TRANSATH: TRANSPORTER PREDICTION VIA ANNOTATION TRANSFER BY HOMOLOGY

Faizah Aplop[1] and Greg Butler[2]
[1]School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, Malaysia
[2] Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
E-Mail: faizah_aplop@umt.edu.my

ABSTRACT

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. This paper presents TransATH, which stands for Transporters via ATH (Annotation Transfer by Homology), a system which automates Saiers 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 performance as defined by a gold standard set of transporters and non-transporters from *S. cerevisiae*. We demonstrate 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.

Keywords: transmembrane transport proteins, annotation transfer by homology, subcellular localization.

INTRODUCTION

This paper deals with computational aspects of the automatic reconstruction of the metabolic pathways of an organism that relate to transport of compounds across membranes. We investigate how to include transport reactions, transporter proteins, and the GPR (Gene-Protein-Reaction) 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.

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.

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.

BACKGROUND

An organism carries out a range of processes, such as

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

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 [1]. The membrane is a phospholipid bilayer. 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 IMPs 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.

Structurally, the eukaryote transmembrane proteins have α-helices that span the membrane [2]. In gram-negative bacteria, there are transmembrane strand proteins that span the membrane with β-strands [3]. These are called transmembrane segments (TMS). 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 [4]. Transporters constitute up to 30% of all cellular proteins [5], and they play important roles in cellular metabolism [6]. Transporters have a high degree of substrate specificity and bind to one or a few substrate molecules [7]. The different forms of molecule transport are (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 [1].

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 Transport Classification scheme, or the substrate category. It would be useful in these three approaches were cross-referenced with each other, and with the protein domains [8]. Here we briefly overview the three schemes.

The International Union of Biochemistry and Molecular Biology (IUBMB) introduced the Transporter Classification System (TC) [9] in June 2001 for classifying membrane transport proteins. Analogous to EC numbers for classifying enzymes, a TC identifier such as TC 2.A.1.1.35 has five components representing

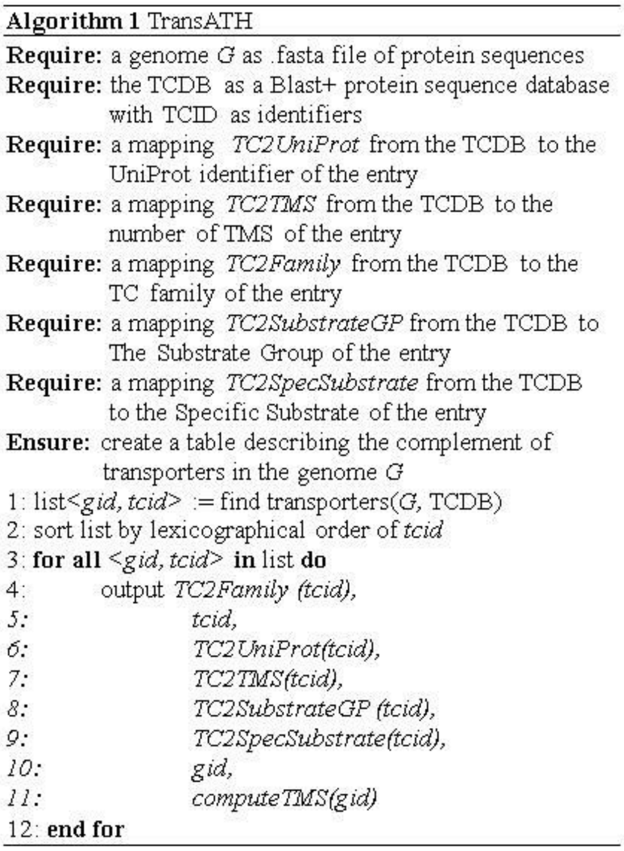
1. the transporter class (TC-class), eg 2;
2. the transporter subclass (TC-Subclass), eg 2.A;
3. the transporter family (TC-Family), eg 2.A.1, which in some cases is a superfamily;
4. the transporter subfamily, eg 2.A.1.1; and
5. the specific transporter (TC-ID), eg 2.A.1.1.35.

