amino acid composition of TMS and non-TMS regions, they could classify four different families with an accuracy of 80%.

Barghash and Helms comparison [BH13] performed a comparison of three different approaches (homology, HMMER, MEME) for predicting substrate category and predicting TC family. They used four substrate categories, metal ions, phosphate, sugar, and amino acid; and 29 TC families, with the most numerous examples. The datasets are from *E. coli*, *S. cerevisiae*, and *A. thaliana*, consisting of the 155, 158, 177, respectively, proteins that had both a substrate annotation and TC family annotation that are experimentally determined.

We summarize the best and worst of their results in Table 21 and Table 22. There are many proteins that are unclassified by their predictors, the overall prediction of TC family is better

	Homology		HMM		MEME	
	Best	Worst	Best	Worst	Best	Worst
P	97.5	54.1	97.5	73.3	100	9.6
R	97.5	62.9	97.5	73.3	100	36.6
F	97.5	55.2	97.5	73.3	100	13.0
U	35.0	0.0	2.5	76.7	28.3	0.0

Table 21: Results Predicting TC Family

The table compares the results of BLAST, HMMER, and MAST for predicting TC family [BH13]. It presents the best and the worst results for each method as determined by F-measure. Abbreviations: P for precision; R for Recall, F for F-measure; and U for Unclassified. All results are given as percentages.

	Homology		HMM		MEME	
	Best	Worst	Best	Worst	Best	Worst
P	95.5	34.9	99.3	51.4	82.9	25.0
R	100	51.5	96.2	51.4	96.7	31.7
F	97.2	35.7	97.2	51.4	87.7	27.3
U	45.7	1.4	45.7	93.1	68.7	0.0

Table 22: Results Predicting Substrate Category

The table compares the results of BLAST, HMMER, and MAST for predicting substrate category [BH13]. It presents the best and the worst results for each method as determined by F-measure. Abbreviations: P for precision; R for Recall, F for F-measure; and U for Unclassified. All results are given as percentages.

than that of substrate category, and homology performs as well if not better than the other two approaches.

Work	Scale	Classifier	Target	Scope	Localization	Dataset	Software
TransAAP [RKP04]	P P	B MC	Transporter TC-Family	B B	NoLoc NoLoc	TransportDB	Web
G-Blast [RS12]	P P	B MC	Transporter TC-ID	All All	NoLoc NoLoc	TCDB	Yes
TransportTP [LBUZ09]	P P	B MC	Transporter TC-Family	All All	NoLoc NoLoc	TransportDB *S. cerevisiae* *A. thaliana*	Web
TrSSP [MCZ14]	P P	B ML	Transporter SubstrateType(7)	All All	NoLoc NoLoc	SwissProt	Web

Table 23: Existing Work on Predicting Transport Proteins

4.3 Case Study

In 2008 MR Andersen [ANN08] published a comprehensive gapless metabolic model of the CBS 513.88 strain of the fungus *Aspergillus niger* widely used for the production of chemicals.

The model was based on extensive review of the literature and comparisons with models of closely related species and strains. The model was gapless so there are no missing reactions and the network is connected.

The genome has 14,156 ORFs. The GENRE contained 1190 unique reactions and identified GPR associations for 871 ORFs. The modeled cellular compartments are extracellular space, cytosol, and mitochondrion. The metabolic reactions numbered 986 of which 131 have no assigned GPR. There were 205 transport reactions of which only 3 have assigned GPR, so 1.46% of transport reaction have assigned GPR, compared to 96.86% for metabolic reactions.

The transport across the cell membrane from extracellular space to cytosol covers 151 transport reactions, while transport across the mitochondrion membrane covered 54 transport reactions. The metabolites transported included nucleotides, amino acids, alcohols, acids, fatty acids, phosphate, urea, aldehydes, sugars, and others (CO2, H2O, O2, H2O2, etc). Of particular interest to us are the sugars. There are 21 sugars in total: disaccharides (trehalose, lactose, maltose), and monosaccharides categorized by the number of carbon atoms as tetrose, pentose (arabinose, ribose, ribulose, xylose, xylulose), and hexose (glucose, galactose, mannose, iditol, sorbose, rhamnose). There are separate transport reactions for the two forms D- and L- of arabinose and xylulose; and separate transport reactions for the open chain and ring forms of glucose: D−glucose, α−D−glucose and β−D−glucose.

Note that for *S. cerevisiae*, the most studied fungi, there are 66 transporters, of which 15 are sugar transporters. Of these 15 there are 5 that are experimentally characterized as sugar transporters, and 3 of the characterized sugar transporters are for a specific sugar, glucose.

4.3.1 A Pathway Tools Reconstruction

We constructed a GENRE for *A. niger* CBS 513.88 with Pathway Tools and the AspGD annotation. Pathway Tools includes the Transport Inference Parser (TIP) [LPK08] which analyses keywords in a gene annotation to assign GPR associations to transport reactions in MetaCyc. The model contained 332 pathways, 1868 metabolic reactions, and 10 transport reactions. There were 1580 ORFs assigned to metabolic reactions, and 41 ORFs assigned to transport reactions. There were 335 holes (31%) in the model.

Pathway Tools models only the extracellular space `[out]` and the cytosol `[in]`. Figure 21 shows the 10 transport reactions of the *A. niger* CBS 513.88 Pathway Tools GENRE.

(1) $NADP^+ + NADH + H^+_{[out]} \longleftrightarrow NAD^+ + NADPH + H^+_{[in]}$
(2) $UDP{-}\alpha{-}D{-}glucose + glucosyl{-}glycogenin_{[in]} \longrightarrow$
$1,4{-}\alpha{-}D{-}glucosylglycogenin + UDP + H^+_{[out]}$
(3) $phospolipid_{[in]} + ATP + H_2O \longleftrightarrow phospolipid_{[out]} + ADP + phosphate + H^+$
(4) $Cu^{2+}_{[in]} + ATP + H_2O \longrightarrow Cu^{2+}_{[out]} + ADP + phosphate + H^+$
(5) $ATP + H^+_{[in]} + H_2O \longleftrightarrow ADP + phosphate + H^+_{[out]}$
(6) $Ca^{2+}_{[out]} + ATP + H_2O \longleftrightarrow Ca^{2+}_{[in]} + ADP + phosphate + H^+$
(7) $ATP + 3H^+_{[in]} + H_2O \longleftrightarrow ADP + phosphate + 4H^+_{[out]}$
(8) $oligopeptide_{[out]} + ATP + H_2O \longleftrightarrow oligopeptide_{[in]} + ADP + phosphate$
(9) $xenobiotic_{[in]} + ATP + H_2O \longleftrightarrow xenobiotic_{[out]} + ADP + phosphate$
(10) $4H^+_{[in]} \longrightarrow 4H^+_{[out]}$

Figure 21: Transport Reactions Predicted by Transport Inference Parser

We investigated the application of existing methods for predicting transporters to our case study; in particular, for the transport of sugar. The results were poor and not in agreement with each other. This contradicts the good results by the authors of the existing work as reported in Section 4.2. Indeed, our approach using homology is competitive with the existing approaches.

We also report results for two important substeps: the predicting of transmembrane segments, and the localization of transporters.

4.3.2 TCDB-Blast— Our G-Blast(v2) Implementation

We modified the G-Blast version 2 implementation of Saier's lab to do more than simply take the top BLAST hit, and calculate the number of TMS using HMMTOP. The details are in Section 4.4. The results here refer to TCDB-Blast, the modified G-Blast(v2) which collects all hits passing a set of thresholds: e-value 1e-20; percent alignment 70%; query coverage 70%; subject coverage 70%; and difference in length of 10%. The standard thresholds for BLAST alignments for the purpose of functional annotation of proteins in general [HPCW11] use percent identity of 70% rather than percent alignment; however, for transmembrane proteins there is less conservation of identity during evolution. After running HMMTOP, we rejected sequences without a TMS.

4.3.3 Sanity Check of Prediction on TCDB

The TCDB dataset as of May 2014 has 11572 transporter sequences. UniProt has 11589 protein sequences tagged with TC-IDs. Out of 11589, 5321 are reviewed sequences (SwissProt), while 6268 are unreviewed. There are discrepancies due to update and synchronization between these two databases.

We ran each predictor against the TCDB. The results are in Table 24. Not surprisingly, the direct homology approach using TCDB-Blast performed best. What is surprising is how many transporters were predicted to be non-transporters by TransportTP and TrSSP. This reinforces the evidence of poor coverage of prediction techniques from Barghash and Helms [BH13].

Predictor	TCDB	Transporter		Non-Transporter	
		No.	Pct	No.	Pct
TCDB-Blast	11572	11218	96.9	354	3.1
TransportTP	11572	5517	47.7	6055	52.3
TrSSP	11572	7528	65.0	4044	35.0

Table 24: Predictions on TCDB

4.3.4 *A niger* CBS 513.88

Each of the systems is run on the genome of *A niger* CBS 513.88. We determined the total number of proteins predicted as transporters (see Table 25) and focused in on those predicted as sugar transporters (see Table 26) either as members of TC family 2.A.1.1 or as transporters of the substrate category `sugar`. Note that the number of transmembrane proteins is 5702 based on those with at least one TMS as determined by HMMTOP, and the number of possible sugar porters is 461, based on those with between 10 and 12 TMS as determined by HMMTOP.

ORFs	MRA	TIP	TB	TP	TR
14067	3	41	565	673	3582

Table 25: Predicted Transporters in the Case Study

The number of ORFs in *A. niger* CBS 513.88 predicted to be transporters by different approaches: MRA, manually by MR Andersen [ANN08]; TIP, Pathway Tools Transport Inference Parser; TB, TCDB-Blast; TP, TransportTP; TR, TrSSP.

ORFs	Motifs		Predictors		
	ST1	ST2	TB	TP	TR
14067	74	65	62	23	482

Table 26: Predicted Sugar Transporters in the Case Study

The number of ORFs in *A. niger* CBS 513.88 predicted to be sugar transporters by different approaches: ST1, Prosite PS00216 Sugar_Transport_1 motif; ST2, Prosite PS00217 Sugar_Transport_2 motif; TB, TCDB-Blast; TP, TransportTP; TR, TrSSP. Note that [ANN08] had 21 unique sugar transport reactions.

4.3.4.1 Topology

We compared the results of two common predictors HMMTOP v2.1 and TMHMM v2.0 of transmembrane helices on two subsets of the TCDB, namely the MFS Superfamily (2.A.1), and the Sugar Porters (Family 2.A.1.1). Table 27 shows the results. As sugar transporters in TCDB all have 12 TMS, HMMTOP is clearly better, confirming the overall best rating for HMMTOP for predicting topology of membrane proteins in a broader comparison of systems [RCL+14].

Helices	8	9	10	11	12
HMMTOP v2.1			3	5	111
TMHMM v2.0	2	3	18	17	79

Table 27: Comparison of HMMTOP and TMHMM on Sugar Porters

4.3.4.2 Localization

For localization of the 62 sugar transporters in *A. niger* CBS 513.88 as predicted by TCDB-Blast, LocTree3 placed 48 in the plasma membrane, 10 in the mitochondrion membrane, and 4 in the vacuole membrane.

4.3.4.3 Sugar Transporters

We compared the TrSSP predictions of the 62 sugar transporters predicted by TCDB-Blast in Table 28. Almost always, TrSSP predicted at least one other substrate category in addition to sugar, generally amino acid and/or anion. In 13 cases (21%), TrSSP did not predict the substrate to be sugar.

TCDB-Blast Prediction				*A. niger*	TrSSP Prediction							
SubFamily TC #	**SubFamily Name**	**TCID**	**Hits**	**SequenceID**	**AA**	**An**	**Ca**	**El**	**Pr/mR**	**Su**	**Ot**	**Uk**
2.A.1.1	Sugar Porter (SP)	2.A.1.1.7	P11636	An01g00820	X					X	X	
				An01g10970	X					X	X	
				An07g06300	X	X				X	X	
				An08g03850	X					X	X	
				An12g01560						X	X	
				An14g04280		X				X	X	
				An14g06890		X				X		
				An15g04270		X				X	X	
				An16g06580	X					X	X	
				An18g01700	X	X				X	X	
		2.A.1.1.10	P15685	An02g02810	X					X	X	
		2.A.1.1.11	P53048	An08g08000	X					X	X	
		2.A.1.1.33	Q8NJ22	An06g02270	X					X	X	
		2.A.1.1.38	P39932	An01g08780	X					X	X	
				An02g00060	X						X	
				An04g08030	X						X	
				An05g02010		X				X	X	
				An07g01260						X	X	
				An18g00040		X				X	X	
				An18g00440			X				X	
		2.A.1.1.39	P49374	An02g00590	X	X				X	X	
				An02g07850	X					X	X	
				An03g01620	X	X				X	X	
				An07g10370							X	
				An08g04040	X					X		
				An11g01100	X	X				X	X	
		2.A.1.1.40	Q64L87	An01g00850	X	X				X	X	
				An04g10090	X	X				X		
				An06g00560	X	X				X	X	
				An07g01310	X						X	
				An11g05280	X	X				X	X	
				An12g05820		X					X	
				An16g06610	X					X	X	
				An18g01760	X	X				X	X	
		2.A.1.1.51	Q2MEV7	An15g00310		X					X	
		2.A.1.1.57	Q8J0V1	An12g07450	X	X				X	X	
		2.A.1.1.58	Q8J0U9	An02g03540	X					X	X	
				An03g02190						X	X	
				An05g01290	X					X	X	
		2.A.1.1.68	A3M0N3	An03g01750	X						X	
				An11g00120	X					X	X	
				An15g03940	X	X				X	X	
		2.A.1.1.69	A1Z264	An14g02700			X				X	
		2.A.1.1.70	Q0ULF7	An15g01500	X	X				X	X	
		2.A.1.1.73	Q5A8J5	An01g14620		X				X		
				An02g11260	X	X				X	X	
				An05g02510	X					X	X	
				An07g06880	X					X	X	
				An09g02930	X					X	X	
				An14g02740	X	X				X		
				An14g03990	X					X	X	
		2.A.1.1.82	Q7SCU1	An12g09270	X	X				X	X	
		2.A.1.1.83	Q7SD12	An03g05320	X						X	
				An04g02790	X					X	X	
				An08g09350	X						X	
				An09g04810	X	X				X	X	
				An13g03250	X						X	
				An14g01600	X					X	X	
				An16g06220	X						X	
		2.A.1.1.96	P38142	An09g04680		X				X		
		2.A.1.1.110	P39924	An08g08520		X				X	X	
		2.A.1.1.117	G4N740	An06g02030		X				X	X	

Table 28: TCDB-Blast Results for Sugar Porters with their TrSSP Substrates Prediction

TrSSP prediction of substrate category for the TCDB-Blast predicted sugar transporters in *A. niger* CBS 513.88. Abbreviations: Ca: Cation, An: Anion, Su: Sugar, El: Electron, Pr/mR: Protein/mRNA, Ot: Other, Uk: Unknown

4.3.5 Transport Prediction on Fungal Genomes

The transport predictors are applied to a number of fungal genomes. Table 29 summarizes the number of predictions by fungal genome and by predictor. Further details of the results for TransportTP are shown in Table 50 in Appendix B; and for TCDB-Blast in Table 52 in Appendix C. In Appendix D in Table 53 is a comparison of the TrSSP results with TCDB-Blast results for the channel/pores transporters. Appendix D also contains Table 54 which shows the usual localization of predicted sugar porters and highlights the unusual predicted localizations by LocTree3 for the fungal genomes.

Genomes	**TCDB-Blast**	**TransportTP**	**TrSSP**
A. fumigatus Af293	448	528	2892
A. nidulans FGSC A4	503	605	3220
A. niger CBS513.88	565	673	3582
A. niger NRRL3	649	701	3758
A. oryzae RIB40	622	784	3754
N. crassa OR74A	311	348	2657
P. chrysosporium RP78	231	338	2692
S. pombe	243	222	1519

Table 29: Summary of Results by TCDB-Blast, TransportTP and TRSSP
Number of transporters predicted by the tools for the eight fungal genomes. TCDB-Blast and TransportTP use a threshold e-value 1e-20. There is no threshold for TrSSP prediction results.

4.3.6 Discussion

From the state of the art, it is clear that neither software nor web services are available for most of the approaches in the literature. Furthermore, the tools are not directly comparable becuase they are solving diverse problems, so one is left with "comparing apples and oranges". In the case study, we evaluate the available tools in the same setting.

Coverage of transporters is poor

The sanity check as presented in Table 24 reveals that both TransportTP and TrSSP recognize less than 65% of the entries in the TCDB as transporters.

Coverage is 97% for TCDB-Blast, which uses sequence similarity against TCDB as its prediction. However, the argument is circular, as we use the TCDB as the benchmark for the sanity check.

Sequence similarity works best

Table 24 and the other tables in Section 4.3 show prediction by sequence similarity to work well, and generally better than other methods. This confirms the results of the Barghash and Helms comparison [BH13] above.

Overprediction by TrSSP

TrSSP [MCZ14] is the most recent work from Zhao's lab. It is their first effort to predict substrate rather than TC family, and it is their first multi-label predictor. Table 25 shows that TrSSP predicts 3583 transporters for *A. niger* CBS 513.88, while TCDB-Blast and TransportTP predict 565 and 673 respectively. These latter numbers are more in line with the consensus that filamentous fungi have some 500 to 800 transporters. Similarly, in Table 26 TrSSP predicts 482 sugar transporters compared to 62 and 23 by TCDB-Blast and TransportTP, respectively, and compared to 74 and 65 by the Prosite motifs *Sugar_Transport_1* and *Sugar_Transport_2* respectively.

On closer inspection in Table 28, TrSSP predicts 3–4 substrates for each sugar transporter identified by TCDB-Blast.

The numbers suggest strongly that TrSSP is overpredicting, maybe by a factor of 4 to 8 times.

Topology prediction does not identify TCID

As transporters are transmembrane proteins, one direct approach to identifying them is to use the number of TMS as an identifying attribute. However, strict reliance on the equality of the number of TMS would miss many cases due to the errors in the prediction of topology, as highlighted in Table 27.

4.4 Automation of Manual Protocol of Saier

This section presents an implementation that automates the protocol for predicting the transporters in a genome used by Saier's lab. The reason for this choice are multifold: the Barghash and Helms comparison [BH13] shows that homology works as well as other approaches in predicting transporters; Milton Saier and the TCDB are the authority on transporters; Saier's lab uses homology; and Saier's lab applies their approach to whole genomes. The protocol used by Saier's lab is as we discerned it to be from their publications.

Our system is named TransATH, which stands for Transporters via ATH (Annotation Transfer by Homology).

4.4.1 The Protocol

Saier's lab has analysed the genomes of several organisms for their complement of transporters [YS12, GNY+13, PVL+14]. Figure 22 shows the protocol that we obtained from the Materials and Methods sections of their papers [YS12, GNY+13, PVL+14].

> [G-Blast] Run blastp against TCDB with e-value 1e-3 and no low complexity filter.
> [G-Blast] Run HMMTOP to determine TMS.
> [TMS check] Use WHAT [ZSJ01] with window size of 19 and angle of 100 degrees to create hydropathy plot.
> [TMS check (Manual)] Check plot and TMS prediction.
> [TMS?] Reject any protein with zero TMS in target or query.
>
> [G-Blast] Run blastp with e-value 0.1.
> [Putative transporters (Manual)] A new hit (query) may be member of new transporter family.
>
> [Beta-barrel proteins] Run BOMP (Beta-barrel integral Outer Membrane Proteins) program (http://services.cbu.uib.no/tools/bomp/handleForm).
> [Manual review] Hits may be putative transporters.

Figure 22: Protocol of Saier Lab

Algorithm 1 shows G-Blast(v2). This is an algorithmic formalization of Saier's protocol in Figure 22. For clarity we make explicit the use of the Blast+ package for BLAST from `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download`.

Algorithm 1 G-Blast(v2)

Require: a genome G as .fasta file of protein sequences
Require: the TCDB as a Blast+ protein sequence database with TCID as identifiers
Require: a mapping $TC2TMS$ from the TCDB to the number of TMS of the entry
Ensure: result is list$<gid, tcid>$ of matches of proteins gid in G with transporters $tcid$

```
function G-Blast(v2)(G, TCDB)
    list<gid, tcid, _, _, _, _, _> := Blast+:blastp(G, TCDB, e-3)
    return list<gid, tcid> where
        ( TC2TMS(tcid) ≠ 0) ∧ (computeTMS(gid) ≠ 0)
Comment We omit searching for putative transporters
Comment We omit searching for beta-barrel transporters
end function
```

Algorithm 2 presents the TransATH algorithm for the implementation of the protocol of Saier's lab for determining the transporters in a given genome. TransATH stands for Transporters via ATH (Annotation Transfer by Homology). Note that Algorithm 2 requires several items of information from the TCDB to be provided. This pre-processing is presented in Algorithm 3. We represent this information as mappings from the TCID to the information, irrespective of whether it is easily available at TCDB or not. The information on topology of a protein can be retrieved from UniProtKB for the entries of SwissProt; in other cases, the information may be computed by HMMTOP. Algorithm 4 presents a utility function find_transporters which calls TCDB-Blast, the BLAST search at the heart of TransATH. Algorithm 5 shows TCDB-Blast, the BLAST search of the TCDB using our choice of thresholds. Algorithm 6 shows the algorithm to determine the topology of a protein, and Algorithm 7 shows the algorithm to determine subcellular localization. Finally Algorithm 8 presents an extended version of TransATH, which includes subcellular localization information.

Algorithm 2 TransATH— Transporters via ATH (Annotation Transfer by Homology)

Require: a genome G as .fasta file of protein sequences
Require: the TCDB as a Blast+ protein sequence database with TCID as identifiers
Require: a mapping $TC2UniProt$ from the TCDB to the UniProt identifier of the entry
Require: a mapping $TC2TMS$ from the TCDB to the number of TMS of the entry
Require: a mapping $TC2Family$ from the TCDB to the TC family of the entry
Require: a mapping $TC2SubstrateGP$ from the TCDB to the Substrate Group of the entry
Require: a mapping $TC2SpecSubstrate$ from the TCDB to the Specific Substrate of the entry
Ensure: creates a table describing the complement of transporters in the genome G

```
list<gid, tcid> := find_transporters(G, TCDB)
sort list by lexicographical order of tcid
for all <gid, tcid> in list do
    output TC2Family(tcid),
           tcid,
           TC2UniProt(tcid),
           TC2TMS(tcid),
           TC2SubstrateGP(tcid),
           TC2SpecSubstrate(tcid),
           gid,
           computeTMS(gid)
end for
```

Algorithm 3 Pre-Processing for TransATH

Require: the TCDB
Require: SwissProt
Ensure: the TCDB as a Blast+ protein sequence database with TCID as identifiers
Ensure: a mapping $TC2UniProt$ from the TCDB to the UniProt identifier of the entry
Ensure: a mapping $TC2TMS$ from the TCDB to the number of TMS of the entry
Ensure: a mapping $TC2Family$ from the TCDB to the TC family of the entry
Ensure: a mapping $TC2SubstrateGP$ from the TCDB to the Substrate Group of the entry
Ensure: a mapping $TC2SpecSubstrate$ from the TCDB to the Specific Substrate of the entry
Ensure: a mapping $TC2Loc$ from the TCDB to the subcellular localization of the entry

```
download data from TCDB website
compute the TCDB Blast+ protein sequence database with TCID identifiers
manually curate list of Substrate Group terms
manually curate list of Specific Substrate terms
for all gid in TCDB and Swissprot do
    retrieve TMS data for gid from SwissProt
    retrieve subcellular localization for gid from SwissProt
end for
for all gid in TCDB without TMS data do
    computeTMS(gid)
end for
for all gid in TCDB without subcellular localization do
    computeLocalization(gid)
end for
```

Algorithm 4 find_transporters

Require: a genome G as .fasta file of protein sequences
Require: the TCDB as a Blast+ protein sequence database with TCID as identifiers
Require: a mapping $TC2TMS$ from the TCDB to the number of TMS of the entry
Ensure: result is list$<gid, tcid>$ of matches of proteins gid in G with transporters $tcid$

```
function FIND_TRANSPORTERS(G, TCDB)
    list<gid, tcid, _, _, _, _, _> := TCDB_BLAST(G, TCDB)
    return list<gid, tcid> where
        ( TC2TMS(tcid) ≠ 0) ∧ (computeTMS(gid) ≠ 0)
end function
```

4.4.2 TCDB-Blast Search

We modified G-Blast(v2), the second version of the G-Blast implementation of Saier's lab to do more than simply take the top BLAST hit. The results here refer to TCDB-Blast, the modified G-Blast(v2) which collects all hits passing a set of thresholds: e-value 1e-20; percent identity 40%; query coverage 70%; subject coverage 70%; and difference in length of 10%, which were selected following the evaluation in Section 4.5. Algorithm 5 shows the main step of the algorithm for the BLAST search of the TCDB.

Algorithm 5 The Algorithm for TCDB-Blast

Require: a genome G as .fasta file of protein sequences
Require: the TCDB as a Blast+ protein sequence database with TCID as identifiers
Ensure: result is list$<gid, tcid, pid, qcov, scov, eval, score>$ of matches $<gid, tcid>$ meeting thresholds, with percent identity pid, query coverage $qcov$, subject coverage $scov$, e-value $eval$, and score $score$

1: **function** TCDB_BLAST(G, TCDB)
2: Set e-value threshold t_{evalue} := 1e-20
3: Set percent identity threshold t_{pid} := 40%
4: Set query coverage threshold t_{qcov} := 70%
5: Set subject coverage threshold t_{scov} := 70%
6: Set difference threshold t_{diff} := 10%
7: list$<gid, tcid, pid, qcov, scov, eval, score>$:= Blast+:blastp(G, TCDB, t_{evalue})
8: **return** list$<gid, tcid, pid, qcov, scov, eval, score>$ **where**
9: $(pid \geq t_{pid}) \wedge (qcov \geq t_{qcov}) \wedge (scov \geq t_{scov}) \wedge$
10: $(|length(gid) - length(tcid)|/max(length(gid), length(tcid)) \leq t_{diff})$
11: **end function**

4.4.3 Topology Step

There are several programs for predicting the topology of membrane proteins. Topology is widely predicted using TMHMM. However, as shown above in Section 4.3.4.1, HMMTOP is superior. In a comparison of nine programs on four TC families [RCL+14], HMMTOP [TS01] is overall the best, performing best for the sugar porters, and performing well for the other families. Also performing well were MEMSAT-SVM [NJ10] and SPOCTOPUS [VBSE08]. Note that Saier's protocol [PVL+14] manually considers hydropathy plots using WHAT [ZSJ01] to correct HMMTOP predictions.

The term *hydropathy*, which means *"strong feeling about water"*, is introduced by Kyte and Doolittle [KD82] in 1982 to refer to the relationship between the hydrophilicity and

hydrophobicity of an amino acid. The hydropathy plot averages across a window to smooth out the values.

A similar tool, the *hydrophobic moment plot* of Eisenberg and co-workers [EWT82, ESKW84], is used in the protocol of UniProt (`http://www.uniprot.org/help/transmem`), which requires agreement of at least two methods from TMHMM, MEMSAT, Phobius and the hydrophobic moment plot method to predict alpha-helical TMS. Phobius is used to resolve conflicts between overlaps in predicted N-terminal signal peptides and transmembrane domains.

Our implementation relied on TM-Coffee [CDTTN12] which computes MSA of transmembrane proteins, to determine the alignment of the TMS regions of the query protein sequence with the the TMS regions of the entry in TCDB. This approach uses the transmembrane proteins in SwissProt as further entries in the MSA.

Algorithm 6 shows our implementation to determine the topology of a protein.

Algorithm 6 computeTMS function for Topology

Require: a protein sequence *gid*

Ensure: result is $<num, topology>$ of the number and topology of TMS of *gid*

```
1: function COMPUTETMS(gid)
2:     <num, topology> := HMMTOP(gid)
3:     msa := TM-Coffee(gid, SwissProt )
4:     adjust list<num, topology> based on the alignment msa
5:     return list<num, topology>
6: end function
```

4.4.4 Localization Step

A widely used tool for subcellular localization in fungi is WoLF PSORT [HPO+07]. It predicts localization to the nucleus, mitochondrion, cytosol, plasma membrane, extracellular region, Golgi, endoplasmic reticulum, peroxisome, vacuole, and several dual localizations. WoLF PSORT does not explicitly separate localizations inside an organelle and localizations in the membrane of an organelle.

A tool for localization prediction that has a comprehensive treatment of placing proteins in membranes of organelles is LocTree3 [GHH+14]. LocTree3 targets 18 sites, including 8 membranes: plasma membrane, nuclear membrane, mitochondrion membrane, ER membrane,

Golgi membrane, vacuole membrane, peroxisome membrane, and chloroplast membrane. LocTree3 achieves an overall accuracy of 80%. Furthermore, LocTree3 is shown to be superior to existing tools, including WoLF PSORT, in the experimental comparison [GHH+14].

Algorithm 7 computeLocalization function

Require: a transmembrane protein sequence *gid*
Ensure: result is the localization of protein *gid*

```
function computeLocalization(gid)
    return LocTree3( gid )
end function
```

Algorithm 8 presents an extended version of Saier's protocol which includes localization information. Although the TCDB does not store localization information, for those entries in SwissProt, the localization can be retrieved using the UniProt identifier of the TCDB entry. In other cases, it can be computed using LocTree3.

Algorithm 8 TransATH Extended Version

Require: a genome G as .fasta file of protein sequences
Require: the TCDB as a Blast+ protein sequence database with TCID as identifiers
Require: a mapping $TC2UniProt$ from the TCDB to the UniProt identifier of the entry
Require: a mapping $TC2TMS$ from the TCDB to the number of TMS of the entry
Require: a mapping $TC2Family$ from the TCDB to the TC family of the entry
Require: a mapping $TC2SubstrateGP$ from the TCDB to the Substrate Group of the entry
Require: a mapping $TC2SpecSubstrate$ from the TCDB to the Specific Substrate of the entry
Require: a mapping $TC2Loc$ from the TCDB to the subcellular localization of the entry
Ensure: creates a table describing the complement of transporters in the genome G

```
list<gid, tcid> := find_transporters(G, TCDB)
sort list by lexicographical order of tcid
for all <gid, tcid> in list do
    output TC2Family(tcid),
           tcid,
           TC2UniProt(tcid),
           TC2TMS(tcid),
           TC2SubstrateGP(tcid),
           TC2SpecSubstrate(tcid),
           TC2Loc(tcid),
           gid,
           computeTMS(gid),
           computeLocalization(gid)
end for
```

4.4.5 Substrate Information

In the application of the protocol [PVL+14], Saier assigns a *Substrate Group* and *Specific Substrate* to each predicted transporter. The categories of Substrate Group that Saier uses are given in Section 2.3.1.2. Such information may be implicit in the descriptions of TCDB entries, and in their related literature, but it is not officially defined in the Transporter Classification, nor is it explicitly accessible on the TCDB website.

For our purposes, this information is captured in a file mapping each TCID to a Substrate Group and to a Specific Substrate, where possible. The mapping is then used to augment the prediction.

4.4.6 Case Study Revisited

To demonstrate our implementation of Saier's protocol we apply it to our case study genome of *A. niger* CBS 513.88 to produce Table 30 that mimics [PVL+14, Table 1]. Table 30 presents the results of TransATH for the *A. niger* CBS 513.88 genome. The table is organised by TC-Family. The columns Family and Family Name contain the TC-Family identifier and its name. The column TCID contains the TCID of the matching TCDB entry predicted by TransATH. The column Hit is the UniProtKB identifier for the matching TCDB entry. The column HTMS contains the number of TMS for the hit. The column Substrate Group contains the name of the group for the substrate transported by the hit, if known. The column Specific Substrate contains the name of the substrate transported by the hit, if known. The column Query is the identifier for the entry in the *A. niger* CBS 513.88 genome. The column QTMS contains the number of TMS for the query.

Table 30: TransATH Results for *A. niger* CBS 513.88

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
1.A. Alpha-type channel-forming proteins and peptides								
1.A.9	the neurotransmitter receptor, cys loop, ligand-gated ion channel (lic) family.	1.A.9.5.2	O95166	1	Anion	Unknown	An07g10020	1
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.1.4	O67997	12	Cation	Ammonia	An08g03200	11
		1.A.11.3.1	P40260	11	Unknown	Unknown	An08g03200	11
		1.A.11.3.2	P41948	11	Unknown	Unknown	An08g03200	11
		1.A.11.3.2	P41948	11	Unknown	Unknown	An14g02390	11
		1.A.11.3.3	Q8NKD5	11	Cation	NH4+	An08g03200	11
		1.A.11.3.3	Q8NKD5	11	Cation	NH4+	An14g02390	11
		1.A.11.3.4	Q96UY0	11	Unknown	Unknown	An08g03200	11
		1.A.11.3.4	Q96UY0	11	Unknown	Unknown	An14g02390	11
		1.A.11.3.5	Q59UP8	11	Cation	NH4+	An08g03200	11
		1.A.11.3.5	Q59UP8	11	Cation	NH4+	An14g02390	11
1.A.17	the calcium-dependent chloride channel (ca-clc) family.	1.A.17.6.4	B0YES0	7	Anion	Unknown	An14g03020	7
		1.A.17.6.4	B0YES0	7	Anion	Unknown	An14g01960	8

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
1.A.23	the small conductance mechanosensitive ion channel (mscs) family.	1.A.23.4.9	F9X0Q3	6	Cation	Ca2+	An15g03150	6
1.A.33	the cation channel-forming heat shock	1.A.33.1.2	P0A6Y8	1	Unknown	Unknown	An11g04180	1
	protein-70 (hsp70) family.	1.A.33.1.2	P0A6Y8	1	Unknown	Unknown	An16g09260	1
		1.A.33.1.3	P08107	1	Cation	Unknown	An11g04180	1
		1.A.33.1.3	P08107	1	Cation	Unknown	An16g09260	1
1.A.46	the anion channel-forming bestrophin (bestrophin) family.	1.A.46.2.2	Q5AXS1	3	Anion	Unknown	An14g05100	3
1.A.56	the copper transporter (ctr) family.	1.A.56.1.10	A9XIK8	3	Cation	Cu2+	An02g11700	3
1.A.77	the mg(2+)/ca(2+) uniporter (mcu) family.	1.A.77.1.5	Q7S4I4	2	Cation	Mg2+,Ca2+	An04g06590	2
1.A.88	the fungal potassium channel (f-kch) family.	1.A.88.1.4	A2QW01	4	Cation	K+	An11g03330	4
1.B. Beta-type Barel porins								
1.B.69	the peroxysomal membrane	1.B.69.1.4	A2R8R0	4	Peptide	Unknown	An16g08040	4
	porin 4 (pxmp4) family.	1.B.69.1.6	B0CP94	4	Unknown	Unknown	An16g08040	4
1.F. Vesicle fusion pores proteins								
1.F.1	the synaptosomal vesicle fusion pore (svf-pore) family.	1.F.1.1.2	P33328	1	Nonselective	Unknown	An12g07570	1
1.H. Paracellular Channels								
1.H.1	the claudin tight junction	1.H.1.4.1	F5H8T9	5	Cation	Unknown	An08g01170	4
	(claudin) family.	1.H.1.4.3	G3XZI4	5	Unknown	Unknown	An07g08960	5
2.A. Carrier-type facilitators								
2.A.1	the major facilitator superfamily (mfs).	2.A.1.1.5	P43581	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.6	P13181	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.7	P11636	12	Monocarboxylate	Quinate:H+	An08g03850	12
		2.A.1.1.8	P30605	12	Unknown	Unknown	An04g00340	12
		2.A.1.1.21	O74969	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.22	O74849	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.22	O74849	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.31	P39004	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.33	Q8NJ22	12	Sugar	Fructose:H+	An15g01500	12
		2.A.1.1.33	Q8NJ22	12	Sugar	Fructose:H+	An06g02270	12
		2.A.1.1.36	Q400D8	12	Unknown	Unknown	An02g03540	12
		2.A.1.1.36	Q400D8	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.36	Q400D8	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.38	P39932	12	Sugar	Glycerol:H+	An14g02740	12
		2.A.1.1.38	P39932	12	Sugar	Glycerol:H+	An09g02930	12
		2.A.1.1.38	P39932	12	Sugar	Glycerol:H+	An14g03990	12
		2.A.1.1.39	P49374	12	Sugar	Glucose	An11g01100	12
		2.A.1.1.39	P49374	12	Sugar	Glucose	An02g00590	12
		2.A.1.1.39	P49374	12	Sugar	Glucose	An03g01620	12
		2.A.1.1.40	Q64L87	12	Sugar	Xylose	An01g00850	12
		2.A.1.1.51	Q2MEV7	12	Sugar	Glucose/Xylose	An15g03940	12
		2.A.1.1.51	Q2MEV7	12	Sugar	Glucose/Xylose	An12g07450	12
		2.A.1.1.57	Q8J0V1	12	Sugar	Monosaccharides	An12g07450	12
		2.A.1.1.57	Q8J0V1	12	Sugar	Monosaccharides	An15g03940	12
		2.A.1.1.58	Q8J0U9	12	Sugar	Glucose:H+	An02g03540	12
		2.A.1.1.58	Q8J0U9	12	Sugar	Glucose:H+	An05g01290	12
		2.A.1.1.58	Q8J0U9	12	Sugar	Glucose:H+	An03g02190	12
		2.A.1.1.67	Q2MDH1	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.67	Q2MDH1	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.68	A3M0N3	12	Sugar	Glucose	An15g03940	12
		2.A.1.1.68	A3M0N3	12	Sugar	Glucose	An12g07450	12
		2.A.1.1.70	Q0ULF7	12	Unknown	Unknown	An15g01500	12
		2.A.1.1.70	Q0ULF7	12	Unknown	Unknown	An06g02270	12
		2.A.1.1.73	Q5A8J5	12	Sugar	Glycerol:H+	An14g02740	12
		2.A.1.1.73	Q5A8J5	12	Sugar	Glycerol:H+	An09g02930	12
		2.A.1.1.73	Q5A8J5	12	Sugar	Glycerol:H+	An14g03990	12
		2.A.1.1.105	P54862	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.108	P32465	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.108	P32465	12	Unknown	Unknown	An02g03540	12
		2.A.1.1.110	P39924	12	Sugar	Hexose	An05g01290	12
		2.A.1.1.111	P23585	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.111	P23585	12	Unknown	Unknown	An03g02190	12
		2.A.1.1.112	Q9P3U6	12	Unknown	Unknown	An05g01290	12
		2.A.1.1.117	G4N740	12	Sugar	Glucose	An15g03940	12
		2.A.1.2.6	P28873	11	Unknown	Unknown	An18g01720	11
		2.A.1.2.16	Q07824	12	Amines	Spermine/Spermidine	An09g03320	12
		2.A.1.2.16	Q07824	12	Amines	Spermine/Spermidine	An18g01150	12
		2.A.1.2.16	Q07824	12	Amines	Spermine/Spermidine	An01g11540	12
		2.A.1.2.17	P38124	12	Specific drug	Fluconazole:H+	An16g02610	12
		2.A.1.2.17	P38124	12	Specific drug	Fluconazole:H+	An18g01720	11
		2.A.1.2.23	Q70WR7	12	Sugar	Fructose	An15g04060	11
		2.A.1.2.35	O94528	12	Cation	Unknown	An18g01720	11
		2.A.1.2.35	O94528	12	Cation	Unknown	An16g02610	12
		2.A.1.2.45	C5E4Z7	12	Unknown	Unknown	An15g04060	11

