BIOINFORMATICS

Vol. 24 ISMB 2008, pages i259–i267
doi:10.1093/bioinformatics/btn180

Annotation-based inference of transporter function

Thomas J. Lee[1,*], Ian Paulsen[2] and Peter Karp[1]

[1]Artificial Intelligence Center, SRI International, Menlo Park, CA, USA and [2]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW, Australia

ABSTRACT

Motivation: We present a method for inferring and constructing transport reactions for transporter proteins based primarily on the analysis of the names of individual proteins in the genome annotation of an organism. Transport reactions are declarative descriptions of transporter activities, and thus can be manipulated computationally, unlike free-text protein names. Once transporter activities are encoded as transport reactions, a number of computational analyses are possible including database queries by transporter activity; inclusion of transporters into an automatically generated metabolic-map diagram that can be painted with omics data to aid in their interpretation; detection of anomalies in the metabolic and transport networks, such as substrates that are transported into the cell but are not inputs to any metabolic reaction or pathway; and comparative analyses of the transport capabilities of different organisms.

Results: On randomly selected organisms, the method achieves precision and recall rates of 0.93 and 0.90, respectively in identifying transporter proteins by name within the complete genome. The method obtains 67.5% accuracy in predicting complete transport reactions; if allowance is made for predictions that are overly general yet not incorrect, reaction prediction accuracy is 82.5%.

Availability: The method is implemented as part of PathoLogic, the inference component of the Pathway Tools software. Pathway Tools is freely available to researchers at non-commercial institutions, including source code; a fee applies to commercial institutions.

Contact: tomlee@ai.sri.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The importance of membrane transport proteins (transporters) to cells is illustrated by the fact that transporters typically make up 5–15% of the total gene content of sequenced organisms. Transporters bring essential nutrients into the cell, and therefore partially determine the environments in which cell growth is possible. They also provide pathogenic bacteria with resistance to antibiotics, and provide cancer cells with resistance to chemotherapies.

This research is motivated by the need to perform symbolic systems biology (Karp, 2001) analyses involving cellular transport systems, such as to compute the answers to the following queries for a given organism: (1) What chemical compounds can the organism import or export? (2) For which cellular metabolic pathways can the organism neither import, nor produce via a metabolic reaction, the precursors required by that pathway? This query identifies incompleteness in our knowledge of the metabolic and transport networks. (3) Which molecules can enter a given cellular compartment based on its known transporter complement? (4) How do the complements of transporter functions differ among two or more organisms? These applications are discussed in more detail in Section 5.

Such analyses demand a computable (ontology based) representation of transporter function. We developed such a representation several years ago as part of the Pathway Tools project (Karp *et al.*, 2002; Krummenacker *et al.*, 2005; Paley and Karp, 2006). In Pathway Tools, transporter functions are represented as *reactions* in which the substrates of the reactions are labeled with the cellular compartments in which those substrates reside.

The curators of the EcoCyc database (DB) have manually populated EcoCyc with 238 transport reactions describing the functions of 214 transporters (Keseler *et al.*, 2007) using the interactive editors within Pathway Tools. However, manually curating transport reactions is very time-consuming. Therefore, the problem addressed in this article is to develop an algorithm that will automatically infer the correct transport reaction for a transport protein given the English functional annotation assigned to the transporter. Our work follows in a long line of research involving the application of natural-language processing techniques to bioinformatics [see, for example, Rzhetsky *et al.* (2004) and the biomedical text mining tracks in the Pacific Symposium on Biocomputing proceedings from 2006–2008] to extract ontology-based descriptions of biological knowledge from natural-language text. However, our work does not involve processing of long texts, but rather of short textual descriptions of transporter functions.

This article describes an algorithm called the Transport Inference Parser (TIP) that identifies the transporter proteins within a genome, that infers the transport reaction(s) catalyzed by these proteins, and that constructs a full ontology-based representation for each transport reaction and protein within a Pathway/Genome DB (PGDB). In addition, TIP infers multimeric transporter complexes, and constructs PGDB objects describing the inferred complexes.

TIP is a component of Pathway Tools, which is a comprehensive symbolic-systems biology software system that supports several use cases in bioinformatics and systems biology. It supports development of organism-specific DBs [also called model-organism DBs (MODs)] that integrate many bioinformatics datatypes, from genomes to pathways, and is in use by many MOD projects including EcoCyc (Keseler *et al.*, 2007), SGD (Nash *et al.*, 2007), Mouse Genome Informatics (Bult *et al.*, 2008), dictyBase (Chisholm *et al.*, 2006) and the BioCyc collection of 370 PGDBs (Caspi *et al.*, 2008). Pathway Tools provides several other computational inferences including prediction of metabolic pathways, prediction of metabolic pathway hole fillers and prediction of operons. It provides scientific visualization services including automatic display of metabolic pathways and full metabolic networks; a genome browser; and display of operons, regulons and full transcriptional regulatory