A superfamily is a large divergent family, in which cases the distant clades are considered families within the larger superfamily. 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) [10]. 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 [11].

TransATH system

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 [12] 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).

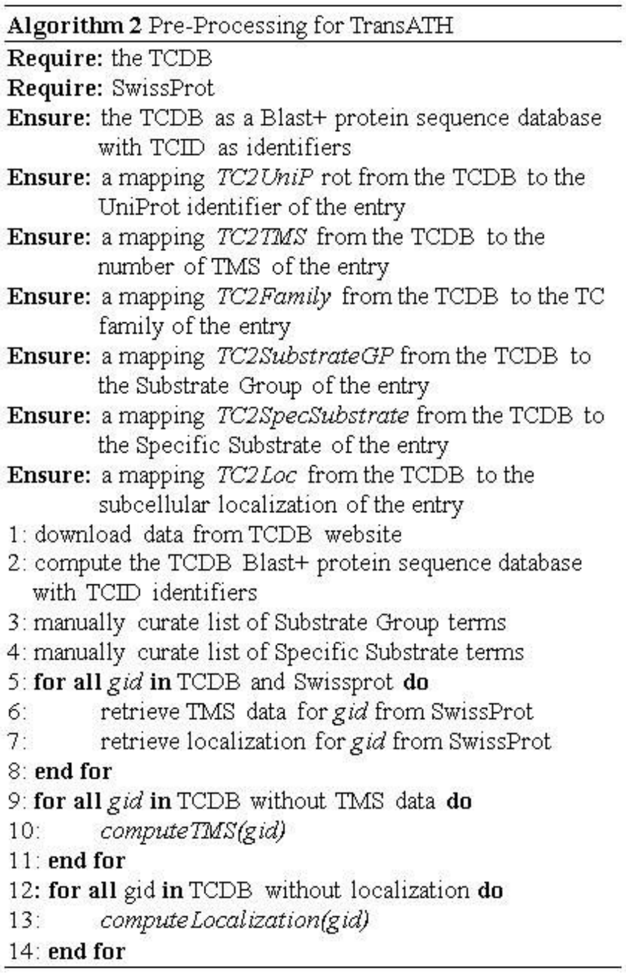
Algorithm 1 presents the TransATH algorithm for the implementation of the protocol of Saier's lab for determining the transporters in a given genome. TransATH stands for Transporters via ATH (Annotation Transfer by Homology). Note that Algorithm 1 requires several items of information to be provided from the TCDB and this pre-processing is presented in Algorithm 2. 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 3 presents a utility function find transporters which calls TCDB-Blast, the BLAST search at the heart of TransATH. Algorithm 4 shows TCDB-Blast, the BLAST search of the TCDB using our choice of thresholds. Algorithm 5 shows the algorithm to determine the topology of a protein, and Algorithm 6 shows the algorithm to determine subcellular localization.



```
Algorithm 1 TransATH
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: create 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
```

We modified G-Blast(v2), the second version of the GBlast 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 an evaluation. Algorithm 4 shows the main step of the algorithm for the BLAST search of the TCDB.

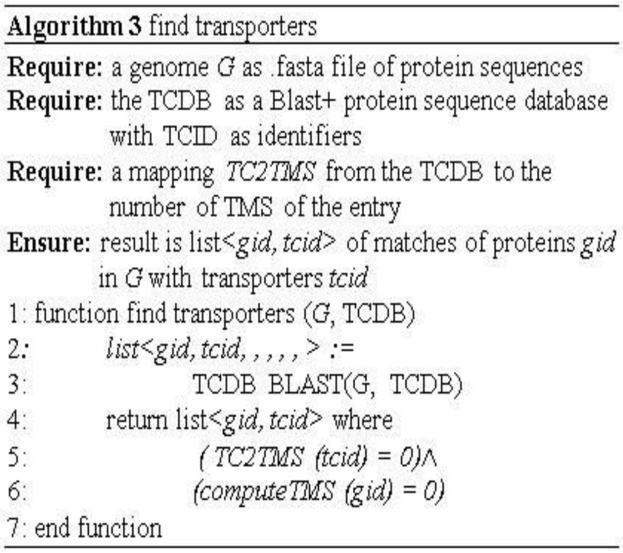
There are several programs for predicting the topology of membrane proteins. Topology is widely predicted using TMHMM. In a comparison of nine programs on four TC families [13], HMMTOP [14] is overall the best, performing best for the sugar porters, and performing well for the other families. Also performing well were MEMSAT-SVM [15] and SPOCTOPUS [16]. Note that Saier's protocol [17] manually considers hydropathy plots using WHAT [18] to correct HMMTOP predictions. The term hydropathy, which means "strong feeling about water", is introduced by Kyte and Doolittle [19] 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. The hydrophobic moment plot of Eisenberg and co-workers [20, 21] is a similar tool used in the UniProt protocol (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.