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
		2.A.1.2.46	C5DX43	12	Unknown	Unknown	An15g04060	11
		2.A.1.2.48	A2QTF4	9	Specific drug	Tetracycline	An09g01910	9
		2.A.1.2.67	P53283	11	Unknown	Unknown	An04g08300	12
		2.A.1.2.77	Q8NKG7	12	Multiple drug	Unknown	An02g09970	12
		2.A.1.2.77	Q8NKG7	12	Multiple drug	Unknown	An17g01070	11
		2.A.1.2.77	Q8NKG7	12	Multiple drug	Unknown	An04g08300	12
		2.A.1.2.77	Q8NKG7	12	Multiple drug	Unknown	An02g03620	12
		2.A.1.2.77	Q8NKG7	12	Multiple drug	Unknown	An08g06980	12
		2.A.1.2.78	B6HIC2	12	Multiple drug	Unknown	An02g09970	12
		2.A.1.2.78	B6HIC2	12	Multiple drug	Unknown	An17g01070	11
		2.A.1.2.85	B6H9Q3	12	Multiple drug	Phenylacetate/penoxyacetate	An04g08300	12
		2.A.1.2.85	B6H9Q3	12	Multiple drug	Phenylacetate/penoxyacetate	An04g07680	12
		2.A.1.2.85	B6H9Q3	12	Multiple drug	Phenylacetate/penoxyacetate	An02g09970	12
		2.A.1.2.85	B6H9Q3	12	Multiple drug	Phenylacetate/penoxyacetate	An02g03620	12
		2.A.1.2.85	B6H9Q3	12	Multiple drug	Phenylacetate/penoxyacetate	An17g01070	11
		2.A.1.2.86	B6HN82	12	Specific drug	Isopenicillin N	An16g00090	12
		2.A.1.2.86	B6HN82	12	Specific drug	Isopenicillin N	An04g08250	12
		2.A.1.2.86	B6HN82	12	Specific drug	Isopenicillin N	An02g03670	12
		2.A.1.2.86	B6HN82	12	Specific drug	Isopenicillin N	An08g10970	12
		2.A.1.3.52	Q08902	14	Cation	NH4+	An08g08220	14
		2.A.1.3.52	Q08902	14	Cation	NH4+	An08g08710	14
		2.A.1.3.52	Q08902	14	Cation	NH4+	An10g00700	14
		2.A.1.3.65	H2E274	14	Multiple drug	Unknown	An12g08620	14
		2.A.1.3.65	H2E274	14	Multiple drug	Unknown	An01g11290	15
		2.A.1.3.65	H2E274	14	Multiple drug	Unknown	An09g00870	13
		2.A.1.3.65	H2E274	14	Multiple drug	Unknown	An01g15000	14
		2.A.1.3.65	H2E274	14	Multiple drug	Unknown	An06g00770	14
		2.A.1.8.5	P22152	12	Anion	Nitrate	An08g05670	12
		2.A.1.8.13	Q8X193	12	Unknown	Unknown	An08g05670	12
		2.A.1.9.7	P25346	13	Organoions	Phospholipid	An16g06190	12
		2.A.1.14.38	P40445	12	Unknown	Unknown	An16g01940	11
		2.A.1.14.38	P40445	12	Unknown	Unknown	An01g11450	11
		2.A.1.14.38	P40445	12	Unknown	Unknown	An08g06430	9
		2.A.1.14.38	P40445	12	Unknown	Unknown	An07g00980	10
		2.A.1.16.1	P39980	15	Siderophore	Ferroxamine	An01g00720	14
		2.A.1.16.7	Q870L2	14	Siderophore	Ferric triacetylfusarinine C	An03g03560	14
		2.A.1.16.7	Q870L2	14	Siderophore	Ferric triacetylfusarinine C	An07g06240	14
		2.A.1.19.38	Q9C101	11	Unknown	Unknown	An12g00940	11
		2.A.1.19.38	Q9C101	11	Unknown	Unknown	An07g07980	12
		2.A.1.58.1	Q5A7S4	10	Sugar	N-acetylglucosamine:H+	An16g09020	12
		2.A.1.58.1	Q5A7S4	10	Sugar	N-acetylglucosamine:H+	An06g02510	11
		2.A.1.58.4	Q01HW9	11	Unknown	Unknown	An06g02510	11
		2.A.1.58.5	C9S7Y7	10	Unknown	Unknown	An09g02880	10
		2.A.1.75.2	E9CYW5	12	Monocarboxylate	Unknown	An14g04560	12
2.A.3	the amino acid-polyamine-organocation (apc) family.	2.A.3.4.1	P19807	12	Amino acid	Choline	An15g01900	12
		2.A.3.4.1	P19807	12	Amino acid	Choline	An09g05010	12
		2.A.3.4.2	Q9Y860	12	Amino acid	GABA	An16g02000	12
		2.A.3.4.2	Q9Y860	12	Amino acid	GABA	An09g02550	12
		2.A.3.4.3	P32837	12	Amino acid	GABA	An14g01850	12
		2.A.3.4.3	P32837	12	Amino acid	GABA	An17g01540	12
		2.A.3.4.6	Q9UT18	12	Amino acid	Thiamin	An02g09790	12
		2.A.3.8.4	P50276	11	Amino acid	Met	An04g03940	12
		2.A.3.10.1	P06775	12	Unknown	Unknown	An13g00840	12
		2.A.3.10.2	P19145	12	Unknown	Unknown	An13g00840	12
		2.A.3.10.4	P04817	12	Amino acid	Arg	An13g03650	12
		2.A.3.10.4	P04817	12	Amino acid	Arg	An09g06730	12
		2.A.3.10.8	P38967	12	Unknown	Unknown	An13g00840	12
		2.A.3.10.10	P32487	12	Amino acid	Arg,His,Lys	An13g03650	12
		2.A.3.10.11	P38971	12	Unknown	Unknown	An13g03650	12
		2.A.3.10.11	P38971	12	Unknown	Unknown	An09g06730	12
		2.A.3.10.13	P53388	12	Amino acid	Unknown	An12g04180	12
		2.A.3.10.13	P53388	12	Amino acid	Unknown	An13g03650	12
		2.A.3.10.17	Q8J266	12	Unknown	Unknown	An12g10130	12
		2.A.3.10.17	Q8J266	12	Unknown	Unknown	An09g00400	11
		2.A.3.10.18	Q8NKC4	13	Amino acid	Unknown	An09g00400	11
		2.A.3.10.18	Q8NKC4	13	Amino acid	Unknown	An05g01740	11
		2.A.3.10.19	P38090	12	Amino acid	Polyamine/Carnitine	An04g00530	12
		2.A.3.10.19	P38090	12	Amino acid	Polyamine/Carnitine	An04g09620	12
		2.A.3.10.20	P43059	12	Unknown	Unknown	An09g06730	12
		2.A.3.10.20	P43059	12	Unknown	Unknown	An13g03650	12
		2.A.3.10.21	Q9URZ4	12	Amino acid	Arg, Lys	An13g00840	12
		2.A.3.10.22	Q2VQZ4	12	Unknown	Unknown	An13g00840	12
		2.A.3.10.23	Q5AG77	12	Amino acid	Arg,Leu,Met,Phe	An13g00840	12
		2.A.3.10.24	Q59YT0	12	Amino acid	Unknown	An13g00840	12
		2.A.3.10.25	Q59WB3	12	Unknown	Unknown	An13g00840	12
		2.A.3.10.26	Q59NZ6	12	Unknown	Unknown	An13g00840	12
		2.A.3.10.28	O60170	12	Amino acid	Arg,Lys	An13g00840	12
2.A.4	Cation diffusion facilitator (CDF) family.	2.A.4.2.2	P20107	6	Cation	Zn^{2+}, Co^{2+}	An15g03900	6
2.A.5	the zinc (zn(2+))-iron (fe(2+)) permease (zip) family.	2.A.5.1.1	P32804	8	Cation	Zn2+	An01g01620	8
		2.A.5.1.1	P32804	8	Cation	Zn2+	An15g07190	8
		2.A.5.1.1	P32804	8	Cation	Zn2+	An01g06690	7

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
2.A.6	Resistance-nodulation-cell division (RND) superfamily.	2.A.6.6.3	Q12200	13	Lipid	Sphingolipid	An11g05000	13
2.A.7	the drug/metabolite transporter (dmt) superfamily.	2.A.7.13.2	Q5A477	9	Nucleotide	GDP-mannose	An17g02140	10
		2.A.7.24.11	Q4WUA9	10	Unknown	Unknown	An03g03820	10
		2.A.7.24.11	Q4WUA9	10	Unknown	Unknown	An01g00340	10
2.A.16	the telurite-resistance/ dicarboxylate transporter (tdt) family.	2.A.16.4.1	A2QYD7	9	Unknown	Unknown	An12g00870	9
		2.A.16.4.2	A3R044	10	Unknown	Unknown	An12g00870	9
		2.A.16.4.3	Q2TJJ2	10	Unknown	Unknown	An12g00870	9
2.A.17	the proton-dependent oligopeptide transporter (pot) family.	2.A.17.2.1	Q9P380	12	Peptide	Unknown	An12g01210	11
		2.A.17.2.1	Q9P380	12	Peptide	Unknown	An08g04600	11
		2.A.17.2.2	P32901	12	Peptide	dipeptide,tripeptide	An12g01210	11
2.A.18	the amino acid/auxin permease	2.A.18.4.1	P38680	11	Amino acid	Unknown	An15g07550	11
		2.A.18.4.1	P38680	11	Amino acid	Unknown	An09g03660	11
		2.A.18.4.1	P38680	11	Amino acid	Unknown	An16g05880	11
		2.A.18.4.2	Q6IT47	11	Amino acid	Unknown	An15g07550	11
		2.A.18.4.2	Q6IT47	11	Amino acid	Unknown	An09g03660	11
		2.A.18.4.2	Q6IT47	11	Amino acid	Unknown	An16g05880	11
		2.A.18.7.1	P36062	11	Unknown	Unknown	An04g02150	11
2.A.19	the ca(2+):cation antiporter (caca) family.	2.A.19.2.2	Q99385	11	Cation	Ca2+,K+	An01g03100	11
		2.A.19.2.2	Q99385	11	Cation	Ca2+,K+	An19g00340	11
		2.A.19.2.8	O59940	10	Cation	Ca2+,K+	An01g03100	11
2.A.21	the solute:sodium symporter (sss) family.	2.A.21.6.1	P33413	15	Amines	Urea,polyamines	An01g03790	15
		2.A.21.6.1	P33413	15	Amines	Urea,polyamines	An18g01360	15
		2.A.21.6.2	Q9FHJ8	15	Unknown	Unknown	An01g03790	15
		2.A.21.6.4	Q59VF2	15	Unknown	Unknown	An01g03790	15
2.A.29	the mitochondrial carrier (mc) family.	2.A.29.1.1	P05141	4	Unknown	Unknown	An18g04220	6
		2.A.29.1.2	P12235	6	Unknown	Unknown	An18g04220	6
		2.A.29.1.3	P04710	4	Unknown	Unknown	An18g04220	6
		2.A.29.1.4	Q8TFA7	4	Unknown	Unknown	An18g04220	6
		2.A.29.1.6	Q8LB08	4	Unknown	Unknown	An18g04220	6
		2.A.29.1.7	P18239	4	Nucleotide	ATP:ADP antiporter	An18g04220	6
		2.A.29.1.8	Q9H0C2	5	Unknown	Unknown	An18g04220	6
		2.A.29.1.9	P18238	6	Unknown	Unknown	An18g04220	6
		2.A.29.1.10	P12236	6	Unknown	Unknown	An18g04220	6
		2.A.29.2.1	P22292	6	Unknown	Unknown	An02g01730	5
		2.A.29.2.2	O89035	2	Unknown	Unknown	An02g01730	5
		2.A.29.2.3	Q06143	6	Dicarboxylate	Unknown	An02g01730	5
		2.A.29.2.5	Q99297	1	Dicarboxylate	Unknown	An08g01370	3
		2.A.29.2.6	Q8SF04	4	Unknown	Unknown	An11g02540	6
		2.A.29.2.7	Q9UBX3	3	Unknown	Unknown	An02g01730	5
		2.A.29.2.8	Q03028	4	Unknown	Unknown	An08g01370	3
		2.A.29.2.10	Q8IB73	6	Dicarboxylate	alpha-ketogluterate	An11g02540	6
		2.A.29.2.11	Q9CR62	5	Unknown	Unknown	An02g01730	5
		2.A.29.2.13	Q02978	6	Unknown	Unknown	An02g01730	5
		2.A.29.4.1	P12234	6	Unknown	Unknown	An02g12070	4
		2.A.29.4.2	Q00325	6	Unknown	Unknown	An02g12070	4
		2.A.29.4.3	P23641	6	Inorganic	phosphate	An01g13600	6
		2.A.29.4.3	P23641	6	Inorganic	phosphate	An02g04160	4
		2.A.29.4.4	P40035	6	Inorganic	phosphate	An02g04160	4
		2.A.29.4.5	Q8VEM8	6	Unknown	Unknown	An02g12070	4
		2.A.29.4.6	Q9FMU6	7	Inorganic	phosphate	An02g12070	4
		2.A.29.5.1	P10566	6	Cation	Fe2+	An06g01730	6
		2.A.29.5.2	P23500	6	Unknown	Unknown	An06g01730	6
		2.A.29.5.3	Q287T7	1	Unknown	Unknown	An06g01730	6
		2.A.29.5.5	Q920G8	1	Unknown	Unknown	An06g01730	6
		2.A.29.5.7	Q9NYZ2	1	Unknown	Unknown	An06g01730	6
		2.A.29.7.3	P38152	4	Dicarboxylate	Tricarboxylate	An11g11230	3
		2.A.29.7.3	P38152	4	Dicarboxylate	Tricarboxylate	An18g00070	2
		2.A.29.7.4	Q7KSQ0	6	Unknown	Unknown	An11g11230	3
		2.A.29.8.2	Q27257	6	Unknown	Unknown	An03g03360	6
		2.A.29.8.4	Q12289	5	Cation	Carnitine	An03g03360	6
		2.A.29.8.11	P38087	6	Unknown	Unknown	An18g05590	2
		2.A.29.8.12	P32331	4	Organic acid	Unknown	An18g05590	2
		2.A.29.9.1	Q01356	3	Amino acid	Unknown	An03g06860	5
		2.A.29.10.4	P38127	4	Nucleotide	Pyrimidine	An14g01860	5
		2.A.29.10.5	P40556	4	Nucleotide	NAD+, pyruvate	An04g01190	4
		2.A.29.10.7	Q9BSK2	6	Unknown	Unknown	An14g01860	5
		2.A.29.13.1	P33303	2	Dicarboxylate	Succinate, fumerate	An04g09030	1
		2.A.29.14.1	O75746	3	Unknown	Unknown	An07g03070	5
		2.A.29.21.1	P38988	5	Nucleotide	Guanine	An07g10010	5
		2.A.29.29.1	Q04013	2	Dicarboxylate	Tricarboxylates	An09g06670	2
		2.A.29.29.1	Q04013	2	Dicarboxylate	Tricarboxylates	An02g11090	5
2.A.39	the nucleobase:cation symporter-1 (ncs1) family.	2.A.39.3.7	Q10279	13	Nucleobase	Uracil:cation	An08g06240	12
2.A.40	the nucleobase:cation symporter-2 (ncs2) family.	2.A.40.4.1	Q07307	12	Nucleobase	Urate, xanthine	An07g01950	15
		2.A.40.4.1	Q07307	12	Nucleobase	Urate, xanthine	An02g00560	13
		2.A.40.4.4	P48777	14	Nucleobase	Purine	An07g01950	15
		2.A.40.4.4	P48777	14	Nucleobase	Purine	An02g00560	13
		2.A.40.7.1	Q7Z8R3	12	Nucleobase	Purine	An13g02390	10

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
2.A.41	the concentrative nucleoside transporter (cnt) family.	2.A.41.2.7	Q874I3	12	Nucleoside	Unknown	An08g10300	13
2.A.43	the lysosomal cystine transporter (lct) family.	2.A.43.2.7	P38279	7	Unknown	Unknown	An09g06510	7
2.A.47	the divalent anion:na(+) symporter (dass) family.	2.A.47.2.1	P25360	10	Unknown	Unknown	An01g03120	11
		2.A.47.2.2	P27514	12	Anion	Phosphate	An01g03120	11
		2.A.47.2.3	P39535	12	Unknown	Unknown	An01g03120	11
2.A.52	the ni(2+)-co(2+) transporter (nicot) family.	2.A.52.1.8	Q7S3L8	7	Cation	Ni2+	An12g04470	8
2.A.53	the sulfate permease (sulp) family.	2.A.53.1.2	P23622	13	Anion	sulphate	An15g04600	15
2.A.55	the metal ion (mn(2+)-iron) transporter (nramp) family.	2.A.55.1.1	P38925	11	Unknown	Unknown	An04g05680	11
		2.A.55.1.2	P38778	10	Unknown	Unknown	An04g05680	11
		2.A.55.1.4	Q10177	11	Cation	Mn2+	An04g05680	11
2.A.59	the arsenical resistance-3 (acr3) family.	2.A.59.1.1	Q06598	10	Unknown	Unknown	An18g03550	10
		2.A.59.1.2	P45946	10	Anion	Unknown	An18g03550	10
2.A.66	the multidrug/oligosaccharidyl-lipid/polysaccharide (mop) flippase superfamily.	2.A.66.1.5	P38767	11	Specific drug	Unknown	An08g07590	12
2.A.67	the oligopeptide transporter (opt) family.	2.A.67.1.1	O14411	19	Peptide	Unknown	An14g05290	15
		2.A.67.1.1	O14411	19	Peptide	Unknown	An11g05350	16
		2.A.67.1.2	P40900	17	Unknown	Unknown	An14g05290	15
		2.A.67.1.2	P40900	17	Unknown	Unknown	An11g05350	16
		2.A.67.1.3	P40897	15	Unknown	Unknown	An16g00810	14
		2.A.67.1.5	O14031	15	Peptide	Glutathione	An16g00810	14
		2.A.67.1.5	O14031	15	Peptide	Glutathione	An14g05290	15
		2.A.67.1.5	O14031	15	Peptide	Glutathione	An11g05350	16
		2.A.67.1.5	O14031	15	Peptide	Glutathione	An11g03640	15
2.A.69	the auxin efflux carrier (aec) family.	2.A.69.2.3	B8MZ51	10	Unknown	Unknown	An01g11100	10
2.A.72	the k(+) uptake permease (kup) family.	2.A.72.3.2	O74724	14	Cation	K+	An02g05630	13
2.A.89	the vacuolar iron transporter (vit) family.	2.A.89.1.1	P47818	5	Unknown	Unknown	An16g03690	5
2.A.96	the acetate uptake transporter (acetr) family.	2.A.96.1.3	Q5B2K4	6	Anion	Acetate	An07g08810	6
		2.A.96.1.3	Q5B2K4	6	Anion	Acetate	An13g02020	7
		2.A.96.1.4	P25613	6	Unknown	Unknown	An13g02020	7
		2.A.96.1.6	O14201	6	Unknown	Unknown	An07g08810	6
		2.A.96.1.7	P32907	6	Unknown	Unknown	An13g02020	7
2.A.105	the mitochondrial pyruvate carrier (mpc) family.	2.A.105.1.1	P53157	2	Monocarboxylates	Pyruvate	An04g02140	2
2.A.108	the iron/lead transporter (ilt) family.	2.A.108.1.1	P40088	7	Cation	Unknown	An01g08950	7
		2.A.108.1.1	P38993	1	Cation	Unknown	An15g05520	1
		2.A.108.1.1	P38993	1	Cation	Unknown	An01g08960	1
		2.A.108.1.1	P40088	7	Cation	Unknown	An16g01130	7
		2.A.108.1.1	P40088	7	Cation	Unknown	An15g05510	7
		2.A.108.1.2	Q9P8U9	7	Cation	Fe2+	An01g08950	7
		2.A.108.1.2	Q9P8U9	7	Cation	Fe2+	An16g01130	7
		2.A.108.1.2	Q9P8U9	7	Cation	Fe2+	An15g05510	7
		2.A.108.1.3	Q9P8U8	7	Cation	Fe2+	An01g08950	7
		2.A.108.1.3	Q9P8U8	7	Cation	Fe2+	An16g01130	7
		2.A.108.1.3	Q9P8U8	7	Cation	Fe2+	An15g05510	7
		2.A.108.1.4	P43561	1	Cation	Unknown	An15g05520	1
		2.A.108.1.4	P43561	1	Cation	Unknown	An01g08960	1
		2.A.108.1.5	Q09919	7	Unknown	Unknown	An01g08950	7
		2.A.108.1.5	Q09919	7	Unknown	Unknown	An16g01130	7
		2.A.108.1.5	Q09919	7	Unknown	Unknown	An15g05510	7
3.A. P-P-bond hydrolysis-driven transporters								
3.A.1	the atp-binding cassette (abc) superfamily.	3.A.1.201.1	P08183	12	Unknown	Unknown	An17g01770	12
		3.A.1.201.3	P21439	12	Unknown	Unknown	An17g01770	12
		3.A.1.201.10	B0Y3B6	12	Multiple drug	Unknown	An17g01770	12
		3.A.1.201.10	B0Y3B6	12	Multiple drug	Unknown	An04g08340	9
		3.A.1.201.16	I0DHH7	12	Unknown	Unknown	An17g01770	12
		3.A.1.201.17	Q9NRK6	6	Unknown	Unknown	An04g07060	6
		3.A.1.201.18	P36619	13	Unknown	Unknown	An04g08340	9
		3.A.1.203.1	P28288	5	Unknown	Unknown	An08g05780	3
		3.A.1.203.3	P33897	4	Unknown	Unknown	An08g05780	3
		3.A.1.203.7	Q9UBJ2	5	Lipid	Unknown	An08g05780	3
		3.A.1.203.7	Q9UBJ2	5	Lipid	Unknown	An01g03680	4
		3.A.1.203.10	I7MJ28	6	Unknown	Unknown	An08g05780	3
		3.A.1.205.1	P33302	15	Unknown	Unknown	An01g12380	12
		3.A.1.205.1	P33302	15	Unknown	Unknown	An15g02930	16
		3.A.1.205.1	P33302	15	Unknown	Unknown	An05g01660	11
		3.A.1.205.1	P33302	15	Unknown	Unknown	An08g03300	11
		3.A.1.205.1	P33302	15	Unknown	Unknown	An08g04500	11
		3.A.1.205.1	P33302	15	Unknown	Unknown	An13g03570	13
		3.A.1.205.1	P33302	15	Unknown	Unknown	An07g01250	14
		3.A.1.205.2	P32568	12	Unknown	Unknown	An01g12380	12
		3.A.1.205.2	P32568	12	Unknown	Unknown	An07g01250	14
		3.A.1.205.2	P32568	12	Unknown	Unknown	An08g03300	11

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
		3.A.1.205.3	Q02785	15	Unknown	Unknown	An07g01250	14
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An01g12380	12
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An05g01660	11
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An15g02930	16
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An13g03570	13
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An08g03300	11
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An08g04500	11
		3.A.1.205.4	P43071	13	Multiple drug	Unknown	An07g01250	14
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An15g02930	16
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An01g12380	12
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An05g01660	11
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An13g03570	13
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An08g03300	11
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An08g04500	11
		3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An07g01250	14
		3.A.1.205.6	Q8X0Z3	14	Unknown	Unknown	An13g03060	11
		3.A.1.205.6	Q8X0Z3	14	Unknown	Unknown	An15g01130	15
		3.A.1.205.7	P78577	11	Multiple drug	Unknown	An13g03060	11
		3.A.1.205.7	P78577	11	Multiple drug	Unknown	An14g03570	14
		3.A.1.205.7	P78577	11	Multiple drug	Unknown	An14g02610	11
		3.A.1.205.7	P78577	11	Multiple drug	Unknown	An11g02110	12
		3.A.1.205.11	P41820	13	Unknown	Unknown	An08g03300	11
		3.A.1.205.11	P41820	13	Unknown	Unknown	An15g02930	16
		3.A.1.205.11	P41820	13	Unknown	Unknown	An05g01660	11
		3.A.1.205.11	P41820	13	Unknown	Unknown	An07g01250	14
		3.A.1.205.11	P41820	13	Unknown	Unknown	An08g04500	11
		3.A.1.205.11	P41820	13	Unknown	Unknown	An01g12380	12
		3.A.1.205.11	P41820	13	Unknown	Unknown	An13g03570	13
		3.A.1.205.11	P41820	13	Unknown	Unknown	An11g02110	12
		3.A.1.205.12	P51533	15	Unknown	Unknown	An15g02930	16
		3.A.1.205.12	P51533	15	Unknown	Unknown	An01g12380	12
		3.A.1.205.12	P51533	15	Unknown	Unknown	An05g01660	11
		3.A.1.205.12	P51533	15	Unknown	Unknown	An08g03300	11
		3.A.1.205.12	P51533	15	Unknown	Unknown	An07g01250	14
		3.A.1.205.12	P51533	15	Unknown	Unknown	An13g03570	13
		3.A.1.205.12	P51533	15	Unknown	Unknown	An08g04500	11
		3.A.1.208.2	Q92887	16	Multiple drug	Organic anion	An03g04060	13
		3.A.1.208.11	P39109	14	Peptide	Bilirubin	An03g04060	13
		3.A.1.208.16	Q10185	16	Unknown	Unknown	An03g04060	13
		3.A.1.208.28	Q9P5N0	12	Unknown	Unknown	An03g04060	13
		3.A.1.208.32	D2WF19	16	Unknown	Unknown	An03g04060	13
		3.A.1.210.1	P40416	5	Cation	Unknown	An08g10600	5
		3.A.1.210.2	Q02592	10	Cation	Glutathione	An07g07500	11
		3.A.1.210.4	O75027	5	Unknown	Unknown	An08g10600	5
		3.A.1.210.7	Q9XUJ1	10	Unknown	Unknown	An08g10600	5
		3.A.1.210.8	Q9LVM1	7	Unknown	Unknown	An08g10600	5
		3.A.1.212.2	P33311	5	Unknown	Unknown	An04g07060	6
3.A.2	the h(+)- or na(+)-translocating f-type,	3.A.2.1.3	P05626	2	Cation	H+	An16g07290	2
	v-type and a-type atpase (f-atpase) superfamily.	3.A.2.2.3	P25515	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.3	P25515	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.3	P32842	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.3	P32842	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.3	P32842	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.3	P32563	9	Unknown	Unknown	An04g05310	7
		3.A.2.2.3	P25515	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.3	P37296	8	Unknown	Unknown	An04g05310	7
		3.A.2.2.4	Q93050	8	Unknown	Unknown	An04g05310	7
		3.A.2.2.5	P59229	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.5	P59227	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.5	P59229	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.5	P59228	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.5	P59227	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.5	P59227	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.5	P59228	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.5	P59229	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.6	P63082	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.6	Q91V37	5	Unknown	Unknown	An15g05730	5
		3.A.2.2.6	P63082	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.6	P63082	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.6	Q9Z1G4	9	Unknown	Unknown	An04g05310	7
		3.A.2.2.6	Q920R6	9	Unknown	Unknown	An04g05310	7
		3.A.2.2.7	G5EDB8	5	Unknown	Unknown	An15g05730	5
		3.A.2.2.7	P34546	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.7	Q21898	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.7	P34546	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.7	Q21898	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.7	P34546	4	Unknown	Unknown	An07g05080	4
		3.A.2.2.7	P30628	7	Unknown	Unknown	An04g05310	7
		3.A.2.2.8	Q4UJ88	4	Unknown	Unknown	An02g08020	4
		3.A.2.2.8	Q4UJ88	4	Unknown	Unknown	An10g00680	4
		3.A.2.2.8	Q4UJ88	4	Unknown	Unknown	An07g05080	4

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
3.A.3	the p-type atpase (p-atpase) superfamily.	3.A.3.1.7	Q2U3D2	10	Cation	Unknown	An14g02290	10
		3.A.3.2.3	P13586	10	Unknown	Unknown	An02g14450	9
		3.A.3.2.6	Q9UUX9	10	Cation	Ca2+	An02g14450	9
		3.A.3.2.7	P16615	11	Cation	Ca2+	An18g06290	10
		3.A.3.2.9	O75185	10	Unknown	Unknown	An02g14450	9
		3.A.3.2.13	P92939	10	Unknown	Unknown	An18g06290	10
		3.A.3.2.19	Q9SY55	12	Unknown	Unknown	An18g06290	10
		3.A.3.2.27	Q9UUY2	10	Cation	Ca2	An08g03090	10
		3.A.3.2.32	Q49LV5	11	Unknown	Unknown	An18g06290	10
		3.A.3.2.35	Q9HDW7	12	Unknown	Unknown	An08g03090	10
		3.A.3.2.36	Q5IH90	10	Unknown	Unknown	An18g06290	10
		3.A.3.2.37	O76974	10	Unknown	Unknown	An18g06290	10
		3.A.3.3.1	P07038	10	Cation	H+	An01g05670	11
		3.A.3.3.1	P07038	10	Cation	H+	An16g05840	10
		3.A.3.3.1	P07038	10	Cation	H+	An09g05950	10
		3.A.3.3.6	P05030	10	Cation	H+	An01g05670	11
		3.A.3.3.6	P05030	10	Cation	H+	An16g05840	10
		3.A.3.3.6	P05030	10	Cation	H+	An09g05950	10
		3.A.3.8.2	P39524	8	Lipid	Phospholipid	An12g04500	10
		3.A.3.8.4	P32660	10	Lipid	Phospholipid	An12g08790	8
		3.A.3.8.4	P32660	10	Lipid	Phospholipid	An09g03160	10
		3.A.3.8.5	Q12675	10	Unknown	Unknown	An12g08790	8
		3.A.3.8.10	Q5KP96	10	Unknown	Unknown	An12g04500	10
		3.A.3.9.1	P13587	10	Unknown	Unknown	An15g01830	10
		3.A.3.9.1	P13587	10	Unknown	Unknown	An09g00690	8
		3.A.3.9.2	P22189	10	Unknown	Unknown	An15g01830	10
		3.A.3.9.2	P22189	10	Unknown	Unknown	An09g00690	8
		3.A.3.9.3	O13398	10	Unknown	Unknown	An09g00690	8
		3.A.3.9.3	O13398	10	Unknown	Unknown	An15g01830	10
		3.A.3.9.4	P78981	10	Unknown	Unknown	An15g01830	10
		3.A.3.9.4	P78981	10	Unknown	Unknown	An09g00690	8
		3.A.3.9.5	B5B9V9	10	Cation	Unknown	An15g01830	10
		3.A.3.9.5	B5B9V9	10	Cation	Unknown	An09g00690	8
		3.A.3.9.6	Q4PI59	10	Unknown	Unknown	An09g00690	8
		3.A.3.9.6	Q4PI59	10	Unknown	Unknown	An15g01830	10
3.A.5	the general secretory pathway (sec) family.	3.A.5.8.1	P32915	12	Protein	Peptide	An03g04340	10
		3.A.5.9.1	Q9H9S3	10	Protein	Protein	An03g04340	10
		3.A.5.9.1	P61619	12	Protein	Protein	An03g04340	10
		3.A.5.9.1	P60059	1	Protein	Protein	An01g11630	1
3.A.8	the mitochondrial protein translocase (mpt) family.	3.A.8.1.1	P39515	4	Protein	Protein	An11g02140	3
		3.A.8.1.1	Q02776	1	Protein	Protein	An07g07880	2
		3.A.8.1.1	P32897	3	Protein	Protein	An02g01360	3
3.A.16	the endoplasmic reticular retrotranslocon (er-rt) family.	3.A.16.1.2	E7NGV2	1	Protein	Protein	An14g00230	2
3.A.19	the tms recognition/insertion complex (trc) family.	3.A.19.1.2	A2QHQ3	3	Protein	Protein	An04g00670	3
3.D. Oxidoreduction-driven transporters								
3.D.1	the h(+) or na(+)-translocating nadh dehydrogenase (ndh) family.	3.D.1.6.1	P42026	2	Unknown	Unknown	An11g08840	1
		3.D.1.6.2	Q7S1I2	1	Cation	H+	An16g02130	1
		3.D.1.6.2	Q02854	3	Cation	H+	An14g00060	2
		3.D.1.6.2	P25710	3	Cation	H+	An06g01390	4
		3.D.1.6.3	Q9FNN5	1	Unknown	Unknown	An04g05640	1
		3.D.1.6.3	Q42577	1	Unknown	Unknown	An11g08840	1
		3.D.1.6.4	Q6V9B2	1	Unknown	Unknown	An04g05640	1
3.D.2	the proton-translocating transhydrogenase (pth) family.	3.D.2.3.1	P11024	16	Cation	Unknown	An02g09810	14
3.D.3	the proton-translocating quinol: cytochrome c reductase (qcr) superfamily.	3.D.3.2.1	P08067	1	Electron	Unknown	An14g04080	1
		3.D.3.3.1	P07143	2	Electron	Unknown	An01g06180	2
8.A. Auxiliary transport proteins								
8.A.27	the cdc50 p-type atpase lipid flippase subunit (cdc50) family.	8.A.27.1.2	P25656	3	Lipid	Unknown	An07g10420	2
9.A. Recognized transporters of known biochemical								
9.A.2	the endomembrane protein-70 (emp70) family.	9.A.2.1.1	E7NFP9	9	Protein	Protein	An06g01200	10
		9.A.2.1.2	Q9LIC2	10	Unknown	Unknown	An06g01200	10
		9.A.2.1.6	Q99805	9	Unknown	Unknown	An06g01200	10
9.A.6	the atp exporter (atp-e) family.	9.A.6.1.1	P36051	14	Nucleotide	ATP	An14g00900	14
9.A.41	the capsular polysaccharide exporter (cps-e) family.	9.A.41.1.1	P44669	1	Unknown	Unknown	An11g04180	1
9.A.54	the lysosomal cobalamin (b12) transporter (l-b12t) family.	9.A.54.1.3	A6QTW5	10	Protein	cobalamin	An16g09150	10
9.B. Putative transport proteins								
9.B.1	the integral membrane caax protease	9.B.1.1.2	Q8RX88	7	Unknown	Unknown	An04g01950	7

Continued on next page

Table 30 – continued from previous page

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
	(caax protease) family.	9.B.1.1.3	P47154	5	Peptide	Unknown	An04g01950	7
		9.B.1.2.2	F9FER0	5	Peptide	Unknown	An14g03420	6
9.B.7	the putative sulfate transporter (cysz) family.	9.B.7.2.3	E2PST1	5	Protein	Unknown	An07g06140	5
9.B.16	the putative ductin channel (ductin) family.	9.B.16.1.1	P23380	4	Unknown	Unknown	An02g08020	4
		9.B.16.1.1	P23380	4	Unknown	Unknown	An10g00680	4
		9.B.16.1.1	P23380	4	Unknown	Unknown	An07g05080	4
		9.B.16.1.2	Q03105	4	Unknown	Unknown	An02g08020	4
		9.B.16.1.2	Q03105	4	Unknown	Unknown	An10g00680	4
		9.B.16.1.2	Q03105	4	Unknown	Unknown	An07g05080	4
9.B.25	the mitochondrial inner/outer membrane fusion (mmf) family.	9.B.25.1.1	P32266	1	Nucleotide	Unknown	An08g04250	1
9.B.26	the regulator of er stress and autophagy tmem208 (tmem208) family.	9.B.26.1.4	K9FAK7	2	Unknown	Unknown	An12g03980	2
9.B.82	endoplasmic reticulum retrieval protein1 (putative heavy metal transporter) (rer1) family.	9.B.82.1.1	P25560	4	Unknown	Unknown	An02g02830	4
		9.B.82.1.2	O15258	4	Unknown	Unknown	An02g02830	4
		9.B.82.1.3	O48670	4	Unknown	Unknown	An02g02830	4
9.B.119	the glycan synthase, fks1 (fks1) family.	9.B.119.1.1	P38631	16	Sugar	Unknown	An06g01550	18
9.B.142	the integral membrane glycosyltransferase family 39 (gt39) family.	9.B.142.3.3	B3S136	13	Unknown	Unknown	An16g08570	13
		9.B.142.3.5	G9P430	13	Sugar	Unknown	An16g08570	13
9.B.143	the 6 tms duf1275/pf06912 (duf1275) family.	9.B.143.5.1	G7XY82	6	Unknown	Unknown	An10g00830	6

4.4.7 The TransATH Web Service

The beta version of TransATH is publicly available and can be accessed at `http://transath.umt.edu.my`. Figure 23 shows the input page for the user to upload a fasta file of protein sequences. The user is able to choose the thresholds for percentage alignment and e-values. For percent alignment the thresholds from 40 for less stringent filtering to over 70 for more stringency. For e-value thresholds there are six choices: 10, e-5, e-10, e-20, e-30 and e-50.

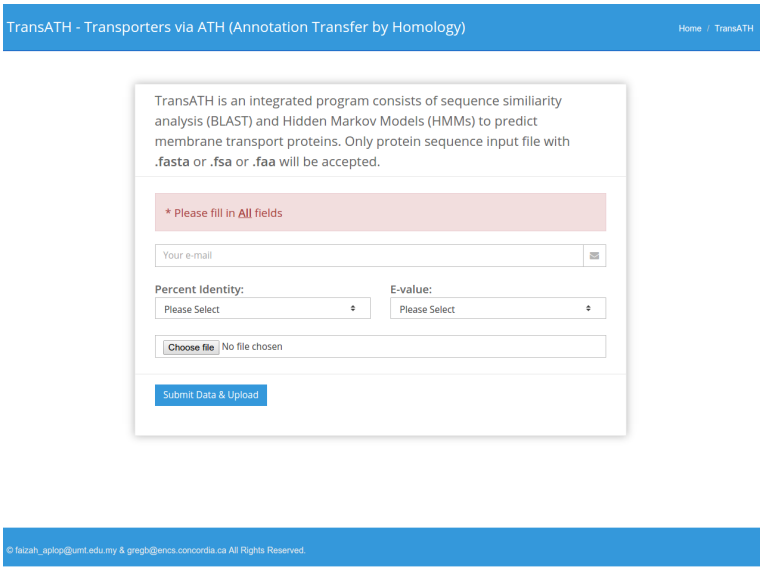
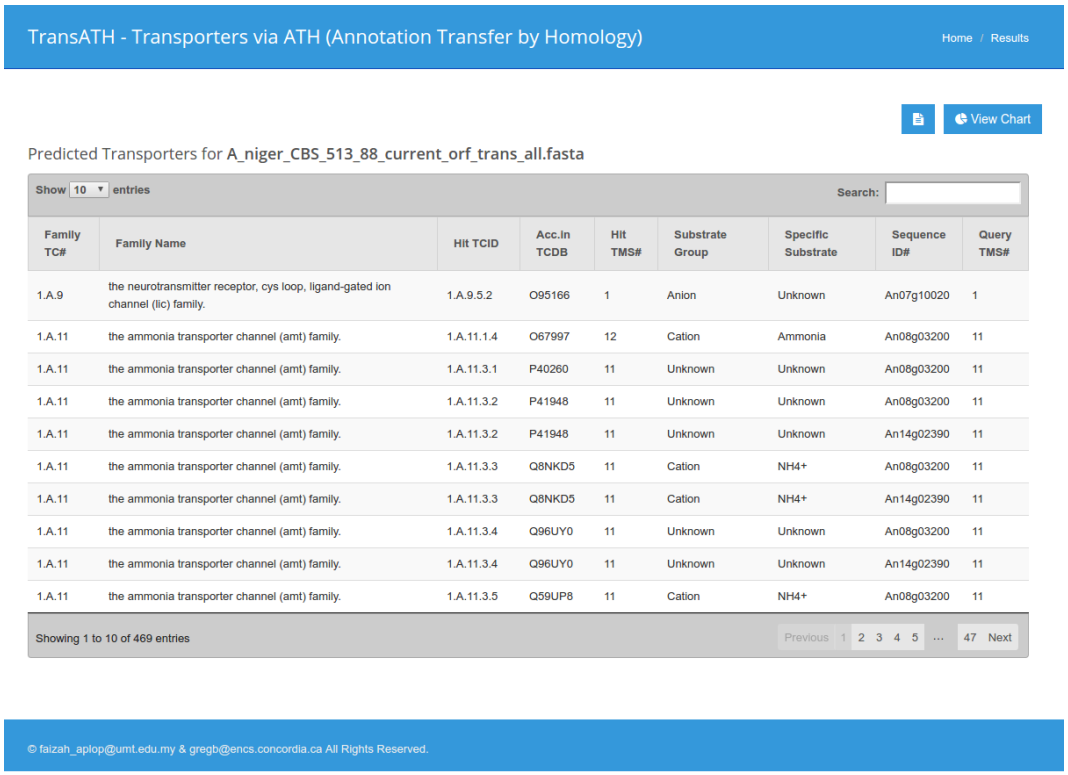


Figure 23: Input Page for TransATH

TransATH takes approximately 80–100 minutes for a typical fungal genome fasta input file of size approximately 10MB using a web server with an 8-core processor, 8GB memory and 45GB of disk space. A link to the result page is generated once TransATH finishes. Figure 24 shows an example of an output page that displays a table of predicted transporters immitating the result by Saier [PVL+14, Table 1]. There are nine columns: *Family TC#, Family Name, Hit TCID, Access in TCDB, Hit TMS#, Substrate Group, Specific Substrate, Sequence ID#* and *Query TMS#*.

The user is able to download the whole table in tsv format by clicking on the first icon at the top right of the output page.

The user can generate a pie chart of the predicted substrate groups by clicking on the *View Chart* icon at the top right of the result page. Figure 25 shows an example. By mousing over the pie chart, the specific slice will be highlighted and the *Percentage Values* box to the left of the chart wlll display the substrate group name with its percentage of the total.



Family TC#	Family Name	Hit TCID	Acc.In TCDB	Hit TMS#	Substrate Group	Specific Substrate	Sequence ID#	Query TMS#
1.A.9	the neurotransmitter receptor, cys loop, ligand-gated ion channel (lic) family.	1.A.9.5.2	O95166	1	Anion	Unknown	An07g10020	1
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.1.4	O67997	12	Cation	Ammonia	An08g03200	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.1	P40260	11	Unknown	Unknown	An08g03200	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.2	P41948	11	Unknown	Unknown	An08g03200	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.2	P41948	11	Unknown	Unknown	An14g02390	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.3	Q8NKD5	11	Cation	NH4+	An08g03200	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.3	Q8NKD5	11	Cation	NH4+	An14g02390	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.4	Q96UY0	11	Unknown	Unknown	An08g03200	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.4	Q96UY0	11	Unknown	Unknown	An14g02390	11
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.3.5	Q59UP8	11	Cation	NH4+	An08g03200	11

Figure 24: Page of Results of TransATH for *A.niger* CBS513.88

This is a beta version of TransATH. To date, there are 467 TCIDs from the TCDB that map to information on their substrate groups and specific substrates. There are 32 substrate groups identified to date, including the *Unknown* group. This preprocessing was done manually for the beta implementation of TransATH. In future we will extract the roughly 4000

entries available in merlin [DRFR15] which were also manually collected from the TCDB. The beta version of the implementation does not use the web services of TM-Coffee and LocTree3 yet. HMMTOP is used to compute the TMS, and localization information is not yet available. Furthermore, the facility to be notified by email does not function yet. The system will in future notify users when jobs complete and provide a link to the result page of the job.

