Notes on the development of an automated predictor for the Transporter Classification Data-Base (TCDB)

June 20, 2015

Contents

1 Classification criteria underlying the TCDB

The TCDB website details a comprehensive classification system for membrane transport proteins known as the Transport Classification (TC) system. The TC system is analogous to the Enzyme Commission (EC) system for classification of enzymes, except that it incorporates both functional and phylogenetic information [3, 2]. Basically the taxonomy is based on the mode of action of the transport activity and the energy coupling mechanism used for the transport. Phylogenetic grouping reflects structure, function, mechanism, and often substrate specificity and therefore provides a reliable secondary basis for classification. Finally substrate specificity and polarity of transport provide a tertiary basis for classification.

Schematically, the basic criteria of classification proposed in the TCDB are the following [1]:

i. transport mode

ii. energy coupling mechanism

iii. phylogenetic grouping

iv. substrates transported

These criteria of classification are reflected in the 5-tier taxonomy, coded by 5 point separated digits V.W.X.Y.Z:

1. W (a number): the first level of the hierarchy, i.e. the class of the transport protein. It corresponds to the most general classes of the taxonomy.
2. V (a letter): the second level of the hierarchy, i.e. the subclass of the transport protein.
3. X (a number): the third level of the hierarchy, i.e. the family (sometimes the superfamily) of the transport protein.
4. Y (a number): the fourth level of the hierarchy, i.e. the subfamily
5. Z (a number): the fifth level of the hierarchy is mostly related to the substrate(s) on which the transport proteins acts.

1.1 Structure of TCDB

TCDB is a 5-tier taxonomy of transport proteins structured according to a tree. We have the following per-level distribution of the 12587 classes (from top to bottom):

1. level 1: 7
2. level 2: 30
3. level 3: 867

4. level 4: 2235
5. level 5: 9448

Table 1: Distribution of the cardinality of the classes. The number of annotated proteins is 12508 belonging to at least 2574 different species.

Number of annotations	**Number of classes**
1	9173
> 1	3414
> 5	1245
> 10	597
> 20	232
> 50	67
> 100	37
> 200	24
> 500	12
> 1000	10

Table 2: Per-level distribution of the per-class TCDB annotations

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Level 1	131	318	1391	1788	3408	3542
Level 2	1	19	56	417	426	3416
Level 3	1	2	6	14.4	12	1661
Level 4	1	1	3	5.6	6	196
Level 5	1	1	1	1.3	1	45

1.2 The levels of the TCDB hierarchy

The different levels articulates the function and the phylogenetic characeristics of the transport proteins across all the living organisms.

1.2.1 The class level

It represents the highest level of classification reflecting the general transport mode as well as the way in which energy is exploited to perform the transport (Tab. 3). Fig. 3 provides a summary of the cardinality of the 7 protein transport categories defined at this level (note that classes 6 and 7 are left undefined for possible future rearrangements of the hierarchy).

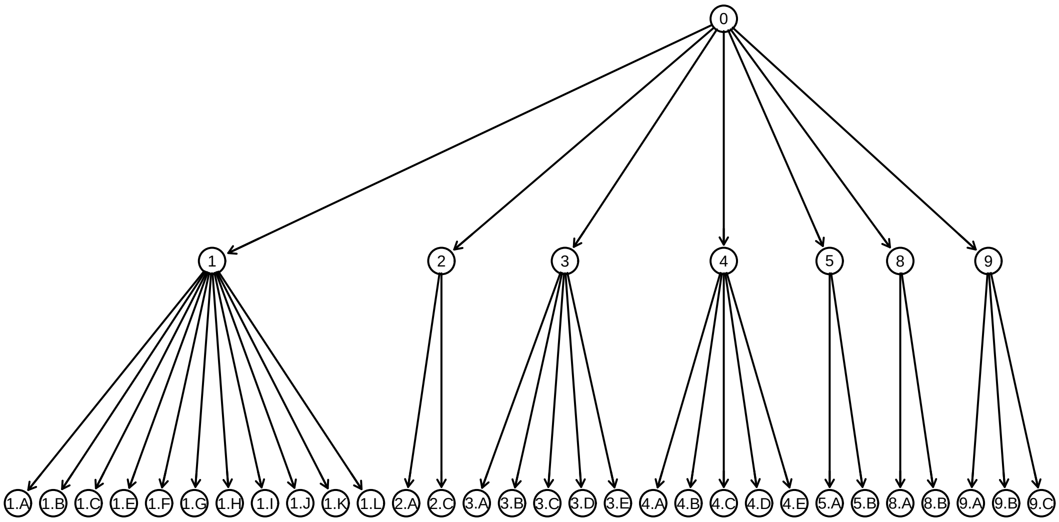

Figure 1: The first two levels of the TDCB taxonomy. 0 is a dummy added node to obtain a tree instead of a forest.