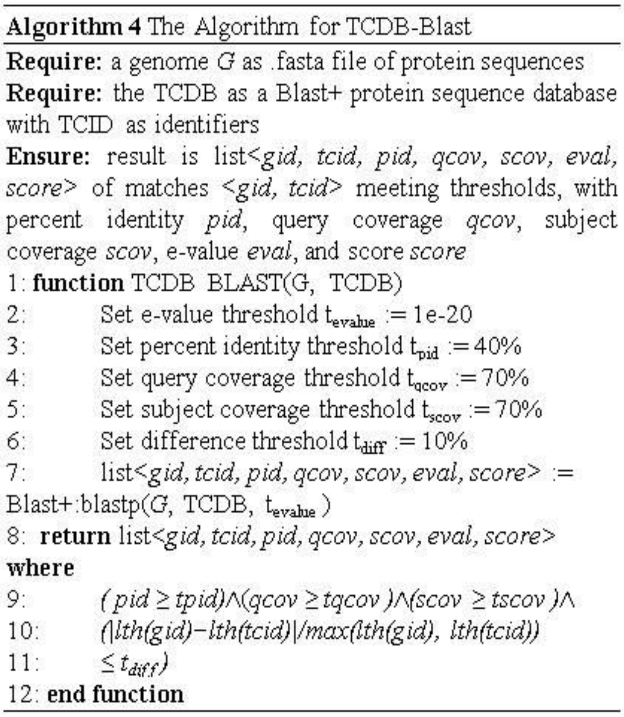
```
Algorithm 2 Pre-Processing for TransATH
Require: the TCDB
Require: SwissProt
Ensure: the TCDB as a Blast+ protein sequence database
        with TCID as identifiers
Ensure: a mapping TC2UniP rot from the TCDB to the
        UniProt identifier of the entry
Ensure: a mapping TC2TMS from the TCDB to 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 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 localization do
        computeLocalization(gid)
end for
```

Our implementation relies on TM-Coffee [22] which computes MSA of transmembrane proteins, to determine the alignment of the TMS regions of the query protein sequence with the TMS regions of the entry in TCDB. This approach uses the transmembrane proteins in SwissProt as further entries in the MSA. Algorithm 5 shows our implementation to determine the topology of a protein.

A widely used tool for subcellular localization in fungi is WoLF PSORT [23]. 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.



```
Algorithm 3 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
```



```
Algorithm 4 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
function TCDB BLAST(G, TCDB)
        Set e-value threshold $t_{evalue}$ := 1e-20
        Set percent identity threshold $t_{pid}$ := 40%
        Set query coverage threshold $t_{qcov}$ := 70%
        Set subject coverage threshold $t_{scov}$ := 70%
        Set difference threshold $t_{diff}$ := 10%
        list<gid, tcid, pid, qcov, scov, eval, score> :=
Blast+:blastp(G, TCDB, $t_{evalue}$ )
  return list<gid, tcid, pid, qcov, scov, eval, score>
where
        $( pid \geq tpid)\wedge(qcov \geq tqcov )\wedge(scov \geq tscov )\wedge$
        $(|lth(gid)-lth(tcid)|/max(lth(gid), lth(tcid))$
        $\leq t_{diff})$
end function
```