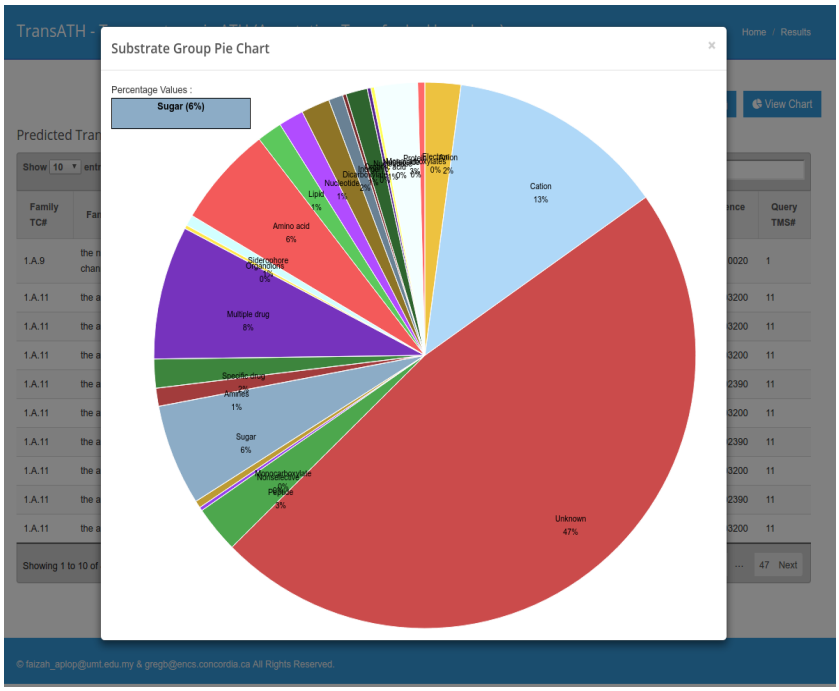


Figure 25: Pie Chart of TransATH Predictions for *A.niger* CBS513.88

4.5 Evaluation

This section addresses two questions. The first question is what is the impact of the choice of thresholds for TCDB-Blast on its performance? In particular, how do our choice of thresholds affect performance relative to G-Blast(v2)? Section 4.5.1 addresses this question using a gold standard set of transporters and non-transporters from *S. cerevisiae*. Section 4.5.2 presents the impact of the choice of thresholds on the genome of *A. niger* CBS 513.88. The second question is how do we evaluate the performance of TransATH? Section 4.5.3 addresses this question.

4.5.1 Thresholds for TCDB-Blast

G-Blast(v2) and TCDB-Blast both use blastp to search the TCDB for hits of protein sequences of a genome. G-Blast(v2) sets an e-value cut-off of e-3 for its main search, and then a lenient cut-off of e-1 when searching for putative transporters. Note that in this thesis the exponent is always base 10, so e-3 is 0.001 which is 10^{-3}. G-Blast(v2) does not apply thresholds to the other parameters. In Section 4.3 TCDB-Blast requires each of the following thresholds to be met: e-value 1e-20; percent alignment 70%; query coverage 70%; subject coverage 70%; and difference in length of 10%.

This section answers the following questions: What is the effect of using other thresholds? How does TCDB-Blast compare to G-Blast(v2)?

For the evaluation we took the gold standard dataset used by [BH13, Table S3] of 177 transporters in *S. cervisiae* that have been experimentally characterized. These were the positive examples in the dataset. A set for negative examples of size 177 was chosen at random from *S. cervisiae* at SGD (`http://www.yeastgenome.org`) taking care to avoid entries in the positive set and transmembrane proteins. The gold standard dataset of positives and negatives was compared against the 11,572 entries of the TCDB as of May 2014.

Table 31 shows the effect of different e-value cut-offs for the blastp search using no other thresholds. The impact of the a more stringent threshold has minimal effect on the number of results for transporters. However, for non-transporters there is a noticeable effect at e-3, e-10, and e-30.

Cut-Off	e-1	e-3	e-5	e-10	e-20	e-30	e-50
Results for Transporters	177	177	176	176	175	174	174
Results for Non-Transporters	37	23	22	17	14	10	9

Table 31: Effect of e-value Cut-off

The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with the given e-value cut-off. No other thresholds were set.

BLAST returns a local alignment. By default the alignment has gaps. Gap to amino acid alignments are ignored in two statistics of interest: percent identity and percent alignment. The percent identity of the alignment is the percentage of the aligned region where the two aligned amino acids are identical. A related statistic is the percent alignment which is the number of amino acid to amino acid alignments (not necessarily identical) divided by the

length of the alignment (including gaps). Table 32 shows the effect of different thresholds for percent alignment. Table 33 shows the effect of different thresholds for percent identity. Clearly there is no impact of the threshold for percent alignment. For percent identity the most noticeable effect on transporters occurs at a threshold of 50%, while for non-transporters there is a large impact at a threshold of 50% and a lesser impact at a threshold of 60%.

Threshold	30	40	50	60	70	80	90
Results for Transporters	177	177	177	177	177	177	177
Results for Non-Transporters	37	37	37	37	37	37	36

Table 32: Effect of Percent Alignment

The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with the given percent alignment threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Threshold	30	40	50	60	70	80	90
Results for Transporters	175	175	170	169	167	162	160
Results for Non-Transporters	23	23	10	6	6	6	6

Table 33: Effect of Percent Identity

The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with the given percent identity threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Query coverage is the percentage of the query sequence that is included in the alignment. Table 34 shows the effect of different thresholds for query coverage. The impact is relatively minor for transporters and non-transporters. There is a noticeable effect for non-transporters at a threshold of 80% coverage.

Threshold	50	60	70	80	90
Results for Transporters	175	174	173	172	172
Results for Non-Transporters	17	16	15	12	12

Table 34: Effect of Coverage Threshold

The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with the given query coverage threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Percent difference is the percentage that the query sequence the subject sequence differ in

length. Table 35 shows the effect of different thresholds for percent difference. The impact is relatively minor for transporters and non-transporters.

Threshold	20	15	10	5
Results for Transporters	177	176	176	175
Results for Non-Transporters	19	18	18	16

Table 35: Effect of Percent Difference Threshold
The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with the given percent difference threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

The effect of each parameter is monotonic: as we make the parameter more stringent we obtain fewer results because more sequences are filtered out. However, there are some changes in thresholds for parameters that have a noticeable effect, mainly on the results for non-transporters than for transporters. Table 31 suggests using a threshold for e-value of e-30 rather than e-20. Table 33 suggests using a threshold for percent identity of 50% or 60% rather than 70%. Table 34 suggests using a threshold for query coverage of 80% rather than 70%. Table 36 and Table 37show the results for different combinations of parameter thresholds. They include the F-measure for each combination:

$$F = 2 * TP/(2 * TP + FP + FN)$$

where TP is the number of true positives, FP the number of false positives, and FN the number of false negatives. Table 36 and Table 37 compare G-Blast(v2) and TCDB-Blast. Table 37 shows the optimal thresholds for TCDB-Blast. The optimal thresholds for TCDB-Blast use 60% as the threshold for percent identity. The other suggested threshold values have no effect on the results. With the optimal thresholds, TCDB-Blast achieves an F-measure of 95.73% which is slightly better than the F-measure of 93.90% achieved by G-Blast(v2).

4.5.2 Thresholds of TCDB-Blast for *A. niger* CBS 513.88

This section explores how the choice of thresholds impacts the results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with different combination of thresholds. The threshold for percent alignment has minimal impact. The threshold for percent identity has a major impact and

G-Blast(v2)				Transporters	Non-Transporters	F-measure
e-value	%ID	QCov	Diff			
e-1	0	0	100	177	37	90.54
e-3	0	0	100	177	23	**93.90**

Table 36: F-Measures for G-Blast(v2) Predictions for Combinations of Thresholds
The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with different combination of thresholds. In this trial neither G-Blast(v2) nor TCDB-Blast removed sequences without transmembrane segments. G-Blast(v2) uses an initial e-value threshold of e-3 for transporters, and then a threshold of e-1 for putative transporters. The table shows the effect of both thresholds. G-Blast(v2) does not explicity constrain percent identity, query coverage, and percent difference, so the table shows the default values for these parameters that do not filter out any alignments. **Bold** indicates the maximum F-measure.

greatly limits the number of results. The remaining thresholds have a gradual impact as they are made more stringent. Table 39 shows the effect of different e-value cut-offs for the blastp search using no other thresholds. Table 40 shows the effect of different thresholds for percent alignment. Table 41 shows the effect of different thresholds for percent identity. Table 42 shows the effect of different thresholds for query coverage. Table 43 shows the effect of different thresholds for percent difference.

Table 38 shows the results for different combinations of parameter thresholds. It highlights the impact of the threshold for percent identity. It suggests a threshold of 40% be used rather than the threshold of 60% found to be optimal in the previous evaluation in Section 4.5.1.

4.5.3 Correctness of TransATH

The methodology used to determine the correctness of the predictions by TransATH in Table 30 was to compare the predictions with the high confidence annotations for transporters in the AspGD database.

The AspGD is a well-curated database. Annotation information is recorded in terms of the Gene Ontology. The curators read the literature in order to assess which evidence code to assign to a Gene Ontology term. The experimental evidence codes of Inferred from Experiment (EXP), Inferred from Direct Assay (IDA), Inferred from Physical Interaction (IPI), Inferred from Mutant Phenotype (IMP), Inferred from Genetic Interaction (IGI), and Inferred from

TCDB-Blast				Transporters	Non-Transporters	F-measure
e-value	%ID	QCov	Diff			
e-20	70	70	10	166	6	95.13
e-20	60	70	10	168	6	**95.73**
e-20	50	70	10	169	8	95.48
e-20	40	70	10	169	9	95.21
e-30	70	70	10	166	6	95.13
e-30	60	70	10	168	6	**95.73**
e-30	50	70	10	169	8	95.48
e-30	40	70	10	169	9	95.21
e-30	70	80	10	166	6	95.13
e-30	60	80	10	168	6	**95.73**
e-30	50	80	10	169	8	95.48
e-30	40	80	10	169	9	95.21

Table 37: F-Measures for Prediction using Combinations of Thresholds
The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with different combination of thresholds. In this trial neither G-Blast(v2) nor TCDB-Blast removed sequences without transmembrane segments. For TCDB-Blast uses default thresholds of e-20, 70%, 70%, and 10% for e-value, percent identity, query coverage, and percent difference, respectively. The effect of modifying the threshold for percent identity is shown in the first block. The effect of using e-30 as the threshold for e-value is shown in the second block. The effect of modifying the threshold for query coverage is shown in the third block. **Bold** indicates the maximum F-measure.

Expression Pattern (IEP) indicate the inference by the curators from the experimental evidence presented in the literature. In addition the team at AspGD has compared the genomes of the *Aspergillus* genomes and other well-curated fungal genomes to create high confidence orthology mappings between the genomes. They use this to assign GO terms based on orthology. Although they assign the evidence code Inferred from Electronic Annotation (IEA) to the GO term, the source indicates the orthologous gene that is experimentally characterized. In addition there are the GO terms with evidence code IEA where the source is an InterPro entry. This indicates an inference because an InterPro domain was located on the sequence.

The TCDB as of May 2014 has 9 entries from *A. niger* CBS 513.88 as shown in Table 44.

The high confidence AspGD annotations for transporters were determined by downloading the `gene_association.aspgd` file from the AspGD website at `http://www.aspgd.org`. The entries pertaining to *A. niger* CBS 513.88 were extracted and cross-referenced with the set

TCDB-Blast				Results
e-value	%ID	QCov	Diff	
e-20	70	70	10	55
e-20	60	70	10	93
e-20	50	70	10	170
e-20	40	70	10	321
e-20	30	70	10	696

Table 38: *A. niger* CBS 513.88 Predictions using Combinations of Thresholds
The number of results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with different combination of thresholds.

Cut-Off	e-1	e-3	e-5	e-10	e-20	e-30	e-50
Results	2803	2108	1866	1576	1295	1124	833

Table 39: Effect of e-value Cut-off
The number of results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with the given e-value cut-off. No other thresholds were set.

Threshold	30	40	50	60	70	80	90
Results	2803	2803	2803	2803	2803	2794	2661

Table 40: Effect of Percent Alignment
The number of results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with the given percent alignment threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Threshold	30	40	50	60	70	80	90
Results	2052	2052	300	124	65	28	13

Table 41: Effect of Percent Identity
The number of results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with the given percent identity threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Threshold	50	60	70	80	90
Results	1593	1447	1291	1117	834

Table 42: Effect of Coverage Threshold
The number of results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with the given query coverage threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Threshold	20	15	10	5
Results	1722	1576	1424	1194

Table 43: Effect of Percent Difference Threshold
The number of results when using blastp to search the 14,067 protein sequences of *A. niger* CBS 513.88 against the 11,572 entries of the TCDB as of May 2014 with the given percent difference threshold. An e-value cut-off of e-1 was used. No other thresholds were set.

Gene	TCID	UniProt	Substrate Group	Specific Substrate
An04g00670	3.A.19.1.2	A2QHQ3	Protein	Protein
An05g01290	2.A.1.1.58	Q8J0U9	Sugar	Glucose:H+
An07g06140	9.B.7.2.3	E2PST1	Protein	Unknown
An07g08960	1.H.1.4.3	G3XZI4	Unknown	Unknown
An09g01910	2.A.1.2.48	A2QTF4	Specific drug	Tetracycline
An11g03330	1.A.88.1.4	A2QW01	Cation	K+
An12g00870	2.A.16.4.1	A2QYD7	Unknown	Unknown
An12g07450	2.A.1.1.57	Q8J0V1	Sugar	Monosaccharides
An16g08040	1.B.69.1.4	A2R8R0	Peptide	Unknown

Table 44: TCDB Entries from *A. niger* CBS 513.88
The table shows information for the 9 TCDB entries that come from *A. niger* CBS 513.88. The Gene column shows the gene identifier in AspGD. The TCID column shows the identifier in the TCDB. The UniProt column shows the identifier in UniProt. The Substrate Group column shows the type of substrate transported, as known by TCDB. The Specific Substrate column shows the specific substrate transported, as known by TCDB. As of May 2014.

of all GO terms in BP (Biological Process) and MF (Molecular Function) in the subtree of GO:0006810(transport) from BP and GO:0005215(transporter activity) from MF. The GO terms with experimental evidence codes and the GO terms that had IEA evidence code and were derived by orthology were extracted to give the final list of high confidence annotations for transporters in *A. niger* CBS 513.88. The list contained 242 GO terms for 190 individual genes. Table 45 shows the information for the 10 genes with expermental evdidence.

From the total 242 GO terms for 190 genes only a few include detail about the substrate being transported. Table 46 shows the 33 GO terms for Molecular Function for 30 genes where information about the substrate being transported is given.

Of the nine genes from *A. niger* CBS 513.88 that are entries in the TCDB as of May 2014, only three have high confidence GO term annotations relating to transport in the AspGD as shown in Table 47.

Gene	GO ID	Description	Code	Source	Domain
An12g07450	GO:0034219	carbohydrate transmembrane transport	IDA	PMID:14717659	P
An12g07450	GO:0034219	carbohydrate transmembrane transport	IMP	PMID:14717659	P
An14g03790	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An11g09910	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An01g03190	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An03g04215	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An12g07570	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An14g00010	GO:0006886	intracellular protein transport	IMP	PMID:11489135	P
An14g00010	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An12g01190	GO:0016192	vesicle-mediated transport	IMP	PMID:24295824	P
An02g08670	GO:0090481	pyrimidine nucleotide-sugar transmembrane transport	IGI		P
An06g00300	GO:0090481	pyrimidine nucleotide-sugar transmembrane transport	IGI		P

Table 45: Transport GO Entries with Experimental Evidence for *A. niger* CBS 513.88 The table shows information for the genes from *A. niger* CBS 513.88 with transport-related GO terms supported by experimental evidence. The Gene column shows the gene identifier in AspGD. The GO ID column shows the Gene Ontology identifier for the GO term. The Description column shows the short description of the GO term. The Code column shows the evidence code for the GO term as curated by AspGD. The Source column shows the source of the evidence. The Domain column shows the GO domain BP(P), MF(F), CC(C) of the GO term. As curated in the AspGD as of 28 March 2016.

For the evaluation TransATH was run at `transath.umt.edu.my` using the thresholds: e-value 1e-20; percent identity 40%; query coverage 70%; subject coverage 70%; and difference in length of 10%. The TCDB as of May 2014 was used. Sequences in the TCDB and in the *A. niger* CBS 513.88 genome without transmembrane segments were filtered out.

In total TransATH returned predictions for 221 sequences in the *A. niger* CBS 513.88 genome. Of these 52 were matches to the 190 genes that had high confidence GO terms related to transport according to AspGD. Another 85 of the 190 genes had blastp hits to TCDB sequences that fell below the thresholds set for this evaluation. A further 20 genes with predictions by TransATH that did not have high confidence GO terms for transport in the AspGD had GO terms for transport inferred from InterPro domain hits in AspGD. In summary 157 of the 221 sequences in the *A. niger* CBS 513.88 genome for which TransATH returned a prediction had good corroborating evidence in the AspGD that they were transporters.

Gene	GO ID	Description	Code	Source	Domain
An01g00720	GO:0042929	ferrichrome transporter activity	IEA	CGD:CAL0000196424	F
An01g03640	GO:0008565	protein transporter activity	IEA	SGD:S000003530	F
An01g08400	GO:0008565	protein transporter activity	IEA	SGD:S000005595	F
An01g14510	GO:0008526	phosphatidylinositol transporter activity	IEA	SGD:S000004372	F
An02g03540	GO:0005358	high-affinity hydrogen:glucose symporter activity	IEA	PomBase:SPBC4B4.08	F
An02g03540	GO:0055054	fructose:proton symporter activity	IEA	PomBase:SPBC4B4.08	F
An02g04260	GO:0008565	protein transporter activity	IEA	SGD:S000003413	F
An02g07570	GO:0015248	sterol transporter activity	IEA	SGD:S000006066	F
An02g13460	GO:0051183	vitamin transporter activity	IEA	SGD:S000003154	F
An03g01800	GO:0005324	long-chain fatty acid transporter activity	IEA	SGD:S000003269	F
An04g01190	GO:0051724	NAD transporter activity	IEA	SGD:S000001268	F
An05g01660	GO:0015244	fluconazole transporter activity	IEA	CGD:CAL0000186516	F
An07g09190	GO:0005324	long-chain fatty acid transporter activity	IEA	SGD:S000000245	F
An08g01030	GO:0008565	protein transporter activity	IEA	SGD:S000001054	F
An10g00500	GO:0008565	protein transporter activity	IEA	SGD:S000007256	F
An11g03640	GO:0015198	oligopeptide transporter activity	IEA	SGD:S000006398	F
An11g05000	GO:0046624	sphingolipid transporter activity	IEA	SGD:S000005927	F
An12g01210	GO:0042937	tripeptide transporter activity	IEA	SGD:S000001801	F
An14g06210	GO:0008526	phosphatidylinositol transporter activity	IEA	SGD:S000005175	F
An15g02930	GO:0015244	fluconazole transporter activity	IEA	CGD:CAL0000186516	F
An15g07460	GO:0042937	tripeptide transporter activity	IEA	CGD:CAL0000191307	F
An15g07510	GO:0042937	tripeptide transporter activity	IEA	CGD:CAL0000200802	F
An15g07510	GO:0042936	dipeptide transporter activity	IEA	CGD:CAL0000200802	F
An16g03590	GO:0008526	phosphatidylinositol transporter activity	IEA	PomBase:SPAC3H8.10	F
An16g03590	GO:0008525	phosphatidylcholine transporter activity	IEA	PomBase:SPAC3H8.10	F
An16g04270	GO:0008565	protein transporter activity	IEA	SGD:S000003690	F
An16g08830	GO:0008565	protein transporter activity	IEA	SGD:S000000375	F
An17g00560	GO:0008565	protein transporter activity	IEA	SGD:S000005658	F
An18g04110	GO:0008565	protein transporter activity	IEA	SGD:S000006046	F
An18g04910	GO:0032217	riboflavin transporter activity	IEA	SGD:S000005833	F
An12g07450	GO:0005358	high-affinity hydrogen:glucose symporter activity	IEA	AspGD:ASPL0000073615	F
An01g11630	GO:0008565	protein transporter activity	IEA	SGD:S000002493	F
An03g04340	GO:0015197	peptide transporter activity	IEA	SGD:S000004370	F

Table 46: Transport GO MF Entries with Substrate Information for *A. niger* CBS 513.88 The table shows GO MF information for the genes from *A. niger* CBS 513.88 with high confidence transport-related GO terms that include information on the substrate. The Gene column shows the gene identifier in AspGD. The GO ID column shows the Gene Ontology identifier for the GO term. The Description column shows the short description of the GO term. The Code column shows the evidence code for the GO term as curated by AspGD. The Source column shows the source of the evidence. The Domain column shows the GO domain BP(P), MF(F), CC(C) of the GO term. As curated in the AspGD as of 28 March 2016.

Gene	GO ID	Description	Code	Source	Domain
An12g07450	GO:0005358	high-affinity hydrogen:glucose symporter activity	IEA	AspGD:ASPL0000073615	F
An02g03540	GO:0005358	high-affinity hydrogen:glucose symporter activity	IEA	PomBase:SPBC4B4.08	F
An02g03540	GO:0055054	fructose:proton symporter activity	IEA	PomBase:SPBC4B4.08	F
An12g00870	GO:0000316	sulfite transport	IEA	AspGD:ASPL0000109974	P

Table 47: Transport GO Entries with TCDB Entries for *A. niger* CBS 513.88
The table shows the available high confidence GO terms in AspGD for the nine TCDB entries from *A. niger* CBS 513.88. The Gene column shows the gene identifier in AspGD. The GO ID column shows the Gene Ontology identifier for the GO term. The Description column shows the short description of the GO term. The Code column shows the evidence code for the GO term as curated by AspGD. The Source column shows the source of the evidence. The Domain column shows the GO domain BP(P), MF(F), CC(C) of the GO term. Note that only 3 of the 9 genes have high confidence GO terms relating to transport. Note that An02g03540 appears to have superceded An05g01290 in the genome. As of 28 March 2016.

For the 30 genes in Table 46 with information on the substrate transported, TransATH returned predictions for 11 of the 30 genes. Another 9 of the 30 genes had blastp hits to TCDB sequences that fell below the thresholds set for this evaluation. Table 48 shows the TransATH predictions for the 11 genes for comparison with the information in Table 46. For 9 of the 11 genes with predictions from TransATH and in Table 46 there is agreement on the substrate transported, while for the other two (An05g01660 and An15g02930) there is agreement at the Substrate Group level if *fluconazole* is considered a *Multiple Drug*.

In conclusion, at the level of predicting transporter versus non-transporter, TransATH was correct at least for 157 of the 221 sequences predicted to be transporters; that is, there was had good corroborating evidence in the AspGD that they were transporters. This is at least 71.0% of the predictions were correct. Keep in mind that 43.7% (6141/14067) of genes in the *A. niger* CBS 513.88 genome have no annotation.

At the level of predicting substrate, TransATH returned predictions for 11 of the 30 genes in Table 46 with information on the substrate transported. For 9 of the 11 there was good agreement on the substrate, and for the other 2 there was plausible evidence that the predictions were correct at the level of Substrate Group.

Family	Family Name	TCID	Hit	HTMS	Substrate Group	Specific Substrate	Query	QTMS
2.A.1	major facilitator superfamily (mfs).	2.A.1.16.1	P39980	15	Siderophore	Ferroxamine	An01g00720	14
3.A.5	general secretory pathway (sec) family.	3.A.5.9.1	P60059	1	Protein	Protein	An01g11630	1
2.A.1	major facilitator superfamily (mfs).	2.A.1.1.36	Q400D8	12	Unknown	Unknown	An02g03540	12
2.A.1	major facilitator superfamily (mfs).	2.A.1.1.58	Q8J0U9	12	Sugar	Glucose:H+	An02g03540	12
2.A.1	major facilitator superfamily (mfs).	2.A.1.1.108	P32465	12	Unknown	Unknown	An02g03540	12
3.A.5	general secretory pathway (sec) family.	3.A.5.8.1	P32915	12	Protein	Peptide	An03g04340	10
3.A.5	general secretory pathway (sec) family.	3.A.5.9.1	Q9H9S3	10	Protein	Protein	An03g04340	10
3.A.5	general secretory pathway (sec) family.	3.A.5.9.1	P61619	12	Protein	Protein	An03g04340	10
2.A.29	mitochondrial carrier (mc) family.	2.A.29.10.5	P40556	4	Nucleotide	NAD+, pyruvate	An04g01190	4
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.1	P33302	15	Unknown	Unknown	An05g01660	11
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.4	P43071	13	Multiple drug	Unknown	An05g01660	11
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An05g01660	11
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.11	P41820	13	Unknown	Unknown	An05g01660	11
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.12	P51533	15	Unknown	Unknown	An05g01660	11
2.A.67	oligopeptide transporter (opt) family.	2.A.67.1.5	O14031	15	Peptide	Glutathione	An11g03640	15
2.A.6	resistance-nodulation-cell division (rnd) superfamily.	2.A.6.6.3	Q12200	13	Lipid	Sphingolipid	An11g05000	13
2.A.17	proton-dependent oligopeptide transporter (pot) family.	2.A.17.2.1	Q9P380	12	Peptide	Unknown	An12g01210	11
2.A.17	proton-dependent oligopeptide transporter (pot) family.	2.A.17.2.2	P32901	12	Peptide	dipeptide, tripeptide	An12g01210	11
2.A.1	major facilitator superfamily (mfs).	2.A.1.1.51	Q2MEV7	12	Sugar	Glucose/Xylose	An12g07450	12
2.A.1	major facilitator superfamily (mfs).	2.A.1.1.57	Q8J0V1	12	Sugar	Monosaccharides	An12g07450	12
2.A.1	major facilitator superfamily (mfs).	2.A.1.1.68	A3M0N3	12	Sugar	Glucose	An12g07450	12
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.1	P33302	15	Unknown	Unknown	An15g02930	16
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.4	P43071	13	Multiple drug	Unknown	An15g02930	16
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.5	P78595	11	Multiple drug	Phospholipid	An15g02930	16
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.11	P41820	13	Unknown	Unknown	An15g02930	16
3.A.1	atp-binding cassette (abc) superfamily.	3.A.1.205.12	P51533	15	Unknown	Unknown	An15g02930	16

Table 48: TransATH Predictions for Genes with Substrate Information

The table shows the TransATH predictions for the genes from *A. niger* CBS 513.88 with information about substrates available from the high confidence GO terms in AspGD related to transport. The columns Family and Family Name contain the TC-Family identifier and its name. The TCID column shows the identifier in the TCDB. The column Query is the identifier for the entry in the *A. niger* CBS 513.88 genome. The column Hit is the UniProtKB identifier for the matching TCDB entry. The columns QTMS and HTMS contain the number of TMS for the query and the hit, respectively, as determined by HMMTOP. The Substrate Group column shows the type of substrate transported, as known by TCDB. The Specific Substrate column shows the specific substrate transported, as known by TCDB. As of 28 March 2016.

4.6 Predicting Specific Substrates

This section explores a number of approaches to solving the problem of predicting the specific substrates transported across a membrane by a given transmembrane transporter protein. The review of the state of the art in Section 4.2 does not discover any predictor for transporters that works at this level of specificity. Furthermore, the known examples of transporters, as illustrated in Section 2.3, show that the specific substrate is determined by only a few residues in the protein sequence. Hence, on the face of this evidence, the task is likely to be difficult.

For this work we focus on sugar porters.

The techniques that hold promise for the exploration are

- multiple sequence alignment (MSA);
- profile HMM;
- identifying clades in phylogenetic trees [Wu15];
- amino acid composition and the various alphabets for amino acids (Section 2.5.3);
- multilevel alphabets [HKMG13]; and
- identifying specificity-determining positions [CC14].

Multiple sequence alignments are at the heart of many techniques to explore protein families. By considering several members of the protein family rather than a single member, or pair of members in a pairwise sequence alignment, the MSA hopes to amplify the signal in the sequences that characterize the family. However, the MSA algorithms do not guarantee an optimal alignment, and they differ in the alignment that they do compute. So the choice of MSA algorithm can play a major role in the effectiveness of the downstream application. The MSA algorithms that we consider are

- Clustal Omega [SWD+11], the latest in the Clustal series of algorithms, which is fast and scaleable, and capable of aligning 10,000 or more sequences;
- MAFFT [KS13], which was used in phylogenetic analysis of the GH10 xylanase enzymes [Wu15];
- AQUA [MCT+10], which considers alignments from MAFFT and MUSCLE [Edg04], refines them with RASCAL [TTP03], assesses the refined alignments with NorMD [TPR+01], and which is used at EMBL to construct eggNOG [PFS+13]; and

- PipeAlign [PBB+03], which is a pipeline — no longer available — for constructing protein families, constructing a MSA, and identifying subfamilies and adjusting the MSA to reflect the distinguishing features of the subfamilies within the family.

For this work the algorithm should produce an alignment consistent with the TMS regions within the full MSA, as the TMS contain the specificity determining residues.

For profile Hidden Markov Models (HMM) we use the HMMER package [Edd98]. We use *hmmbuild* to train HMMs, given a MSA, and subsequently use *hmmscan* to scan protein sequences against trained HMMs. In their work on TransportTP, Li et al (2009) [LBUZ09] only build HMMs for TC families of size at least 5. However, they only achieve precision and recall greater than 70% for families of size greater than 15.

In the phylogenetic analysis of the GH10 xylanase enzymes [Wu15] a Maximum Likelihood tree is constructed using RAxML [Sta14] with a bootstrap value of 1000 to estimate branch support. Subfamilies are based on the topology of the phylogenetic tree requiring boostrap support of at least 55%. Subfamilies are validated by considering average percent identity of pairwise alignments within a subfamily and between subfamilies. This analysis is done by aligning the catalytic domains of the enzymes only, and not the full protein sequence. For transporters, that is, sugar porters, we consider the alignment of the Prosite *Sugar_-Transport_1* domain.

Section 2.5.3 introduces the many variations on amino acid composition and the various alphabets for amino acids. Let us be precise about the definition of each of those we consider, and let $C(x)$ denote a generic amino acid composition function for a protein sequence x. The composition functions that we use are as follows. The *length* function L is defined as

$$L(x) = |x|, \text{the number of amino acids in } x. \tag{1}$$

The *amino acid composition* function AAC is a vector of length 20 defined as

$$AAC(x)[a] = |\{i : a = x[i]\}|/L(x). \tag{2}$$

The *pair amino acid composition* function $PAAC$ is a vector of length 400 defined as

$$PAAC(x)[aa'] = |\{i : aa' = x[i..i+1]\}|/(L(x) - 1). \tag{3}$$

Helms work [SCH10] shows that there is no gain to be had by considering the more complicated variations of amino acid composition. However, their work [SH12] obtains a 10% improvement by considering the amino acid composition of the TMS and non-TMS regions of the protein individually. Let us define, for an amino acid composition function C and a protein x, $C_{TMS}(x)$ as the value of C when restricted to the TMS segments of x, and $C_{\overline{TMS}}(x)$ as the value of C when restricted to the non-TMS segments of x.

It is convenient to record the number of TMS in the protein, so define

$$TMS(x) = \text{number of transmembrane segments in } x. \tag{4}$$

When we need to be precise, we will indicate the method M such as HMMTOP or TMHMM used to determine the TMS and denote this as

$$TMS_M(x) = \text{number of transmembrane segments in } x \text{ as computed by method } M, \tag{5}$$

and use $TMS(x)$ to be the number of TMS as curated in SwissProt.

The feature vector that we consider, based on the lessons from Helms work, is

$$L(x).TMS(x).AAC(x).PAAC_{TMS}(x).PAAC_{\overline{TMS}}(x) \tag{6}$$

where . is the concatenation operator. This is a vector of length 822.

An alphabet in Section 2.5.3 is a translation function t from the set of amino acids A to a set of symbols S. The translation may be applied to the protein sequence x to yield a sequence $t(x)$ of symbols. The composition of $t(x)$ in terms of the frequency of each symbol s, or each pair of symbols ss', can be determined directly from $t(x)$ or by "translating" the composition vector of x. Let $C_{t:A\to S}(x)$ denote the composition of $t(x)$, then

$$C_{t:A\to S}(x)[s] = \sum_{a\in t^{-1}(s)} C(x)[a]. \tag{7}$$

As a shorthand, we will write $C_t(x)$ or $C(t(x))$.

In 2013 Hod et al. [HKMG13] introduced the concept of a *multilevel alphabet* to protein sequence analysis from the field of signal processing. They solved a difficult problem of finding short motifs by encoding several alphabets for amino acids with information on secondary

structure and surface accessibility into a single alphabet, and then applying MEME to the translated sequences, in order to find the motifs. For transporters, the TMS represent the secondary structure, and Helms work has shown the importance of using properties of both the TMS and non-TMS regions of the protein sequence. So the approach of Hod et al. appears to be a way to generalize Helms work to use amino acid composition and various alphabets together.

For families of enzymes, there is much success at determining which positions and residues within the catalytic domain are the active site, based on knowledge from 3D structures of enzymes and enzymes bound to their ligands (substrates). Many prediction methods for the specificity-determining positions and specificity-determining residues exist [CC14], including recent work predicting detailed enzyme function [NNM14]. There is no predictor specifically designed for transporters; however, the survey [CC14] does compare existing predictors on a dataset of transporters, amongst its numerous comparisons. Unfortunately, their comparison does not reveal any significant difference in performance between the predictors. Hence, any predictor is as good a choice as the next for our exploration. Of course, these methods are strongly dependent on the MSA.

4.7 A New Computational Framework

This section presents a proposal for a way forward for the prediction of transport that attempts to cope with a number of inherent problems to the field. The problems are

- The Transporter Classification (TC) and the TCDB are the official collectors and describers of transporters. As such they act as the final arbiters of knowledge about transporters. However, the state of the TCDB does not provide vital information for GENRE such as database fields for substrates and transporter reactions.

- GENREs and researchers work with transporters in terms of the specific substrates that they transport, the mechanism of transport, and the localization of the transport reaction. These two perspectives, namely TC and substrate, need to be reconciled. In particular, there needs to be an official standard for naming substrates, and classes of substrates. This role could be filled by ChEBI. Furthermore, there needs to be a standard identifier for transport reactions; a role which might be taken up by the TC, or BioPAX, or BiGG.

- The datasets are small, as experimentally characterized transporters are small in number, and their number is very variable across the different TC families.
- The task is hierarchical. One reason for this hierarchy is the need to aggregate data on specific substrates in order to have a dataset for a "group" of substrates that is sufficiently large for the purpose of machine learning. A second reason is that biology organizes knowledge hierarchically as a way to deal with complexity. A third reason is the need to summarize the knowledge on all the transporters in a genome; this may involve information on a subset of 500 genes in a genome of 12,000 genes.
- The task is multi-label classification. That is, a transporter may facilitate the movement of more than a single substrate. We have examples of sugar transporters which transport four substrates, although with different levels of efficiency.

Therefore the challenge is to predict as much as we can about the transporters in a genome, as precisely and as reliably as we can, given the available data or knowledge about transporters. So the machine learning problem

1. adapts to the amount of available data (and its predictive power);
2. measures reliability of predictions, so it can determine whether the available data is sufficient for this purpose;
3. seeks to make a prediction that is as precise as possible (in the hierarchy), given the need to be reliable;
4. seeks to include multiple labels in the prediction, where possible, in recognition that this is a multi-label classification task; and
5. identifies those niches amongst the space of transporters where the available data supports precise and reliable prediction.

Given a suitable framework, then our ability to make predictions should improve as the dataset of experimentally characterized transporters increases.

There are several hierarchies related to the framework. There is the protein family organization in the Transporter Classification (TC) of transporter versus non-transporter, TC superfamily, TC family, and TC subfamily. There are the various groupings of specific substrates that could be organized into hierarchies; for example, sugar, monosaccharide, pentose,

arabinose, D-arabinose. There is the hierarchy in the Gene Ontology terms for transporters, which individually captures mechanism, substrate, and localization. A part of the GO hierarchy mirrors a substrate hierarchy. The GO terms cross-reference to ChEBI when they specify a substrate.

Hierarchical multi-label classification [VSS+08] is often transformed into other tasks [SJF11] or performed incrementally [CBGZ06]. However, hierarchical multi-label classification can be performed directly using traditional machine learning techniques such as genetic algorithms [CBdC12], neural networks [CBDC14], decision trees [VSS+08], SVM [RSSST06], and ensembles [ZSK14].

One inspiration for the proposed framework in this section is the history of hierarchical multi-label classification for predicting gene function where it has been occasionally used in the context of a single hierarchy, such as FunCat from MIPs, or directed-acyclic graph, such as the Gene Ontology [BST06, SVS+10, BCFdC13, SCMD13, FF+14].

However, in our task there are multiple hierarchies, which may complicate the classification problem. Nevertheless, another inspiration is the recent harmonization effort [CVP+15] of the TCDB, GO, and Pfam which illustrates how to relate the hierarchies. This effort should help address the difficulty of comparing predictors that target the TC with those that target substrates.

The framework takes a relational view of the available dataset and the properties of the transporters. The framework is a new "twist" on the feature vector approach of TransportTP. TransportTP adopts a somewhat complicated hybrid approach where its algorithm is a series of phases. It uses amino acid composition, Pfam domains, and GO terms amongst the features. The feature space can be structured as a relational space, and relations can represent the associations between the various hierarchies.

In this dataspace, requirement (5) needs to identify a niche, which we call a *transporter cluster* T, that is as specific as possible, given the available data, and that is a group of related transport proteins. The classification task for the transporter cluster T is to extract characterizing features of T in order to be able to classify query proteins into the cluster T. Sometimes the cluster may be a substrate category, sometimes a TC family, sometimes a TC subfamily, and maybe a specific substrate.

The proposed computational framework for the transporter prediction problem is *multi-hierarchical multi-label classification* using a *relational* dataspace.

For the solution of the classification problem, a proposed way to proceed is to first identify a transporter cluster T, and then develop a profile HMM classifier from a suitable multiple sequence alignment MSA of the protein sequences in the cluster T. One would want the MSA to conserve the topology by aligning TMS with TMS. One would also want the MSA to align specificity-determining residues so that information on those positions and residues are incorporated into the profile HMM, even if one did not explicitly run a specificity-determining residue method.

There are many clustering techniques that one might apply to identify a transporter cluster. From the relational dataspace representation, one is able to transform the representation into a feature vector, a network, or relations, and thus apply techniques from data mining, graph mining, and machine learning to identify clusters.

4.7.1 The Relational Dataspace

The dataspace represents the knowledge about a set of proteins, their properties, and their classifications. The classifications of interest are

- the Transporter Classification (TC); and
- the Gene Ontology (GO).

Important properties for proteins are

- the Pfam domains of the protein;
- the TMS of the protein; and
- the amino acid composition of the protein;

in particular, but the list of properties is open-ended. The proteins of interest are those with curated information, such as

- the proteins in SwissProt;
- the proteins in TCDB; and
- the proteins in genomes that are well-curated.

Most of the proteins of interest will be in SwissProt, but some will be in UniProtKB and unreviewed.

The information is modeled as relations.

Proteins are represented by their UniProt Identifier *pid*.

Transporter Classification information from the TCDB is encoded as a set of relations:

TC(pid, TCID, TC_subfamily, TC_family, TC_superfamily)
TCsubstrate(pid, TCID, substrate_group, specific_substrate)

These entries will define the "standard" names for substrates and define the hierarchy for one level of substrates, as well as the hierarchy of TC families.

Gene Ontology information is encoded as a set of relations:

GOTransport(GOterm)
GONonTransport(GOTerm)
GOaspect(GOterm, BP|MF|CC)
GOparent(GOterm, GOterm)
GOroot(GOterm)

which captures the GO hierarchy, which is a DAG allowing for multiple parents; the root terms of the hierarchy; the aspect to which the term belongs; and whether the GO term is associated with transporters only; or is clearly indicative of a non-transporter.

The GO annotation from SwissProt and other curated databases is represented by

GO(pid, GOterm, evidenceCode, source)

Pfam Domain information is encoded as a set of relations:

PfamTransport(PfamID)
PfamNonTransport(PfamID)

which captures those Pfam domains which are only associated with transporters, or never associated with transporters. The relation

Pfam(pid, PfamID, start, end)

records the existence of a Pfam domain at the [*start*..*end*] position of a protein.

Transmembrane Segment information is encoded by the relations:

TMSnumber(pid, count, source)
TMS(pid, start, end, source)

record the number of TMS for a protein, and the existence of a TMS at the [*start..end*] position of a protein, according to the tool *source*, or according to Swissprot as the source.

Amino Acid Composition is recorded in the relations:

AALength(pid, protein_length)
AATMS(pid, TMS_count)
AA_AAC(pid, AAC_vector)
AA_PAAC(pid, PAAC_vector)
AA_PAAC_TMS(pid, PAAC_TMS_vector)
AA_PAAC_notTMS(pid, PAAC_notTMS_vector)

which are the components of the 822 dimensional vector selected in Section 4.6,

$$L(x).TMS(x).AAC(x).PAAC_{TMS}(x).PAAC_{\overline{TMS}}(x)$$

together with $PAAC(x)$.

Substrates as they are grouped or organized into a hierarchy need to be captured in the dataspace. This information needs to include, and be consistent with, the information used in TCDB as represented in the *TCsubstrate* relation above. A standard set of names or identifiers need to be assigned to the substrates and the groupings. The relations are

SubstrateID(substrateName)
SubstrateParent(SubstrateName, SubstrateName_of_Parent)
SubstrateRoot(SubstrateName)

Note that a "substrate" is either a specific substrate, a grouping of substrates, or a class of substrates. For human readers, there could be a second argument providing a brief text description in the *SubstrateName* relation.

4.8 Conclusion

In this chapter we investigate the issue of including transport reactions, transporter proteins, and the GPR associations for transport in the reconstruction of metabolic pathways. To clarify the state of the art in that area, we develop a scheme to describe and compare the different approaches. This is necessary so that we can show that the existing work of predicting transport proteins actually is diverse and incomparable. We use a case study to get a deeper understanding of the existing work, and to compare them in a practical setting using a fungal genome of interest. In Section 4.4 we automate a protocol for determining the transporters in a genome that is used in the lab of Milton Saier, who develops the Transporter Classification and maintains the TCDB. In Section 4.6 we explore how to predict specific substrates of transporters. This is a very difficult problem, so we do not find a solution. Based on our experience, in Section 4.7 we propose a framework for the overall problem of predicting transporters, which includes the problem of determining specific substrates.

The scheme to describe and compare existing methods for predicting transporters allowed us to perform a meaningful analysis of the state of the art. This guided our case study that applied existing techniques to the fungal genome of *A. niger* CBS 513.88 for which there is a manually created and curated GENRE available.