1.2.2 The subclass level

This level provides a further specification of the main mode of action and energy mechanism underlying the transporter. For instance in the case of primary active transporters(class 3) refers to the energy source used to drive transport, i.e. chemical (3.A P-P bond hydrolisis, 3.B decarboxylation-driven, 3.C methyltransfer-drive), electrical (3.D oxidoreduction-driven), and solar (3.E light absorption-driven), or in the case of Class 1 (channels and pores) fundamental structural differences lead to different subclasses such as proteins characterized by α-elical portions, ubiquitously found i all organisms(1.A α- Type channels) and proteins whose transmembrane potions are characterized by β-strands found usually in Bacteria (1.B *beta*-Barrel porins).

1.2.3 The family level

A phylogenetic family of transporters includes members that function by a single transport mode and energy coupling mechanism, although a variety of substrates may be transported [1]. It should be noted that this level corresponds

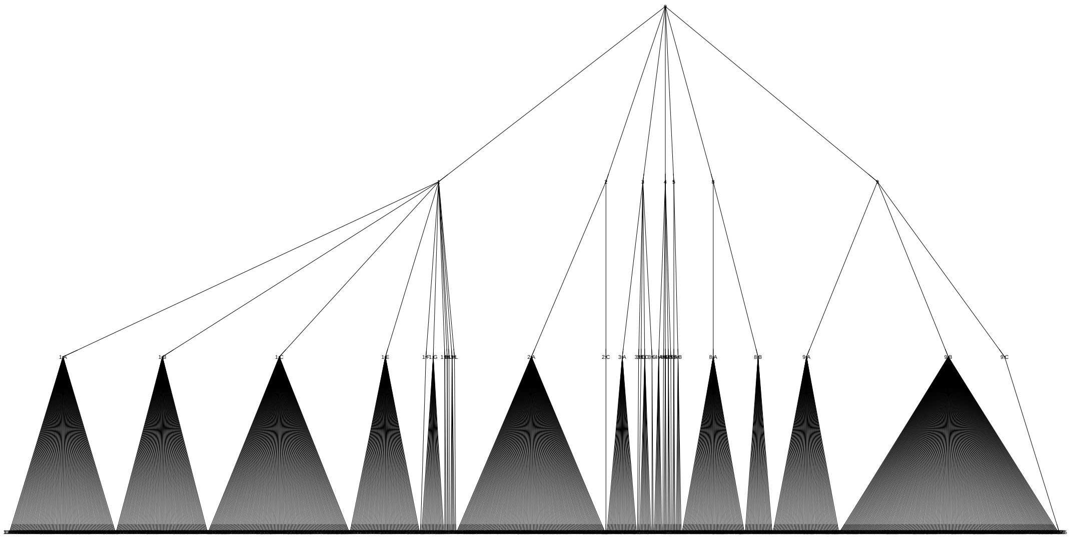

Figure 2: The first three levels of the TDCB taxonomy

to phylogenetically related proteins, whose evolution reflects common modes of transports and common ways to exploit energy to realize transport functions. Families are based on (limited) sequence or motif similarity, similar protein size, topology and structure (when available).

Some of these families are actually large superfamilies with more than a thousand currently sequenced members (e.g., the voltage-gated ion channel (VIC) family – TC 1.A.1) Others are very small families with only one or a few currently sequenced members.

1.2.4 The subfamily level

This level corresponds to a further phylogenetic separation inside families that may correspond also to slightly different functions or actions on different substrates. In some cases they could be also families, when the previous level corresponds to a superfamily. We could try to predict till to this level.