Algorithm 5 computeTMS function for Topology

```
Require: a protein sequence gid
Ensure: result is <num, topology> of the number and topology of TMS of gid
function computeTMS(gid)
    <num, topology> := HMMTOP(gid)
    msa := TM-Coffee(gid, SwissProt )
    adjust list<num, topology> based on msa
    return list<num, topology>
end function
```

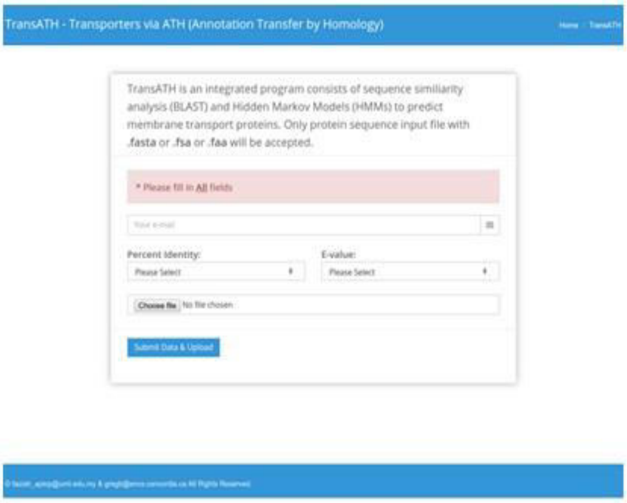
A tool for localization prediction that has a comprehensive treatment of placing proteins in membranes of organelles is LocTree3 [24]. LocTree3 targets 18 sites, including 8 membranes: plasma membrane, nuclear membrane, mitochondrion membrane, ER membrane, Golgi membrane, vacuole membrane, peroxisome membrane, and chloroplast membrane. LocTree3 achieves an overall accuracy of 80%. Furthermore, in the experimental comparison [24], LocTree3 is shown to be superior to existing tools, including WoLF PSORT.

Algorithm 6 computeLocalization function

```
Require: a transmembrane protein sequence gid
Ensure: result is the localization of protein gid
function computeLocalization(gid)
    return LocTree3(gid)
end function
```

We have 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.

The beta version of TransATH is publicly available and can be accessed at http://transath.umt.edu.my. Figure-1 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. While for e-value thresholds, there are six choices: 10, e-5, e-10, e-20, e-30 and e-50.


TransATH - Transporters via ATH (Annotation Transfer by Homology)
Home / TransATH

TransATH is an integrated program consists of sequence similiarity analysis (BLAST) and Hidden Markov Models (HMMs) to predict membrane transport proteins. Only protein sequence input file with .fasta or .fsa or .faa will be accepted.

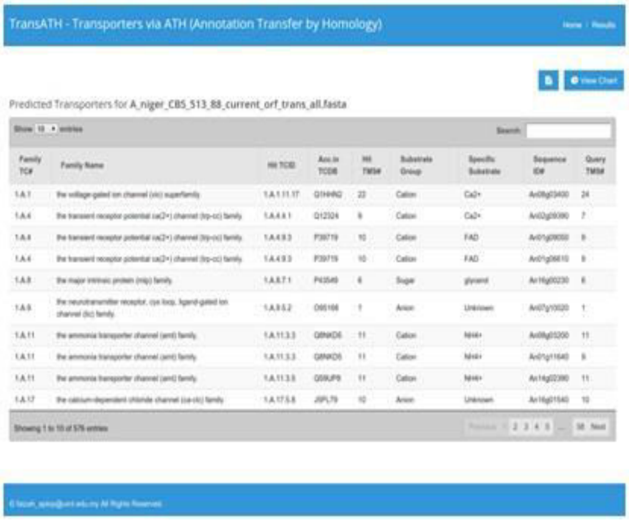
* Please fill in All fields

Your e-mail
Percent Identity:
Please Select
E-value:
Please Select
Choose file
No file chosen
Submit Data & Upload
© faizah_aziz@umt.edu.my & gregb@encs.concordia.ca All Rights Reserved

Figure-1. 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-2 shows an example of an output page that displays a table of predicted transporters imitating the result by Saier [17, 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-3 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 will display the substrate group name with its percentage of the total.