This study reveals several issues:

- the disjointedness of the field with little connection between those that use the Transporter Classification (TC) as their target for prediction, and those that use the chemical substrates being transported as their target for prediction;
- the limited coverage of the predictors, due to the small size of available Gold Standard datasets for transport; and
- the inability of the techniques to predict the specific substrate, or specific collection of substrates, that is transported across the membrane by the transport protein, even though they could identify the type of substrate in some cases.

In Section 4.4 we automate a protocol of Saier's lab for determining the transporters in a genome, and applied the implementation to the fungal genome of *A. niger* CBS 513.88. This included determining localization, and improvements in predicting transmembrane segments (TMS) of a protein.

In Section 4.6 we explore how to predict specific substrates of transporters. Section 4.6 shows just how difficult the problem is, as we explore a number of approaches in order to address the problem, but we come up short. We do not find a solution to the problem of predicting specific substrates.

In Section 4.7 we propose a framework for the overall problem of predicting transporters, which includes the problem of determining specific substrates. From our perspective, it clearly identifies the issues of how to best proceed given the amount of experimental evidence for transporters, and how to harmonize the different points of view. It is however, only a proposal, and not a worked solution.

Chapter 5

Conclusion

This chapter concludes the thesis. It recaps the thesis work, and presents a summary of challenges addressed, the progress made, and the current state of the art. It also presents the contributions of our work, the limitations of our work, and potential future directions for this work.

This thesis deals with computational aspects of the automatic reconstruction of the metabolic pathways of an organism. It is motivated by the critical role of genome-scale network reconstructions (GENREs) of metabolism in systems biology, and the significant impact of systems biology on biology today, especially in industrial applications.

Chapter 2 contains the background material that is important to the understanding of this dissertation. Key are the Gene-Protein-Reaction (GPR) associations that are the units of the metabolic pathway reconstructions. They relate the central dogma of biology that genes through the processes of transcription and translation produce proteins, and these proteins in turn carry out the functional roles of the cell, including the enzymatic reactions of metabolism and the transport reactions across membranes.

In Chapter 3, through a review of the state of the art and case studies with fungal genomes, we investigate the reconstruction of metabolic pathways and the obstacles to full automation of the process. The first constribution of the thesis is to identify those obstacles and identify the issues preventing automation.

In Chapter 4 we investigate the issue of including transport reactions, transporter proteins, and the GPR associations for transport in the reconstruction of metabolic pathways. To

clarify the state of the art in that area, we develop a scheme to describe and compare the different approaches. This is necessary so that we can show that the existing work of predicting transport proteins actually is diverse and incomparable. We use a case study to get a deeper understanding of the existing work, and to compare them in a practical setting using a fungal genome of interest. In Section 4.4 we automate a protocol for determining the transporters in a genome that is used in the lab of Milton Saier, who develops the Transporter Classification and maintains the TCDB. In Section 4.6 we explore how to predict specific substrates of transporters. This is a very difficult problem, so we do not find a solution. Based on our experience, in Section 4.7 we propose a framework for the overall problem of predicting transporters, which includes the problem of determining specific substrates.

The chapter is organized as follows: Section 5.1 presents the contributions of our work; Section 5.2 discusses the limitations of our work; and Section 5.3 offers some directions for future work. For transparency, Section 5.4 points ot very late-breaking work that is directly relevant to this thesis.

5.1 Contributions

Contribution 1: Identification of issues in the reconstruction of metabolic networks.

The issues for eukaryotes in particular are the need to model a cell's internal organelles, predict localization of proteins, and predict transport proteins with their specific substrate and membrane localization.

The issues identified are as follows.

- The reference template approaches are dependent on the body of existing knowledge, and the effort to manually curate the scientific literature to extract that knowledge and encode it in public databases.

- The evaluation of methods is difficult when applied to new genomes. Internal validation of the model can be measured in terms of numbers of pathways, reactions, and GPR associations to indicate coverage, and by the number of holes to indicate completeness. Further internal validation requires constructing a systems biology model so one can apply flux balance analysis for atoms, charges, energy, etc. External validation requires

the scientist to make predictions from the model and then to validate those predictions in the wet lab; this is not expertise available usually to the developer of algorithms.

- The validation of methods for *de novo* discovery of pathways is difficult, even for model organisms. Internal validation shows that the pathways are sound in terms of the chemical transformation of compounds, but external validation of the existence of the pathway in the organism requires extensive wet lab work.

- Even with gap filling, there are typically many holes in the resulting reconstruction. Most approaches to gap-filling do not make use of gene expression data, which today can be readily available even for non-model organisms through RNA-Seq.

- The widely available and widely used tools are biased towards prokaryotes. In particular, they do not model cell compartments such as mitochondrion, Golgi, peroxisome, ER, vacuole, or lysosome in their reconstructions.

- Transport reactions are often an afterthought in the modeling of the cell, despite the fact that the reconstruction needs to view the cell as a closed system importing and exporting compounds to its surroundings in order to perform internal validation.

Contribution 2: A scheme to describe and compare existing methods for predicting transporters.

The scheme allowed us to perform a meaningful analysis of the state of the art. This guided our case study that applied existing techniques to the fungal genome of *A. niger* CBS 513.88 for which there was a manually created and curated GENRE available.

This study reveals several issues:

- the disjointedness of the field with little connection between those that use the Transporter Classification (TC) as their target for prediction, and those that use the chemical substrates being transported as their target for prediction;

- the limited coverage of the predictors, due to the small size of available Gold Standard datasets for transport; and

- the inability of the techniques to predict the specific substrate, or specific collection of substrates, that is transported across the membrane by the transport protein, even though they could identify the type of substrate in some cases.

A paper describing this work appeared at the 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology in Niagara Falls:

> Faizah Aplop and Greg Butler, On predicting transport proteins and their substrates for the reconstruction of metabolic networks, Proceedings of the 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2015, Niagara Falls, ON, Canada, August 12–15, 2015.

Contribution 3: Automation of a protocol used in Saier's lab for the determination of transporters for an organism. This included determining localization, and improvements in predicting transmembrane segments (TMS) of a protein.

In Section 4.4 we automate a protocol of Saier's lab for determining the transporters in a genome, and applied the implementation to the fungal genome of *A. niger* CBS 513.88.

Contribution 4: Exploration of techniques to predict the specific substrates transported by a transporter.

In Section 4.6 we explore how to predict specific substrates of transporters. This is a very difficult problem, so we do not find a solution.

Contribution 5: A proposed framework for the overall problem of predicting transporters, which includes the problem of determining specific substrates.

Based on our experience, in Section 4.7 we propose a framework for the overall problem of predicting transporters, which includes the problem of determining specific substrates.

5.2 Limitations

In Chapter 4 we demonstrate an implementation to automate the protocol used in Saier's lab. This is beta version software that is available at `transath.umt.edu.my`. The documentation is lacking.

In Chapter 4 we demonstrate the difficult nature of predicting the specific substrates that are transported by a transport protein. Section 4.6 shows just how difficult the problem is, as we explore a number of approaches in order to address the problem, but we come up short. We do not find a solution to the problem of predicting specific substrates.

The framework in Section 4.7 is a proposal for the problem of predicting transport. From our perspective, it clearly identifies the issues of how to best proceed given the amount of experimental evidence for transporters, and how to harmonize the different points of view. It is however, only a proposal, and not a worked solution.

5.3 Future Directions

In Section 4.7 we propose a framework for the problem of predicting transport proteins. This includes harmonizing the different schemes from TC, GO, Pfam, and substrates. The framework is a roadmap for moving ahead.

The techniques in Section 4.6 should be revisited now and then as more experimental data is collected.

The first future direction is to cluster the sequences of the TCDB using any one of the available approaches such as MCL (Markov Clustering) [VD00] which is widely used for clustering protein families, and Transitivity Clustering [Wit10] which computes a hierarchical clustering. Ideally the clusters would match the TC classification of Superfamily, Family and Subfamily. For each cluster, one could compute an MSA and then construct a HMM to act as a classifier for the cluster and as predictors for TC-Family and TC-Subfamily.

The second future direction is to attempt to predict the sites in the protein sequence that are responsble for the substrate specificity of the transporter. One should then investigate whether the properties of the amino acids at these sites can be used to predict the substrate. From known examples it is likely that the sites are located in the TMS regions of the protein, and the number of important sites is small. Therefore a multiple sequence alignment algorithm which preseves TMS regions, such as TM-Coffee [CDTTN12], would be a good choice. The MSA could then be processed by JDet [MGMR$^+$12] to determine the specificity-determining sites. A predictor for the specific substrate transported could be based on the amino acids at these sites using the various alphabets in Table 6 or the multilevel alphabet encoding of Hod et al. [HKMG13].

Hod et al. [HKMG13] use the secondary structure of the protein in their multilevel alphabet encoding. For transporters this could be generalized to record the location relative to the start or end of a TMS rather than simply TMS versus non-TMS. For substrate specificity of transporters this level of precision in location seems to be important. Therefore a third

future direction would be to combine the information on amino acid composition in the 822-dimensional vector of Section 4.6 with the various amino acid alphabets in Table 6 and with this encoding of location relative to the TMS apply the approach of Hod et al. [HKMG13].

The fourth future direction would be to construct the relational dataspace described in Section 4.7 and explore available machine learning approaches. Two candidates from clustering would be MCL (Markov Clustering) [VD00] and Transitivity Clustering [Wit10].

5.4 Postscript

On 6 April 2015, the PhD work of Oscar Dias at University of Minho in Portugal was published:

> Oscar Dias, Miguel Rocha, Eugénio C Ferreira and Isabel Rocha, Reconstructing genome-scale metabolic models with *merlin*, *Nucleic Acids Research*, 43(8): 3899–3910, 2015.

The *merlin* system is a robust implementation for the automatic reconstruction of metabolic networks that has the features that we identified in this thesis as lacking in existing systems, and necessary for the investigation of fungal genomes. The *merlin* system handles eukaryote genomes, and includes the determination of transport Gene-Protein-Reaction associations, as well as localization of reactions across a number of compartments: mitochondrion, endoplasmic reticulum (ER), and Golgi apparatus.

In *merlin*, transport proteins are predicted based on the existence of TMS as predicted by TMHMM, and by similarity to entries in TCDB using the Smith-Waterman algorithm. The association of transport reactions and specific substrates for the predicted transport proteins is taken from a manually curated database of some 4000 TCDB entries.

In *merlin*, localization is determined using PSORTb 3.0 for prokaryotes, and WoLF PSORT for eukaryotes. These tools predict localizations in organelles, and the definition of organelle for these tools includes the membrane. In *merlin*, localization in membranes of an organelle is assumed for the proteins predicted to be transport proteins.

The *merlin* paper emphasises that no other system exists for the reconstruction of metabolic pathways with these three features, namely, predicts transport GPR; models localization; and handles genomes of eukaryotes.

The *merlin* software is available as open source Java code.

Note that *merlin* adopts different strategies to the steps of predicting transport, and to predicting localization than we do in this work. In particular, the prediction of transport is conditional upon TMS as predicted by TMHMM. Section 4.3 shows that TMHMM is not always accurate, and this work develops a better approach. For localization, we adopt LocTree3. LocTree3 has demonstrated superiority to WoLF PSORT, and LocTree3 directly predicts localization to membranes.

As in *merlin*, we map from TC entries to substrates; in our case, substrate group and specific substrate. However, we do not identify a transport reaction, which *merlin* does.

Our automated approach, as in *merlin*, builds on identfying similar sequences in TCDB. However, we recognize that this is limiting in that it does not discover novel transporters. Therefore we investigate other means of predicting substrates in Section 4.6.

Appendix A

Sugar Porters

This appendix presents information on the known sugar transporters in the TCDB. Table 49 lists the members of TC-Subfamily 2.A.1.1 which are the sugar porters. The column ID contains the identifier, which includes the UniProtKB identifier as welll as the TCID. The column Description contains the nam of the transporter. The column Organismal Type contains the type of organism from which the protein comes. The column Status indicates whether the UniProtKB entry is reviewed or not.

Table 49: Sugar Porter Subfamily in TCDB as of May 2014

ID	Description	Organismal Type	Status
gnl|TC-DB|P0AEP1|2.A.1.1.1	Galactose-proton symporter	Bacteria	Reviewed
gnl|TC-DB|P0AE24|2.A.1.1.2	Arabinose-proton symporter	Bacteria	Reviewed
gnl|TC-DB|P0AGF4|2.A.1.1.3	D-xylose-proton symporter	Bacteria	Reviewed
gnl|TC-DB|P21906|2.A.1.1.4	Glucose facilitated diffusion protein	Bacteria	Reviewed
gnl|TC-DB|P43581|2.A.1.1.5	Hexose transporter HXT10	Yeast	Reviewed
gnl|TC-DB|P13181|2.A.1.1.6	Galactose transporter (Galactose permease)	Yeast	Reviewed
gnl|TC-DB|P11636|2.A.1.1.7	Quinate permease (Quinate transporter)	Fungi	Reviewed
gnl|TC-DB|P30605|2.A.1.1.8	Myo-inositol transporter 1	Yeast	Reviewed
gnl|TC-DB|P07921|2.A.1.1.9	Lactose permease	Yeast	Reviewed
gnl|TC-DB|P15685|2.A.1.1.10	Maltose permease MAL6T	Yeast	Reviewed
gnl|TC-DB|P53048|2.A.1.1.11	General alpha-glucoside permease	Yeast	Reviewed
gnl|TC-DB|Q07647|2.A.1.1.12	Glucose transporter type 3	Animals	Reviewed

Continued on next page

Table 49 – continued from previous page

ID	Description	Organismal Type	Status
gnl|TC-DB|P22732|2.A.1.1.13	Solute carrier family 2, facilitated glucose transporter member 5	Animals	Reviewed
gnl|TC-DB|P15686|2.A.1.1.14	H(+)/hexose cotransporter 1	Plants	Reviewed
gnl|TC-DB|P95908|2.A.1.1.15	Sugar Transporter	Archaea	Unreviewed
gnl|TC-DB|Q01441|2.A.1.1.16	Membrane transporter D2	Protozoa	Reviewed
gnl|TC-DB|P10870|2.A.1.1.17	High-affinity glucose transporter SNF3	Protozoa	Reviewed
gnl|TC-DB|Q06222|2.A.1.1.18	Glucose transporter 2A	Yeast	Reviewed
gnl|TC-DB|Q12300|2.A.1.1.19	High-affinity glucose transporter RGT2	Yeast	Reviewed
gnl|TC-DB|Q01440|2.A.1.1.20	Membrane transporter D1	Protozoa	Reviewed
gnl|TC-DB|O74969|2.A.1.1.21	High-affinity glucose transporter ght2 (Hexose transporter 2)	Yeast	Reviewed
gnl|TC-DB|O74849|2.A.1.1.22	High-affinity fructose transporter ght6 (Hexose transporter 6)	Yeast	Reviewed
gnl|TC-DB|Q92339|2.A.1.1.23	High-affinity gluconate transporter ght3 (Hexose transporter 3)	Yeast	Reviewed
gnl|TC-DB|O97467|2.A.1.1.24	Hexose Transporter 1	Protozoa	Unreviewed
gnl|TC-DB|Q96QE2|2.A.1.1.25	Proton myo-inositol co-transporter (Hmit)	Animals	Reviewed
gnl|TC-DB|O34718|2.A.1.1.26	Metabolite Transport Protein	Bacteria	Reviewed
gnl|TC-DB|P42417|2.A.1.1.27	Myo-inositol transport protein	Bacteria	Reviewed
gnl|TC-DB|P11166|2.A.1.1.28	The erythrocyte/brain hexose facilitator, Gtr1 or Glut1	Animals	Reviewed
gnl|TC-DB|P11168|2.A.1.1.29	Glucosamine/glucose uniporter, Glut-2	Animals	Reviewed
gnl|TC-DB|P32467|2.A.1.1.30	Low affinity glucose transporter HXT4 (LGT1)	Yeast	Reviewed
gnl|TC-DB|P39004|2.A.1.1.31	High affinity hexose transporter HXT6	Yeast	Reviewed
gnl|TC-DB|P15729|2.A.1.1.32	Glucose transport protein	Bacteria	Reviewed
gnl|TC-DB|Q8NJ22|2.A.1.1.33	Hexose transporter (Similarity)	Yeast	Unreviewed
gnl|TC-DB|Q8VZ80|2.A.1.1.34	H+ symporter, AtPLT5	Plants	Reviewed
gnl|TC-DB|Q7BEC4|2.A.1.1.35	Glucose transport protein GlcP	Bacteria	Unreviewed
gnl|TC-DB|Q400D8|2.A.1.1.36	Putative low affinity glucose transporter MstE	Fungi	Unreviewed
gnl|TC-DB|Q6PXP3|2.A.1.1.37	Intestinal facilitative glucose transporter 7	Animals	Reviewed
gnl|TC-DB|P39932|2.A.1.1.38	Sugar transporter STL1	Yeast	Reviewed
gnl|TC-DB|P49374|2.A.1.1.39	High-affinity glucose transporter	Yeast	Reviewed
gnl|TC-DB|Q64L87|2.A.1.1.40	Xylhp (Fragment)	Yeast	Unreviewed

Continued on next page

Table 49 – continued from previous page

ID	Description	Organismal Type	Status
gnl|TC-DB|O52733|2.A.1.1.41	D-xylose-proton symporter	Bacteria	Reviewed
gnl|TC-DB|Q8G3X1|2.A.1.1.42	D-Glucose-proton symporter	Bacteria	Unreviewed
gnl|TC-DB|A0ZXK6|2.A.1.1.43	Monosaccharide transporter	Fungi	Unreviewed
gnl|TC-DB|Q9BYW1|2.A.1.1.44	Solute carrier family 2, facilitated glucose transporter member 11	Animals	Reviewed
gnl|TC-DB|Q8L6Z8|2.A.1.1.45	D-xylose-proton symporter-like 1	Plants	Reviewed
gnl|TC-DB|Q9JIF3|2.A.1.1.46	Solute carrier family 2, facilitated glucose transporter member 8	Animals	Reviewed
gnl|TC-DB|Q5ERC7|2.A.1.1.47	Glucose transporter 9b	Animals	Unreviewed
gnl|TC-DB|Q9LNV3|2.A.1.1.48	Sugar transport protein 2	Plants	Reviewed
gnl|TC-DB|Q39228|2.A.1.1.49	Sugar transport protein 4	Plants	Reviewed
gnl|TC-DB|Q94AZ2|2.A.1.1.50	Sugar transport protein 13	Plants	Reviewed
gnl|TC-DB|Q2MEV7|2.A.1.1.51	Glucose/xylose symporter 1	Yeast	Unreviewed
gnl|TC-DB|Q26579|2.A.1.1.52	Glucose transport protein	Animals	Unreviewed
gnl|TC-DB|Q8NTX0|2.A.1.1.53	Myo-Inositol upatake porter, IoIT1	Bacteria	Unreviewed
gnl|TC-DB|Q8NL90|2.A.1.1.54	Myo-Inositol uptake porter, IoIT2	Actinobacteria	Unreviewed
gnl|TC-DB|P96710|2.A.1.1.55	L-Arabinose-proton symporter AraE	Bacteria	Reviewed
gnl|TC-DB|Q9SFG0|2.A.1.1.56	High affinity Monosaccharides: H+ symporter, Stp6	Plants	Reviewed
gnl|TC-DB|Q8J0V1|2.A.1.1.57	High affinity glucose:H+ symporter, MstA	Fungi	Unreviewed
gnl|TC-DB|Q8J0U9|2.A.1.1.58	Low affinity glucose:H+ symporter, MstC	Fungi	Unreviewed
gnl|TC-DB|O95528|2.A.1.1.59	The glucose transporter, GLUT10	Animals	Reviewed
gnl|TC-DB|P23586|2.A.1.1.60	The major hexose transporter, Htr1	Plants	Reviewed
gnl|TC-DB|Q9FMX3|2.A.1.1.61	High affinity Monosaccharides transporter, STP11	Plants	Reviewed
gnl|TC-DB|O23492|2.A.1.1.62	High affinity plasma membrane myoinositol-specific H+ symporter, INT4	Plants	Reviewed
gnl|TC-DB|Q9C757|2.A.1.1.63	Low affinity inositol	Plants	Reviewed
gnl|TC-DB|B1PM37|2.A.1.1.64	The hexose transporter, Hxs1	Yeast	Unreviewed
gnl|TC-DB|A0QZX3|2.A.1.1.65	Glucose permease GlcP	Bacteria	Unreviewed
gnl|TC-DB|Q8VZR6|2.A.1.1.66	The tonoplast H+:inositol transporter 1, Int1	Plants	Reviewed
gnl|TC-DB|Q2MDH1|2.A.1.1.67	Glucose/xylose facilitator 1 Gxf1	Yeast	Unreviewed
gnl|TC-DB|A3M0N3|2.A.1.1.68	Glucose transporter/sensor RGT2	Yeast	Unreviewed
gnl|TC-DB|A1Z264|2.A.1.1.69	Sugar & polyol transporter 1, SPT1	Red Algae	Unreviewed

Continued on next page

Table 49 – continued from previous page

ID	Description	Organismal Type	Status
gnl|TC-DB|Q0ULF7|2.A.1.1.70	MFS permease	Fungi	Unreviewed
gnl|TC-DB|B1PLM1|2.A.1.1.71	Hexose (glucose) transporter, GT4 (D2)	Trypanosomatidae	Unreviewed
gnl|TC-DB|Q9NRM0|2.A.1.1.72	Human SLC2A9a and SLC2A9b isoform mediate electrogenic transport of urate	Animals	Reviewed
gnl|TC-DB|Q5A8J5|2.A.1.1.73	Glycerol uptake permease, STL1	Yeast	Unreviewed
gnl|TC-DB|Q926Q9|2.A.1.1.74	The putative L-rhamnose porter, RhaY	Firmicutes, Actinobacteria	Unreviewed
gnl|TC-DB|Q9XIH7|2.A.1.1.75	The fructose/xylose:H+ symporter, PMT1	Plants	Reviewed
gnl|TC-DB|O76486|2.A.1.1.76	Glucose transporter GT1	Eukaryota	Unreviewed
gnl|TC-DB|O61059|2.A.1.1.77	D-glucose/D-ribose transporter LmGT2	Protozoa	Unreviewed
gnl|TC-DB|O61060|2.A.1.1.78	Glucose transporter LmGT3	Protozoa	Unreviewed
gnl|TC-DB|Q1XF07|2.A.1.1.79	Putative polyol transporter PLT4	Plants	Unreviewed
gnl|TC-DB|P14672|2.A.1.1.80	Solute carrier family 2, facilitated glucose transporter member 4, SLC2A4	Animals	Reviewed
gnl|TC-DB|Q0SE66|2.A.1.1.81	Glucose uptake porter, Glcp	Bacteria	Unreviewed
gnl|TC-DB|Q7SCU1|2.A.1.1.82	The cellobiose/cellodextrin transporter, Cdt-1	Fungi	Unreviewed
gnl|TC-DB|Q7SD12|2.A.1.1.83	The cellobiose/cellodextrin transporter, Cdt-2	Fungi	Unreviewed
gnl|TC-DB|Q96290|2.A.1.1.84	Monosaccharide-sensing protein 1, TMT1/TMT2 glucose/sucrose:H+ antiporter	Plants	Reviewed
gnl|TC-DB|Q8LPQ8|2.A.1.1.84	Monosaccharide-sensing protein 2, TMT1/TMT2 glucose/sucrose:H+ antiporter	Plants	Reviewed
gnl|TC-DB|A8KB28|2.A.1.1.85	Slc2A10 (Glut10) facilitative glucose transporter	Animals	Unreviewed
gnl|TC-DB|H9BPB6|2.A.1.1.86	Facilitative glucose transporter 1, GLUT1	Animals	Unreviewed
gnl|TC-DB|Q8TD20|2.A.1.1.87	Solute carrier family 2, facilitated glucose transporter member 12, SLC2A12	Animals	Reviewed
gnl|TC-DB|Q9UGQ3|2.A.1.1.88	Solute carrier family 2 facilitated glucose transporter member 6, SLC2A6	Animals	Reviewed

Continued on next page

Table 49 – continued from previous page

ID	Description	Organismal Type	Status
gnl|TC-DB|Q9NY64|2.A.1.1.89	Solute carrier family 2 facilitated glucose transporter member 8, SLC2A8	Animals	Reviewed
gnl|TC-DB|Q8TDB8|2.A.1.1.90	Solute carrier family 2 facilitated glucose transporter member 14, SLC2A14	Animals	Reviewed
gnl|TC-DB|P11169|2.A.1.1.91	Solute carrier family 2 facilitated glucose transporter member 3, SLC2A3	Animals	Reviewed
gnl|TC-DB|P38055|2.A.1.1.92	Inner membrane metabolite transport protein ydjE	Bacteria	Reviewed
gnl|TC-DB|P53142|2.A.1.1.93	Vacuolar protein sorting-associated protein 73, VPS73	Fungi	Reviewed
gnl|TC-DB|Q12407|2.A.1.1.94	Putative metabolite transport protein, YDL199C	Fungi	Reviewed
gnl|TC-DB|Q46909|2.A.1.1.95	Inner membrane metabolite transport protein, YgcS	Bacteria	Reviewed
gnl|TC-DB|P38142|2.A.1.1.96	Probable metabolite transport protein, YBR241C	Fungi	Reviewed
gnl|TC-DB|O04036|2.A.1.1.97	Sugar transporter ERD6	Plants	Reviewed
gnl|TC-DB|Q9FRL3|2.A.1.1.98	Sugar transporter ERD6-like 6, At1g75220	Plants	Reviewed
gnl|TC-DB|A1Z8N1|2.A.1.1.99	Facilitated trehalose transporter, Tret1-1	Animals	Reviewed
gnl|TC-DB|P43562|2.A.1.1.100	Probable metabolite transport protein, YFL040W	Fungi	Reviewed
gnl|TC-DB|Q04162|2.A.1.1.101	Probable metabolite transport protein, YDR387C	Fungi	Reviewed
gnl|TC-DB|Q56ZZ7|2.A.1.1.102	Plastidic glucose transporter 4, At5g16150	Plants	Reviewed
gnl|TC-DB|Q0WWW9|2.A.1.1.103	D-xylose-proton symporter-like 3, At5g59250	Plants	Reviewed
gnl|TC-DB|P30606|2.A.1.1.104	Myo-inositol transporter 2, ITR2	Fungi	Reviewed
gnl|TC-DB|P54862|2.A.1.1.105	Hexose transporter HXT11 (LGT3)	Fungi	Reviewed
gnl|TC-DB|P46333|2.A.1.1.106	Probable metabolite transport protein, CsbC	Bacili	Reviewed
gnl|TC-DB|P54854|2.A.1.1.107	Hexose transporter HXT15	Fungi	Reviewed
gnl|TC-DB|P32465|2.A.1.1.108	Low-affinity glucose transporter HXT1	Fungi	Reviewed
gnl|TC-DB|P42833|2.A.1.1.109	Hexose transporter HXT14	Fungi	Reviewed

Continued on next page

Table 49 – continued from previous page

ID	Description	Organismal Type	Status
gnl|TC-DB|P39924|2.A.1.1.110	Hexose transporter HXT13	Fungi	Reviewed
gnl|TC-DB|P23585|2.A.1.1.111	High-affinity glucose transporter HXT2	Fungi	Reviewed
gnl|TC-DB|Q9P3U6|2.A.1.1.112	High-affinity glucose transporter ght1	Yeast	Reviewed
gnl|TC-DB|P37514|2.A.1.1.113	Putative metabolite transport protein yyaJ	Bacili	Reviewed
gnl|TC-DB|P31679|2.A.1.1.114	Putative metabolite transport protein yaaU	Bacteria	Reviewed
gnl|TC-DB|P76230|2.A.1.1.115	Putative metabolite transport protein ydjK	Bacteria	Reviewed
gnl|TC-DB|C4B4V9|2.A.1.1.116	L-arabinose transporter, araE	Actinobacteria	Unreviewed
gnl|TC-DB|G4N740|2.A.1.1.117	Glucose transporter rco-3/MoST1	Fungi	Unreviewed
gnl|TC-DB|Q97xw7|2.A.1.1.118	Mfs porter of 435 aas	Crenarchaea	Unreviewed

Appendix B

TransportTP Results

This appendix presents the results of TransportTP on each of the eight fungal genomes in our study. Table 50 presents the number of proteins in each fungi that matches a given TCID. The table is organised by TC-Family. The columns Family and family Name contain the TC-Family identifier and its name. The column TCID contains the TCID of the TCDB entry predicted to be in a fungi predicted by TransportTP. Only those identifiers predicted in at least one fungi occur in this column. The last 8 columns contain the number of transporters in each fungi. The column headings indicate the fungi using the following code: **Aaf**:*A.fumigatus Af293*, **Ani**:*A. nidulans*, **Anc**:*A.niger CBS513.88*, **Ann**:*A. niger NRRL3*, **Aor**: *A. oryzae*, **Ncr**:*N. crassa*, **Pch**:*P. chrysosporium RP78*, **Spo**:*S. pombe*.

Table 50: TransportTP Results for Fungal Genomes

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
1.A.1.	The Voltage-gated Ion Channel (VIC) Superfamily	1.A.1.11.10	-	-	-	-	-	-	-	1
		1.A.1.11.17	1	1	1	1	1	1	1	-
		1.A.1.7.1	1	-	1	1	1	-	-	-
1.A.11.	The Ammonia Transporter Channel (Amt) Family	1.A.11.3.1	1	1	1	1	1	1	1	-
		1.A.11.3.2	-	-	1	1	-	1	-	1
		1.A.11.3.3	2	2	1	1	3	1	1	1
		1.A.11.3.4	-	-	-	-	-	1	-	-
1.A.33.	The Cation Channel-forming Heat Shock Protein-70 (Hsp70) Family	1.A.33.1.2	1	1	1	1	1	-	-	-
		1.A.33.1.3	-	-	-	-	-	-	1	-
1.A.35.	The CorA Metal Ion Transporter (MIT) Family	1.A.35.2.1	1	1	1	1	2	1	1	2
		1.A.35.2.2	1	1	1	1	1	1	-	-
		1.A.35.5.1	1	2	1	1	2	1	1	1
1.A.4.	The Transient Receptor Potential Ca2+ Channel (TRP-CC) Family	1.A.4.4.1	1	-	-	-	-	-	-	-
		1.A.4.7.1	-	-	2	4	1	-	-	-
		1.A.4.7.2	-	-	1	-	2	-	-	-
1.A.56.	The Copper Transporter (Ctr) Family	1.A.56.1.10	-	-	-	-	-	2	-	-
		1.A.56.1.5	-	1	1	-	-	-	-	2
		1.A.56.1.6	-	-	-	-	-	-	-	1

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
1.A.8.	The Major Intrinsic Protein (MIP) Family	1.A.8.6.1	-	-	-	-	-	-	1	-
		1.A.8.6.2	1	1	1	1	1	1	1	-
		1.A.8.7.1	2	3	2	2	2	-	3	1
		1.A.8.9.3	-	1	-	-	-	-	1	-
		1.A.8.9.4	-	-	1	1	1	-	-	-
2.A.1.	The Major Facilitator Superfamily (MFS)	2.A.1.1.1	1	-	1	1	1	1	1	-
		2.A.1.1.2	-	-	-	-	1	1	1	-
		2.A.1.1.3	1	-	1	1	1	-	-	-
		2.A.1.1.5	-	1	1	1	1	-	-	-
		2.A.1.1.6	-	-	1	1	1	-	-	-
		2.A.1.1.7	1	1	3	3	2	1	1	-
		2.A.1.1.8	1	1	1	1	1	-	1	1
		2.A.1.1.9	2	2	1	1	1	2	1	-
		2.A.1.1.10	1	1	1	1	1	1	-	-
		2.A.1.1.11	1	2	1	1	1	1	-	-
		2.A.1.1.12	-	-	-	-	-	1	-	-
		2.A.1.1.14	1	-	1	1	1	-	-	-
		2.A.1.1.18	1	-	-	1	-	-	-	-
		2.A.1.1.19	1	1	1	1	1	-	-	-
		2.A.1.1.21	1	-	1	1	1	-	-	1
		2.A.1.1.22	-	-	1	1	-	-	-	1
		2.A.1.1.23	-	-	-	-	-	-	-	1
		2.A.1.1.30	-	1	1	1	-	-	-	-
		2.A.1.1.31	1	-	-	-	1	-	-	-
		2.A.1.1.33	2	1	1	1	1	1	1	-
		2.A.1.1.34	-	1	1	1	1	-	-	-
		2.A.1.1.36	-	1	-	-	-	-	1	-
		2.A.1.1.38	2	2	2	2	2	1	1	-
		2.A.1.1.39	1	2	1	1	1	1	1	-
		2.A.1.1.40	1	1	1	1	1	1	1	-
		2.A.1.1.43	1	1	1	1	1	-	1	-
		2.A.1.1.49	-	-	1	-	1	-	-	-
		2.A.1.1.51	1	1	2	2	1	1	-	-
		2.A.1.1.55	-	1	-	-	-	-	-	-
		2.A.1.1.57	1	1	1	1	1	1	1	-
		2.A.1.1.58	1	1	1	1	1	1	1	-
		2.A.1.1.60	1	-	1	1	-	-	-	-
		2.A.1.1.63	-	-	-	-	-	1	-	-
		2.A.1.1.64	1	-	-	-	-	1	-	-
		2.A.1.2.1	-	1	1	1	-	-	-	1
		2.A.1.2.2	1	1	1	1	1	1	-	-
		2.A.1.2.6	1	1	1	1	1	1	-	-
		2.A.1.2.7	-	1	-	-	-	-	-	-
		2.A.1.2.16	2	2	3	3	4	1	2	3
		2.A.1.2.17	1	1	1	1	1	1	1	-
		2.A.1.2.23	1	1	1	1	1	1	1	-
		2.A.1.2.31	1	1	1	-	1	1	1	-
		2.A.1.2.33	1	1	1	1	1	1	-	1
		2.A.1.2.35	1	1	2	3	2	1	1	2
		2.A.1.2.36	1	1	1	1	1	-	1	-
		2.A.1.3.1	1	1	1	1	1	1	1	1
		2.A.1.3.8	1	-	-	-	-	-	-	-
		2.A.1.3.11	-	-	-	-	1	-	-	-
		2.A.1.3.15	-	-	1	1	1	-	-	-
		2.A.1.3.29	1	1	1	1	1	1	1	1
		2.A.1.3.30	-	-	-	-	1	-	-	-
		2.A.1.7.2	1	1	1	1	1	1	1	-
		2.A.1.8.5	1	1	1	1	1	1	1	-
		2.A.1.9.1	1	1	1	1	1	-	1	1

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
		2.A.1.9.2	-	1	-	1	-	1	-	-
		2.A.1.9.3	-	-	1	1	-	-	-	-
		2.A.1.12.2	1	1	1	1	1	1	1	-
		2.A.1.13.2	-	1	1	1	-	-	-	-
		2.A.1.13.3	-	-	-	-	1	-	-	-
		2.A.1.13.4	1	2	1	1	2	1	1	-
		2.A.1.14.11	2	2	1	1	2	2	2	2
		2.A.1.14.12	1	1	1	1	1	1	-	-
		2.A.1.14.17	1	1	1	1	3	2	1	1
		2.A.1.14.18	-	-	1	1	1	-	1	-
		2.A.1.14.19	1	1	1	1	1	1	-	1
		2.A.1.14.20	1	1	1	-	1	-	1	1
		2.A.1.14.3	2	2	2	2	2	1	2	-
		2.A.1.14.4	2	2	3	2	2	1	1	1
		2.A.1.14.8	-	1	1	1	1	-	1	-
		2.A.1.14.9	-	1	-	1	-	1	-	-
		2.A.1.16.1	1	1	1	1	1	1	1	1
		2.A.1.16.2	1	1	1	1	1	-	-	1
		2.A.1.16.3	1	1	1	1	1	-	-	1
		2.A.1.16.4	1	1	1	1	1	-	-	-
		2.A.1.22.1	-	-	-	-	-	1	-	-
		2.A.1.24.1	-	-	1	1	-	1	-	-
		2.A.1.25.1	1	1	1	1	1	1	1	1
		2.A.1.28.2	-	1	1	1	-	-	-	-
		2.A.1.48.1	-	1	1	1	-	-	-	-
		2.A.1.48.2	1	-	1	1	1	1	1	1
		2.A.1.48.3	2	1	2	3	1	1	2	1
		2.A.1.48.4	2	1	2	1	2	2	1	2
		2.A.1.58.1	1	1	1	1	1	1	1	1
2.A.2.	The Glycoside-Pentoside-Hexuronide (GPH):Cation Symporter Family	2.A.2.6.1	2	2	4	4	2	3	2	1
2.A.3.	The Amino Acid-Polyamine-Organocation (APC) Family	2.A.3.1.2	1	1	1	1	1	-	1	-
		2.A.3.10.10	-	1	1	1	1	1	-	1
		2.A.3.10.11	1	-	-	-	-	-	-	-
		2.A.3.10.13	1	1	1	1	1	1	1	1
		2.A.3.10.14	1	-	-	-	1	1	-	1
		2.A.3.10.17	-	1	-	-	1	1	1	-
		2.A.3.10.18	2	-	1	1	1	1	-	-
		2.A.3.10.19	1	1	1	1	1	1	-	-
		2.A.3.10.2	1	1	1	1	2	1	-	-
		2.A.3.10.21	1	1	1	1	1	1	-	4
		2.A.3.10.22	-	-	1	1	1	-	-	1
		2.A.3.10.3	1	1	1	1	1	1	-	1
		2.A.3.10.4	1	1	1	1	1	-	-	-
		2.A.3.10.8	-	1	-	-	-	-	-	-
		2.A.3.4.1	1	1	2	2	2	1	1	-
		2.A.3.4.2	4	2	3	2	2	2	1	-
		2.A.3.4.3	1	1	1	1	1	2	2	3
		2.A.3.4.6	1	1	1	1	1	-	1	2
		2.A.3.8.1	-	1	1	1	-	1	-	-
		2.A.3.8.15	-	1	-	-	-	-	1	-
		2.A.3.8.2	1	-	1	1	1	-	-	-
		2.A.3.8.4	1	1	1	1	1	1	1	-
2.A.4.	The Cation Diffusion Facilitator (CDF) Family	2.A.4.2.2	2	2	2	1	1	2	1	1
		2.A.4.4.1	1	1	1	2	1	-	-	-
		2.A.4.4.2	-	1	-	-	-	1	-	-
		2.A.4.4.5	-	-	-	-	-	-	-	1
		2.A.4.5.1	2	1	1	2	1	5	1	-

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
2.A.5.	The Zinc (Zn2+)-Iron (Fe2+) Permease (ZIP) Family	2.A.5.1.1	2	2	3	4	6	4	3	1
		2.A.5.4.3	1	-	-	-	-	1	-	-
		2.A.5.4.4	-	1	1	1	1	-	1	1
2.A.6.	The Resistance-Nodulation-Cell Division (RND) Superfamily	2.A.6.6.1	-	-	1	1	-	-	-	-
		2.A.6.6.3	-	-	-	-	-	1	-	-
2.A.7.	The Drug/Metabolite Transporter (DMT) Superfamily	2.A.7.10.1	1	-	1	1	1	1	-	1
		2.A.7.10.2	-	-	-	-	-	-	1	-
		2.A.7.11.1	-	-	-	-	-	-	-	1
		2.A.7.12.4	-	1	-	-	-	-	-	-
		2.A.7.12.7	1	-	-	1	1	1	-	-
		2.A.7.12.8	-	-	1	-	-	-	1	1
		2.A.7.12.9	-	-	-	-	-	-	1	-
		2.A.7.13.1	-	1	-	-	-	1	-	-
		2.A.7.13.2	1	1	1	1	1	-	1	1
		2.A.7.16.1	-	-	-	-	-	-	1	-
		2.A.7.16.2	-	-	-	-	-	-	1	-
		2.A.7.24.1	-	-	-	-	-	1	1	1
		2.A.7.24.6	1	1	1	1	1	-	-	-
		2.A.7.9.1	-	-	-	-	-	1	-	-
		2.A.7.9.4	1	-	-	-	-	1	-	-
2.A.9.	The Cytochrome Oxidase Biogenesis (Oxa1) Family	2.A.9.1.1	1	1	-	-	1	-	-	2
2.A.16.	The Telurite-resistance/Dicarboxylate Transporter (TDT) Family	2.A.16.2.1	2	3	3	4	2	1	-	1
		2.A.16.4.1	2	1	1	1	2	-	2	4
2.A.17.	The Proton-dependent Oligopeptide Transporter (POT) Family	2.A.17.2.1	2	2	1	2	4	-	-	1
		2.A.17.2.2	2	1	1	1	2	2	1	-
2.A.18.	The Amino Acid/Auxin Permease (AAAP) Family	2.A.18.4.1	3	6	2	2	2	1	1	-
		2.A.18.4.2	3	1	3	3	3	1	1	-
		2.A.18.5.2	1	1	1	1	1	1	1	-
		2.A.18.6.2	-	-	-	-	-	-	1	-
		2.A.18.6.6	1	1	1	1	-	1	1	1
		2.A.18.7.1	2	2	2	2	1	2	1	1
2.A.19.	The Ca2+:Cation Antiporter (CaCA) Family	2.A.19.2.1	-	-	-	-	-	-	1	-
		2.A.19.2.2	4	3	4	4	4	9	1	-
		2.A.19.2.4	-	-	-	-	1	-	-	1
		2.A.19.4.4	1	1	1	1	1	1	-	1
		2.A.19.7.1	1	1	1	1	1	1	1	1
2.A.20.	The Inorganic Phosphate Transporter (PiT) Family	2.A.20.2.1	1	1	1	1	1	2	-	-
		2.A.20.2.2	2	2	-	-	1	-	-	-
2.A.21.	The Solute:Sodium Symporter (SSS) Family	2.A.21.6.1	3	4	4	4	3	1	3	3
2.A.22.	The Neurotransmitter:Sodium Symporter (NSS) Family	2.A.22.3.1	-	1	-	-	-	-	-	-
		2.A.22.6.3	-	-	-	-	1	-	-	-
2.A.23.	The Dicarboxylate/Amino Acid:Cation (Na+ or H+) Symporter (DAACS) Family	2.A.23.2.3	-	1	-	-	-	-	-	-
2.A.29.	The Mitochondrial Carrier (MC) Family	2.A.29.1.5	-	-	-	-	-	1	-	-
		2.A.29.1.5	-	-	-	-	-	-	1	-
		2.A.29.1.7	1	-	1	-	1	-	-	-
		2.A.29.1.7	-	1	-	1	-	-	1	1
		2.A.29.2.1	1	-	1	-	1	2	-	-
		2.A.29.2.1	-	1	-	1	-	-	-	-
		2.A.29.2.3	1	-	1	-	1	1	-	-
		2.A.29.2.3	-	1	-	1	-	-	1	-
		2.A.29.2.5	1	-	1	-	-	1	-	-
		2.A.29.2.5	-	1	-	1	-	-	1	1
		2.A.29.2.8	-	-	-	-	1	-	-	-
		2.A.29.2.9	-	-	-	-	-	-	1	-
		2.A.29.4.1	-	-	-	-	-	-	1	-
		2.A.29.4.3	1	-	1	-	1	1	-	-