[*] To whom correspondence should be addressed

Table 1. Example inputs and outputs to TIP

Input: protein function	**Output: inferred transport reaction**
predicted ATP transporter of cyanate	$cyanate_{[extracellular]} + H_2O + ATP = cyanate + phosphate + ADP$
putative phosphate ABC transporter, ATP binding subunit	$phosphate_{[extracellular]} + H_2O + ATP = 2phosphate + ADP$
putative potassium channel	$K^{+}_{[extracellular]} = K^{+}$
sodium/proline symporter	$Na^{+}_{[extracellular]} + \textsc{l}\text{-}proline_{[extracellular]} = Na^{+} + \textsc{l}\text{-}proline$
lactose transport system permease protein	$H^{+}_{[extracellular]} + lactose^{+}_{[extracellular]} = H^{+} + lactose$

The inputs to TIP consist of transporter function names present in protein objects in PGDBs, plus other information available in the genome and its annotation. The outputs of TIP consist of structured descriptions of transporter function in the form of transport reactions.

networks. It supports visual analysis of omics datasets, such as painting omics data onto diagrams of the full metabolic network, full regulatory network and full genome. It supports comparative analyses of PGDBs and analysis of biological networks such as identifying choke points (potential anti-microbial drug targets) in metabolic networks.

Although the annotation of membrane transporters from genome sequence has been extensively studied, we are not aware of any past work to generate transport reaction equations from free-text transporter annotations. Our work fills this fundamental gap and will be of interest to researchers in genomics; metabolic modeling; and development of flux-balance models (Feist *et al.*, 2007; Mo *et al.*, 2007), where transporters are represented by reactions, and it is critical that all transport reactions be included in the model.

2 APPROACH

The algorithm described in this article addresses several related problems:

(1) Given the full set of monomeric proteins P of an organism, identify the subset P_T of P that are transport proteins.

(2) For each monomeric transporter in P_T, infer the one or more multimeric complexes that the transporter is involved in.

(3) For each monomeric transporter or transporter complex in P_T, infer the one or more transport reactions that the transporter facilitates.

(4) Populate a PGDB with objects describing the inferred transport reactions and multimeric complexes.

We consider all information provided as part of an annotated genome sequence to be fair game toward solving these problems, such as the protein sequences, and the functional annotations of the transporters, which are typically provided in the Genbank `/product` field. Our method relies principally on textual analysis of the natural-language descriptions of transporter functions that are found in annotated genomes. Why do we perform this textual analysis rather than simply inferring the transport reaction directly from the protein sequence? The reasons are that inferring every aspect of a transporter function from its sequence in a completely automated fashion is a very hard problem. Many genome centers annotate transporter functions with oversight from a person who is skilled in sequence analysis. We believe that replacing those expert annotations with transporter functions that are inferred automatically would reduce their quality; therefore, our approach is to build on the existing genome annotation.

Table 1 shows example inputs and outputs for TIP. In Table 1, we write transport reaction substrate compartments as bracketed subscripts. Omission of the compartment both in this table and in a PGDB, implies the default compartment: the cytoplasm. Within a PGDB, transport reactions are represented as a single PGDB object. One attribute of the object stores the reactants; a second attribute stores the products. Each substrate (reactant or product) can be labeled with the identities of the one or more compartments in which the substrate occurs using a PGDB construct called an *annotation*, which is simply a way of attaching labeled information to an attribute value.

Our approach to transport function prediction for a PGDB consists of solving the following sequence of subproblems:

(1) Starting with the full set of monomeric proteins of an organism as defined in the PGDB, identify the transport proteins P_T.

(2) For each transport protein T:
   (a) identify the reaction substrates of T;
   (b) determine an energy coupling for T (e.g. is T a passive channel, or an ATP-driven transporter, or a sodium-driven symporter?)
   (c) assign a compartment to each substrate of T;
   (d) identify and construct transporter complexes for T; and
   (e) construct a full PGDB reaction for T.

The TIP algorithm is discussed in the following section. See Supplementary Material for data supporting the algorithm.

3 METHODS

Development of the rules underlying TIP was guided by Paulsen's expert knowledge of transporters in general, and of the key pieces of knowledge one must have to characterize the function of a transporter: its primary substrate, carrier substrate(s) (if any), energy coupling mechanism and directionality. Rule development was further guided by a review of hundreds of example transporter function descriptions from many genomes. TIP has undergone an iterative development process whereby we have run it on many genomes, and manually adjusted the algorithm to improve its performance on transporter functions that it did not properly interpret, such as by extending the lists of indicator and counterindicator keywords.