TransATH - Transporters via ATH (Annotation Transfer by Homology)
Home / Results
View Chart

Predicted Transporters for A_niger_CBS_513_88_current_orf_trans_all.fasta

Family TC#	Family Name	Hit TCID	Acc in TCDB	Hit TMS#	Substrate Group	Specific Substrate	Sequence ID#	Query TMS#
1.A.1	the voltage-gated ion channel (vic) superfamily	1.A.1.11.17	Q1HHN2	22	Cation	Ca2+	An08g03400	24
1.A.4	the transient receptor potential ca(2+) channel (trp-cc) family	1.A.4.4.1	Q12324	6	Cation	Ca2+	An02g09090	7
1.A.4	the transient receptor potential ca(2+) channel (trp-cc) family	1.A.4.9.3	P39719	10	Cation	FAD	An01g09050	8
1.A.4	the transient receptor potential ca(2+) channel (trp-cc) family	1.A.4.9.3	P39719	10	Cation	FAD	An01g06610	8
1.A.8	the major intrinsic protein (mip) family	1.A.8.7.1	P43549	6	Sugar	glycerol	An16g00230	6
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.11.3.3	Q8NKD5	11	Cation	NH4+	An08g05200	11
1.A.11	the ammonia transporter channel (amt) family	1.A.11.3.3	Q8NKD5	11	Cation	NH4+	An01g11640	8
1.A.11	the ammonia transporter channel (amt) family	1.A.11.3.5	Q59UP8	11	Cation	NH4+	An14g02390	11
1.A.17	the calcium-dependent chloride channel (ca-clc) family	1.A.17.5.8	J9PL79	10	Anion	Unknown	An16g01540	10

Showing 1 to 10 of 576 entries

Previous 1 2 3 4 5 … 58 Next
© faizah_aziz@umt.edu.my All Rights Reserved

Figure-2. Page of results of TransATH for A.niger CBS513.88

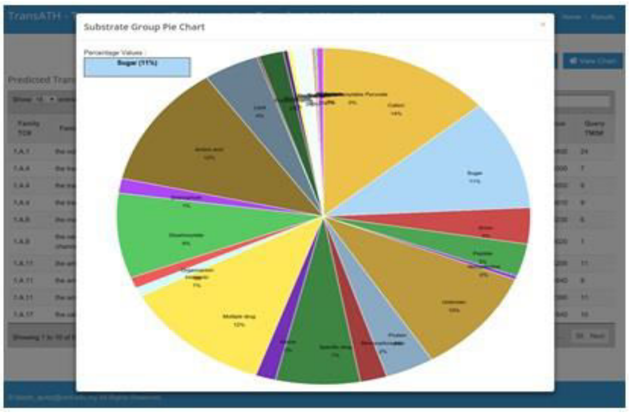


Figure-3. Pie chart of TransATH predictions for A.niger CBS513.88.

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 pre-processing was done manually for the beta implementation of TransATH. In future we will extract the roughly 4000 entries available in merlin [25] 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.

SETTING PARAMETER THRESHOLDS

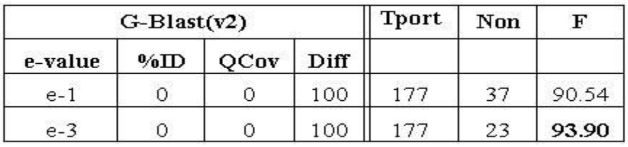
For the evaluation we took the gold standard dataset used by [12, Table S3] of 177 transporters in S. *cerevisiae* that have been experimentally characterized. These were the positive examples in the dataset. A set for negative examples of size 177 was chosen at random from S. *cerevisiae* at SGD (http://www.yeastgenome.org) 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.

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-1 and Table-2 show 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-1 and Table-2 compare G-Blast(v2) and TCDB-Blast. Table-2 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).

Table-1. F-measures for G-blast(v2) predictions for combinations of thresholds.