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
		2.A.29.4.3	-	1	-	1	-	-	1	-
		2.A.29.4.4	1	-	1	-	1	2	-	-
		2.A.29.4.4	-	1	-	1	-	-	-	1
		2.A.29.5.1	1	-	1	-	1	1	-	-
		2.A.29.5.1	-	1	-	1	-	-	-	1
		2.A.29.5.3	1	-	1	-	1	-	-	-
		2.A.29.5.3	-	1	-	1	-	-	1	-
		2.A.29.6.1	1	-	1	-	1	1	-	-
		2.A.29.6.1	-	1	-	1	-	-	-	-
		2.A.29.7.3	1	-	1	-	2	1	-	-
		2.A.29.7.3	-	1	-	1	-	-	1	1
		2.A.29.8.2	-	-	-	-	-	1	-	-
		2.A.29.8.2	-	-	-	-	-	-	1	-
		2.A.29.8.3	-	1	-	-	-	-	-	-
		2.A.29.8.4	1	-	1	-	1	1	-	-
		2.A.29.8.4	-	1	-	1	-	-	1	-
		2.A.29.8.5	1	-	1	-	1	1	-	-
		2.A.29.8.5	-	1	-	1	-	-	-	1
		2.A.29.9.1	1	-	1	-	1	1	-	-
		2.A.29.9.1	-	1	-	1	-	-	1	-
		2.A.29.10.2	-	-	1	-	-	-	-	-
		2.A.29.10.2	-	1	-	1	-	-	1	-
		2.A.29.10.3	1	-	-	-	1	1	-	-
		2.A.29.12.1	1	-	1	-	1	1	-	-
		2.A.29.12.1	-	1	-	1	-	-	1	1
		2.A.29.13.1	1	-	1	-	1	1	-	-
		2.A.29.13.1	-	1	-	1	-	-	1	-
		2.A.29.14.1	1	-	-	-	-	1	-	-
		2.A.29.14.1	-	1	-	-	-	-	1	-
		2.A.29.14.2	-	-	1	-	1	-	-	-
		2.A.29.14.2	-	-	-	1	-	-	-	-
		2.A.29.15.1	1	-	1	-	1	1	-	-
		2.A.29.15.1	-	1	-	1	-	-	1	1
		2.A.29.16.1	1	-	1	-	1	1	-	-
		2.A.29.16.1	-	-	-	1	-	-	1	-
		2.A.29.17.1	1	-	1	-	-	1	-	-
		2.A.29.17.1	-	1	-	1	-	-	1	-
		2.A.29.17.2	-	-	-	-	-	-	-	1
		2.A.29.18.1	-	-	1	-	-	1	-	-
		2.A.29.18.1	-	-	-	1	-	-	1	1
		2.A.29.18.2	-	-	-	-	-	-	-	1
		2.A.29.20.1	-	-	-	-	-	-	1	-
		2.A.29.21.1	1	-	1	-	1	1	-	-
		2.A.29.21.1	-	1	-	1	-	-	1	1
		2.A.29.22.1	1	-	-	-	-	-	-	-
		2.A.29.22.1	-	1	-	1	-	-	-	-
		2.A.29.23.2	1	-	1	-	1	1	-	-
		2.A.29.23.2	-	1	-	1	-	-	1	1
		2.A.29.27.1	1	-	1	-	1	1	-	-
		2.A.29.27.1	-	1	-	1	-	-	1	1
		2.A.29.28.1	-	1	-	-	-	-	-	-
		2.A.29.29.1	1	-	1	-	1	1	-	-
		2.A.29.29.1	-	1	-	1	-	-	1	1
		2.A.29.30.1	1	-	1	-	2	1	-	-
		2.A.29.30.1	-	1	-	1	-	-	1	1
2.A.30.	The Cation-Chloride Cotransporter (CCC) Family	2.A.30.2.1	1	-	-	-	-	-	-	1
		2.A.30.4.2	-	1	-	-	-	-	-	-
		2.A.30.5.1	-	-	-	-	-	1	-	-

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
		2.A.30.5.2	-	-	1	1	1	-	-	-
2.A.31.	The Anion Exchanger (AE) Family	2.A.31.3.2	2	2	3	3	2	2	1	1
2.A.36.	The Monovalent Cation:Proton Antiporter-1 (CPA1) Family	2.A.36.2.1	1	1	1	1	1	1	1	1
		2.A.36.4.1	1	1	1	2	1	1	-	-
		2.A.36.4.2	-	3	-	-	1	1	-	-
		2.A.36.4.3	-	-	1	-	-	-	-	1
		2.A.36.4.4	2	1	1	1	1	-	1	1
2.A.37.	The Monovalent Cation:Proton Antiporter-2 (CPA2) Family	2.A.37.4.1	1	1	1	1	1	1	2	1
2.A.38.	The K+ Transporter (Trk) Family	2.A.38.2.1	-	-	-	-	-	1	-	-
		2.A.38.2.2	1	-	1	1	-	1	-	1
		2.A.38.2.3	-	-	1	1	1	-	1	-
		2.A.38.2.4	1	2	1	1	3	-	-	1
		2.A.38.2.5	1	1	-	-	-	-	1	-
2.A.39.	The Nucleobase:Cation Symporter-1 (NCS1) Family	2.A.39.2.1	-	-	1	-	1	-	-	-
		2.A.39.2.2	-	-	1	2	-	-	-	-
		2.A.39.2.3	1	-	1	1	-	-	-	-
		2.A.39.2.4	2	4	1	1	1	2	2	-
		2.A.39.3.1	1	2	2	2	1	-	1	2
		2.A.39.3.2	1	-	-	-	1	1	-	-
		2.A.39.3.3	-	1	1	1	1	-	-	1
		2.A.39.4.1	-	-	-	-	1	-	-	-
2.A.40.	The Nucleobase:Cation Symporter-2 (NCS2) Family	2.A.40.4.1	-	1	1	1	-	-	-	-
		2.A.40.5.1	1	1	1	1	1	1	1	1
2.A.41.	The Concentrative Nucleoside Transporter (CNT) Family	2.A.41.2.7	1	1	1	1	1	1	1	-
		2.A.41.3.1	1	1	1	1	1	-	-	1
2.A.43.	The Lysosomal Cystine Transporter (LCT) Family	2.A.43.1.1	1	1	1	1	-	1	-	-
		2.A.43.3.1	1	1	1	1	1	1	1	-
2.A.44.	The Formate-Nitrite Transporter (FNT) Family	2.A.44.1.1	-	-	-	-	-	1	-	-
		2.A.44.2.1	1	1	-	1	2	-	-	-
2.A.47.	The Divalent Anion:Na+ Symporter (DASS) Family	2.A.47.2.1	-	-	-	-	-	1	-	-
		2.A.47.2.2	1	1	1	1	1	-	1	1
2.A.49.	The Chloride Carrier/Channel (ClC) Family	2.A.49.1.2	3	3	-	3	-	3	1	2
		2.A.49.1.3	-	-	1	-	1	-	1	-
		2.A.49.2.2	-	-	-	-	1	-	-	-
		2.A.49.2.3	-	-	1	-	1	1	-	-
2.A.50.	The Glycerol Uptake (GUP) Family	2.A.50.1.1	-	-	-	-	-	-	1	1
2.A.52.	The Ni2+-Co2+ Transporter (NiCoT) Family	2.A.52.1.1	1	1	1	1	1	1	-	-
		2.A.52.1.3	-	-	-	-	-	-	-	1
2.A.53.	The Sulfate Permease (SulP) Family	2.A.53.1.2	1	1	2	2	1	2	1	2
		2.A.53.1.3	1	-	-	-	1	2	1	-
		2.A.53.1.7	-	-	1	1	-	-	-	-
		2.A.53.11.1	1	1	1	1	1	-	-	-
		2.A.53.2.6	-	-	-	-	-	-	-	1
		2.A.53.2.8	-	1	-	-	-	-	-	-
		2.A.53.7.1	1	1	1	1	1	1	1	1
2.A.54.	The Mitochondrial Tricarboxylate Carrier (MTC) Family	2.A.54.1.1	1	1	1	1	2	1	-	1
2.A.55.	The Metal Ion (Mn2+-iron) Transporter (Nramp) Family	2.A.55.1.1	-	-	-	-	-	2	-	-
		2.A.55.1.2	1	1	1	1	1	-	-	1
		2.A.55.1.3	-	-	-	-	-	-	2	-
2.A.57.	The Equilibrative Nucleoside Transporter (ENT) Family	2.A.57.3.1	1	1	1	1	1	-	-	-
2.A.59.	The Arsenical Resistance-3 (ACR3) Family	2.A.59.1.1	-	-	-	-	-	-	1	-
		2.A.59.1.2	3	1	1	1	1	1	-	-
2.A.66.	The Multidrug/Oligosaccharidyl-lipid /Polysaccharide (MOP) Flippase Superfamily	2.A.66.1.15	1	1	1	1	1	1	1	1
		2.A.66.1.5	1	1	1	1	1	2	1	2
		2.A.66.3.1	1	-	1	1	1	-	1	-

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
2.A.67.	The Oligopeptide Transporter (OPT) Family	2.A.67.1.1	1	1	2	1	2	1	5	-
		2.A.67.1.2	2	1	3	3	5	1	5	2
		2.A.67.1.3	1	-	1	1	-	1	1	1
		2.A.67.1.4	-	1	1	1	-	1	-	-
		2.A.67.3.1	-	-	-	-	-	-	1	-
2.A.72.	The K+ Uptake Permease (KUP) Family	2.A.72.2.1	-	-	-	-	-	-	1	-
		2.A.72.3.2	-	-	1	1	1	1	-	-
		2.A.72.3.4	-	-	-	-	-	-	1	-
2.A.89.	The Vacuolar Iron Transporter (VIT) Family	2.A.89.1.1	2	2	2	2	2	1	1	-
2.A.94.	The Phosphate Permease (Pho1) Family	2.A.94.1.2	-	-	1	1	-	1	-	-
3.A.1.	The ATP-binding Cassette (ABC) Superfamily	3.A.1.106.1	-	-	-	-	-	-	1	-
		3.A.1.120.1	-	-	-	-	-	-	-	1
		3.A.1.120.5	-	-	-	-	-	1	-	1
		3.A.1.121.2	1	1	1	1	1	-	-	-
		3.A.1.121.4	1	1	1	1	1	1	2	2
		3.A.1.201.1	1	2	1	1	2	1	-	-
		3.A.1.201.2	1	-	1	1	2	1	-	-
		3.A.1.201.3	1	-	1	-	-	2	1	1
		3.A.1.201.5	1	1	-	-	-	-	1	-
		3.A.1.201.6	-	1	1	1	1	-	-	-
		3.A.1.201.7	1	-	-	-	1	-	1	1
		3.A.1.201.9	1	-	1	1	1	1	-	-
		3.A.1.202.1	-	-	1	1	-	-	-	-
		3.A.1.203.1	1	1	1	1	1	1	2	-
		3.A.1.203.3	1	1	1	1	1	1	-	-
		3.A.1.204.2	-	-	-	-	-	-	1	-
		3.A.1.204.3	-	-	-	-	-	1	-	-
		3.A.1.204.4	-	-	1	-	-	-	2	-
		3.A.1.204.5	-	1	1	1	2	1	1	-
		3.A.1.204.6	-	-	-	-	1	-	-	-
		3.A.1.204.7	1	1	1	1	1	1	-	-
		3.A.1.205.1	1	1	2	1	1	-	-	-
		3.A.1.205.10	1	-	-	-	-	-	-	-
		3.A.1.205.11	1	1	1	1	1	-	1	1
		3.A.1.205.2	1	1	1	1	1	-	1	-
		3.A.1.205.3	1	-	-	-	-	-	-	-
		3.A.1.205.4	1	1	1	1	-	1	-	-
		3.A.1.205.6	2	1	1	1	1	1	1	-
		3.A.1.205.7	1	1	1	1	2	1	-	-
		3.A.1.206.1	1	-	1	1	-	1	-	-
		3.A.1.208.1	-	-	-	1	-	-	-	-
		3.A.1.208.10	-	-	1	-	-	1	-	-
		3.A.1.208.11	1	1	-	1	2	-	1	-
		3.A.1.208.12	2	1	1	1	1	1	1	1
		3.A.1.208.13	-	-	-	-	1	1	1	1
		3.A.1.208.14	1	1	1	1	1	-	1	-
		3.A.1.208.15	-	-	-	-	2	-	1	-
		3.A.1.208.16	-	1	1	1	1	-	1	1
		3.A.1.208.17	-	-	1	1	-	-	1	-
		3.A.1.208.18	-	-	-	-	1	-	-	-
		3.A.1.208.2	1	1	1	1	-	-	-	-
		3.A.1.208.3	-	1	1	2	1	-	1	-
		3.A.1.208.4	1	-	1	1	2	-	-	-
		3.A.1.208.5	1	1	-	1	1	1	2	-
		3.A.1.208.6	-	-	1	1	1	-	-	-
		3.A.1.208.7	1	-	-	-	-	1	1	-
		3.A.1.208.8	-	1	1	-	1	2	-	-
		3.A.1.208.9	1	1	1	1	1	-	-	-

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
		3.A.1.210.1	1	1	1	1	1	1	1	1
		3.A.1.210.2	1	1	1	1	1	1	-	1
		3.A.1.210.6	-	-	-	-	-	-	1	-
		3.A.1.211.2	1	1	-	-	-	1	-	-
		3.A.1.211.5	-	-	-	-	1	-	-	-
		3.A.1.212.1	1	1	1	1	1	1	1	-
		3.A.1.212.2	-	-	-	-	-	-	-	1
3.A.2.	The H+- or Na+-translocating F-type, V-type and A-type ATPase (F-ATPase) Superfamily	3.A.2.1.3	2	3	2	4	2	3	2	2
		3.A.2.2.3	3	5	4	4	4	3	3	3
		3.A.2.2.4	1	1	1	1	1	2	1	2
3.A.3.	The P-type ATPase (P-ATPase) Superfamily	3.A.3.1.1	1	-	2	1	1	-	1	-
		3.A.3.1.3	-	1	-	-	-	-	-	-
		3.A.3.1.4	1	-	-	-	-	-	-	-
		3.A.3.10.1	1	1	1	1	1	1	1	1
		3.A.3.13.1	1	-	1	-	-	-	-	-
		3.A.3.14.1	-	1	-	1	1	1	1	-
		3.A.3.15.1	-	-	-	-	-	-	-	1
		3.A.3.17.1	-	-	-	-	-	-	-	1
		3.A.3.2.1	-	1	-	-	-	1	-	-
		3.A.3.2.10	1	1	1	1	1	-	-	-
		3.A.3.2.11	-	-	-	-	-	-	1	-
		3.A.3.2.14	-	1	1	1	1	-	-	-
		3.A.3.2.15	1	-	-	-	1	1	-	-
		3.A.3.2.19	-	-	-	-	-	-	1	-
		3.A.3.2.2	1	1	1	1	1	-	-	1
		3.A.3.2.3	-	-	-	-	-	-	1	-
		3.A.3.2.5	-	-	-	-	-	-	-	1
		3.A.3.2.6	1	1	1	1	1	1	-	-
		3.A.3.2.7	1	1	1	1	1	1	-	-
		3.A.3.3.1	1	1	1	1	1	2	-	1
		3.A.3.3.6	1	1	1	1	1	-	-	-
		3.A.3.3.7	-	-	-	-	-	-	1	-
		3.A.3.5.14	1	-	-	-	1	1	1	-
		3.A.3.5.17	1	-	-	-	-	-	-	1
		3.A.3.5.3	-	-	-	-	-	-	1	-
		3.A.3.5.8	-	1	1	1	1	1	-	-
		3.A.3.5.9	1	1	1	1	1	1	1	-
		3.A.3.8.1	1	1	-	-	1	1	1	1
		3.A.3.8.2	2	1	1	1	1	1	1	1
		3.A.3.8.4	-	1	-	-	1	1	1	1
		3.A.3.8.5	1	-	1	1	-	1	1	1
		3.A.3.8.6	-	-	-	-	-	-	-	1
		3.A.3.9.1	-	-	-	-	-	1	-	-
		3.A.3.9.2	1	1	-	-	1	1	-	1
		3.A.3.9.3	1	1	1	1	1	1	-	-
		3.A.3.9.4	1	-	1	-	-	-	-	-
3.A.5.	The General Secretory Pathway (Sec) Family	3.A.5.8.1	1	1	1	1	1	1	-	-
		3.A.5.9.1	-	-	-	-	-	-	1	-
3.A.8.	The Mitochondrial Protein Translocase (MPT) Family	3.A.8.1.1	1	1	1	1	1	1	-	1
3.D.1.	The Proton-translocating NADH Dehydrogenase (NDH) Family	3.D.1.6.2	10	11	6	6	12	5	5	-
		3.D.1.6.4	-	-	-	-	-	-	-	1
3.D.2.	The Proton-translocating Transhydrogenase (PTH) Family	3.D.2.3.1	-	-	-	-	1	-	1	-
3.D.3.	The Proton-translocating Quinol:Cytochrome c Reductase (QCR) Superfamily	3.D.3.3.1	3	3	2	2	3	2	2	3
3.D.4.	The Proton-translocating Cytochrome Oxidase (COX) Superfamily	3.D.4.7.1	-	-	-	-	-	-	-	1

Continued on next page

Table 50 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Ann	Aor	Ncr	Pch	Spo
		3.D.4.8.1	4	4	1	1	4	1	2	3
3.E.1.	The Ion-translocating Microbial Rhodopsin (MR) Family	3.E.1.4.2	-	1	1	1	1	1	-	-
		3.E.1.4.3	1	-	-	-	-	-	-	-
		3.E.1.5.1	-	-	-	-	-	-	4	1

Appendix C

TCDB-Blast Results

This appendix presents the results of TCDB-Blast on the eight fungal genomes in our study.

C.1 TCDB-Blast Results for *A. niger* CBS 513.88

This section presents detailed statistics for TCDB-Blast when run on the *A. niger* CBS 513.88 genome. Table 51 presents the statistics of each alignment. The table is organised by TC-Family. The columns Family and Family Name contain the TC-Family identifier and its name. The column Query is the identifier for the entry in the *A. niger* CBS 513.88 genome. The column Hit is the UniProtKB identifier for the matching TCDB entry. The column TCID contains the TCID of the matching TCDB entry predicted by TCDB-Blast. The columns QTMS and HTMS contain the number of TMS for the query and the hit, respectively, as determined by HMMTOP. The last four columns contain the statistics for the blastp alignment between the query and the hit: %ID is the percent identity, QCov is the query coverage, SCov is the subject coverage (in this case the subject is the TCDB hit), Diff is the percent difference of the lengths of the query and hit, and eVal is the e-value.

Table 51: TCDB-Blast Results for *A. niger* CBS513.88

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
1.A.9	the neurotransmitter receptor, cys loop, ligand-gated ion channel (lic) family.	An07g10020	O95166	1.A.9.5.2	1	1	59.48	98	99	1	e-48
1.A.11	the ammonia transporter channel (amt) family.	An08g03200	O67997	1.A.11.1.4	11	12	43.31	86	94	9	e-84
		An08g03200	P40260	1.A.11.3.1	11	11	47.87	88	86	3	e-136
		An08g03200	P41948	1.A.11.3.2	11	11	51.84	96	92	4	e-156
		An14g02390	P41948	1.A.11.3.2	11	11	46.32	89	84	5	e-118
		An08g03200	Q8NKD5	1.A.11.3.3	11	11	63.89	95	96	0	0
		An14g02390	Q8NKD5	1.A.11.3.3	11	11	45.11	88	88	1	e-121
		An08g03200	Q96UY0	1.A.11.3.4	11	11	63.90	88	89	2	0
		An14g02390	Q96UY0	1.A.11.3.4	11	11	46.25	84	85	1	e-117
		An08g03200	Q59UP8	1.A.11.3.5	11	11	52.62	88	88	0	e-148
		An14g02390	Q59UP8	1.A.11.3.5	11	11	47.62	89	88	1	e-133
1.A.17	the calcium-dependent chloride channel (ca-clc) family.	An14g03020	B0YES0	1.A.17.6.4	7	7	75.14	99	99	0	0
		An14g01960	B0YES0	1.A.17.6.4	8	7	43.56	89	89	0	0
1.A.23	the small conductance mechanosensitive ion channel (mscs) family.	An15g03150	F9X0Q3	1.A.23.4.9	6	6	55.44	85	77	9	0
1.A.33	the cation channel-forming heat shock protein-70 (hsp70) family.	An11g04180	P0A6Y8	1.A.33.1.2	1	1	47.55	91	96	5	e-172
		An16g09260	P0A6Y8	1.A.33.1.2	1	1	44.48	99	95	4	e-156
		An11g04180	P08107	1.A.33.1.3	1	1	60.79	90	95	5	0
		An16g09260	P08107	1.A.33.1.3	1	1	56.86	97	93	4	0
1.A.46	the anion channel-forming bestrophin (bestrophin) family.	An14g05100	Q5AXS1	1.A.46.2.2	3	3	69.07	94	96	2	e-176
1.A.56	the copper transporter (ctr) family.	An02g11700	A9XIK8	1.A.56.1.10	3	3	47.33	91	82	9	e-40
1.A.77	the mg(2+)/ca(2+) uniporter (mcu) family.	An04g06590	Q7S4I4	1.A.77.1.5	2	2	44.96	80	79	2	e-101
1.A.88	the fungal potassium channel (f-kch) family.	An11g03330	A2QW01	1.A.88.1.4	4	4	95.49	100	100	0	0
1.B.69	the peroxysomal membrane porin 4 (pxmp4) family.	An16g08040	A2R8R0	1.B.69.1.4	4	4	100.00	100	100	0	e-160
		An16g08040	B0CP94	1.B.69.1.6	4	4	41.59	99	101	3	e-50
1.F.1	the synaptosomal vesicle fusion pore (svf-pore) family.	An12g07570	P33328	1.F.1.1.2	1	1	55.32	79	82	3	e-30
1.H.1	the claudin tight junction (claudin) family.	An08g01170	F5H8T9	1.H.1.4.1	4	5	46.64	87	90	3	e-78
		An07g08960	G3XZI4	1.H.1.4.3	5	5	80.91	100	100	0	0
2.A.1	the major facilitator superfamily (mfs).	An05g01290	P43581	2.A.1.1.5	12	12	43.34	87	87	1	e-117
		An03g02190	P13181	2.A.1.1.6	12	12	40.57	87	85	3	e-127
		An08g03850	P11636	2.A.1.1.7	12	12	55.58	100	100	0	0
		An04g00340	P30605	2.A.1.1.8	12	12	40.56	98	92	7	e-114
		An03g02190	O74969	2.A.1.1.21	12	12	41.60	85	90	5	e-135
		An03g02190	O74849	2.A.1.1.22	12	12	40.59	86	89	4	e-138
		An05g01290	O74849	2.A.1.1.22	12	12	40.12	92	93	1	e-125
		An05g01290	P39004	2.A.1.1.31	12	12	40.29	90	86	5	e-121
		An15g01500	Q8NJ22	2.A.1.1.33	12	12	57.72	85	84	2	0
		An06g02270	Q8NJ22	2.A.1.1.33	12	12	44.33	102	101	1	e-148
		An02g03540	Q400D8	2.A.1.1.36	12	12	71.28	101	100	0	0
		An03g02190	Q400D8	2.A.1.1.36	12	12	52.15	92	91	1	0
		An05g01290	Q400D8	2.A.1.1.36	12	12	50.68	95	91	4	0
		An14g02740	P39932	2.A.1.1.38	12	12	47.16	99	93	6	e-169
		An09g02930	P39932	2.A.1.1.38	12	12	44.47	96	89	7	e-152
		An14g03990	P39932	2.A.1.1.38	12	12	41.12	97	91	7	e-145
		An11g01100	P49374	2.A.1.1.39	12	12	47.01	93	91	2	e-156
		An02g00590	P49374	2.A.1.1.39	12	12	43.19	101	100	1	e-160
		An03g01620	P49374	2.A.1.1.39	12	12	40.84	102	95	7	e-129

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
		An01g00850	Q64L87	2.A.1.1.40	12	12	41.70	88	96	9	e-126
		An15g03940	Q2MEV7	2.A.1.1.51	12	12	51.63	93	94	1	e-162
		An12g07450	Q2MEV7	2.A.1.1.51	12	12	44.05	95	97	2	e-122
		An12g07450	Q8J0V1	2.A.1.1.57	12	12	91.13	100	100	0	0
		An15g03940	Q8J0V1	2.A.1.1.57	12	12	45.98	92	92	1	e-128
		An02g03540	Q8J0U9	2.A.1.1.58	12	12	89.20	96	92	4	0
		An05g01290	Q8J0U9	2.A.1.1.58	12	12	53.95	96	89	7	0
		An03g02190	Q8J0U9	2.A.1.1.58	12	12	53.50	95	91	4	0
		An03g02190	Q2MDH1	2.A.1.1.67	12	12	41.33	94	96	2	e-137
		An05g01290	Q2MDH1	2.A.1.1.67	12	12	40.41	99	98	1	e-118
		An15g03940	A3M0N3	2.A.1.1.68	12	12	50.39	97	97	0	e-170
		An12g07450	A3M0N3	2.A.1.1.68	12	12	42.20	94	95	0	e-114
		An15g01500	Q0ULF7	2.A.1.1.70	12	12	69.15	90	87	3	0
		An06g02270	Q0ULF7	2.A.1.1.70	12	12	42.21	93	92	2	e-118
		An14g02740	Q5A8J5	2.A.1.1.73	12	12	53.16	95	93	2	0
		An09g02930	Q5A8J5	2.A.1.1.73	12	12	49.43	100	96	3	e-172
		An14g03990	Q5A8J5	2.A.1.1.73	12	12	46.59	96	94	3	e-163
		An03g02190	P54862	2.A.1.1.105	12	12	40.35	93	91	2	e-130
		An05g01290	P32465	2.A.1.1.108	12	12	42.62	87	83	5	e-122
		An02g03540	P32465	2.A.1.1.108	12	12	42.44	85	84	2	e-129
		An05g01290	P39924	2.A.1.1.110	12	12	40.12	92	89	4	e-120
		An05g01290	P23585	2.A.1.1.111	12	12	41.76	94	94	0	e-125
		An03g02190	P23585	2.A.1.1.111	12	12	40.31	92	94	3	e-132
		An05g01290	Q9P3U6	2.A.1.1.112	12	12	40.04	91	88	3	e-121
		An15g03940	G4N740	2.A.1.1.117	12	12	47.71	95	92	4	e-161
		An18g01720	P28873	2.A.1.2.6	11	11	42.02	83	84	1	e-115
		An09g03320	Q07824	2.A.1.2.16	12	12	43.60	91	88	4	e-129
		An18g01150	Q07824	2.A.1.2.16	12	12	40.49	87	90	3	e-116
		An01g11540	Q07824	2.A.1.2.16	12	12	40.17	84	82	3	e-121
		An16g02610	P38124	2.A.1.2.17	12	12	40.76	84	86	2	e-105
		An18g01720	P38124	2.A.1.2.17	11	12	40.13	81	85	4	e-114
		An15g04060	Q70WR7	2.A.1.2.23	11	12	60.19	89	84	6	0
		An18g01720	O94528	2.A.1.2.35	11	12	50.21	82	89	7	e-161
		An16g02610	O94528	2.A.1.2.35	12	12	46.58	81	85	6	e-133
		An15g04060	C5E4Z7	2.A.1.2.45	11	12	60.23	89	82	8	0
		An15g04060	C5DX43	2.A.1.2.46	11	12	57.39	89	86	4	0
		An09g01910	A2QTF4	2.A.1.2.48	9	9	90.80	100	100	0	0
		An04g08300	P53283	2.A.1.2.67	12	11	54.36	106	95	10	0
		An02g09970	Q8NKG7	2.A.1.2.77	12	12	69.74	78	83	6	0
		An17g01070	Q8NKG7	2.A.1.2.77	11	12	43.86	81	74	8	e-110
		An04g08300	Q8NKG7	2.A.1.2.77	12	12	42.79	83	82	1	e-129
		An02g03620	Q8NKG7	2.A.1.2.77	12	12	41.84	78	85	9	e-119
		An08g06980	Q8NKG7	2.A.1.2.77	12	12	40.86	77	83	7	e-113
		An02g09970	B6HIC2	2.A.1.2.78	12	12	72.99	85	93	9	0
		An17g01070	B6HIC2	2.A.1.2.78	11	12	41.45	81	76	6	e-104
		An04g08300	B6H9Q3	2.A.1.2.85	12	12	78.79	99	97	2	0
		An04g07680	B6H9Q3	2.A.1.2.85	12	12	76.34	98	96	2	0
		An02g09970	B6H9Q3	2.A.1.2.85	12	12	41.44	79	84	6	e-113
		An02g03620	B6H9Q3	2.A.1.2.85	12	12	40.72	77	84	8	e-105
		An17g01070	B6H9Q3	2.A.1.2.85	11	12	40.00	80	73	9	e-92
		An16g00090	B6HN82	2.A.1.2.86	12	12	56.94	100	96	3	0
		An04g08250	B6HN82	2.A.1.2.86	12	12	45.49	92	92	1	e-135
		An02g03670	B6HN82	2.A.1.2.86	12	12	44.99	94	96	3	e-141
		An08g10970	B6HN82	2.A.1.2.86	12	12	42.51	101	96	5	e-125
		An08g08220	Q08902	2.A.1.3.52	14	14	57.89	89	92	4	0
		An08g08710	Q08902	2.A.1.3.52	14	14	51.45	101	100	0	e-180
		An10g00700	Q08902	2.A.1.3.52	14	14	40.87	99	102	3	e-127

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
		An12g08620	H2E274	2.A.1.3.65	14	14	49.20	97	100	3	0
		An01g11290	H2E274	2.A.1.3.65	15	14	46.93	89	90	1	e-157
		An09g00870	H2E274	2.A.1.3.65	13	14	44.76	90	88	2	e-138
		An01g15000	H2E274	2.A.1.3.65	14	14	44.55	89	90	0	e-146
		An06g00770	H2E274	2.A.1.3.65	14	14	42.38	84	85	1	e-123
		An08g05670	P22152	2.A.1.8.5	12	12	65.42	100	100	1	0
		An08g05670	Q8X193	2.A.1.8.13	12	12	56.20	99	101	1	0
		An16g06190	P25346	2.A.1.9.7	12	13	51.24	97	93	4	e-169
		An16g01940	P40445	2.A.1.14.38	11	12	43.18	88	91	2	e-147
		An01g11450	P40445	2.A.1.14.38	11	12	43.07	90	99	9	e-150
		An08g06430	P40445	2.A.1.14.38	9	12	40.75	92	99	7	e-139
		An07g00980	P40445	2.A.1.14.38	10	12	40.39	91	95	5	e-132
		An01g00720	P39980	2.A.1.16.1	14	15	41.61	101	93	8	e-145
		An03g03560	Q870L2	2.A.1.16.7	14	14	57.56	90	90	1	0
		An07g06240	Q870L2	2.A.1.16.7	14	14	41.19	100	95	5	e-154
		An12g00940	Q9C101	2.A.1.19.38	11	11	46.51	86	90	5	e-150
		An07g07980	Q9C101	2.A.1.19.38	12	11	40.98	87	92	5	e-125
		An16g09020	Q5A7S4	2.A.1.58.1	12	10	47.19	95	95	0	e-151
		An06g02510	Q5A7S4	2.A.1.58.1	11	10	43.37	90	82	10	e-111
		An06g02510	Q01HW9	2.A.1.58.4	11	11	41.76	94	92	2	e-81
		An09g02880	C9S7Y7	2.A.1.58.5	10	10	40.17	76	79	4	e-67
		An14g04560	E9CYW5	2.A.1.75.2	12	12	50.91	99	103	4	0
2.A.3	the amino acid-polyamine-organocation (apc) family.	An15g01900	P19807	2.A.3.4.1	12	12	51.25	93	85	8	e-161
		An09g05010	P19807	2.A.3.4.1	12	12	48.80	97	89	8	e-163
		An16g02000	Q9Y860	2.A.3.4.2	12	12	69.88	96	96	0	0
		An09g02550	Q9Y860	2.A.3.4.2	12	12	54.81	99	98	0	0
		An14g01850	P32837	2.A.3.4.3	12	12	43.61	91	86	5	e-140
		An17g01540	P32837	2.A.3.4.3	12	12	43.00	93	85	8	e-131
		An02g09790	Q9UT18	2.A.3.4.6	12	12	45.23	97	96	1	e-154
		An04g03940	P50276	2.A.3.8.4	12	11	53.97	92	88	4	0
		An13g00840	P06775	2.A.3.10.1	12	12	43.21	98	95	3	e-160
		An13g00840	P19145	2.A.3.10.2	12	12	51.85	102	99	3	0
		An13g03650	P04817	2.A.3.10.4	12	12	45.26	92	88	4	e-161
		An09g06730	P04817	2.A.3.10.4	12	12	45.22	92	85	8	e-156
		An13g00840	P38967	2.A.3.10.8	12	12	44.31	98	96	1	e-149
		An13g03650	P32487	2.A.3.10.10	12	12	43.41	106	98	8	e-167
		An13g03650	P38971	2.A.3.10.11	12	12	43.19	100	99	1	e-151
		An09g06730	P38971	2.A.3.10.11	12	12	42.97	95	91	5	e-151
		An12g04180	P53388	2.A.3.10.13	12	12	50.27	98	90	7	e-169
		An13g03650	P53388	2.A.3.10.13	12	12	41.19	89	83	7	e-119
		An12g10130	Q8J266	2.A.3.10.17	12	12	40.72	98	96	2	e-115
		An09g00400	Q8J266	2.A.3.10.17	11	12	40.55	73	79	7	e-95
		An09g00400	Q8NKC4	2.A.3.10.18	11	13	78.92	72	76	6	0
		An05g01740	Q8NKC4	2.A.3.10.18	11	13	42.66	86	89	3	e-142
		An04g00530	P38090	2.A.3.10.19	12	12	45.89	95	92	3	e-155
		An04g09620	P38090	2.A.3.10.19	12	12	41.15	97	90	8	e-136
		An09g06730	P43059	2.A.3.10.20	12	12	43.64	97	92	5	e-134
		An13g03650	P43059	2.A.3.10.20	12	12	43.16	93	92	1	e-131
		An13g00840	Q9URZ4	2.A.3.10.21	12	12	45.92	101	101	1	e-166
		An13g00840	Q2VQZ4	2.A.3.10.22	12	12	48.30	91	99	8	e-152
		An13g00840	Q5AG77	2.A.3.10.23	12	12	48.44	98	99	1	e-179
		An13g00840	Q59YT0	2.A.3.10.24	12	12	57.04	97	97	1	0
		An13g00840	Q59WB3	2.A.3.10.25	12	12	45.22	100	97	4	e-146
		An13g00840	Q59NZ6	2.A.3.10.26	12	12	45.17	92	93	2	e-152
		An13g00840	O60170	2.A.3.10.28	12	12	44.44	88	89	2	e-144
2.A.4	the cation diffusion facilitator (cdf) family.	An15g03900	P20107	2.A.4.2.2	6	7	42.54	93	91	2	e-103

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
2.A.5	the zinc (zn(2+))-iron	An01g01620	P32804	2.A.5.1.1	8	8	58.47	105	97	7	e-141
	(fe(2+)) permease (zip)	An15g07190	P32804	2.A.5.1.1	8	8	55.71	104	98	6	e-136
	family.	An01g06690	P32804	2.A.5.1.1	7	8	43.09	89	83	7	e-76
2.A.6	the resistance-nodulation-cell	An11g05000	Q12200	2.A.6.6.3	13	13	40.21	97	106	8	0
	division (rnd) superfamily.										
2.A.7	the drug/metabolite	An17g02140	Q5A477	2.A.7.13.2	10	9	68.08	81	83	3	e-138
	transporter (dmt)	An03g03820	Q4WUA9	2.A.7.24.11	10	10	67.40	96	94	2	0
	superfamily.	An01g00340	Q4WUA9	2.A.7.24.11	10	10	52.58	97	89	8	e-147
2.A.16	the telurite-resistance/	An12g00870	A2QYD7	2.A.16.4.1	9	9	96.68	100	100	0	0
	dicarboxylate transporter	An12g00870	A3R044	2.A.16.4.2	9	10	50.44	88	91	4	e-111
	(tdt) family.	An12g00870	Q2TJJ2	2.A.16.4.3	9	10	77.29	92	87	5	0
2.A.17	the proton-dependent	An12g01210	Q9P380	2.A.17.2.1	11	12	42.25	98	92	6	e-164
	oligopeptide transporter	An08g04600	Q9P380	2.A.17.2.1	11	12	40.98	90	89	1	e-124
	(pot) family.	An12g01210	P32901	2.A.17.2.2	11	12	46.81	94	91	3	e-177
2.A.18	the amino acid/auxin	An15g07550	P38680	2.A.18.4.1	11	11	56.72	95	93	1	e-168
	permease (aaap) family.	An09g03660	P38680	2.A.18.4.1	11	11	52.72	90	90	0	e-143
		An16g05880	P38680	2.A.18.4.1	11	11	48.16	92	92	0	e-129
		An15g07550	Q6IT47	2.A.18.4.2	11	11	64.86	99	100	0	0
		An09g03660	Q6IT47	2.A.18.4.2	11	11	52.89	92	94	2	e-148
		An16g05880	Q6IT47	2.A.18.4.2	11	11	52.26	94	96	2	e-152
		An04g02150	P36062	2.A.18.7.1	11	11	42.32	82	88	8	e-127
2.A.19	the ca(2+):cation antiporter	An01g03100	Q99385	2.A.19.2.2	11	11	51.00	93	98	5	e-119
	(caca) family.	An19g00340	Q99385	2.A.19.2.2	11	11	41.03	95	90	5	e-75
	(caca) family.	An01g03100	O59940	2.A.19.2.8	11	10	53.69	91	89	2	e-132
2.A.21	the solute:sodium symporter	An01g03790	P33413	2.A.21.6.1	15	15	46.82	102	94	7	0
	(sss) family.	An18g01360	P33413	2.A.21.6.1	15	15	40.25	96	89	7	e-159
		An01g03790	Q9FHJ8	2.A.21.6.2	15	15	40.74	95	93	2	e-141
		An01g03790	Q59VF2	2.A.21.6.4	15	15	45.68	95	89	6	0
2.A.29	the mitochondrial carrier	An18g04220	P05141	2.A.29.1.1	6	4	48.29	92	98	7	e-85
	(mc) family.	An18g04220	P12235	2.A.29.1.2	6	6	47.26	92	98	7	e-84
		An18g04220	P04710	2.A.29.1.3	6	4	66.02	97	100	3	e-152
		An18g04220	Q8TFA7	2.A.29.1.4	6	4	64.95	91	94	3	e-135
		An18g04220	Q8LB08	2.A.29.1.6	6	4	58.56	92	95	4	e-122
		An18g04220	P18239	2.A.29.1.7	6	4	73.74	93	93	0	e-161
		An18g04220	Q9H0C2	2.A.29.1.8	6	5	49.17	95	96	1	e-89
		An18g04220	P18238	2.A.29.1.9	6	6	71.04	93	97	4	e-156
		An18g04220	P12236	2.A.29.1.10	6	6	47.12	92	99	7	e-84
		An02g01730	P22292	2.A.29.2.1	5	6	40.56	91	91	0	e-66
		An02g01730	O89035	2.A.29.2.2	5	2	45.45	88	96	9	e-76
		An02g01730	Q06143	2.A.29.2.3	5	6	49.29	89	94	5	e-86
		An08g01370	Q99297	2.A.29.2.5	3	1	55.19	101	100	1	e-120
		An11g02540	Q8SF04	2.A.29.2.6	6	4	43.45	92	98	6	e-63
		An02g01730	Q9UBX3	2.A.29.2.7	5	3	45.82	88	96	8	e-74
		An08g01370	Q03028	2.A.29.2.8	3	4	56.54	100	99	2	e-113
		An11g02540	Q8IB73	2.A.29.2.10	6	6	43.94	92	91	1	e-76
		An02g01730	Q9CR62	2.A.29.2.11	5	5	43.06	92	92	0	e-68
		An02g01730	Q02978	2.A.29.2.13	5	6	41.61	91	91	0	e-67
		An02g12070	P12234	2.A.29.4.1	4	6	48.08	76	79	4	e-83
		An02g12070	Q00325	2.A.29.4.2	4	6	47.39	76	79	4	e-82
		An01g13600	P23641	2.A.29.4.3	6	6	57.93	92	93	1	e-115
		An02g04160	P23641	2.A.29.4.3	4	6	41.78	95	94	1	e-72
		An02g04160	P40035	2.A.29.4.4	4	6	63.05	96	98	2	e-136
		An02g12070	Q8VEM8	2.A.29.4.5	4	6	48.78	76	80	6	e-90
		An02g12070	Q9FMU6	2.A.29.4.6	4	7	55.17	84	85	1	e-111
		An06g01730	P10566	2.A.29.5.1	6	6	48.63	92	93	2	e-88
		An06g01730	P23500	2.A.29.5.2	6	6	45.70	95	99	5	e-85
		An06g01730	Q287T7	2.A.29.5.3	6	1	40.98	96	92	4	e-68