One of the most challenging aspects of this project was the fact that different genome annotation centers use different styles and conventions in the phrasing of transporter annotations. That is, the same transporter function is expressed in many different ways in different genomes. Therefore, our development of TIP involved running it on genomes from multiple genome centers, and from a taxonomically diverse set of organisms. Overall, the rules used by TIP are specific to transporter functions, but it may be possible to use a similar strategy for converting free-text descriptions of other types of proteins into ontology-based descriptions.

3.1 Identify transport proteins

To identify transporters, all proteins of the PGDB are filtered based on their annotation. Each protein whose annotation contains an *indicator keyword or phrase* indicative of transport function (e.g. 'transport', 'channel', 'permease') is designated as a transporter, unless it contains a *counterindicator keyword or phrase* (e.g. 'regulator'). A counterindicator indicates that the annotation is likely to be difficult to analyze without sophisticated parsing techniques. For example, such annotations sometimes refer to a transport function of another protein, rather than describing the function of its protein.

Each transporter is classified as either a *high-quality* or a *low quality* prediction. A low-quality prediction is one in which there is incomplete or ambiguous information in the annotation. Proteins whose annotations contain an *ambiguity keyword or phrase* (e.g. 'resistance') are considered low quality, as are those with annotations that exceed a threshold number of words (12), because we observed that textual analysis of long annotations is error prone. One type of error is improperly identifying non-transporters as transporters because their annotation happens to contain transporter indicator keywords.

3.2 Identify reaction substrate(s)

The protein annotation text is parsed and analyzed to identify one or more compounds that are the substrates of the transport reaction. The set of small-molecular weight chemical compounds and compound classes within the MetaCyc DB (Caspi *et al.*, 2006) is used as a dictionary for compound identification. MetaCyc contains an extensive set of synonyms for the names of metabolites.

All transporters have at least one *primary* substrate that crosses a cell membrane. Most transporters have a single primary substrate (for example, 'probable phosphate transporter'). Other transporters have multiple primary substrates, due to loose substrate specificity (for example, 'cytosine/purine/uracil/thiamine/allantoin permease family protein' or 'magnesium and cobalt transport protein corA, putative'). If no primary substrate can be found, the transporter is considered a low-quality prediction.

Secondary transporters have an additional *carrier* substrate, for example, for many transporters the transport of the primary substrate is driven by the proton gradient between the interior and exterior of the cell. The carrier substrate defines the energy-coupling mechanism of the transport process. The carrier substrate may be explicitly named in the annotation (for example, 'sodium:sulfate symporter transmembrane domain protein') or it may be unspecified. Finally, a transport reaction may have one or more *auxiliary* substrates. For example, ATP-driven transporters include the compounds ATP, ADP and H_2O as auxiliary substrates; these are not determined by parsing the annotation for substrate names, but are implied by the energy-coupling mechanism.

A simple search of a compound DB for words that occur in the annotation will yield many false positives (for example, 'as'), which are filtered using an exception list of compounds that do not occur biologically, or do not typically occur in transport reactions.

A substrate name in the annotation may not correspond exactly to the chemical names in MetaCyc. Substrate names are canonicalized for DB matching by removing spaces, hyphens and other punctuation; by converting to lower case; by converting plurals to singular form; and by stripping off affixes (for example, '- specific', '- transporting'). Certain elemental forms are converted to their ionic form (for example, 'hydrogen' to 'H+').

Some substrates are specified as classes of compounds, rather than individual compounds (for example, 'amino acid transporter'). This is not a difficult complication, because MetaCyc includes both compound classes and individual compounds. However, since many classes have multi-word names, word sequences as well as individual words must be considered.

It is assumed that the first substrate or a group of textually contiguous substrates found in the annotation are the substrates in the transport reaction; subsequent substrates are ignored. For example, in the annotation 'Ca+2 transporter; possible Mg+2 transporter' only 'Ca+2' would be detected as a substrate. In practice, this simplifying assumption leads to few incorrect substrate predictions.

3.3 Determine energy coupling mechanism

In a PGDB, the energy coupling of a transport reaction determines the class of reaction within the Pathway Tools reaction ontology under which that reaction object is created. The couplings currently handled by TIP are

- Ion channel: ions passively diffuse through the transporter. The reaction equation is $ion_{[extracellular]} = ion_{[cytosol]}$.
- Secondary transporters: a secondary (carrier) substrate is co-transported with the primary substrate. If substrates are transported in the same direction the reaction equation is $primary_{[extracellular]} + carrier_{[extracellular]} = primary_{[cytosol]} + carrier_{[cytosol]}$. If substrates are transported in the opposite direction the reaction equation is $primary_{[extracellular]} + carrier_{[cytosol]} = primary_{[cytosol]} + carrier_{[extracellular]}$.
- ATP-dependent (ATP): the hydrolysis of ATP provides energy to transport a primary substrate that is a solute. The reaction equation is $H_2O + ATP + solute_{[extracellular]} = phosphate + ADP + solute_{[cytosol]}$.
- Phosphenolpyruvate-dependent phosphotransferase system (PTS): phosphenolpyruvate is used as an energy source and phosphate donor to transport and phosphorylate a primary substrate that is a sugar. The reaction equation is $phosphoenolpyruvate + sugar_{[extracellular]} = pyruvate + phosphorylated\text{-}sugar_{[cytosol]}$.
- Unknown: a catch-all class used if a more specific determination cannot be made. The reaction equation is $substrate_{[extracellular]} = substrate_{[cytosol]}$.