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

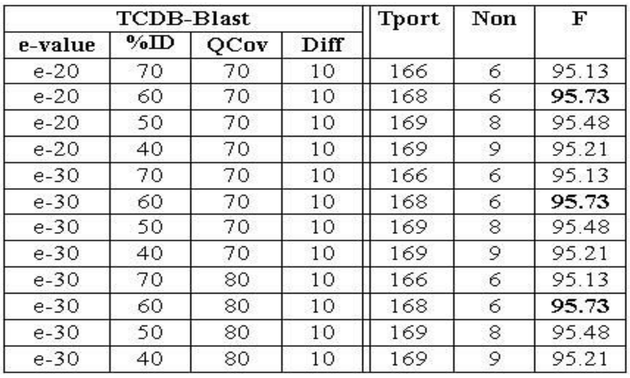
The number of results when using blastp to search the 354 protein sequences of the gold standard dataset consisting of 177 transporters and 177 non-transporters against the 11,572 entries of the TCDB as of May 2014 with different combination of thresholds. In this trial neither G-Blast(v2) nor TCDB-Blast 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 explicitly 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.

TransATH CORRECTNESS

The methodology used to determine the correctness of the predictions by TransATH on *A. niger* CBS513.88 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 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.

Table-2. F-measures for prediction using combinations of thresholds.



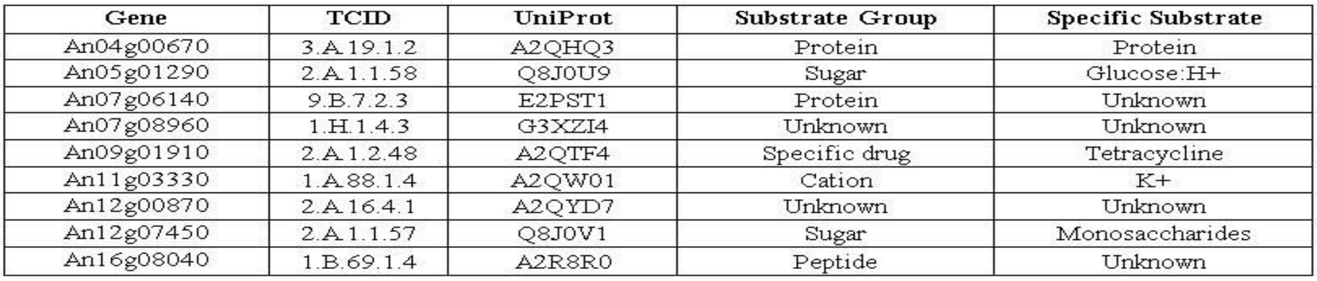
TCDB-Blast				Tport	Non	F
e-value	%ID	QCov	Diff			
e-20	70	70	10	166	6	95.13
e-20	60	70	10	168	6	**95.73**
e-20	50	70	10	169	8	95.48
e-20	40	70	10	169	9	95.21
e-30	70	70	10	166	6	95.13
e-30	60	70	10	168	6	**95.73**
e-30	50	70	10	169	8	95.48
e-30	40	70	10	169	9	95.21
e-30	70	80	10	166	6	95.13
e-30	60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

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.

The TCDB as of May 2014 has 9 entries from A. *niger* CBS 513.88 as shown in Table 3. As mentioned before, TCDB is a curated database that incorporates functional and phylogenetic information of membrane transport proteins, which are organized according to TC systems. The information in TCDB gathered from many species, domains, kingdoms or phylum. In our study, the 9 entries mentioned in Table-3 representing 7 different TC-superfamilies were identified belong to A. *niger* CBS 513.88 of AspGD.

According to the rules of TC systems, the last digit of TCID represents the substrates or range of substrates being transported. The last two columns of Table-3 display substrates information. Some might be putative, too vague or general, or even unknown. This kind of information is only stored as unknown in TCDB.