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
		An06g01730	Q920G8	2.A.29.5.5	6	1	40.89	91	86	6	e-66
		An06g01730	Q9NYZ2	2.A.29.5.7	6	1	41.45	95	90	6	e-68
		An11g11230	P38152	2.A.29.7.3	3	4	47.65	101	100	1	e-85
		An18g00070	P38152	2.A.29.7.3	2	4	40.15	91	90	1	e-59
		An11g11230	Q7KSQ0	2.A.29.7.4	3	6	40.14	96	90	7	e-62
		An03g03360	Q27257	2.A.29.8.2	6	6	40.61	70	73	4	e-41
		An03g03360	Q12289	2.A.29.8.4	6	5	40.41	75	75	1	e-50
		An18g05590	P38087	2.A.29.8.11	2	6	45.67	99	91	8	e-82
		An18g05590	P32331	2.A.29.8.12	2	4	50.34	95	94	1	e-91
		An03g06860	Q01356	2.A.29.9.1	5	3	53.73	98	89	9	e-111
		An14g01860	P38127	2.A.29.10.4	5	4	43.93	99	92	7	e-88
		An04g01190	P40556	2.A.29.10.5	4	4	40.69	81	85	5	e-67
		An14g01860	Q9BSK2	2.A.29.10.7	5	6	41.05	93	101	8	e-63
		An04g09030	P33303	2.A.29.13.1	1	2	62.54	94	95	1	e-131
		An07g03070	O75746	2.A.29.14.1	5	3	43.08	93	95	2	e-156
		An07g10010	P38988	2.A.29.21.1	5	5	71.67	95	98	3	e-154
		An09g06670	Q04013	2.A.29.29.1	2	2	66.45	96	97	1	e-143
		An02g11090	Q04013	2.A.29.29.1	5	2	52.05	93	93	0	e-105
2.A.39	the nucleobase:cation symporter-1 (ncs1) family.	An08g06240	Q10279	2.A.39.3.7	12	13	45.64	96	93	4	e-157
2.A.40	the nucleobase:cation symporter-2 (ncs2) family.	An07g01950	Q07307	2.A.40.4.1	15	12	59.53	95	96	1	0
		An02g00560	Q07307	2.A.40.4.1	13	12	46.46	81	89	8	e-156
		An07g01950	P48777	2.A.40.4.4	15	14	75.23	94	94	0	0
		An02g00560	P48777	2.A.40.4.4	13	14	45.21	84	90	7	e-148
		An13g02390	Q7Z8R3	2.A.40.7.1	10	12	67.37	74	74	1	0
2.A.41	the concentrative nucleoside transporter (cnt) family.	An08g10300	Q874I3	2.A.41.2.7	13	12	42.49	98	96	1	e-150
2.A.43	the lysosomal cystine transporter (lct) family.	An09g06510	P38279	2.A.43.2.7	7	7	42.95	99	105	6	e-74
2.A.47	the divalent anion:na(+) symporter (dass) family.	An01g03120	P25360	2.A.47.2.1	11	10	40.85	73	69	5	e-160
		An01g03120	P27514	2.A.47.2.2	11	12	40.23	106	105	1	0
		An01g03120	P39535	2.A.47.2.3	11	12	40.79	72	72	0	e-157
2.A.52	the ni(2+)-co(2+) transporter (nicot) family.	An12g04470	Q7S3L8	2.A.52.1.8	8	7	54.39	77	83	7	e-123
2.A.53	the sulfate permease (sulp) family.	An15g04600	P23622	2.A.53.1.2	15	13	48.11	101	100	1	0
2.A.55	the metal ion (mn(2+)-iron) transporter (nramp) family.	An04g05680	P38925	2.A.55.1.1	11	11	50.82	75	75	0	e-144
		An04g05680	P38778	2.A.55.1.2	11	10	51.93	72	75	5	e-139
		An04g05680	Q10177	2.A.55.1.4	11	11	52.63	73	80	9	e-145
2.A.59	the arsenical resistance-3 (acr3) family.	An18g03550	Q06598	2.A.59.1.1	10	10	40.12	91	84	8	e-75
		An18g03550	P45946	2.A.59.1.2	10	10	46.33	92	99	7	e-95
2.A.66	the multidrug/oligosaccharidyl-lipid/polysaccharide (mop) flippase superfamily.	An08g07590	P38767	2.A.66.1.5	12	11	45.47	73	78	7	e-122
2.A.67	the oligopeptide transporter (opt) family.	An14g05290	O14411	2.A.67.1.1	15	19	43.12	100	98	2	0
		An11g05350	O14411	2.A.67.1.1	16	19	41.89	101	100	1	0
		An14g05290	P40900	2.A.67.1.2	15	17	43.40	99	97	2	0
		An11g05350	P40900	2.A.67.1.2	16	17	41.65	100	99	1	0
		An16g00810	P40897	2.A.67.1.3	14	15	43.41	93	90	3	0
		An16g00810	O14031	2.A.67.1.5	14	15	50.07	93	85	9	0
		An14g05290	O14031	2.A.67.1.5	15	15	47.74	95	86	9	0
		An11g05350	O14031	2.A.67.1.5	16	15	45.55	94	86	9	0
		An11g03640	O14031	2.A.67.1.5	15	15	41.67	89	83	6	e-177
2.A.69	the auxin efflux carrier (aec) family.	An01g11100	B8MZ51	2.A.69.2.3	10	10	72.04	101	99	2	0

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
2.A.72	the k(+) uptake permease (kup) family.	An02g05630	O74724	2.A.72.3.2	13	14	54.48	99	91	8	0
2.A.89	the vacuolar iron transporter (vit) family.	An16g03690	P47818	2.A.89.1.1	5	5	46.15	74	69	8	e-48
2.A.96	the acetate uptake transporter (acetr) family.	An07g08810	Q5B2K4	2.A.96.1.3	6	6	69.84	88	85	3	e-116
		An13g02020	Q5B2K4	2.A.96.1.3	7	6	61.34	79	80	1	e-96
		An13g02020	P25613	2.A.96.1.4	7	6	45.95	74	78	6	e-49
		An07g08810	O14201	2.A.96.1.6	6	6	40.64	76	72	5	e-44
		An13g02020	P32907	2.A.96.1.7	7	6	45.05	74	79	6	e-49
2.A.105	the mitochondrial pyruvate carrier (mpc) family.	An04g02140	P53157	2.A.105.1.1	2	2	59.81	86	82	4	e-40
2.A.108	the iron/lead transporter (ilt) family.	An01g08950	P40088	2.A.108.1.1	7	7	50.00	81	77	4	e-106
		An15g05520	P38993	2.A.108.1.1	1	1	48.51	98	95	2	0
		An01g08960	P38993	2.A.108.1.1	1	1	48.21	95	92	4	0
		An16g01130	P40088	2.A.108.1.1	7	7	46.33	89	84	5	e-96
		An15g05510	P40088	2.A.108.1.1	7	7	43.33	95	89	7	e-102
		An01g08950	Q9P8U9	2.A.108.1.2	7	7	54.69	83	84	2	e-117
		An16g01130	Q9P8U9	2.A.108.1.2	7	7	50.14	90	91	0	e-109
		An15g05510	Q9P8U9	2.A.108.1.2	7	7	46.24	99	98	1	e-113
		An01g08950	Q9P8U8	2.A.108.1.3	7	7	51.94	80	81	1	e-112
		An16g01130	Q9P8U8	2.A.108.1.3	7	7	47.49	89	89	0	e-104
		An15g05510	Q9P8U8	2.A.108.1.3	7	7	46.41	96	95	1	e-115
		An15g05520	P43561	2.A.108.1.4	1	1	46.63	96	96	0	0
		An01g08960	P43561	2.A.108.1.4	1	1	43.87	99	97	1	e-171
		An01g08950	Q09919	2.A.108.1.5	7	7	50.47	83	81	3	e-110
		An16g01130	Q09919	2.A.108.1.5	7	7	45.92	93	89	4	e-100
		An15g05510	Q09919	2.A.108.1.5	7	7	43.90	87	83	5	e-92
3.A.1	the atp-binding cassette (abc) superfamily.	An17g01770	P08183	3.A.1.201.1	12	12	41.10	48	99	1	0
		An17g01770	P21439	3.A.1.201.3	12	12	40.03	49	99	1	0
		An17g01770	B0Y3B6	3.A.1.201.10	12	12	79.79	49	94	6	0
		An04g08340	B0Y3B6	3.A.1.201.10	9	12	58.24	99	95	4	0
		An17g01770	I0DHH7	3.A.1.201.16	12	12	40.19	46	97	2	0
		An04g07060	Q9NRK6	3.A.1.201.17	6	6	44.84	75	80	6	e-162
		An04g08340	P36619	3.A.1.201.18	9	13	42.19	44	97	5	0
		An08g05780	P28288	3.A.1.203.1	3	5	41.87	82	88	7	e-159
		An08g05780	P33897	3.A.1.203.3	3	4	44.13	89	85	5	0
		An08g05780	Q9UBJ2	3.A.1.203.7	3	5	44.20	89	85	5	0
		An01g03680	Q9UBJ2	3.A.1.203.7	4	5	40.26	84	92	10	e-162
		An08g05780	I7MJ28	3.A.1.203.10	3	6	41.95	83	81	2	e-163
		An01g12380	P33302	3.A.1.205.1	12	15	48.91	92	94	2	0
		An15g02930	P33302	3.A.1.205.1	16	15	48.57	96	95	1	0
		An05g01660	P33302	3.A.1.205.1	11	15	46.72	101	100	1	0
		An08g03300	P33302	3.A.1.205.1	11	15	46.02	45	94	4	0
		An08g04500	P33302	3.A.1.205.1	11	15	45.41	97	94	3	0
		An13g03570	P33302	3.A.1.205.1	13	15	45.09	100	98	2	0
		An07g01250	P33302	3.A.1.205.1	14	15	42.39	104	99	5	0
		An01g12380	P32568	3.A.1.205.2	12	12	41.01	94	96	2	0
		An07g01250	P32568	3.A.1.205.2	14	12	40.48	96	92	4	0
		An08g03300	P32568	3.A.1.205.2	11	12	40.35	44	90	3	0
		An07g01250	Q02785	3.A.1.205.3	14	15	40.51	90	86	5	0
		An01g12380	P43071	3.A.1.205.4	12	13	51.30	92	95	2	0
		An05g01660	P43071	3.A.1.205.4	11	13	50.91	100	99	0	0
		An15g02930	P43071	3.A.1.205.4	16	13	48.73	100	100	1	0
		An13g03570	P43071	3.A.1.205.4	13	13	46.99	99	97	2	0
		An08g03300	P43071	3.A.1.205.4	11	13	45.30	44	92	3	0
		An08g04500	P43071	3.A.1.205.4	11	13	44.48	100	98	2	0
		An07g01250	P43071	3.A.1.205.4	14	13	43.40	99	95	4	0

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
		An15g02930	P78595	3.A.1.205.5	16	11	49.40	96	95	1	0
		An01g12380	P78595	3.A.1.205.5	12	11	49.22	92	95	3	0
		An05g01660	P78595	3.A.1.205.5	11	11	49.12	99	99	0	0
		An13g03570	P78595	3.A.1.205.5	13	11	45.65	100	99	1	0
		An08g03300	P78595	3.A.1.205.5	11	11	44.61	45	92	3	0
		An08g04500	P78595	3.A.1.205.5	11	11	43.73	96	95	2	0
		An07g01250	P78595	3.A.1.205.5	14	11	42.59	102	98	4	0
		An13g03060	Q8X0Z3	3.A.1.205.6	11	14	40.65	45	90	8	0
		An15g01130	Q8X0Z3	3.A.1.205.6	15	14	40.31	88	88	1	0
		An13g03060	P78577	3.A.1.205.7	11	11	78.66	97	96	0	0
		An14g03570	P78577	3.A.1.205.7	14	11	66.09	96	97	0	0
		An14g02610	P78577	3.A.1.205.7	11	11	47.50	100	96	5	0
		An11g02110	P78577	3.A.1.205.7	12	11	42.78	92	96	4	0
		An08g03300	P41820	3.A.1.205.11	11	13	44.68	45	91	5	0
		An15g02930	P41820	3.A.1.205.11	16	13	43.75	93	91	3	0
		An05g01660	P41820	3.A.1.205.11	11	13	42.77	95	93	2	0
		An07g01250	P41820	3.A.1.205.11	14	13	42.43	98	92	6	0
		An08g04500	P41820	3.A.1.205.11	11	13	41.67	97	93	4	0
		An01g12380	P41820	3.A.1.205.11	12	13	41.21	93	93	1	0
		An13g03570	P41820	3.A.1.205.11	13	13	40.57	97	94	3	0
		An11g02110	P41820	3.A.1.205.11	12	13	40.01	43	90	3	0
		An15g02930	P51533	3.A.1.205.12	16	15	46.99	99	94	5	0
		An01g12380	P51533	3.A.1.205.12	12	15	46.30	96	94	2	0
		An05g01660	P51533	3.A.1.205.12	11	15	44.94	106	101	4	0
		An08g03300	P51533	3.A.1.205.12	11	15	44.65	46	92	7	0
		An07g01250	P51533	3.A.1.205.12	14	15	44.39	93	86	8	0
		An13g03570	P51533	3.A.1.205.12	13	15	42.46	105	99	6	0
		An08g04500	P51533	3.A.1.205.12	11	15	42.37	100	94	6	0
		An03g04060	Q92887	3.A.1.208.2	13	16	41.21	82	82	0	0
		An03g04060	P39109	3.A.1.208.11	13	14	48.97	100	102	2	0
		An03g04060	Q10185	3.A.1.208.16	13	16	45.40	98	102	4	0
		An03g04060	Q9P5N0	3.A.1.208.28	13	12	42.01	97	102	5	0
		An03g04060	D2WF19	3.A.1.208.32	13	16	46.91	100	101	2	0
		An08g10600	P40416	3.A.1.210.1	5	5	56.73	95	99	5	0
		An07g07500	Q02592	3.A.1.210.2	11	10	43.19	89	98	10	0
		An08g10600	O75027	3.A.1.210.4	5	5	52.29	85	81	4	0
		An08g10600	Q9XUJ1	3.A.1.210.7	5	10	43.97	80	72	10	e-157
		An08g10600	Q9LVM1	3.A.1.210.8	5	7	54.56	83	83	1	0
		An04g07060	P33311	3.A.1.212.2	6	5	47.97	75	76	2	0
3.A.2	the h(+)- or na(+)-translocating f-type, v-type and a-type atpase (f-atpase) superfamily.	An16g07290	P05626	3.A.2.1.3	2	2	45.79	88	88	0	e-65
		An02g08020	P25515	3.A.2.2.3	4	4	71.70	99	99	1	e-75
		An10g00680	P25515	3.A.2.2.3	4	4	63.87	95	97	2	e-64
		An07g05080	P32842	3.A.2.2.3	4	4	58.06	77	76	1	e-43
		An02g08020	P32842	3.A.2.2.3	4	4	50.31	100	98	2	e-49
		An10g00680	P32842	3.A.2.2.3	4	4	49.35	94	94	0	e-44
		An04g05310	P32563	3.A.2.2.3	7	9	46.96	101	102	1	0
		An07g05080	P25515	3.A.2.2.3	4	4	45.14	89	90	1	e-35
		An04g05310	P37296	3.A.2.2.3	7	8	42.33	106	101	4	0
		An04g05310	Q93050	3.A.2.2.4	7	8	42.31	101	103	2	0
		An02g08020	P59229	3.A.2.2.5	4	4	55.03	93	90	3	e-50
		An10g00680	P59227	3.A.2.2.5	4	4	54.73	90	90	0	e-47
		An10g00680	P59229	3.A.2.2.5	4	4	54.73	90	89	1	e-47
		An10g00680	P59228	3.A.2.2.5	4	4	54.73	90	90	1	e-47
		An02g08020	P59227	3.A.2.2.5	4	4	53.16	98	96	2	e-50
		An07g05080	P59227	3.A.2.2.5	4	4	48.39	77	76	1	e-35
		An07g05080	P59228	3.A.2.2.5	4	4	48.39	77	75	2	e-35
		An07g05080	P59229	3.A.2.2.5	4	4	48.39	77	75	2	e-35

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
		An02g08020	P63082	3.A.2.2.6	4	4	62.84	92	95	4	e-60
		An15g05730	Q91V37	3.A.2.2.6	5	5	61.44	77	75	2	e-58
		An10g00680	P63082	3.A.2.2.6	4	4	60.14	90	95	5	e-55
		An07g05080	P63082	3.A.2.2.6	4	4	52.42	77	80	4	e-35
		An04g05310	Q9Z1G4	3.A.2.2.6	7	9	42.61	101	102	1	0
		An04g05310	Q920R6	3.A.2.2.6	7	9	40.09	101	103	2	0
		An15g05730	G5EDB8	3.A.2.2.7	5	5	59.28	84	78	7	e-54
		An02g08020	P34546	3.A.2.2.7	4	4	56.38	93	93	0	e-50
		An02g08020	Q21898	3.A.2.2.7	4	4	54.36	93	88	5	e-42
		An10g00680	P34546	3.A.2.2.7	4	4	53.02	91	93	2	e-46
		An10g00680	Q21898	3.A.2.2.7	4	4	52.70	90	88	3	e-38
		An07g05080	P34546	3.A.2.2.7	4	4	50.00	77	77	1	e-32
		An04g05310	P30628	3.A.2.2.7	7	7	40.18	105	99	6	0
		An02g08020	Q4UJ88	3.A.2.2.8	4	4	48.32	93	90	2	e-35
		An10g00680	Q4UJ88	3.A.2.2.8	4	4	46.62	90	90	1	e-33
		An07g05080	Q4UJ88	3.A.2.2.8	4	4	42.74	77	75	2	e-25
3.A.3	the p-type atpase (p-atpase) superfamily.	An14g02290	Q2U3D2	3.A.3.1.7	10	10	72.55	92	96	4	0
		An02g14450	P13586	3.A.3.2.3	9	10	48.78	96	104	8	0
		An02g14450	Q9UUX9	3.A.3.2.6	9	10	55.26	99	99	0	0
		An18g06290	P16615	3.A.3.2.7	10	11	53.26	100	97	3	0
		An02g14450	O75185	3.A.3.2.9	9	10	42.31	97	105	8	0
		An18g06290	P92939	3.A.3.2.13	10	10	48.64	102	97	5	0
		An18g06290	Q9SY55	3.A.3.2.19	10	12	52.58	98	99	1	0
		An08g03090	Q9UUY2	3.A.3.2.27	10	10	51.17	86	89	3	0
		An18g06290	Q49LV5	3.A.3.2.32	10	11	52.85	99	98	1	0
		An08g03090	Q9HDW7	3.A.3.2.35	10	12	44.92	79	88	10	0
		An18g06290	Q5IH90	3.A.3.2.36	10	10	49.13	102	94	8	0
		An18g06290	O76974	3.A.3.2.37	10	10	46.31	100	97	3	0
		An01g05670	P07038	3.A.3.3.1	11	10	71.85	89	93	4	0
		An16g05840	P07038	3.A.3.3.1	10	10	47.93	90	97	7	0
		An09g05950	P07038	3.A.3.3.1	10	10	44.99	90	100	10	0
		An01g05670	P05030	3.A.3.3.6	11	10	66.35	87	91	5	0
		An16g05840	P05030	3.A.3.3.6	10	10	47.43	92	99	8	0
		An09g05950	P05030	3.A.3.3.6	10	10	45.14	90	100	10	0
		An12g04500	P39524	3.A.3.8.2	10	8	53.96	99	99	0	0
		An12g08790	P32660	3.A.3.8.4	8	10	52.53	86	83	3	0
		An09g03160	P32660	3.A.3.8.4	10	10	48.12	99	90	10	0
		An12g08790	Q12675	3.A.3.8.5	8	10	51.08	86	81	6	0
		An12g04500	Q5KP96	3.A.3.8.10	10	10	57.09	84	86	2	0
		An15g01830	P13587	3.A.3.9.1	10	10	47.67	91	96	5	0
		An09g00690	P13587	3.A.3.9.1	8	10	46.64	103	98	4	0
		An15g01830	P22189	3.A.3.9.2	10	10	49.11	88	97	10	0
		An09g00690	P22189	3.A.3.9.2	8	10	48.11	101	102	1	0
		An09g00690	O13398	3.A.3.9.3	8	10	50.44	98	95	3	0
		An15g01830	O13398	3.A.3.9.3	10	10	49.57	91	97	6	0
		An15g01830	P78981	3.A.3.9.4	10	10	50.70	87	95	9	0
		An09g00690	P78981	3.A.3.9.4	8	10	49.17	98	98	0	0
		An15g01830	B5B9V9	3.A.3.9.5	10	10	54.97	88	92	5	0
		An09g00690	B5B9V9	3.A.3.9.5	8	10	54.02	100	95	5	0
		An09g00690	Q4PI59	3.A.3.9.6	8	10	43.27	100	93	7	0
		An15g01830	Q4PI59	3.A.3.9.6	10	10	40.84	97	99	2	0
3.A.5	the general secretory pathway (sec) family.	An03g04340	P32915	3.A.5.8.1	10	12	65.06	100	100	0	0
		An03g04340	Q9H9S3	3.A.5.9.1	10	10	66.38	97	97	0	0
		An03g04340	P61619	3.A.5.9.1	10	12	66.31	98	99	0	0
		An01g11630	P60059	3.A.5.9.1	1	1	52.24	96	99	3	e-22
3.A.8	the mitochondrial protein translocase (mpt) family.	An11g02140	P39515	3.A.8.1.1	3	4	74.48	94	92	3	e-74
		An07g07880	Q02776	3.A.8.1.1	2	1	42.70	70	76	8	e-72

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
		An02g01360	P32897	3.A.8.1.1	3	3	42.26	83	76	9	e-41
3.A.16	the endoplasmic reticular retrotranslocon (er-rt) family.	An14g00230	E7NGV2	3.A.16.1.2	2	1	50.24	77	84	9	e-66
3.A.19	the tms recognition/insertion complex (trc) family.	An04g00670	A2QHQ3	3.A.19.1.2	3	3	100.00	93	93	0	e-135
3.D.1	the h(+) or na(+)-translocating nadh dehydrogenase (ndh) family.	An11g08840	P42026	3.D.1.6.1	1	2	74.05	72	73	1	e-93
		An16g02130	Q7S1I2	3.D.1.6.2	1	1	64.17	101	99	2	e-49
		An14g00060	Q02854	3.D.1.6.2	2	3	59.66	94	93	1	e-71
		An06g01390	P25710	3.D.1.6.2	4	3	47.42	99	97	3	e-61
		An04g05640	Q9FNN5	3.D.1.6.3	1	1	75.66	84	85	2	0
		An11g08840	Q42577	3.D.1.6.3	1	1	73.38	70	71	0	e-86
		An04g05640	Q6V9B2	3.D.1.6.4	1	1	72.58	85	87	2	0
3.D.2	the proton-translocating transhydrogenase (pth) family.	An02g09810	P11024	3.D.2.3.1	14	16	49.17	100	100	1	0
3.D.3	the proton-translocating quinol:cytochrome c reductase (qcr) superfamily.	An14g04080	P08067	3.D.3.2.1	1	1	57.38	77	85	10	e-79
		An01g06180	P07143	3.D.3.3.1	2	2	64.14	79	81	2	e-120
8.A.27	the cdc50 p-type atpase lipid flippase subunit (cdc50) family.	An07g10420	P25656	8.A.27.1.2	2	3	46.93	89	92	3	e-118
9.A.2	the endomembrane protein-70 (emp70) family.	An06g01200	E7NFP9	9.A.2.1.1	10	9	42.86	105	101	4	e-176
		An06g01200	Q9LIC2	9.A.2.1.2	10	10	41.19	99	100	1	e-165
		An06g01200	Q99805	9.A.2.1.6	10	9	40.18	102	99	3	e-167
9.A.6	the atp exporter (atp-e) family.	An14g00900	P36051	9.A.6.1.1	14	14	41.04	97	105	8	0
9.A.41	the capsular polysaccharide exporter (cps-e) family.	An11g04180	P44669	9.A.41.1.1	1	1	41.57	79	86	8	e-122
9.A.54	the lysosomal cobalamin (b12) transporter (l-b12t) family.	An16g09150	A6QTW5	9.A.54.1.3	10	10	51.31	99	97	2	0
9.B.1	the integral membrane caax protease (caax protease) family.	An04g01950	Q8RX88	9.B.1.1.2	7	7	43.26	93	100	7	e-117
		An04g01950	P47154	9.B.1.1.3	7	5	45.09	98	99	1	e-137
		An14g03420	F9FER0	9.B.1.2.2	6	5	50.17	91	99	8	e-93
9.B.7	the putative sulfate transporter (cysz) family.	An07g06140	E2PST1	9.B.7.2.3	5	5	100.00	100	100	0	0
9.B.16	the putative ductin channel (ductin) family.	An02g08020	P23380	9.B.16.1.1	4	4	60.38	99	100	1	e-59
		An10g00680	P23380	9.B.16.1.1	4	4	57.42	95	97	3	e-52
		An07g05080	P23380	9.B.16.1.1	4	4	57.26	77	78	2	e-39
		An02g08020	Q03105	9.B.16.1.2	4	4	60.39	96	100	4	e-61
		An10g00680	Q03105	9.B.16.1.2	4	4	56.49	94	100	6	e-55
		An07g05080	Q03105	9.B.16.1.2	4	4	52.42	77	81	5	e-36
9.B.25	the mitochondrial inner/outer membrane fusion (mmf) family.	An08g04250	P32266	9.B.25.1.1	1	1	42.66	95	99	4	0
9.B.26	the regulator of er stress and autophagy tmem208 (tmem208) family.	An12g03980	K9FAK7	9.B.26.1.4	2	2	65.41	76	78	3	e-57
9.B.82	endoplasmic reticulum retrieval protein1 (putative heavy metal transporter) (rer1) family.	An02g02830	P25560	9.B.82.1.1	4	4	45.45	93	94	1	e-53
		An02g02830	O15258	9.B.82.1.2	4	4	54.76	89	86	4	e-63
		An02g02830	O48670	9.B.82.1.3	4	4	52.27	93	92	1	e-59
9.B.119	the glycan synthase, fks1 (fks1) family.	An06g01550	P38631	9.B.119.1.1	18	16	62.17	95	96	1	0
9.B.142	the integral membrane	An16g08570	B3S136	9.B.142.3.3	13	13	55.44	92	98	7	0

Continued on next page

Table 51 – continued from previous page

Family	Family Name	Query	Hit	TCID	QTMS	HTMS	%ID	QCov	SCov	Diff	eVal
	glycosyltransferase family 39 (gt39) family.	An16g08570	G9P430	9.B.142.3.5	13	13	76.84	96	95	0	0
9.B.143	the 6 tms duf1275/pf06912 (duf1275) family.	An10g00830	G7XY82	9.B.143.5.1	6	6	91.16	100	100	0	e-167

C.2 TCDB-Blast Results for Fungal Genomes

Table 52 presents the number of proteins in each fungi that matches a given TCID. The table is organised by TC-Family. The columns Family and Family Name contain the TC-Family identifier and its name. The column TCID contains the TCID of the TCDB entry predicted to be in a fungi. Only those identifiers predicted in at least one fungi occur in this column. The last 8 columns contain the number of transporters in each fungi as predicted by TCDB-Blast. The column headings indicate the fungi using the following code: **Aaf**:*A.fumigatus Af293*, **Ani**:*A. nidulans*, **Anc**:*A.niger CBS513.88*, **Ann**:*A. niger NRRL3*, **Aor**: *A. oryzae*, **Ncr**:*N. crassa*, **Pch**:*P. chrysosporium RP78*, **Spo**:*S. pombe*.

Table 52: TCDB-Blast Results for Fungal Genomes

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
1.A.1	the voltage-gated ion channel (vic) superfamily.	1.A.1.11.23	-	-	-	-	-	-	-	1
1.A.4	the transient receptor potential ca(2+) channel (trp-cc) family.	1.A.4.10.1	-	-	-	-	-	-	-	1
		1.A.4.9.2	-	-	-	-	-	-	-	1
1.A.8	the major intrinsic protein (mip) family.	1.A.8.6.1	-	-	-	1	-	-	-	-
		1.A.8.6.2	-	-	-	1	-	-	-	-
		1.A.8.6.3	-	-	-	1	-	-	-	-
		1.A.8.6.4	1	1	-	1	1	-	-	-
		1.A.8.7.1	-	-	-	-	-	-	-	1
		1.A.8.9.4	-	-	-	1	-	-	-	-
1.A.9	the neurotransmitter receptor, cys loop, ligand-gated ion channel (lic) family.	1.A.9.5.2	1	1	1	1	1	1	1	1
1.A.11	the ammonia transporter channel (amt) family.	1.A.11.1.4	1	1	1	1	1	1	-	-
		1.A.11.3.1	1	2	1	1	1	2	2	1
		1.A.11.3.2	2	3	2	2	2	3	2	2
		1.A.11.3.3	2	3	2	3	2	3	2	2
		1.A.11.3.4	2	2	2	2	2	3	2	2
		1.A.11.3.5	2	2	2	2	2	3	2	2
1.A.14	the testis-enhanced gene transfer (tegt) family.	1.A.14.3.2	-	-	-	-	-	1	-	-
1.A.16	the formate-nitrite transporter (fnt) family.	1.A.16.2.2	-	1	-	1	1	1	-	-
1.A.17	the calcium-dependent chloride channel (ca-clc) family.	1.A.17.5.5	-	1	-	-	-	-	-	-
		1.A.17.6.2	-	-	-	-	-	-	-	1
		1.A.17.6.4	2	2	2	2	2	1	-	-
1.A.23	the small conductance mechanosensitive ion channel (mscs) family.	1.A.23.4.9	1	1	1	1	1	-	-	-
1.A.33	the cation channel-forming heat shock protein-70 (hsp70) family.	1.A.33.1.2	2	2	2	2	3	3	3	4
		1.A.33.1.3	2	2	2	2	3	3	3	4
1.A.35	the cora metal ion transporter (mit) family.	1.A.35.2.3	-	-	-	-	-	-	-	1
		1.A.35.5.5	-	-	-	-	-	-	-	1
1.A.43	the camphor resistance (crcb) family.	1.A.43.2.3	-	-	-	-	-	-	-	2
		1.A.43.2.6	-	-	-	-	-	1	-	-
1.A.46	the anion channel-forming bestrophin (bestrophin) family.	1.A.46.2.1	1	1	-	1	1	1	-	-
		1.A.46.2.2	-	1	1	1	1	-	-	-
1.A.55	the synaptic vesicle-associated ca(2+) channel, flower (flower) family.	1.A.55.4.1	1	1	-	-	1	1	-	1
1.A.56	the copper transporter (ctr) family.	1.A.56.1.10	-	-	1	-	1	2	-	-
		1.A.56.1.5	-	-	-	-	-	-	-	2
		1.A.56.1.6	-	-	-	-	-	-	-	1
1.A.77	the mg(2+)/ca(2+) uniporter (mcu) family.	1.A.77.1.5	1	-	1	1	1	1	-	-

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
1.A.81	the low affinity ca(2+) channel (lacc) family.	1.A.81.4.1	-	-	-	-	-	1	-	-
		1.A.81.5.1	1	1	-	2	1	1	-	-
1.A.88	the fungal potassium channel (f-kch) family.	1.A.88.1.4	-	-	1	-	1	-	-	-
		1.A.88.1.6	-	-	-	-	-	1	-	-
1.B.69	the peroxysomal membrane porin 4 (pxmp4) family.	1.B.69.1.1	-	-	-	1	-	-	-	-
		1.B.69.1.4	1	1	1	1	1	1	-	-
		1.B.69.1.6	1	1	1	1	1	1	-	-
		1.B.69.1.7	-	1	-	1	-	-	-	-
1.C.47	the insect/fungal defensin (insect/fungal defensin) family.	1.C.47.1.8	-	2	-	-	-	-	-	-
1.F.1	the synaptosomal vesicle fusion pore (svf-pore) family.	1.F.1.1.2	2	1	1	1	1	-	1	1
1.H.1	the claudin tight junction (claudin) family.	1.H.1.4.1	1	-	1	1	1	1	-	-
		1.H.1.4.3	1	1	1	1	1	-	-	-
		1.H.1.4.5	-	-	-	1	-	-	-	-
2.A.1	the major facilitator superfamily (mfs).	2.A.1.1.104	-	-	-	-	-	-	-	2
		2.A.1.1.105	2	3	1	3	-	-	-	4
		2.A.1.1.107	1	3	-	2	-	1	-	1
		2.A.1.1.108	2	3	2	3	2	1	-	5
		2.A.1.1.10	-	-	-	1	-	1	-	-
		2.A.1.1.110	1	3	1	1	1	1	-	-
		2.A.1.1.111	2	3	2	3	2	1	-	6
		2.A.1.1.112	1	3	1	1	2	1	-	8
		2.A.1.1.117	2	4	1	3	1	4	1	-
		2.A.1.1.11	-	1	-	1	-	2	-	-
		2.A.1.1.21	1	3	1	2	1	1	1	8
		2.A.1.1.22	1	3	2	1	2	-	-	8
		2.A.1.1.23	-	-	-	-	-	-	-	8
		2.A.1.1.30	2	3	-	2	-	1	-	3
		2.A.1.1.31	1	3	1	2	2	1	-	2
		2.A.1.1.33	1	1	2	1	2	-	-	-
		2.A.1.1.36	3	3	3	3	3	2	1	7
		2.A.1.1.38	3	2	3	4	3	1	-	-
		2.A.1.1.39	1	3	3	3	3	1	1	-
		2.A.1.1.40	1	2	1	1	2	1	1	-
		2.A.1.14.17	1	1	-	1	-	-	1	1
		2.A.1.14.18	1	1	-	1	-	-	1	-
		2.A.1.14.19	1	1	-	1	-	-	-	2
		2.A.1.14.20	-	-	-	1	-	-	-	2
		2.A.1.14.38	3	5	4	5	4	3	3	-
		2.A.1.14.4	-	-	-	-	-	-	-	1
		2.A.1.1.51	3	5	2	3	2	3	1	-
		2.A.1.1.57	3	4	2	3	2	3	1	-
		2.A.1.1.58	3	3	3	3	3	2	1	7
		2.A.1.1.5	2	3	1	3	1	-	-	5
		2.A.1.16.1	1	1	1	1	1	-	-	-
		2.A.1.1.64	-	-	-	-	-	3	-	-
		2.A.1.16.5	-	-	-	-	-	-	-	1
		2.A.1.16.6	-	-	-	-	-	-	-	1
		2.A.1.1.67	2	3	2	3	2	1	-	4
		2.A.1.16.7	2	4	2	2	2	-	-	-
		2.A.1.1.68	3	5	2	3	2	3	1	-
		2.A.1.1.6	1	3	1	2	-	1	-	3
		2.A.1.1.70	1	1	2	1	2	-	-	-
		2.A.1.1.73	3	2	3	4	3	1	1	-
		2.A.1.1.7	1	3	1	2	2	1	-	-
		2.A.1.1.82	1	1	-	-	-	1	-	-
		2.A.1.1.83	3	1	-	2	-	1	3	-

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
		2.A.1.1.8	-	-	1	-	1	-	-	2
		2.A.1.19.38	-	1	2	1	2	-	2	1
		2.A.1.19.48	-	2	-	1	-	1	1	-
		2.A.1.2.16	2	4	3	3	3	1	-	4
		2.A.1.2.17	-	-	2	-	2	-	-	1
		2.A.1.2.1	-	-	-	-	-	-	-	3
		2.A.1.2.23	-	1	1	-	-	-	-	-
		2.A.1.2.33	-	-	-	-	-	-	-	1
		2.A.1.2.35	-	-	2	1	3	-	3	2
		2.A.1.24.2	-	-	-	-	-	-	-	1
		2.A.1.2.45	-	1	1	-	-	-	-	-
		2.A.1.2.46	-	1	1	1	1	-	-	-
		2.A.1.2.48	-	-	1	-	-	-	-	-
		2.A.1.2.59	-	-	-	-	-	-	-	3
		2.A.1.2.66	-	-	-	-	-	-	1	-
		2.A.1.2.67	1	-	1	-	-	-	2	-
		2.A.1.2.6	-	-	1	-	1	-	-	1
		2.A.1.2.76	-	-	-	-	-	-	-	3
		2.A.1.2.77	4	2	5	5	6	2	5	1
		2.A.1.2.78	1	2	2	2	1	1	4	-
		2.A.1.2.79	1	2	-	1	1	-	-	-
		2.A.1.2.85	2	-	5	2	3	1	4	1
		2.A.1.2.86	2	6	4	7	5	1	1	-
		2.A.1.3.52	2	1	3	2	3	1	-	1
		2.A.1.3.65	3	2	5	8	7	3	-	-
		2.A.1.48.2	-	-	-	-	1	-	-	-
		2.A.1.48.3	-	-	-	-	-	-	-	1
		2.A.1.48.4	-	-	-	-	-	-	-	1
		2.A.1.58.1	3	2	2	2	2	1	1	1
		2.A.1.58.4	3	1	1	1	1	-	-	1
		2.A.1.58.5	1	1	1	-	2	1	-	-
		2.A.1.75.2	1	1	1	1	1	1	-	-
		2.A.1.8.13	2	2	1	1	1	-	-	-
		2.A.1.8.5	2	2	1	1	1	-	-	-
		2.A.1.9.10	-	-	-	-	-	-	1	2
		2.A.1.9.1	1	-	-	-	1	1	3	-
		2.A.1.9.2	1	-	-	-	1	1	3	-
		2.A.1.9.7	-	-	1	-	1	-	1	-
		2.A.1.9.8	-	-	-	-	-	-	1	4
2.A.2	the glycoside-pentoside-hexuronide (gph):cation symporter family.	2.A.2.6.1	-	-	-	-	-	-	-	1
2.A.3	the amino acid-polyamine-organocation (apc) family.	2.A.3.10.10	2	-	1	2	2	-	-	-
		2.A.3.10.11	2	-	2	2	2	-	-	-
		2.A.3.10.13	1	1	2	2	2	-	1	-
		2.A.3.10.14	-	-	-	-	-	1	-	1
		2.A.3.10.15	-	1	-	1	-	-	-	-
		2.A.3.10.16	-	-	-	-	-	-	-	1
		2.A.3.10.17	-	1	2	3	-	2	2	-
		2.A.3.10.18	1	-	2	2	2	-	-	-
		2.A.3.10.19	1	4	2	2	2	1	-	-
		2.A.3.10.1	1	1	1	1	1	2	-	2
		2.A.3.10.20	2	-	2	2	2	-	-	-
		2.A.3.10.21	1	1	1	2	1	1	-	8
		2.A.3.10.22	1	1	1	2	1	-	-	8
		2.A.3.10.23	1	1	1	1	1	1	-	8
		2.A.3.10.24	1	1	1	2	1	2	-	6
		2.A.3.10.25	1	1	1	1	1	2	-	2
		2.A.3.10.26	1	1	1	1	1	-	-	2

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
		2.A.3.10.28	1	1	1	2	1	-	-	8
		2.A.3.10.2	2	1	1	1	1	3	-	7
		2.A.3.10.4	2	-	2	2	2	-	-	-
		2.A.3.10.6	-	-	-	-	-	-	-	1
		2.A.3.10.7	-	-	-	1	-	-	-	-
		2.A.3.10.8	1	1	1	1	1	1	-	5
		2.A.3.10.9	-	-	-	-	-	1	-	-
		2.A.3.4.1	-	-	2	2	2	-	-	-
		2.A.3.4.2	2	1	2	-	2	-	-	-
		2.A.3.4.3	2	1	2	2	2	-	-	-
		2.A.3.4.6	1	-	1	1	1	-	-	2
		2.A.3.8.4	1	1	1	1	1	-	-	-
2.A.4	the cation diffusion facilitator (cdf) family.	2.A.4.2.1	-	-	-	-	-	-	1	-
		2.A.4.2.2	-	-	1	-	-	-	1	-
2.A.5	the zinc (zn(2+))-iron (fe(2+)) permease (zip) family.	2.A.5.1.1	2	1	3	4	4	-	1	-
		2.A.5.1.8	-	-	-	-	-	-	-	1
		2.A.5.5.4	-	-	-	-	-	-	-	1
2.A.6	the resistance-nodulation-cell division (rnd) superfamily.	2.A.6.6.3	-	-	1	-	-	1	-	-
2.A.7	the drug/metabolite transporter (dmt) superfamily.	2.A.7.12.4	-	-	-	-	-	-	-	1
		2.A.7.12.5	-	-	-	-	-	-	-	1
		2.A.7.12.7	-	-	-	-	-	-	1	1
		2.A.7.12.8	-	-	-	-	-	-	-	1
		2.A.7.12.9	-	-	-	-	-	-	1	1
		2.A.7.13.1	-	1	-	-	-	-	1	1
		2.A.7.13.2	1	1	1	1	1	1	-	1
		2.A.7.16.3	-	-	-	-	-	-	-	1
		2.A.7.24.11	1	1	2	1	2	1	-	-
		2.A.7.24.7	-	-	-	-	-	1	-	-
		2.A.7.25.2	-	1	-	1	1	-	-	-
		2.A.7.25.5	-	-	-	1	-	-	-	-
		2.A.7.25.6	1	1	-	1	1	2	-	-
		2.A.7.25.7	1	1	-	1	1	-	-	-
		2.A.7.25.9	-	1	-	-	-	-	-	-
		2.A.7.32.4	-	-	-	-	-	1	-	-
		2.A.7.9.17	-	-	-	-	1	-	-	-
		2.A.7.9.18	-	-	-	-	-	-	-	1
2.A.16	the telurite-resistance/dicarboxylate transporter (tdt) family.	2.A.16.2.1	-	-	-	-	-	-	-	2
		2.A.16.4.1	3	-	1	2	1	-	-	-
		2.A.16.4.2	3	-	1	2	1	-	-	-
		2.A.16.4.3	2	-	1	2	1	-	-	-
2.A.17	the proton-dependent oligopeptide transporter (pot) family.	2.A.17.2.1	2	1	2	2	2	-	1	1
		2.A.17.2.2	1	1	1	1	1	-	-	1
2.A.18	the amino acid/auxin permease (aaap) family.	2.A.18.4.1	1	1	3	3	3	1	-	-
		2.A.18.4.2	1	1	3	3	3	1	-	-
		2.A.18.6.10	1	-	-	-	1	-	-	-
		2.A.18.7.1	-	1	1	-	1	-	-	1
		2.A.18.7.3	-	-	-	-	-	-	-	1
2.A.19	the ca(2+):cation antiporter (caca) family.	2.A.19.2.2	-	1	2	2	2	-	-	1
		2.A.19.2.8	1	1	1	1	1	-	-	1
2.A.20	the inorganic phosphate transporter (pit) family.	2.A.20.2.1	1	2	-	2	-	2	-	-
		2.A.20.2.2	1	2	-	2	-	2	-	-
2.A.21	the solute:sodium symporter (sss) family.	2.A.21.6.1	1	1	2	1	2	1	-	1
		2.A.21.6.2	1	1	1	1	1	1	1	1
		2.A.21.6.3	1	1	-	1	-	1	1	1
		2.A.21.6.4	1	1	1	1	1	1	-	1
		2.A.21.6.5	-	-	-	-	-	1	-	-
2.A.29	the mitochondrial carrier (mc) family.	2.A.29.10.1	1	-	-	-	-	-	-	-