Table 3: Characteristics of the most general classes (first level) of the hierarchy

t **Code**	**Name of the class**	**Short description**
1	Channels/Pores	catalyze facilitated diffusion (by an energy-independent process) by passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism
2	Electrochemical Potential driven Transporters	utilize a carrier-mediated process not directly linked to a form of energy other than chemiosmotic energy
3	Primary Active Transporters	use a primary source of energy (chemical, electrical or solar) to drive active transport of a solute against a concentration gradient
4	Group Translocators	involves a combined chemical and vectorial reaction where the transported substrate is modified during the transport process
5	Transmembrane Electron Carriers	systems that catalyze electron flow across a biological membrane, from donors to acceptors
8	Accessory Factors Involved in Transport	proteins that in some way facilitate transport but do not participate directly in transport
9	Incompletely Characterized Transport Systems	Transporters of unknown classification

1.2.5 The most specific level

This level delineates the substrate or range of substrates transported as well as the polarity of transport. Any two proteins in the same subfamily that transport the same substrate(s) using the same mechanism are given the same TC number, regardless of whether they are orthologs (i.e., arose in distinct organisms by speciation) or paralogs (i.e., arose within a single organism by gene duplication). The predictions at this level seem quite difficult since we have usually only 1 protein annotated at this level, and when we have more proteins annotated for a class as this level, this means that they belong to the same transport system or a complex exploiting a specific transport function with a specific mode of action and energy coupling and on specific substrates.

1.3 Remarks on the TCDB annotations

1.3.1 Meaning of the multiple annotations for the most specific classes.

At level 5 the class having 45 annotations is 3.D.1.6.1: The animal H+-translocating NADH dehydrogenase (NDH) complex (Table 2). This is a complex of the mi-

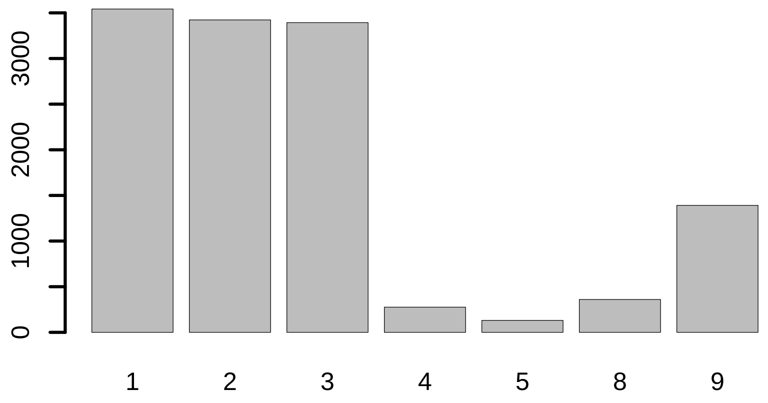

Figure 3: Cardinality of the most general classes (Level 1)

tochondrial inner membrane involved in the respiratory chain. This complex is composed by 45 subunits and the 45 annotations correspond exactly to the 45 protein subunits.

This is true also for 3.D.1.6.2: The fungal H+ translocating NADH dehydrogenase (NDH) complex having 31 subunits and 31 annotations, for 3.D.1.6.2: The green algal H+ translocating NADH dehydrogenase (NDH) complex having 33 subunits and 33 annotations, and for 1.I.1.1.1: The Nuclear Pore Complex (NPC) in Yeast, a transport system with 35 components corresponding to 35 different proteins.

More in general, of the 9845 classes at the fifth level, 985 have more than 1 annotation. I checked by sampling 3.A.1.135.5: The heterodimeric ABC transporter that has 2 annotations corresponding to the 2 components systems (Thermatoga maritima), belonging to the the ATP-binding Cassette (ABC) Superfamily. The 1.F.1.1.1 is a 10 component system: The Synaptosomal Vesicle Fusion Pore (SVF-Pore) having just 10 annotations.

Hence the classes at the fifth level having more than 1 annotation are likely those composed by different subunits/components included in the same transport system. This could be checked more in detail but I guess that this is the situation

1.3.2 Proteins annotated to more than 1 leaf of the TCDB

Of the 12508 annotated proteins, only 28 have more than 5 annotations (corresponding to an unique path from the root to a leaf of the TCDB tree). Most of them differs only at the fifth level (e.g. D4ZJA6 Sodium-type flagellar protein MotY and Q8EAG6 Sodium-type flagellar protein MotX are both part of the complex 1.A.30.1.5: The H+-driven flagellar motor complexT and 1.A.30.1.5: The Na+-driven flagellar motor complex and for these reason they have a double annotation). In other case the annotation can be different also at subfamily level (e.g. Q03PY- Energy-coupling factor transporter ATP-binding protein EcfA1) annotated to both 3.A.1.28.2 and 3.A.1.26.9 since these different Folate-tranporters use the same energy coupling factor), or also at class level (e.g. O24303 annotated to 1.A.18.1.1 Protein import-related anion-selective channel but also as component of the complex 3.A.9.1.1 Chloroplast envelope protein translocase (CEPT)).