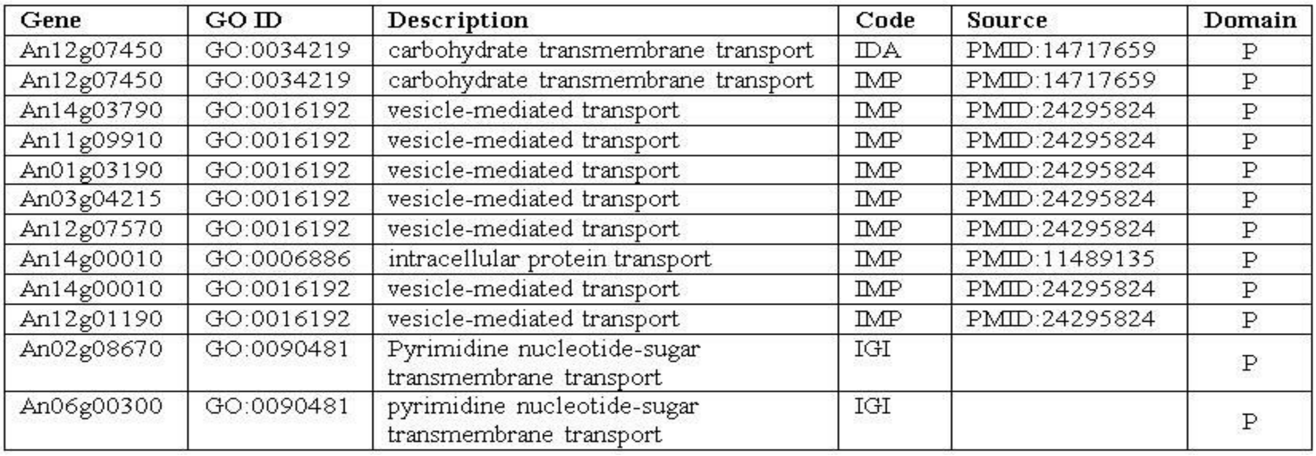
Table-3. TCDB entries from A. *niger* CBS 513.88.



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

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.

Table-4. Transport GO entries with experimental evidence for A. *niger* CBS 513.88.



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

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.

The high confidence AspGD annotations for transporters were determined by downloading the gene association.aspgd file from the AspGD web-site at http://www.aspgd.org. The entries pertaining to *A. niger* CBS 513.88 were extracted and cross-referenced with the set 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 4 shows the information for the 10 genes with experimental evidence.

From the total 242 GO terms for 190 genes only a few include detail about the substrate being transported. 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. For the evaluation TransATH was run at *http://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.

For the 30 genes with information on the substrate transported, TransATH returned predictions for 11of the 30 genes. Another 9 of the 30 genes had blastp hits to TCDB sequences that fell below the thresh-olds set for this evaluation. For 9 of the 11 genes with predictions from TransATH there is agreement on the substrate transported, while for the other two (An05g01660 and An15g02930) there is agreement at the Substrate Group level.

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 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.

CONCLUSIONS

This paper presents TransATH, a system which automates Saiers protocol and includes the computation of subcellular localization and improves the computation of transmembrane segments. TransATH predicts transporters via annotation transfer by homology using the TCDB database of known transporters.

To determine optimal performance as defined by a gold standard set of transporters and non-transporters from *S. cerevisiae,* the choice of thresholds for the parameters of TransATH is investigated. We demonstrate 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.

ACKNOWLEDGEMENTS

This work was done at Concordia University as part of the PhD thesis of Faizah Aplop under the supervision of Greg Butler. This work was funded in part by Universiti Malaysia Terengganu, Malaysia and the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] F. A. Kuypers, "Cell membranes," in Medical Cell Biology, 3rd ed., S. R. Goodman, Ed. Amsterdam Boston: Elsevier/Academic Press, 2008, ch. 2, pp. 27–57.

[2] S. H. White and W. C. Wimley, "Membrane protein folding and stability: physical principles," Annual Review of Biophysics and Biomolecular Structure, vol. 28, no. 1, pp. 319–365, 1999.

[3] G. Schulz, "Transmembrane beta-barrel proteins," Advances in Protein Chemistry, vol. 63, pp. 47–70, 2003.

[4] D. Sadava, D. M. Hillis, H. C. Heller, and M. Berenbaum, Life: The Science of Biology, 9th ed. Gordonsville, Va: W. H. Freeman & Co, 2009.