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
		2.A.29.10.2	1	1	-	-	-	-	1	-
		2.A.29.10.4	1	1	1	1	1	1	-	1
		2.A.29.10.5	-	-	1	-	1	-	-	-
		2.A.29.10.6	-	-	-	-	-	-	1	-
		2.A.29.10.7	-	-	1	1	1	-	-	-
		2.A.29.1.10	1	1	1	1	1	1	2	1
		2.A.29.1.1	1	1	1	1	1	1	2	1
		2.A.29.12.4	-	-	-	-	-	1	1	1
		2.A.29.1.2	1	1	1	1	1	1	2	1
		2.A.29.13.1	1	-	1	1	1	1	1	-
		2.A.29.1.3	1	1	1	1	1	1	2	1
		2.A.29.14.1	1	1	1	1	1	-	1	-
		2.A.29.1.4	1	1	1	1	1	1	2	1
		2.A.29.15.1	-	1	-	1	1	1	1	1
		2.A.29.16.3	-	-	-	-	-	-	-	1
		2.A.29.1.6	1	1	1	1	1	1	2	1
		2.A.29.1.7	1	1	1	1	1	1	2	1
		2.A.29.18.1	-	-	-	-	-	-	1	-
		2.A.29.18.3	-	-	-	-	-	-	1	-
		2.A.29.1.8	1	1	1	1	1	1	2	1
		2.A.29.1.9	1	1	1	1	1	1	2	1
		2.A.29.2.10	1	1	1	1	1	2	-	-
		2.A.29.2.11	1	1	1	-	1	1	-	-
		2.A.29.21.1	1	1	1	1	1	1	1	1
		2.A.29.2.13	1	1	1	-	1	1	-	-
		2.A.29.2.1	1	-	1	-	1	-	-	-
		2.A.29.2.2	1	1	1	-	1	-	1	-
		2.A.29.23.4	-	1	-	-	-	1	1	-
		2.A.29.2.3	1	1	1	-	1	-	1	-
		2.A.29.2.5	1	1	1	1	1	1	-	1
		2.A.29.2.6	1	1	1	1	1	-	-	-
		2.A.29.2.7	1	1	1	-	1	-	-	-
		2.A.29.2.8	1	1	1	1	1	1	-	1
		2.A.29.29.1	1	1	2	2	2	1	-	1
		2.A.29.4.1	-	1	1	1	1	2	-	-
		2.A.29.4.2	-	1	1	1	1	2	-	-
		2.A.29.4.3	2	1	2	3	2	1	1	-
		2.A.29.4.4	1	1	1	1	1	1	1	1
		2.A.29.4.5	1	1	1	1	1	2	-	-
		2.A.29.4.6	1	1	1	1	1	2	-	-
		2.A.29.5.1	1	1	1	1	1	1	1	1
		2.A.29.5.2	1	1	1	1	1	1	1	1
		2.A.29.5.3	1	-	1	1	1	1	1	-
		2.A.29.5.5	-	-	1	1	1	1	1	-
		2.A.29.5.7	1	1	1	1	1	1	1	-
		2.A.29.7.3	1	1	2	2	2	1	1	1
		2.A.29.7.4	-	-	1	-	1	-	-	-
		2.A.29.8.11	1	1	1	1	1	1	1	1
		2.A.29.8.12	1	1	1	1	1	1	1	1
		2.A.29.8.1	-	-	-	-	-	-	1	-
		2.A.29.8.2	-	-	1	-	1	-	1	-
		2.A.29.8.3	-	-	-	-	-	-	1	-
		2.A.29.8.4	-	-	1	-	1	1	1	-
		2.A.29.9.1	-	-	1	1	1	1	1	-
2.A.31	the anion exchanger (ae) family.	2.A.31.3.2	-	-	-	-	-	1	1	-
2.A.36	the monovalent cation:proton antiporter-1 (cpa1) family.	2.A.36.4.3	-	-	-	-	-	-	-	1
		2.A.36.4.5	-	-	-	-	-	-	-	1
2.A.38	the k(+) transporter (trk) family.	2.A.38.2.2	-	-	-	-	-	1	-	-

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
2.A.39	the nucleobase:cation symporter-1 (ncs1) family.	2.A.39.2.1	1	1	-	1	1	-	-	-
		2.A.39.2.3	1	1	-	1	1	-	-	-
		2.A.39.2.4	1	1	-	1	1	-	-	-
		2.A.39.3.1	-	-	-	-	-	-	-	1
		2.A.39.3.7	1	1	1	2	1	1	1	2
2.A.40	the nucleobase:cation symporter-2 (ncs2) family.	2.A.40.4.1	1	2	2	1	2	1	1	1
		2.A.40.4.4	1	2	2	1	2	1	1	1
		2.A.40.7.1	1	1	1	1	1	3	1	1
		2.A.40.7.3	-	1	-	1	1	2	1	1
2.A.41	the concentrative nucleoside transporter (cnt) family.	2.A.41.2.7	1	1	1	1	1	1	-	-
2.A.43	the lysosomal cystine transporter (lct) family.	2.A.43.2.7	-	-	1	-	-	-	-	-
		2.A.43.4.1	-	-	-	-	-	-	-	1
2.A.47	the divalent anion:na(+) symporter (dass) family.	2.A.47.2.1	-	-	1	1	1	-	-	-
		2.A.47.2.2	-	-	1	1	1	-	-	-
		2.A.47.2.3	-	-	1	1	1	-	-	-
2.A.49	the chloride carrier/channel (clc) family.	2.A.49.1.2	1	1	-	-	1	1	1	1
		2.A.49.1.3	1	1	-	-	1	1	-	-
2.A.50	the glycerol uptake (gup) family.	2.A.50.1.1	-	-	-	-	-	2	1	-
2.A.52	the ni(2+)-co(2+) transporter (nicot) family.	2.A.52.1.3	-	1	-	-	-	1	-	1
		2.A.52.1.8	-	1	1	-	-	1	-	1
2.A.53	the sulfate permease (sulp) family.	2.A.53.1.1	1	-	-	-	-	1	1	2
		2.A.53.1.2	1	1	1	1	2	1	1	2
		2.A.53.3.10	-	1	-	-	-	-	-	-
		2.A.53.3.7	1	1	-	-	-	-	-	1
2.A.54	the mitochondrial tricarboxylate carrier (mtc) family.	2.A.54.1.4	1	1	-	2	1	1	-	1
2.A.55	the metal ion (mn(2+)-iron) transporter (nramp) family.	2.A.55.1.1	1	1	1	1	1	2	-	1
		2.A.55.1.2	1	1	1	1	1	1	-	1
		2.A.55.1.3	-	-	-	-	-	-	-	1
		2.A.55.1.4	-	1	1	1	1	1	-	1
2.A.59	the arsenical resistance-3 (acr3) family.	2.A.59.1.1	-	-	1	-	1	-	1	-
		2.A.59.1.2	3	-	1	-	1	-	-	-
2.A.66	the multidrug/oligosaccharidyl-lipid/polysaccharide (mop) flippase superfamily.	2.A.66.1.5	-	1	1	-	1	-	-	-
2.A.67	the oligopeptide transporter (opt) family.	2.A.67.1.1	3	2	2	4	4	1	2	2
		2.A.67.1.2	2	1	2	2	4	2	2	1
		2.A.67.1.3	1	-	1	2	1	-	1	1
		2.A.67.1.5	3	1	4	2	5	2	-	2
2.A.69	the auxin efflux carrier (aec) family.	2.A.69.2.3	1	1	1	1	1	1	-	-
2.A.72	the k(+) uptake permease (kup) family.	2.A.72.3.2	-	-	1	1	1	1	-	-
2.A.85	the aromatic acid exporter (arae) family.	2.A.85.3.1	-	-	-	-	-	-	-	1
		2.A.85.3.5	-	-	-	-	-	-	1	-
2.A.89	the vacuolar iron transporter (vit) family.	2.A.89.1.1	1	-	1	-	1	-	-	-
		2.A.89.3.8	-	-	-	-	-	-	-	1
2.A.96	the acetate uptake transporter (acetr) family.	2.A.96.1.3	1	2	2	1	2	1	-	1
		2.A.96.1.4	1	-	1	-	1	1	-	-
		2.A.96.1.6	1	1	1	1	2	-	-	1
		2.A.96.1.7	1	-	1	-	1	1	-	-
		2.A.96.2.1	1	1	-	-	-	-	-	-
2.A.97	the mitochondrial inner membrane k(+)/h(+) and ca(2+)/h(+) exchanger (letm1) family.	2.A.97.1.2	-	1	-	-	-	-	-	-
2.A.105	the mitochondrial pyruvate carrier (mpc) family.	2.A.105.1.1	1	-	1	-	1	1	1	2
		2.A.105.1.2	-	-	-	-	-	-	1	1
		2.A.105.1.4	-	-	-	-	-	-	1	-
2.A.106	the ca(2+):h(+) antiporter-2 (caca2) family.	2.A.106.2.4	-	-	-	-	-	-	-	2
2.A.108	the iron/lead transporter (ilt) family.	2.A.108.1.1	2	-	5	2	5	1	-	-
		2.A.108.1.2	1	-	3	1	3	-	-	1
		2.A.108.1.3	1	-	3	1	3	-	-	-
		2.A.108.1.4	1	-	2	1	2	-	-	-

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
		2.A.108.1.5	1	-	3	1	3	-	-	1
		2.A.108.1.7	1	-	-	1	-	-	1	-
3.A.1	the atp-binding cassette (abc) superfamily.	3.A.1.201.10	3	2	2	3	2	1	2	-
		3.A.1.201.11	1	1	-	2	1	1	2	-
		3.A.1.201.16	1	-	1	1	1	-	-	-
		3.A.1.201.17	1	1	1	1	1	1	-	1
		3.A.1.201.18	1	2	1	3	1	1	1	1
		3.A.1.201.1	1	-	1	-	1	1	1	-
		3.A.1.201.3	-	-	1	-	1	-	-	-
		3.A.1.203.10	1	-	1	1	1	-	-	-
		3.A.1.203.1	1	-	1	1	1	-	-	-
		3.A.1.203.3	1	-	1	1	1	1	1	-
		3.A.1.203.7	1	2	2	1	2	1	1	-
		3.A.1.204.9	-	-	-	-	-	-	1	-
		3.A.1.205.11	6	7	8	8	10	1	2	1
		3.A.1.205.12	6	7	7	7	10	1	-	-
		3.A.1.205.1	5	7	7	7	9	1	-	1
		3.A.1.205.2	3	2	3	2	4	1	2	1
		3.A.1.205.3	2	1	1	1	1	-	1	-
		3.A.1.205.4	6	6	7	7	10	1	1	1
		3.A.1.205.5	6	7	7	7	10	1	1	1
		3.A.1.205.6	1	2	2	1	2	2	2	-
		3.A.1.205.7	3	4	4	7	4	1	-	-
		3.A.1.206.2	-	-	-	-	-	-	-	1
		3.A.1.208.11	1	1	1	1	1	1	1	2
		3.A.1.208.12	-	-	-	1	-	-	-	-
		3.A.1.208.16	1	1	1	1	1	1	1	2
		3.A.1.208.18	-	-	-	-	-	1	-	-
		3.A.1.208.27	-	-	-	-	-	1	-	-
		3.A.1.208.28	1	1	1	1	1	1	1	2
		3.A.1.208.2	-	-	1	-	1	-	-	-
		3.A.1.208.32	1	1	1	1	1	1	1	2
		3.A.1.208.8	-	-	-	-	-	1	-	-
		3.A.1.210.11	-	-	-	-	-	-	1	-
		3.A.1.210.1	1	1	1	1	1	1	1	1
		3.A.1.210.2	-	-	1	2	1	-	-	1
		3.A.1.210.3	-	-	-	-	-	-	1	-
		3.A.1.210.4	1	1	1	1	1	1	1	1
		3.A.1.210.6	-	-	-	2	-	-	-	1
		3.A.1.210.7	1	1	1	1	1	-	-	1
		3.A.1.210.8	1	1	1	1	1	1	1	1
		3.A.1.210.9	-	-	-	-	-	-	2	1
		3.A.1.212.1	-	-	-	-	-	-	-	1
		3.A.1.212.2	1	1	1	1	1	1	-	1
3.A.2	the h(+)- or na(+)-translocating f-type, v-type and a-type atpase (f-atpase) superfamily.	3.A.2.1.2	-	1	-	1	1	-	-	-
		3.A.2.1.3	1	2	1	2	1	1	-	2
		3.A.2.1.4	-	1	-	1	1	1	-	-
		3.A.2.2.3	7	9	8	9	8	7	5	7
		3.A.2.2.4	1	1	1	1	1	-	2	-
		3.A.2.2.5	6	6	8	8	8	6	6	6
		3.A.2.2.6	5	5	6	5	6	3	5	3
		3.A.2.2.7	5	5	7	7	7	5	5	4
		3.A.2.2.8	2	-	3	3	3	2	2	2
3.A.3	the p-type atpase (p-atpase) superfamily.	3.A.3.10.13	-	-	-	-	-	-	-	1
		3.A.3.10.19	-	-	-	-	-	-	1	1
		3.A.3.10.1	-	-	-	-	-	-	1	1
		3.A.3.10.2	1	1	-	1	1	-	1	1
		3.A.3.10.3	1	1	-	1	1	1	1	1

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
		3.A.3.10.7	1	-	-	-	-	-	1	-
		3.A.3.10.8	-	-	-	-	-	1	-	-
		3.A.3.1.11	-	-	-	-	-	-	1	-
		3.A.3.1.1	-	-	-	-	-	-	3	-
		3.A.3.1.4	-	-	-	-	-	-	1	-
		3.A.3.1.6	-	-	-	-	-	-	1	-
		3.A.3.1.7	2	-	1	2	2	-	-	-
		3.A.3.2.13	1	1	1	1	1	1	1	-
		3.A.3.2.16	-	-	-	-	-	-	1	1
		3.A.3.2.17	1	1	-	1	1	1	1	-
		3.A.3.2.19	1	1	1	1	1	1	1	-
		3.A.3.2.27	1	2	1	-	1	1	-	1
		3.A.3.2.2	1	-	-	-	-	1	-	-
		3.A.3.2.32	1	1	1	1	1	1	1	-
		3.A.3.2.34	-	-	-	-	-	-	1	1
		3.A.3.2.35	2	3	1	1	2	-	-	1
		3.A.3.2.36	1	1	1	1	1	1	1	-
		3.A.3.2.37	1	1	1	1	1	1	1	-
		3.A.3.2.3	-	-	1	-	-	1	1	1
		3.A.3.2.5	-	-	-	-	-	-	1	1
		3.A.3.2.6	1	1	1	1	1	1	-	-
		3.A.3.2.7	1	1	1	1	1	1	1	-
		3.A.3.2.9	-	-	1	-	-	1	1	1
		3.A.3.3.1	3	2	3	3	4	1	-	2
		3.A.3.3.6	3	2	3	3	4	1	-	2
		3.A.3.3.7	-	-	-	-	-	-	1	-
		3.A.3.3.8	-	-	-	-	-	-	1	-
		3.A.3.3.9	-	-	-	-	-	-	1	-
		3.A.3.5.14	1	-	-	1	-	-	-	-
		3.A.3.5.29	-	-	-	-	-	-	-	1
		3.A.3.8.10	1	1	1	1	1	1	1	1
		3.A.3.8.13	-	-	-	-	-	-	1	1
		3.A.3.8.1	-	-	-	-	-	-	1	1
		3.A.3.8.2	1	1	1	1	1	1	-	1
		3.A.3.8.4	1	1	2	1	2	1	-	-
		3.A.3.8.5	1	1	1	1	2	1	-	-
		3.A.3.8.6	-	1	-	-	-	-	1	1
		3.A.3.8.8	-	-	-	-	-	-	1	1
		3.A.3.9.1	3	3	2	3	2	3	-	1
		3.A.3.9.2	3	3	2	3	2	3	-	1
		3.A.3.9.3	3	3	2	3	2	3	-	1
		3.A.3.9.4	3	3	2	3	2	3	-	1
		3.A.3.9.5	3	3	2	3	2	3	-	1
		3.A.3.9.6	3	2	2	3	2	3	-	1
3.A.5	the general secretory pathway (sec) family.	3.A.5.8.1	1	1	1	1	1	1	1	1
		3.A.5.9.1	2	3	3	2	3	3	2	2
3.A.8	the mitochondrial protein translocase (mpt) family.	3.A.8.1.1	3	2	3	1	3	2	2	3
3.A.16	the endoplasmic reticular retrotranslocon (er-rt) family.	3.A.16.1.2	-	1	1	-	1	-	-	1
3.A.19	the tms recognition/insertion complex (trc) family.	3.A.19.1.2	1	1	1	1	1	1	-	-
3.D.1	the h(+) or na(+)-translocating nadh dehydrogenase (ndh) family.	3.D.1.2.1	1	1	-	1	-	-	1	-
		3.D.1.6.1	1	-	1	1	-	1	-	-
		3.D.1.6.2	9	7	3	7	3	4	1	-
		3.D.1.6.3	1	1	2	-	1	-	-	-
		3.D.1.6.4	1	1	1	1	1	-	1	-
		3.D.1.7.1	-	-	-	1	-	-	1	-
3.D.2	the proton-translocating transhydrogenase (pth) family.	3.D.2.3.1	1	-	1	-	1	1	1	-

Continued on next page

Table 52 – continued from previous page

Family	Family Name	TCID	Aaf	Ani	Anc	Aor	Ann	Ncr	Pch	Spo
3.D.3	the proton-translocating quinol:cytochrome c reductase (qcr) superfamily.	3.D.3.2.1	1	2	1	2	1	1	1	2
		3.D.3.3.1	2	2	1	2	1	1	1	2
3.D.4	the proton-translocating cytochrome oxidase (cox) superfamily.	3.D.4.11.1	2	1	-	1	-	-	-	3
		3.D.4.3.1	-	-	-	1	-	-	-	-
		3.D.4.6.1	1	1	-	1	-	-	-	1
		3.D.4.6.2	1	2	-	2	-	-	-	2
		3.D.4.7.1	2	1	-	1	-	-	-	3
		3.D.4.8.1	3	3	-	3	-	-	-	3
3.E.1	the ion-translocating microbial rhodopsin (mr) family.	3.E.1.4.2	1	-	-	-	-	1	-	-
		3.E.1.4.3	1	-	-	-	-	1	-	-
		3.E.1.5.1	-	-	-	-	-	-	4	-
8.A.13	the tetratricopeptide repeat (tpr1) family.	8.A.13.1.1	-	-	-	-	-	-	-	1
8.A.27	the cdc50 p-type atpase lipid flippase subunit (cdc50) family.	8.A.27.1.1	-	-	-	1	-	-	-	-
		8.A.27.1.2	1	1	1	1	1	1	-	2
		8.A.27.1.5	1	1	-	-	-	-	-	-
8.A.40	the tetraspanin (tetraspanin) family.	8.A.40.2.1	-	-	-	-	-	1	-	-
8.A.41	the strech-activated calcium channel auxiliary protein, mid1 (mid1) family.	8.A.41.1.6	-	-	-	-	-	1	-	-
		8.A.41.1.7	-	-	-	-	-	-	-	1
9.A.2	the endomembrane protein-70 (emp70) family.	9.A.2.1.1	1	1	1	1	1	1	1	-
		9.A.2.1.2	1	-	1	1	1	1	-	-
		9.A.2.1.4	-	-	-	-	1	1	-	-
		9.A.2.1.6	1	1	1	1	1	1	-	-
9.A.6	the atp exporter (atp-e) family.	9.A.6.1.1	-	1	1	-	1	1	-	1
9.A.26	the lipid-translocating exporter (lte) family.	9.A.26.1.3	1	-	-	-	-	-	-	-
9.A.27	the non-classical protein exporter (ncpe) family.	9.A.27.1.3	-	-	-	-	-	-	-	1
9.A.41	the capsular polysaccharide exporter (cps-e) family.	9.A.41.1.1	2	2	1	1	2	2	2	3
9.A.54	the lysosomal cobalamin (b12) transporter (l-b12t) family.	9.A.54.1.2	-	-	-	1	-	1	-	-
		9.A.54.1.3	1	1	1	1	1	1	-	-
9.A.62	the aaa-atpase, bcs1 (bcs1) family.	9.A.62.1.1	1	-	-	1	1	1	1	1
9.B.1	the integral membrane caax protease (caax protease) family.	9.B.1.1.1	-	-	-	1	-	-	-	-
		9.B.1.1.2	1	1	1	1	1	1	-	-
		9.B.1.1.3	1	1	1	1	1	1	1	-
		9.B.1.2.2	1	1	1	1	1	1	-	-
9.B.7	the putative sulfate transporter (cysz) family.	9.B.7.2.3	-	-	1	-	-	-	-	-
9.B.12	the sensitivity to sodium or salt stress-induced hydrophobic peptide (sna) family.	9.B.12.2.2	1	1	-	-	1	-	1	1
9.B.16	the putative ductin channel (ductin) family.	9.B.16.1.1	2	3	3	3	3	2	2	2
		9.B.16.1.2	2	3	3	3	3	2	2	2
9.B.20	the putative mg(2+) transporter-c (mgtc) family.	9.B.20.1.3	1	-	-	-	-	-	-	-
9.B.25	the mitochondrial inner/outer membrane fusion (mmf) family.	9.B.25.1.1	1	-	1	-	1	1	-	-
9.B.26	the regulator of er stress and autophagy tmem208 (tmem208) family.	9.B.26.1.4	1	1	1	-	1	-	-	-
9.B.82	endoplasmic reticulum retrieval protein1 (putative heavy metal transporter) (rer1) family.	9.B.82.1.1	1	1	1	1	1	1	1	1
		9.B.82.1.2	1	1	1	1	1	1	1	1
		9.B.82.1.3	1	1	1	1	1	1	1	1
9.B.119	the glycan synthase, fks1 (fks1) family.	9.B.119.1.1	1	1	1	1	1	1	1	4
9.B.131	the post-gpi attachment protein (p-gap2) family.	9.B.131.1.1	1	1	-	1	1	1	-	1
9.B.135	the membrane trafficking yip (yip) family.	9.B.135.1.1	-	-	-	-	-	-	-	1
9.B.142	the integral membrane glycosyltransferase family 39 (gt39) family.	9.B.142.3.3	1	1	1	1	1	1	1	1
		9.B.142.3.5	1	1	1	1	1	1	1	1
9.B.143	the 6 tms duf1275/pf06912 (duf1275) family.	9.B.143.5.1	1	1	1	2	1	-	-	-
		9.B.143.5.2	-	-	-	-	-	-	1	-
9.B.158	the 4 tms putative dmt2 (dmt2) family.	9.B.158.1.8	-	1	-	-	-	-	-	-
		9.B.158.1.9	-	1	-	-	-	1	-	-

Appendix D

TCDB-Blast Results with Substrates and Localization

This appendix presents the results detailing predictions of substrates and localization.

D.1 TCDB-Blast Results with TrSSP Predictions

This section considers the TC-Family 1.A of channels and pores in the TCDB. Table 53 presents the predictions of TrSSP for those proteins in the eight fungal genomes that TCDB-Blast predicts to belong to TC-Family 1.A. The columns Family and Family Name contain the TC-Family identifier and its name. The column TCID contains the TCID of the TCDB entry predicted TCDB-Blast. The column Hit is the UniProtKB identifier for the matching TCDB entry. The column Query is the identifier for the entry in the fungal genome. The last 8 columns indicate the substrate groups predicted by TrSSP for the query protein: **AA**: Amino acid, **An**: Anion, **Ca**: Cation, **El**: Electron, **Pr/mR**: Protein/mRNA, **Su**: Sugar, **Ot**: Other, **NA**: no prediction was made by TrSSP.

Table 53: TCDB-Blast Results for Channels/Pores with Substrate Prediction

Family	Family Name	TCID	Hit	Query	AA	An	Ca	El	Pr/mR	Su	Ot	NA
1.A.1	The voltage-gated ion channel (vic) superfamily.	1.A.1.11.17	Q1HHN2	An08g03400			X					
				NRRL3_10990			X					
				NCU02762T0			X					
		1.A.1.11.23	O14234	SPAC6F6.01		X	X					
1.A.11	The ammonia transporter channel (amt) family.	1.A.11.3.1	P40260	AN7463	X						X	
				NCU03257T0	X	X	X				X	
				$jgi\|Phchr1\|134974\|e\_gww2.11.183.1$			X				X	
		1.A.11.3.2	P41948	SPAC664.14	X		X				X	
		1.A.11.3.3	Q8NKD5	Afu1g10930	X		X				X	
				AN1181	X	X	X				X	
				AN10097		X	X				X	
				An08g03200			X				X	
				An01g11640	X	X	X				X	
				AO090038000314			X				X	
				AO090023000411	X	X	X				X	
				NRRL3_10976			X				X	
				NRRL3_02582	X	X	X				X	
				NCU01065T0		X	X				X	
				NCU06613T0	X	X	X			X	X	
				$jgi\|Phchr1\|121517\|e\_gwh2.5.220.1$		X	X				X	
				SPCPB1C11.01		X	X				X	
		1.A.11.3.5	Q59UP8	Afu5g11020	X	X	X				X	
				AN0209	X	X	X			X	X	
				An14g02390	X	X	X			X	X	
				AO090026000749	X	X	X			X	X	
				NRRL3_00794	X	X	X			X	X	
				NCU05843T0	X	X	X				X	
				SPAC2E1P3.02C	X	X	X				X	
1.A.14	The testis-enhanced gene transfer (tegt) family.	1.A.14.3.3	A2VCJ6	$jgi\|Phchr1\|133598\|e\_gww2.6.475.1$	X		X	X		X		
1.A.16	The formate-nitrite transporter (fnt) family.	1.A.16.2.1	P35839	AO090038000194	X	X	X			X	X	
		1.A.16.2.2	Q5AST3	AN8647	X	X	X			X	X	
				AO090012000169	X	X				X	X	
				NRRL3_02998	X	X	X			X	X	
				NCU00758T0	X	X				X	X	
1.A.17	The calcium-dependent chloride channel (caclc) family.	1.A.17.5.5	G2Y513	Afu1g02130								X
				AN2880								X
				NCU00789T0								X
				$jgi\|Phchr1\|124528\|e\_gwh2.12.23.1$								X
				$jgi\|Phchr1\|35406\|gww2.1.47.1$								X
				SPBC354.08C		X					X	
		1.A.17.5.8	J5PL79	Afu5g10920		X		X				
				AN0229		X	X				X	
				An16g01540								X

Continued on next page

Table 53 – continued from previous page

Family	Family Name	TCID	Hit	Query	AA	An	Ca	El	Pr/mR	Su	Ot	NA
				An14g03660	X	X	X				X	
				AO090010000441			X					
				NRRL3_07300								X
				NRRL3_00911		X	X				X	
				NCU06986T0								X
				NCU06986T1								X
				$jgi\|Phchr1\|126382\|e\_$ gwh2.7.45.1								X
		1.A.17.5.9	Q9SY14	An01g06130								X
				NRRL3_02124								X
				SPAC2G11.09								X
		1.A.17.6.2	B6JZY0	SPBC691.05C								X
		1.A.17.6.4	B0YES0	Afu4g02970		X		X				
				Afu4g03330								X
				AN2477								X
				AN7165								X
				An14g03020		X						
				An14g01960								X
				AO090012000168		X						
				AO090011000165								X
				NRRL3_00851		X						
				NRRL3_00758								X
				NCU08273T0		X	X					
				$jgi\|Phchr1\|137491\|e\_$ gww2.3.132.1								X
1.A.23	The small conductance mechanosensitive ion channel (mscs) family.	1.A.23.4.9	F9X0Q3	Afu2g15000								X
				AN7571								X
				An15g03150								X
				AO090012000418								X
				NRRL3_03830								X
1.A.33	The cation channel-forming heat shock protein-70 (hsp70) family.	1.A.33.1.2	P0A6Y8	SPAC664.11								X
		1.A.33.1.3	P08107	Afu2g04620			X	X	X			
				Afu1g07440				X	X			
				AN2062				X	X			
				An11g04180				X	X		X	
				An16g09260								X
				AO090012000995			X	X	X			
				NRRL3_09797				X	X		X	
				NRRL3_06609								X
				NCU03982T0				X	X			
				NCU09602T0			X	X	X			
				NCU05269T0				X	X		X	
				NCU05269T1				X	X		X	
				NCU02075T0			X		X			
				$jgi\|Phchr1\|123502\|e\_$ gwh2.1.247.1			X	X	X			
				$jgi\|Phchr1\|131983\|e\_$ gww2.9.92.1			X	X	X			
				SPCC1739.13			X	X	X			
1.A.35	The cora metal ion transporter (mit) family.	1.A.35.2.3	O13657	SPBC27B12.12C								X

Continued on next page

Table 53 – continued from previous page

Family	Family Name	TCID	Hit	Query	AA	An	Ca	El	Pr/mR	Su	Ot	NA
		1.A.35.5.5	Q02783	AN7826					X			
				AO090011000032			X		X			
				SPBC25H2.08C								X
1.A.4	The transient receptor potential ca(2+) channel (trp-cc) family.	1.A.4.10.1	O94543	SPCC1322.03								X
		1.A.4.4.1	Q12324	Afu3g13490			X					
				AN3155			X					
				An02g09390			X					
				AO090012000784			X					
				NRRL3_05486			X					
				NCU16725T0		X						
				$jgi\|Phchr1\|138594\|e\_$ gww2.8.86.1		X	X					
		1.A.4.9.1	Q08967	AN1950		X					X	
		1.A.4.9.2	Q09917	SPAC1F7.03		X					X	
		1.A.4.9.3	P39719	Afu4g13340		X					X	
				An01g09050		X					X	
				An01g06610		X					X	
				AO090001000726		X					X	
				AO090005000355		X					X	
				AO090009000239		X					X	
				AO090038000415		X					X	
				NRRL3_02376		X					X	
				NRRL3_02161		X					X	
				NCU05785T0		X					X	
1.A.43	The camphor resistance (crcb) family.	1.A.43.2.3	S9W181	SPBPB8B6.06C	X	X				X	X	
				SPAC977.11	X	X				X	X	
		1.A.43.2.6	Q7SB51	NCU06262T0			X					
1.A.46	The anion channel-forming bestrophin (bestrophin) family.	1.A.46.2.1	Q5BB29	Afu5g06660								X
				AN2251								X
				AO090701000199								X
				NRRL3_06477								X
				NCU09677T0								X
		1.A.46.2.2	Q5AXS1	AN6909		X						
				An14g05100		X						
				AO090113000012		X	X					
				NRRL3_01019		X						
1.A.55	The synaptic vesicle-associated ca(2+) channel flower family.	1.A.55.4.1	B8N1Q6	Afu3g10600				X		X		
				AN11770						X		
				NRRL3_06946						X		
				NCU04760T0		X				X		
				SPBC32F12.12C		X					X	
1.A.56	The copper transporter (ctr) family.	1.A.56.1.10	A9XIK8	AN2934		X	X					
				An02g11700		X	X			X	X	
				NRRL3_05315		X	X					
				NCU03281T0			X					
				NCU03281T1		X	X				X	
		1.A.56.1.4	Q06686	Afu2g03730		X	X				X	
		1.A.56.1.5	O94722	SPCC1393.10		X	X					
		1.A.56.1.5	Q9P7F9	SPAC1142.05		X	X			X		

Continued on next page

Table 53 – continued from previous page

Family	Family Name	TCID	Hit	Query	AA	An	Ca	El	Pr/mR	Su	Ot	NA
		1.A.56.1.6	Q9USV7	SPBC23G7.16			X					
1.A.77	The mg(2+)/ca(2+) uniporter (mcu) family.	1.A.77.1.5	Q7S4I4	Afu4g10310								X
				An04g06590								X
				AO090003001191								X
				NRRL3_07719								X
				NCU08166T0								X
1.A.8	The major intrinsic protein (mip) family.	1.A.8.18.1	E3UN01	AO090010000024		X	X				X	
		1.A.8.18.3	E3UMZ5	NRRL3_01299		X				X	X	
		1.A.8.6.4	H6B4G1	Afu4g03390		X				X	X	
			H6B4G1	AN10902			X				X	
			H6B4G1	NRRL3_00798						X		
		1.A.8.7.1	P43549	An16g00230			X					
				NRRL3_07402								X
				SPAC977.17								X
		1.A.8.9.1	P47862	jgi—Phchr1—138875—e_-gww2.8.316.1		X					X	
		1.A.8.9.4	Q6ZXT4	AO090010000705	X						X	
1.A.81	The low affinity ca(2+) channel (lacc) family.	1.A.81.3.2	Q5A4M8	AN4615						X	X	
				An07g06530						X	X	
				AO090011000512		X	X			X		
				NRRL3_04731						X	X	
		1.A.81.4.1	A7UX97	NCU10610T0		X				X	X	
		1.A.81.5.1	I3VPY1	Afu3g09060	X		X					
				AN3036	X	X					X	
				AO090103000234			X				X	
				AO090005001364	X		X			X	X	
				NRRL3_07102	X	X	X				X	
				NCU02219T0	X		X					
1.A.88	The fungal potassium channel (f-kch) family.	1.A.88.1.4	A2QW01	An11g03330								X
				NRRL3_09876								X
		1.A.88.1.6	Q9P5J0	NCU03928T0								X
1.A.9	The neurotransmitter receptor cys loop ligand-gated ion channel (lic) family.	1.A.9.5.2	O95166	Afu1g07470			X	X				
				AN5131			X	X				
				An07g10020			X	X				
				AO090012000997			X	X	X			
				NRRL3_05018			X	X				
				NCU01545T0			X	X				
				jgi—Phchr1—122422—e_-gwh2.1.1122.1			X	X	X			
				SPBP8B7.24C		X	X					
1.B.69	The peroxysomal membrane porin 4 (pxmp4) family.	1.B.69.1.4	A2R8R0	Afu8g04780								X
				AN1483								X
				An16g08040								X
				AO090005000669								X
				NRRL3_06705								X
				NCU00828T0								X
1.C.47	The insect/fungal defensin family.	1.C.47.1.8	B1NJ41	AN11510			X					

Continued on next page

Table 53 – continued from previous page

Family	Family Name	TCID	Hit	Query	AA	An	Ca	El	Pr/mR	Su	Ot	NA
				AN5046			X					
1.F.1	The synaptosomal vesicle fusion pore (svf-pore) family.	1.F.1.1.2	P33328	Afu6g02920				X	X		X	
				AN8769					X		X	
				An12g07570				X	X		X	
				AO090012000430				X	X		X	
				NRRL3_03138				X	X		X	
				$jgi\|Phchr1\|134289\|e\_gww2.12.354.1$								X
				SPAC6G9.11				X	X			
			Q04338	Afu4g10710								X
				AN1973								X
				An04g05980								X
				AO090003001144								X
				NRRL3_07766								X
				SPBC3B9.10								X
1.H.1	The claudin tight junction family.	1.H.1.4.1	F5H8T9	An08g01170								X
				AO090012000911							X	
				NRRL3_10815								X
				NCU03601T0								X
		1.H.1.4.3	G3XZI4	Afu6g07470		X					X	
				AN5213							X	
				An07g08960							X	
				AO090005001554		X					X	
				AO090026000374								X
				NRRL3_04918							X	
		1.H.1.4.5	Q2TX92	AO090010000235			X				X	
1.I.1	The nuclear pore complex (npc) family.	1.I.1.1.1	P38181	AO090005000465								X
				NCU04463T0								X
			P39685	Afu3g05500								X
				AN3454								X
				An11g11140								X
				AO090020000021								X
				NRRL3_09229								X
				NCU10747T0								X
				SPBC29A10.07								X

D.2 TCDB-Blast Results with LocTree3 Predictions

This section considers the localization of the members of the TC-Superfamily 2.A.1, the MFS Superfamily, as predicted by TCDB-Blast for the eight fungal genomes in our study. The localizations are the predictions of LocTree3. Only those sequences in unusual localizations for the given TCID are listed. The usual localization is defined to be the most common localization predicted for sequences with the given TCID. Table 54 presents the predictions of LocTree3 for those proteins in the eight fungal genomes that TCDB-Blast predicts to belong to TC-Superfamily 2.A.1 and that have an unusual localization. The columns Subfamily and Subfamily Name contain the TC-Subfamily identifier and its name. The column TCID contains the TCID of the TCDB entry predicted TCDB-Blast. The column Hit is the UniProtKB identifier for the matching TCDB entry. The column Query is the identifier for the entry in the fungal genome. The column Location (Usual) contains the usual localization for the given TCID. The column Location (Unusual) contains the localization predicted for the given query sequence in the fungal genome, where a horizontal line makes the start of a new unusual localization and a blank inidcates a continuance of the unusual localization. *Mem.* is short for membrane.

Table 54: Usual and Unsual Location of MFS Superfamily 2.A.1.

Subfamily	Subfamily Name	TCID	Hit	Query	Location (Usual)	Location (Unusual)
2.A.1.1	Sugar Porters (SP)	2.A.1.1.7	P11636	AN1109	Plasma Mem.	Vacuole Mem.
				An07g06300		Mito Mem.
				An15g04270		
				AO090113000088		
				AO090001000641		
				NRRL3_03902		
				NRRL3_04711		
		2.A.1.1.10	P15685	AO090011000064	Plasma Mem.	Mito Mem.
				NCU07861T0		
		2.A.1.1.38	P39932	AN2584	Plasma Mem.	Mito Mem.
				AN3115		
				An04g08030		
				NRRL3_07609		
				Afu4g14610		Vacuole Mem.
				AN5067		
		2.A.1.1.39	P49374	An07g10370	Plasma Mem.	Vacuole Mem.
				An08g04040		
				NRRL3_05043		
				AO090010000063		Mito Mem.
		2.A.1.1.40	Q64L87	An16g06610	Plasma Mem.	Vacuole Mem.
				AO090001000381		
		2.A.1.1.57	Q8J0V1	AO090023000340	Plasma Mem	Mito Mem.
		2.A.1.1.68	A3M0N3	Afu6g14590	Plasma Mem	Vacuole Mem.
		2.A.1.1.73	Q5A8J5	Afu5g01080	Plasma Mem.	Mito Mem.
				AN9168		Vacuole Mem.