In any case considering that we have only 28 proteins annotated to 2 different leaves (mostly strictly related) we could consider predictions of annotations along single paths of the TCDB tree.

2 The data

An R function is available to automatically parse any tcdb text file to automatically extract all the available fields (SwissProt AC, TCDB code, description of the protein, gene name, etc). A library of R function is available for parsing the tcdb text files, construct annotations table, construct the taxonomy tree, providing some basic statistics on the TCDB annotations.

2.1 Taxonomy data

Data of the TCDB tree are just available as graphNEL R classes or as edges in plain text files.

2.2 The annotation data

These data are just available in tabular form in both .rda R compressed files and plain text files.

2.3 The input feature data

2.3.1 BLAST-based data

BLAST all vs all with TCDB proteins –¿ symmetric similarity matrix (data just available)

For unannotated proteins BLAST against all the TCDB proteins to obtain a feature vector.

2.3.2 InterPro feature data

This data for SwissProt proteins can be extracted for UniProtKB text files.

We need to choose which features should be included.

2.3.3 TransMembrane Segments (TMS) data

Available in some way from the TCDB (unfortunately there are not text files available but only data on a per-protein basis on the TCDB web-site). TMS data could be extracted from FASTA sequences using the HMMTOP program or other similar programs. For a given protein we have the number of TMSs: this simple monodimensional data is useful since it can characterize classes at high level. For instance

2.3.4 PDB-based data

We need to extract proper features from the available data (Su and Jim could you provide them?).

2.3.5 Taxonomic data

Considering that this taxonomy is constructed on functional, but also on phylogenetic bases, I guess that this information about the taxa will be useful. At family level some classes are characteristic or exclusive for Bacteria or Eukarya, and at subfamily or at the most specific level we are very close to the species. We should decide which level of detail could be useful for predictions at family or subfamily level. For instance in [1] the categories of organisms belonging to each specific class are details at a very coarse level: Bacteria, Archaea, Eukarya, Fungi, Protozoans, Plants, Animals, etc.

2.3.6 Other features

The size range (number of residues) seem to characterize different families. Hence a simple feature that could be added is for instance a binary vector with entries corresponding to the number of residues: the first entry is set to 1 if we have less than 100 residues, the second if the residue are between 100 and 200 and so on. There are problems with dimeric or multi-meric proteins (that is complexes having more than 1 subunit), that may have very large number of residues (even 6000): we could provide a further feature collection the number of subunits o the transporter system (when o curse this data is available ...).

Another feature is the substrate on which a transporter acts. Of course this is a quite unlikely available information for uncharacterized transport proteins, but if available could be useful. To this end we need to characterize the substrates, using, for instance, the simple taxonomy provided in Table 4 of [1].

3 Methods

3.1 Proposed methods

The taxonomy follows a tree-structure, according to the TCDB. Each annotation is articulated along a single path from the root to the leaves: in practice usually each annotated protein belongs to 5 classes, by following a unique path from the root to a specific leaf at the fifth level of the hierarchy

We could work on three (partially related) research lines:

1. Tree structured output methods
2. Hierarchical ensemble methods
3. Semi-supervised flat network-based methods (e.g. RANKS).

3.2 Related work

TO DO

3.3 Baseline methods

Surely we should use BLAST methods (e.g. using the best-hit approach against the TCDB, in terms of score or p-value). Considering the way the TCDB i constructed, I guess that BLAST will be a strong baseline. As baseline we could also use some flat machine learning-based methods. Flat machine learning-based methods could be also used as base learners with two-steps Hierarchical ensemble methods.

References

[1] M.H. Saier. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, 64(2):354–411, 2000.

[2] Milton H. Saier, Vamsee S. Reddy, Dorjee G. Tamang, and Ake Vastermark. The transporter classification database. *Nucleic Acids Research*, 42(D1):D251–D258, 2014.

[3] Milton H. Saier, Can V. Tran, and Ravi D. Barabote. Tcdb: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Research*, 34(suppl 1):D181–D186, 2006.