[5] H. J. Sharpe, T. J. Stevens, and S. Munro, “A comprehensive comparison of transmembrane domains reveals organelle-specific properties,” Cell, vol. 142, no. 1, pp. 158–169, 2010.

[6] Q. Ren and I. T. Paulsen, “Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes,” PLOS Computational Biology, vol. 1, no. 3, pp. 0190–0201, 2005.

[7] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, “Proteinstructure and function,” in Molecular Cell Biology, 4th ed. New York: W.H. Freeman & Co, 2000, ch. 3, pp. 78–83.

[8] Z. Chiang, A. Vastermark, M. Punta, P. C. Coggill, J. Mistry, R. D. Finn, and M. H. Saier, “The complexity, challenges and benefits of comparing two transporter classification systems in TCDB and Pfam,” Briefings in Bioinformatics, p. bbu053, 2015.

[9] W. Busch and M. H. Saier, Jr, “The IUBMB-Endorsed Transporter Classification System,” Molecular Biotechnology, vol. 27, pp. 253–262, 2004.

[10] M. H. Saier, Jr, M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan, “The Transporter Classification Database: recent advances,” Nucleic Acids Research, vol. 37, pp. D274–D278, 2008.

[11] M. Saier Jr, V. Reddy, D. Tamang, and A. V¨

[12] astermark, “The Transporter Classification Database.” Nucleic Acids Research, vol. 42, no. Database issue, pp. D251–8, 2014.

A. Barghash and V. Helms, “Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs,” BMC Bioinformatics, vol. 14, no. 1, p. 343, 2013.

[13] Reddy, J. Cho, S. Ling, V. Reddy, M. Shlykov, and M. 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, vol. 24, no. 3, pp. 161–190, 2014.

[14] G. E. Tusnady and I. Simon, “The HMMTOP transmembrane topology prediction server,” Bioinformatics, vol. 17, no. 9, pp. 849–850, 2001.

[15] T. Nugent and D. T. Jones, “Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm,” PLoS Computational Biology, vol. 6, no. 3, p.e1000714, 2010.

[16] H. Viklund, A. Bernsel, M. Skwark, and A. Elofsson, “SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology,” Bioinformatics, vol. 24, no. 24, pp. 2928–2929, 2008.

[17] P. Paparoditis, A. V¨astermark, A. J. Le, J. A Fuerst, and M. 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, vol. 1838, no. 1, pp. 193–215, 2014.

[18] Y. Zhai and M. 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, vol. 3, no. 2, pp. 285–286, 2001.

[19] J. Kyte and R. F. Doolittle, “A simple method for displaying the hydropathic character of a protein,” Journal of molecular biology, vol. 157, no. 1, pp. 105–132, 1982.

[20] D. Eisenberg, R. Weiss, and T. Terwilliger, “The helical hydrophobic moment: a measure of the amphiphilicity of a helix.” Nature, vol. 299, no. 5881, pp. 371–374, 1982.

[21] 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, vol. 179, no. 1, pp. 125–142, 1984.

[22] J.-M. Chang, P. Di Tommaso, J.-F. Taly, and C. Notredame, “Accurate multiple sequence alignment of transmembrane proteins with psicoffee,” BMC Bioinformatics, vol. 13, no. 4, p. 1, 2012.

[23] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. Adams-Collier, and K. Nakai, “WoLF PSORT: protein localization predictor,” Nucleic Acids Research, vol. 35, no. suppl 2, pp. W585–W587, 2007.

[24] T. Goldberg, M. Hecht, T. Hamp, T. Karl,G. Yachdav, N. Ahmed, U. Altermann, P. Angerer, S. Ansorge, K. Balasz *et al.*, “LocTree3 prediction of localization,” Nucleic Acids Research, vol. 42, no. W1, pp. W350–W355, 2014.

[25] O. Dias, M. Rocha, E. C. Ferreira, and I. Rocha,“Reconstructing genome-scale metabolic models with merlin,” Nucleic Acids Research, p. gkv294, 2015.