Continued on next page

Table 54 – continued from previous page

Subfamily	Subfamily Name	TCID	Hit	Query	Location (Usual)	Location (Unusual)
		2.A.1.1.82	Q7SCU1	Afu1g17310	Plasma Mem.	Mito Mem.
				NCU00809T0		
				AN6831		Vacuole Mem.
		2.A.1.1.83	Q7SD12	Afu8g04480	Mito Mem.	Vacuole Mem.
				An09g04810		
				AO090166000089		
		2.A.1.1.117	G4N740	An06g02030	Plasma Mem.	Mito Mem.
				NRRL3_11786		
2.A.1.2	The Drug: H+ Antiporter-1	2.A.1.2.1	P33532*	SPAC17A2.01	Plasma Mem.	ER Mem.
				SPCC965.13		
				AO090012000612		
				NRRL3_07427		
				An18g00480		
				AO090009000046		
				NRRL3_10205		
				NRRL3_03393		
				AN5540		
				AO090023000700		
				Afu8g04702		
				Afu5g00430		
				AN10207		
				AN5763		
				AN3398		
				An07g05880		
				NRRL3_10675		
				NRRL3_02942		
				NRRL3_06692		
				NRRL3_04678		
				SPBC1271.10C		
				An03g05750		
				An16g02610		
				NRRL3_07101		
				NRRL3_08364		
				$jgi\|Phchr1\|140622\|e_$gww2.17.40.1		
				$jgi\|Phchr1\|126074\|e_$gwh2.13.144.1		
				$jgi\|Phchr1\|138927\|e_$gww2.8.161.1		
				AO090011000474		Mito Mem.
				NRRL3_05523		
		2.A.1.2.58	Q8RWN2	Afu1g06440	Plasma Mem.	ER Mem.
				Afu1g13800		
				Afu2g11420		
				Afu2g11580		
				Afu2g16860		
				Afu3g01120		
				Afu3g01890		
				Afu3g02060		
				Afu3g02780		
				Afu5g02700		
				Afu6g13780		
				Afu7g04900		
				An02g01100		
				An02g03620		
				An02g09970		
				An02g13050		
				An02g14470		
				An03g00320		
				An04g08250		

Continued on next page

Table 54 – continued from previous page

Subfamily	Subfamily Name	TCID	Hit	Query	Location (Usual)	Location (Unusual)
				AN3270		
				AN5369		
				AN5559		
				AN6451		
				AN6477		
				AN6942		
				AN7972		
				AN0732		
				AN10036		
				AN10152		
				An07g00300		
				An08g06980		
				An08g08560		
				An08g10970		
				An09g02210		
				An09g02580		
				An09g05070		
				An11g07300		
				An11g08990		
				An11g09140		
				An12g02800		
				An13g03610		
				An15g04580		
				An16g01040		
				An16g01660		
				An18g00700		
				NRRL3_00171		
				NRRL3_00195		
				NRRL3_00202		
				NRRL3_00407		
				NRRL3_01291		
				NRRL3_02997		
				NRRL3_03925		
				NRRL3_03930		
				NRRL3_04250		
				NRRL3_05458		
				NRRL3_05968		
				NRRL3_06167		
				NRRL3_07288		
				NRRL3_07345		
				NRRL3_07535		
				NRRL3_07590		
				NRRL3_08755		
				NRRL3_08973		
				NRRL3_09414		
				NRRL3_09421		
				NRRL3_09550		
				NRRL3_10222		
				NRRL3_10595		
				NRRL3_11027		
				NRRL3_11278		
				NRRL3_11761		
				AO090001000704		
				AO090003000523		
				AO090003000563		
				AO090003001037		
				AO090005000054		

Continued on next page

Table 54 – continued from previous page

Subfamily	Subfamily Name	TCID	Hit	Query	Location (Usual)	Location (Unusual)
				AO090005000991		
				AO090010000036		
				AO090010000105		
				AO090010000160		
				AO090010000186		
				AO090011000014		
				AO090011000049		
				AO090011000413		
				AO090012000288		
				AO090012000494		
				AO090020000544		
				AO090023000405		
				AO090026000005		
				AO090026000193		
				AO090026000247		
				AO090026000485		
				AO090102000049		
				AO090102000135		
				AO090102000388		
				AO090103000346		
				AO090113000138		
				AO090113000181		
				AO090138000118		
				NCU00306T0		
				NCU06519T0		
				$jgi\|Phchr1\|133216\|e\_gww2.1.267.1$		
				$jgi\|Phchr1\|140008\|e\_gww2.2.378.1$		
				$jgi\|Phchr1\|26770\|gwh2.2.173.1$		
				$jgi\|Phchr1\|122451\|e\_gwh2.1.419.1$		
				SPAC11D3.05		
				SPBC530.02		
				SPCC794.04C		
				Afu1g03730		Vacuole Mem.
				AN4019		
				AO090003000971		
				AO090023000061		
				NCU06341T0		
				$jgi\|Phchr1\|131654\|e\_gww2.26.8.1$		
				SPCC330.07C		
				SPCC613.01		
				SPCC613.02		
				SPCC757.11C		
2.A.1.3	The Drug:H+ Antiporter-2	2.A.1.3.32*	Q9ZGB6	An03g01790	Vacuole Mem.	Plasma Mem.
				NRRL3_08650		
		2.A.1.3.33	O32182	An01g01245	Plasma Mem.	Vacuole Mem.
				An07g00060		
				AO090012000158		
				NRRL3_04232		
				NCU09978T0		
				$jgi\|Phchr1\|136833\|e\_gww2.7.170.1$		
				$jgi\|Phchr1\|37613\|gww2.4.148.1$		
		2.A.1.3.47*	Q9C1B3	NRRL3_00256	Plasma Mem.	Vacuole Mem.
		2.A.1.3.65	H2E274	Afu1g16910	Vacuole Mem.	Plasma Mem.
				Afu3g14720		
				Afu6g14640		
				An01g11290		
				An02g02780		

Continued on next page

Table 54 – continued from previous page

Subfamily	Subfamily Name	TCID	Hit	Query	Location (Usual)	Location (Unusual)
				An04g06250		
				AN11217		
				AN11821		
				An11g08620		
				An12g08620		
				An16g01590		
				AN3491		
				AN3884		
				AN7200		
				AO090001000543		
				AO090003001490		
				AO090010000407		
				AO090023000039		
				AO090026000199		
				AO090026000577		
				AO090038000038		
				AO090701000567		
				NCU00711T0		
				NCU00857T0		
				NCU03789T0		
				NCU09458T0		
				NCU09458T1		
				NCU09640T0		
				NRRL3_03067		
				NRRL3_06038		
				NRRL3_07295		
				NRRL3_07740		
				NRRL3_08967		
				$jgi\|Phchr1\|122125\|e\_gwh2.9.150.1$		
2.A.1.7	Fucose: H+ Symporter	2.A.1.7.1	P11551	AN5742	Plasma Mem.	ER Mem.
				NRRL3_10670		
				An18g06310		GA Mem.
		2.A.1.7.13	Q08280	AO090011000241	Plasma Mem.	GA Mem.
				AO090308000019		
				NRRL3_04481		
2.A.1.8	Nitrate/Nitrite Porter (NNP)	2.A.1.8.13	Q8X193	AN0399	Plasma Mem.	Vacuole Mem.
2.A.1.9	Phosphate: H+ Symporter (PHS)	2.A.1.9.2	Q7RVX9	An04g04240	Plasma Mem.	ER Mem.
				NRRL3_07894		
				$jgi\|Phchr1\|125289\|e\_gwh2.27.9.1$		
				$jgi\|Phchr1\|128372\|e\_gwh2.4.612.1$		
		2.A.1.9.7*	P25346	AN5549	Plasma Mem.	ER Mem.
				AN2864		
				An11g02600		
				An02g08180		
				AO090003000167		
				NRRL3_09931		
				NRRL3_05607		
				$jgi\|Phchr1\|4504\|fgenesh1\_pg.$		
				C_scaffold_7000317		
				SPBC1271.09		
		2.A.1.13.19	Q08268	AN4481	ER Mem.	Plasma Mem.
				NRRL3_04850		
				NCU06167T0		
				NCU16370T0		
		2.A.1.13.4	Q08777	An11g08190	ER Mem.	Plasma Mem.
				An11g07630		
				AO090023000881		

Continued on next page

Table 54 – continued from previous page

Subfamily	Subfamily Name	TCID	Hit	Query	Location (Usual)	Location (Unusual)
				NRRL3_07645 NCU05089T0		
2.A.1.14	An:Ca Symporter (ACS)	2.A.1.14.11	P53322	NRRL3_03334	Plasma Mem.	Mito Mem.
		2.A.1.14.3	P70786	AN9000	Plasma Mem.	Mito Mem.
		2.A.1.14.36	Q07904	AO090010000742 NRRL3_09657 $jgi\|Phchr1\|133152\|e$_gww2.1.449.1	Plasma Mem.	Mito Mem.
		2.A.1.14.37	P39709	AN3066 AN4107	Plasma Mem.	Mito Mem.
2.A.1.16	Siderophore-Iron Transporter (SIT)	2.A.1.16.1	P39980	An01g00720 NRRL3_01644	Plasma Mem.	Vacuole Mem.
		2.A.1.16.5	O94607	AO090001000692	Plasma Mem.	Vacuole Mem.
		2.A.1.16.6	O74395	AN7485 SPBC4F6.09	Vacuole Mem.	ER Mem.
				AO090009000061		Plasma Mem.
		2.A.1.16.7	Q870L2	AN3160 An03g03560 AO090023000049 NRRL3_08534	Vacuole Mem.	Plasma Mem.
2.A.1.19	Organic Ca Transporter (OCT)	2.A.1.19.38	Q9C101	NRRL3_08676 NRRL3_09127 NRRL3_03935 $jgi\|Phchr1\|138989\|e$_gww2.8.136.1	Plasma Mem.	ER Mem.
				AO090012000051		Peroxisome Mem.
		2.A.1.19.48	Q0CZ13	AO090026000209	Plasma Mem.	Vacuole Mem.
2.A.1.25	Peptide-Acetyl-Coenzyme A Transporter (PAT)	2.A.1.25.1	O00400	AN4836 AO090020000192	ER Mem.	Plasma Mem.
2.A.1.48	Vacuolar Basic AA Transporter (V-BAAT)	2.A.1.48.3	Q09752	AO090102000036	Vacuole Mem.	Plasma Mem.
2.A.1.58	N-Acetylglucosamine Transporter (NAG-T)	2.A.1.58.1	Q5A7S4	Afu1g00440 AN1427 AN8127 An16g09020 NRRL3_06628	ER Mem.	GA Mem.
		2.A.1.58.5	C9S7Y7	Afu3g15000 AO090124000021	ER Mem.	GA Mem.

Bibliography

[ABB+08] Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, Folker Meyer, Gary J Olsen, Robert Olson, Andrei L Osterman, Ross A Overbeek, Leslie K McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics*, 75(9), 2008.

[ABGW94] Stephen F. Altschul, Mark S. Boguski, Warren Gish, and John C. Wootton. Issues in searching molecular sequence databases. *Nature Genetics*, 6(2):119–129, 1994.

[ADD+12] Ramy K. Aziz, Scott Devoid, Terrence Disz, Robert A. Edwards, Christopher S. Henry, Gary J. Olsen, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D. Pusch, Rick L. Stevens, Veronika Vonstein, and Fangfang Xia. SEED servers: High-performance access to the SEED Genomes, Annotations, and Metabolic Model. *PLOS One*, 7(10):1–10, 2012.

[AG96] Stephen F. Altschul and Warren Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 1996.

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Euagene W. Myers, and David J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[ALS+13] Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen. The RAVEN toolbox and its use for generating

a genome-scale metabolic model for *Penicillium chrysogenum*. *PLOS Computational Biology*, 9:1–16, 2013.

[AMS+97] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[Anf73] Christian B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[ANN08] Mikael Rørdam Andersen, Michael Lynge Nielsen, and Jens Nielsen. Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Molecular Systems Biology*, 4(1), 2008.

[ARC+54] Christian B. Anfinsen, Robert R. Redfield, Warren L. Choate, Juanita Page, and William R. Carroll. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *Journal of Biological Chemistry*, 207(1):201–210, 1954.

[AS06] Tero Aittokallio and Benno Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255, 2006.

[BA00] Amos Bairoch and Rolf Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, 2000.

[Bai00] Amos Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1):304–305, 2000.

[BB01] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, second edition, 2001.

[BCD04] Ewan Birney, Michele Clamp, and Richard Durbin. GeneWise and GenomeWise. *Genome Research*, 14(5):988–995, 2004.

[BCFdC13] Rodrigo C Barros, Ricardo Cerri, Alex A Freitas, and André CPLF de Carvalho. Probabilistic clustering for hierarchical multi-label classification of protein functions. In *Machine Learning and Knowledge Discovery in Databases*, pages 385–400. Springer, 2013.

[BH13] Ahmad Barghash and Volkhard Helms. Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs. *BMC Bioinformatics*, 14(1):343, 2013.

[BI] Broad Institute. FungiCyc — an organism-specific database of metabolic pathways, compounds and reactions. `http://fungicyc.broadinstitute.org/`. Accessed in Jan, 2012.

[BN05] Irina Borodina and Jens Nielsen. From genomes to *in silico* cells via metabolic networks. *Current Opinion in Biotechnology*, 16(3):350–355, 2005.

[BS04] Wolfgang Busch and Milton H. Saier, Jr. The IUBMB-Endorsed Transporter Classification System. *Molecular Biotechnology*, 27:253–262, 2004.

[BST06] Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[C+14] UniProt Consortium et al. UniProt: a hub for protein information. *Nucleic Acids Research*, page gku989, 2014.

[CAD+10] Ron Caspi, Tomer Altman, Joseph M. Dale, Kate Dreher, Carol A. Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S. Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G. Shearer, Peifen Zhang, and Peter D. The MetaCyc database of metabolic pathways and enzymes and the Biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 36:D623–D631, 2010.

[CAI+14] Gustavo C Cerqueira, Martha B Arnaud, Diane O Inglis, Marek S Skrzypek, Gail Binkley, Matt Simison, Stuart R Miyasato, Jonathan Binkley, Joshua Orvis, Prachi Shah, et al. The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Research*, 42(D1):D705–D710, 2014.

[CBBWT61] Francis H. C. Crick, Leslie Barnett, Sydney Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192:1227–1232, 1961.

[CBdC12] Ricardo Cerri, Rodrigo C Barros, and Andre CPLF de Carvalho. A genetic algorithm for hierarchical multi-label classification. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 250–255. ACM, 2012.

[CBDC14] Ricardo Cerri, Rodrigo C Barros, and André CPLF De Carvalho. Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, 80(1):39–56, 2014.

[CBGZ06] Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Incremental algorithms for hierarchical classification. *The Journal of Machine Learning Research*, 7:31–54, 2006.

[CBS05] Michael P. Cary, Gary D. Bader, and Chris Sander. Pathway information for systems biology. *FEBS Letters*, 579:1815–1820, 2005.

[CC14] Abhijit Chakraborty and Saikat Chakrabarti. A survey on prediction of specificity-determining sites in proteins. *Briefings in Bioinformatics*, page bbt092, 2014.

[CDTTN12] Jia-Ming Chang, Paolo Di Tommaso, Jean-François Taly, and Cedric Notredame. Accurate multiple sequence alignment of transmembrane proteins with psi-coffee. *BMC Bioinformatics*, 13(4):1, 2012.

[Cho00] Kuo-Chen Chou. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical and Biophysical Research Communications*, 278(2):477–483, 2000.

[CJ10] Ludovic Cottret and Fabien Jourdan. Graph methods for the investigation of metabolic networks in parasitology. *Parasitology*, 137(9):1393–1407, 2010.

[CK08] Kwangmin Choi and Sun Kim. ComPath: comparative enzyme analysis and annotation in pathway/subsystem contexts. *BMC Bioinformatics*, 9(145):1–15, 2008.

[COHA+10] Marija Cvijovic, Roberto Olivares-Hernandez, Rasmus Agren, Niklas Dahr, Wanwipa Vongsangnak, Intawat Nookaew, Kiran Raosaheb Patil, and Jens

Nielsen. BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic Acid Research*, 38:W144–W149, 2010.

[COLG11] Shu-An Chen, Yu-Yen Ou, Tzong-Yi Lee, and M Michael Gromiha. Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties. *Bioinformatics*, 27(15):2062–2067, 2011.

[CRCFK03] Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut, and Daniel Kahn. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research*, 31(22):6633–6639, 2003.

[CVP+15] Zachary Chiang, Ake Vastermark, Marco Punta, Penelope C Coggill, Jaina Mistry, Robert D Finn, and Milton H Saier. The complexity, challenges and benefits of comparing two transporter classification systems in TCDB and Pfam. *Briefings in Bioinformatics*, page bbu053, 2015.

[DCP+10] Emek Demir, Michael P. Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter DEustachio, Carl Schaefer, Joanne Luciano, Frank Schacherer, Irma Martinez-Flores, Zhenjun Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C. Lopez-Fuentes, Huaiyu Mi, Elgar Pichler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha Rajasimha, Ranjani Ramakrishnan, Imran Shah Mustafa Syed, Nadia Anwar, Ozgun Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Augustin Luna, Peter Murray-Rust, Eric Neumann, Oliver Reubenacker, Matthias Samwald, Martijn Van Iersel, Sarala Wimalaratne, Keith Allen, Burk Braun, Michelle Whirl-Carrillo, Kei-Hoi Cheung, Kam Dahlquist, Andrew Finney, Marc Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, David Kane, Shiva Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Ren, Nigam Shah, Margot Sunshine, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H. Buetow, Andrey Rzhetsky, Vincent Schachte, Bruno S Sobral, , Ugur Dogrusoz, Shannon McWeeney, Mirit Aladjem, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le Novre, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edgar Wingender, Peter D. Karp, Chris Sander, and Gary D.

Bader. The BioPAX community standard for pathway data sharing. *Computational Biology, Nature Biotechnology*, 28(9):935–1308, 2010.

[DGHW03] Yves Deville, David Gilbert, Jacques Van Helden, and Shoshana J. Wodak. An overview of data models for the analysis of biochemical pathways. *Briefings in Bioinformatics*, 4(3):246–259, 2003.

[DHP04] Natalie C. Duarte, Markus J. Herrgard, and Bernhard Ø. Palsson. Reconstruction and validation of *Saccharomyces cerevisiae iND750*, a fully compartmentalized genome-scale metabolic model. *Genome research*, 14:1298–1309, 2004.

[DPK10] J.M. Dale, L. Popescu, and P.D. Karp. Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics*, 11(15):1–14, 2010.

[DRFR15] Oscar Dias, Miguel Rocha, Eugénio C Ferreira, and Isabel Rocha. Reconstructing genome-scale metabolic models with *merlin*. *Nucleic Acids Research*, page gkv294, 2015.

[Edd98] Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[Edg04] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[ESKW84] D Eisenberg, E Schwarz, M Komaromy, and R Wall. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology*, 179(1):125–142, 1984.

[EWT82] D Eisenberg, RM Weiss, and TC Terwilliger. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, 299(5881):371–374, 1982.

[FBS+14] Alexander Farwick, Stefan Bruder, Virginia Schadeweg, Mislav Oreb, and Eckhard Boles. Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. *Proceedings of the National Academy of Sciences*, 111(14):5159–5164, 2014.

[FF+14] Fabio Fabris, Alex Freitas, et al. Dependency network methods for hierarchical multi-label classification of gene functions. In *Proceedings of the 2014 IEEE*

Symposium on Computational Intelligence and Data Mining, pages 241–248. IEEE, 2014.

[FFF+03] Jochen Forster, Iman Famili, Patrick Fu, Bernhard Ø. Palsson, and Jen Nielsen. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13:244–253, 2003.

[FK11] L Ferrer and P D Karp. Discovering novel subsystems using comparative genomics. *Bioinformatics*, 27(18):2478–85, 2011.

[FP08] Adam M Feist and Bernhard Ø Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature biotechnology*, 26(6):659–667, 2008.

[FPG10] David A. Fell, Mark G. Poolman, and Albert Gevorgyan. Building and analyzing genome-scale metabolic models. *Biochemical Society Transactions*, 38(5):1197–1201, 2010.

[FS11] B Teusink F Santos, J Boele. A practical guide to genome-scale metabolic models and their analysis. *Methods Enzymol*, 500:509–32, 2011.

[FST05] Christof Francke, Roland J Siezen, and Bas Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in microbiology*, 13(11):550–558, 2005.

[GHH+14] Tatyana Goldberg, Maximilian Hecht, Tobias Hamp, Timothy Karl, Guy Yachdav, Nadeem Ahmed, Uwe Altermann, Philipp Angerer, Sonja Ansorge, Kinga Balasz, et al. LocTree3 prediction of localization. *Nucleic Acids Research*, 42(W1):W350–W355, 2014.

[GK04] Michelle L. Green and Peter D. Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Systems Biology*, 76(5):1–16, 2004.

[GK06] M. L. Green and P. D. Karp. The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, 34(13):3687–3697, 2006.

[GK07] M L Green and P D Karp. Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics*, 23(13):i205–11, 2007.

[GNY+13] Ilya Getsin, Gina H Nalbandian, Daniel C Yee, Ake Vastermark, Philipp CG Paparoditis, Vamsee S Reddy, and Milton H Saier. Comparative genomics of transport proteins in developmental bacteria: *Myxococcus xanthus* and *Streptomyces coelicolor*. *BMC Microbiology*, 13(1):279, 2013.

[GO14] MM Gromiha and YY Ou. Bioinformatics approaches for functional annotation of membrane proteins. *Briefings in Bioinformatics*, 15(2):155–168, 2014.

[Gro] Saier Lab Bioinformatics Group. Transporter Classification Database. `http://www.tcdb.org/`. Accessed in March, 2013.

[GY08] M Michael Gromiha and Yukimitsu Yabuki. Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics*, 9(1):135, 2008.

[HCL+07] Wolfgang Huber, Vincent J Carey, Li Long, Seth Falcon, and Robert Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8(suppl 6)(S8), 2007.

[HdMD+13] Janna Hastings, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, Gareth Owen, Steve Turner, Mark Williams, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(D1):D456–D463, 2013.

[HFS+03] Michael Hucka, Andrew Finney, Herbert M Sauro, Hamid Bolouri, John C Doyle, Hiroaki Kitano, Adam P Arkin, Benjamin J Bornstein, Dennis Bray, Athel Cornish-Bowden, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[HKMG13] Ronit Hod, Refael Kohen, and Yael Mandel-Gutfreund. Searching for protein signatures using a multilevel alphabet. *Proteins: Structure, Function, and Bioinformatics*, 81(6):1058–1068, 2013.

[HPCW11] B J Haas, M D Pearson, C A Cuomo, and J R Wortman. Approaches to fungal genome annotation. *Mycology*, 2(3):118–141, 2011.

[HPO+07] Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, CJ Adams-Collier, and Kenta Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, 35(suppl 2):W585–W587, 2007.

[HR14] Joshua J Hamilton and Jennifer L Reed. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental microbiology*, 16(1):49–59, 2014.

[HSMM+15] Rachael P Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J Martin, and Claire O'Donovan. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, 2015.

[HWGW02] Jacques Van Helden, Lorenz Wernisch, David Gilbert, and Shoshana Wodak. Graph-based analysis of metabolic networks. In *Bioinformatics and Genome Analysis*, pages 245–274. Springer-Verlag, Germany, 2002.

[HZCS09] Christopher S. Henry, Jenifer F. Zinner, Matthew P. Cohoon, and Rick L. Stevens. iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biology*, 10(6):R69 – R69.15, 2009.

[IUB] IUBMB. International Union of Biochemistry and Molecular Biology. `http://www.iubmb.org/`. Accessed in April, 2011.

[Kar] Peter D. Karp. Pathway Tools. `http://bioinformatics.ai.sri.com/ptools/release-notes.html`. Accessed in October, 2011.

[KBM+94] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235(5):1501–1531, 1994.

[KCVSZ+11] Ingrid M. Keseler, Julio Collado-Vides, Alberto Santos-Zavaleta, Martin Peralta-Gil, Socorro Gama-Castro, Luis Muniz-Rascado, Cesar Bonavides-Martinez, Suzanne Paley, Markus Krummenacker, Tomer Altman, Pallavi Kaipa1, Aaron Spaulding, John Pacheco, Mario Latendresse, Carol Fulcher,

Malabika Sarker, Alexander G. Shearer, Amanda Mackie, Ian Paulsen, Robert P. Gunsalus, and Peter D. Karp. EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Research*, 39:D583–D590, 2011.

[KD82] Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132, 1982.

[KL] Kanehisa Laboratories. Kegg: Kyoto Encyclopedia of Genes and Genomes. `http://www.genome.jp/kegg/`. Accessed in Feb, 2013.

[KLC11] Peter D. Karp, Mario Latendresse, and Ron Caspi. The Pathway Tools pathway prediction algorithm. *Standards In Genomic Sciences*, 5:424–429, 2011.

[KPK+09] P.D. Karp, S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I.M. Keseler, and R. Caspi. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 11(1):40–79, 2009.

[KPR02] P.D. Karp, S. Paley, and P. Romero. The Pathway Tools software. *Bioinformatics*, 18:S225–S232, 2002.

[KR93] Peter D. Karp and Monica Riley. Representations of metabolic knowledge. *Proceedings of First International Conference on Intelligent Systems for Molecular Biology*, pages 207–215, 1993.

[KS13] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[KTY+13] Masaaki Kotera, Yasuo Tabei, Yoshihiro Yamanishi, Toshiaki Tokimatsu, and Susumu Goto. Supervised *de novo* reconstruction of metabolic pathways from metabolome-scale compound sets. *Bioinformatics*, 29(13):i135–i144, 2013.

[Kuy08] Frans A. Kuypers. Cell membranes. In Steven R. Goodman, editor, *Medical Cell Biology*, chapter 2, pages 27–57. Elsevier/Academic Press, Amsterdam Boston, third edition, 2008.

[Lag93] Rosario Lagunas. Sugar transport in *Saccharomyces cerevisiae*. *FEMS Microbiology Reviews*, 104:229–242, 1993.

[LBUZ09] Haiquan Li, Vagner A Benedito, Michael K Udvardi, and Patrick Xuechun Zhao. TransportTP: A two-phase classification approach for membrane transporter prediction and characterization. *BMC Bioinformatics*, 10(418):1–13, 2009.

[LBZ+00] Harvey Lodish, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. Protein structure and function. In *Molecular Cell Biology*, chapter 3, pages 78–83. W.H. Freeman & Co, New York, fourth edition, 2000.

[LDZ08] Haiquan Li, Xinbin Dai, and Xuechun Zhao. A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. *Bioinformatics*, 24(9):1129–1136, 2008.

[LHC+06] HH Lin, LY Han, CZ Cai, ZL Ji, and YZ Chen. Prediction of transporter family from protein sequence by support vector machine approach. *Proteins: Structure, Function, and Bioinformatics*, 62(1):218–231, 2006.

[Loi12] Nicolas Loira. *Scaffold-based Reconstruction Method for Genome-Scale Metabolic Models*. PhD thesis, Université Sciences et Technologies-Bordeaux I, 2012.

[LPK08] Thomas J. Lee, Ian Paulsen, and Peter Karp. Annotation-based inference of transporter function. *Critical Reviews in Biochemistry and Molecular Biology*, 24:i259–i267, 2008.

[LZS15] Nicolas Loira, Anna Zhukova, and David James Sherman. Pantograph: A template-based method for genome-scale metabolic model reconstruction. *Journal of Bioinformatics and Computational Biology*, 13(02):1550006, 2015.

[MBZC+13] Aron Marchler-Bauer, Chanjuan Zheng, Farideh Chitsaz, Myra K. Derbyshire, Lewis Y. Geer, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Christopher J. Lanczycki, Fu Lu, Shennan Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Roxanne A. Yamashita, Dachuan

Zhang, and Stephen H. Bryant. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Research*, 41(D1):D348–D352, 2013.

[MCT+10] Jean Muller, Christopher J Creevey, Julie D Thompson, Detlev Arendt, and Peer Bork. AQUA: automated quality improvement for multiple sequence alignments. *Bioinformatics*, 26(2):263–265, 2010.

[MCZ14] Nitish K. Mishra, Junil Chang, and Patrick X. Zhao. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS One*, 9(6):1–14, 2014.

[MGD+] Arnaud MB, Cerquiera GC, Inglis DO, Skrzypek MS, Binkley J, Shah P, Wymore F, Binkley G, Miyasato SR, Simison M, Wortman JR, and Sherlock G. Aspergillus genome database. `http://www.aspergillusgenome.org/`. Accessed in December, 2012.

[MGMR+12] Thilo Muth, Juan A García-Martín, Antonio Rausell, David Juan, Alfonso Valencia, and Florencio Pazos. Jdet: interactive calculation and visualization of function-related conservation patterns in multiple sequence alignments and structures. *Bioinformatics*, 28(4):584–586, 2012.

[MIO+07] Yuki Moriya, Masumi Itoh, Shujiro Okuda, Akiyasu C Yoshizawa, and Minoru Kanehisa. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35(suppl 2):W182–W185, 2007.

[MLP+04] Diego Martinez, Luis F Larrondo, Nik Putnam, Maarten D Sollewijn Gelpke, Katherine Huang, Jarrod Chapman, Kevin G Helfenbein, Preethi Ramaiya, J Chris Detter, Frank Larimer, et al. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain rp78. *Nature biotechnology*, 22(6):695–700, 2004.

[MNP14] Jonathan Monk, Juan Nogales, and Bernhard Ø Palsson. Optimizing genome-scale network reconstructions. *Nature biotechnology*, 32(5):447–452, 2014.

[MPH09] Aziz Mithani, Gail M. Preston, and Jotun Hein. Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, 25(14):1831–1832, 2009.

[NEF+06] Richard A Notebaart, Frank HJ Van Enckevort, Christof Francke, Roland J Siezen, and Bas Teusink. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics*, 7(296), 2006.

[NJ10] Timothy Nugent and David T Jones. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Computational Biology*, 6(3):e1000714, 2010.

[NNM14] Chioko Nagao, Nozomi Nagano, and Kenji Mizuguchi. Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. *PloS One*, 9(1):e84623, 2014.

[OCG10] Yu-Yen Ou, Shu-An Chen, and M Michael Gromiha. Classification of transporters using efficient radial basis function networks with position-specific scoring matrices and biochemical properties. *Proteins: Structure, Function, and Bioinformatics*, 78(7):1789–1797, 2010.

[Ohn70] Susumu Ohno. *Evolution by Gene Duplication*. Springer, New York, 1970.

[OLP+00] Ross Overbeek, Niels Larsen, Gordon D. Pusch, Mark D Souza, Evgeni Selkov, Jr, Nikos Kyrpides, Michael Fonstein, Natalia Maltsev, and Evgeni Selkov. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28(1):123–125, 2000.

[OP10] Jeffrey D. Orth and Bernhard Ø. Palsson. Systematizing the generation of missing metabolic knowledge. *Biotechnology and Bioengineering*, 107(3):403–412, 2010.

[ÖSF+10] Gabriel Östlund, Thomas Schmitt, Kristoffer Forslund, Tina Köstler, David N. Messina, Sanjit Roopra, Oliver Frings, and Erik L. L. Sonnhammer. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38(suppl 1):D196–D203, 2010.

[OYH+08] Shujiro Okuda, Takuji Yamada, Masami Hamajima, Masumi Itoh, Toshiaki Katayama, Peer Bork, Susumu Goto, and Minoru Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Research*, 36(suppl 2):W423–W426, 2008.

[Pal08] Bernhard Ø. Palsson. *Systems Biology: Properties of Reconstructed Network*. Cambridge, New York, 2008.

[Pal11] Bernhard Ø. Palsson. *Systems Biology: Simulation of Dynamic Network States*. Cambridge University Press, United Kingdom, 2011.

[PBB+03] Frederic Plewniak, Laurent Bianchetti, Yann Brelivet, Annaick Carles, Frederic Chalmel, Odile Lecompte, Thiebaut Mochel, Luc Moulinier, Arnaud Muller, Jean Muller, et al. PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Research*, 31(13):3829–3832, 2003.

[PCE+12] Marco Punta, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L. L. Sonnhammer, Sean R. Eddy, Alex Bateman, and Robert D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, 2012.

[PFH+12] German Plata, Tobias Fuhrer, Tzu-Lin Hsiao, Uwe Sauer, and Dennis Vitkup. Global probabilistic annotation of metabolic networks enables enzymes discovery. *Nature Chemical Biology*, 8:848–854, 2012.

[PFS+13] Sean Powell, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldón, Thomas Rattei, Chris Creevey, Michael Kuhn, et al. eggNOG v4. 0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, page gkt1253, 2013.

[PHH07] Yungki Park, Sikander Hayat, and Volkhard Helms. Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics*, 8(1):302, 2007.

[PJ01] Marco Pagni and C. Victor Jongeneel. Making sense of score statistics for sequence alignments. *Briefings in Bioinformatics*, 2(1):51–67, 2001.

[PK03] Keun-Joon Park and Minoru Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656–1663, 2003.

[PL04] Jure Piškur and Rikke B. Langkjaer. Yeast genome sequencing: the power of comparative genomics. *Molecular Microbiology*, 53(2):381–389, 2004.

[PLS+04] Christelle Pommié, Séverine Levadoux, Robert Sabatier, Gérard Lefranc, and Marie-Paule Lefranc. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *Journal of Molecular Recognition*, 17(1):17–32, 2004.

[PPS+05] Luca Pireddu, Brett Poulin, Duane Szafron, Paul Lu, and David S. Wishart. The pathway analyst–automated metabolic pathway prediction. In *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Computational Intelligence in Bioinformatics and Computational Biology*, pages 1–8, 2005.

[PRU10] Esa Pitkanen, Juho Rousu, and Esko Ukkonen. Computational methods for metabolic reconstruction. *Current Opinion in Biotechnology*, 21:70–77, 2010.

[PSLG06] Luca Pireddu, Duane Szafron, Paul Lu, and Russel Greiner. The Path–A metabolic pathway prediction web server. *Nucleic Acids Research*, 34:W714–W719, 2006.

[PSM+11] Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Saldatos, Sophie Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData Mining*, 4(10), 2011.

[PSMW05] John W. Pinney, Martin W. Shirley, Glenn A. McConkey, and David R. Westhead. 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, 2005.

[PVL+14] Philipp Paparoditis, Åke Västermark, Andrew J Le, John A Fuerst, and Milton H Saier. Bioinformatic analyses of integral membrane transport proteins encoded within the genome of the planctomycetes species, *Rhodopirellula baltica*. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1838(1):193–215, 2014.

[Ray06] Soumya Raychaudhuri. *Computational Text Analysis for Functional Genomics and Bioinformatics*. Oxford, New York, 2006.

[RB09] Ute Roessner and Jairus Bowne. What is metabolomics all about. *Beyond Darwin: The Future of Molecular Biology*, 46(5):363–365, 2009.

[RCL+14] Abhinay Reddy, Jaehoon Cho, Sam Ling, Vamsee Reddy, Maksim Shlykov, and Milton H Saier. Reliability of nine programs of topological predictions and their application to integral membrane channel and carrier proteins. *Journal of Molecular Microbiology and Biotechnology*, 24(3):161–190, 2014.

[RCP07] Qinghu Ren, Kaixi Chen, and Ian T. Paulsen. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Research*, 35:D274–D279, 2007.

[RGM+12] R Reyes, D Gamermann, Arnau Montagud, D Fuente, J Triana, JF Urchueguia, and P Fernández de Córdoba. Automation on the generation of genome-scale metabolic models. *Journal of Computational Biology*, 19(12):1295–1306, 2012.

[RKP04] Qinghu Ren, Katherine H Kang, and Ian T Paulsen. TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Research*, 32(suppl 1):D284–D288, 2004.

[RLL06] Fernando Rodrigues, Paula Ludovico, and Cecília Leão. Sugar metabolism in yeast: an overview of aerobic and anaerobic glucose metabolism. In *The Yeast Handbook*, pages 101–121. Springer, Berlin, 2006.

[Roe12] Ute Roessner. *Metabolomics*. InTech, Croatia, 2012.

[RP] Reactome Project. Reactome. `http://www.reactome.org/`. Accessed in February, 2014.

[RP05] Qinghu Ren and Ian T. Paulsen. Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLOS Computational Biology*, 1(3):0190–0201, 2005.

[RS12] Vamsee S Reddy and Milton H Saier. BioV Suite — a collection of programs for the study of transport protein evolution. *FEBS Journal*, 279(11):2036–2046, 2012.

[RSSST06] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *The Journal of Machine Learning Research*, 7:1601–1626, 2006.

[Sch03] GE Schulz. Transmembrane beta-barrel proteins. *Advances in Protein Chemistry*, 63:47–70, 2003.

[SCH10] Nadine S Schaadt, Jan Christoph, and Volkhard Helms. Classifying substrate specificities of membrane transporters from *Arabidopsis thaliana*. *Journal of Chemical Information and Modeling*, 50(10):1899–1905, 2010.

[SCM14] Rajib Saha, Anupam Chowdhury, and Costas D Maranas. Recent advances in the reconstruction of metabolic models and integration of omics data. *Current Opinion in Biotechnology*, 29:39–45, 2014.

[SCMD13] Daniela Stojanova, Michelangelo Ceci, Donato Malerba, and Saso Dzeroski. Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. *BMC Bioinformatics*, 14(1):285, 2013.

[SCP+13] Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research*, 41(D1):D764–D772, 2013.

[SH12] NS Schaadt and V Helms. Functional classification of membrane transporters and channels based on filtered TM/non-TM amino acid composition. *Biopolymers*, 97(7):558–567, 2012.

[SHHB09] David Sadava, David M. Hillis, H. Craig Heller, and May Berenbaum. *Life: The Science of Biology*. W. H. Freeman & Co, Gordonsville, Va, 9 edition, 2009.

[SJF11] Carlos N Silla Jr and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72, 2011.

[SJRTV14] MH Saier Jr, VS Reddy, DG Tamang, and A Västermark. The Transporter Classification Database. *Nucleic Acids Research*, 42(Database issue):D251–8, 2014.

[SLBG] Saier Lab Bioinformatics Group. Transporter Classification Database. `http://www.tcdb.org/`. Accessed in February, 2013.

[SPCP10] Jan Schellenberger, Junyoung O. Park, Tom M. Conrad, and Bernhard Ø. Palsson. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(213):1–10, 2010.

[SPGGC+13] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, et al. RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(D1):D203–D213, 2013.

[SQF+11] Jan Schellenberger, Richard Que, Ronan MT Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nature protocols*, 6(9):1290–1307, 2011.

[SRIa] Stanford Research Institute. BioCyc database collection. `http://biocyc.org/`. Accessed in Jan, 2010.

[SRIb] Stanford Research Institute. MetaCyc knowledgebase. `http://metacyc.org/`. Accessed in Jan, 2010.

[SSM10] Hayley J. Sharpe, Tim J. Stevens, and Sean Munro. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell*, 142(1):158–169, 2010.

[SSM+11] Neil Swainston, Kieran Smallbone, Pedro Mendes, Douglas Kell, and Norman Paton. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *J Integr Bioinform*, 8(2):186, 2011.

[Sta14] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[STB05] Milton H. Saier, Jr, Can V. Tran, and Ravi D. Barabote. TCDB: The Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Research*, 34:D181–D186, 2005.

[SUa] Stanford University. Saccharomyces genome database. `http://www.http://www.yeastgenome.org`. Accessed in November, 2012.

[SUb] Stanford University. YeastCyc biochemical pathways. `http://pathway.yeastgenome.org/`. Accessed in February, 2011.

[SUS07] R Sharan, I Ulitsky, and R Shamir. Network-based prediction of protein function,. *Mol Systems Biol*, 3:88, 2007.

[SVS+10] Leander Schietgat, Celine Vens, Jan Struyf, Hendrik Blockeel, Dragi Kocev, and Sašo Džeroski. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics*, 11(1):2, 2010.

[SWD+11] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539, 2011.

[SYC09] Shih-Yi Chao. Graph theory and analysis of biological data in computational biology. In Kankesu Jayanthakumaran, editor, *Advanced Technologies*, chapter 7. Birkhäuser, Austria, 2009.

[SYN+08] Milton H. Saier, Jr, Ming Ren Yen, Keith Noto, Dorjee G. Tamang, and Charles Elkan. The Transporter Classification Database: recent advances. *Nucleic Acids Research*, 37:D274–D278, 2008.

[SZ04] Jibin Sun and An-Ping Zeng. IdentiCS - identification of coding sequence and *in silico* reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics*, 5(112), 2004.

[TFfIoG] The Fellowship for Interpretation of Genomes. The SEED. `http://www.theseed.org`. Accessed in November, 2012.

[The00] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Natural Genetics*, 25(1):25–29, 2000.

[TP10] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010.

[TPR+01] Julie D Thompson, Frédéric Plewniak, Raymond Ripp, Jean-Claude Thierry, and Olivier Poch. Towards a reliable objective function for multiple sequence alignments. *Journal of Molecular Biology*, 314(4):937–951, 2001.

[TS01] Gabor E Tusnady and Istvan Simon. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850, 2001.

[TTP03] Julie D. Thompson, Jean-Claude Thierry, and Olivier Poch. RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, 19(9):1155–1161, 2003.

[VBSE08] Håkan Viklund, Andreas Bernsel, Marcin Skwark, and Arne Elofsson. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, 24(24):2928–2929, 2008.

[VD00] Stijn Van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, The Netherlands, 2000.

[VSS+08] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

[Wal06] Sharon Walker. Biotechnology demystified. *McGraw-Hill*, 2006.

[Wit10] Tobias Wittkop. *Transitivity Clustering: Clustering biological data by unraveling hidden transitive substructures*. PhD thesis, Bielefeld University, Germany, 2010.

[WMC+06] Jianmin Wu, Xizeng Mao, Tao Cai, Jingchu Luo, and Liping Wei. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Research*, 34(suppl 2):W720–W724, 2006.

[WP99] Todd C. Wood and William R. Pearson. Evolution of protein sequences and structures. *Journal of Molecular Biology*, 291(4):977–995, 1999.

[Wu15] Sherry Wu. Comprehensive bioinformatic analysis of glycoside hydrolase family 10 proteins. Master's thesis, Concordia University, Montreal, Canada, January 2015.

[WW99] Stephen H White and William C Wimley. Membrane protein folding and stability: physical principles. *Annual Review of Biophysics and Biomolecular Structure*, 28(1):319–365, 1999.

[XMH+11] Chen Xie, Xizeng Mao, Jiaju Huang, Yang Ding, Jianmin Wu, Shan Dong, Lei Kong, Ge Gao, Chuan-Yun Li, and Liping Wei. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Research*, 39(suppl 2):W316–W322, 2011.

[YOOG05] Yuzhen Ye, Andrei Osterman, Ross Overbeek, and Adam Godzik. Automatic detection of subsystem/pathway variant in genome analysis. *Bioinformatics*, 21:i478–i486, 2005.

[YS12] Jiwon Youm and Milton H Saier. Comparative analyses of transport proteins encoded within the genomes of *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *Biochimica et Biophysica Acta (BBA) — Biomembranes*, 1818(3):776–797, 2012.

[ZSJ01] Yufeng Zhai and Milton H Saier Jr. A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. *Journal of Molecular Microbiology and Biotechnology*, 3(2):285–286, 2001.

[ZSK14] Lingfeng Zhang, Shishir K Shah, and Ioannis A Kakadiaris. Fully associative ensemble learning for hierarchical multi-label classification. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.