To identify the energy coupling, the following rules are applied to the predicted primary substrate(s) and annotation:

(1) TIP is configured to treat a few primary substrates as being indicative of a particular coupling. For example, the transport of 'protoheme' indicates an ATP transporter because the only transporters known for protoheme are ATP driven.[1]

(2) If no coupling is indicated by the primary substrate, the annotation is searched for the presence of keywords indicating an energy coupling. For example, 'channel' indicates an ion channel transporter, 'carrier' indicates a secondary transporter and 'atp-binding' indicates an ATP transporter.

(3) If no coupling clues are present, the reaction is put into a default class indicating that the coupling mechanism is unknown.

For secondary transporters, the carrier substrate must be determined. The following rules are applied in order:

(1) If one of the substrates identified is either a proton or a sodium ion, it is designated as the carrier, and the remaining substrates become primary substrates (for example, 'sodium:glutamate symporter').

(2) If there are exactly two substrates, one is chosen arbitrarily as the carrier, and the other becomes the primary substrate.

(3) The carrier is assumed to be a proton; all other substrates become primary substrates (for example, 'amino acid transport system carrier protein' implies a proton carrier and 'amino acid' as the one primary substrate).

[1] If multiple substrates have been predicted, it is possible for this rule to make conflicting predictions about the coupling. However, given the few (6) substrates that imply a coupling, conflicts do not occur in practice.

3.4 Assign a compartment to each substrate

A complete transport reaction includes the designation of the initial and final cell compartments of each primary substrate. Our method searches the annotation for keywords indicating the initial or final compartment of the primary substrate(s). For example, ‘uptake’ indicates transport into the cell, and ‘export’ or ‘efflux’ indicates transport out of the cell.

For transporters with a carrier substrate, certain keywords indicate transport direction relative to the carrier. For example, ‘symport’ indicates that both the carrier and the primary substrate start and end in the same compartment. The keyword ‘antiport’ indicates an exchange between two compartments.

In the absence of more specific information, it is assumed that transport of the primary substrate (and carrier substrate if present) is into the cytoplasm.

Our current implementation of TIP is limited in that it is oriented toward bacteria, and does not perform well on eukaryotic compartments. In particular, it does not detect eukaryotic compartment names, nor does it attempt to assign eukaryotic compartments to substrates by other means. In addition, even for bacteria the implementation currently assumes a Gram-positive cell type, meaning it does not distinguish the inner membrane from the outer membrane.

3.5 Identify and construct transporter complexes

Many transporters are multimeric systems of several protein monomers. In these systems, the proteins comprise a transport complex that catalyzes the transport reaction. TIP infers the grouping of a set of individual transport monomer proteins into a complex (and constructs the representation of the complex in the PGDB) if the monomers satisfy all the following rules:

(1) The predicted substrate (or set of substrates) for each monomer in a complex must be identical.

(2) The predicted energy coupling must be ATP or PTS.

(3) The genes of all proteins within a multimeric complex must share a common operon.

Rule (1) enforces a requirement of consistency among the monomeric annotations with respect to substrates. Rule (2) is in place because other types of transporters are less commonly found as multimers. Rule (3) reflects that the co-occurrence of transporter monomers within a single operon is a strong indicator of multimers. Note that this implies that a genome with no operons will have no predicted multimers. Thus, no multimers are inferred in eukaryotes, and multimer inference is attempted only for PGDBs that contain operons. Operons can be predicted by the Pathway Tools operon predictor (Romero and Karp, 2004) or defined manually using Pathway Tools editors. We believe these rules will result in few false positive predictions, but acknowledge they will fail to identify a significant number of multimeric transporters.

3.6 Construct full compartmentalized reaction

At this point in the method, all predictions have been made. Each predicted transport reaction is added to the PGDB. One reaction is created for each predicted primary substrate.

The primary substrate is added to the list of reactants of the reaction, and is annotated with its initial compartment; it is then added to the list of products of the reaction, annotated with its final compartment. If present, the carrier substrate is added as a reactant and a product, annotated with the appropriate compartment depending on whether it is a symport or an antiport reaction. Each auxiliary substrate implied by the coupling (for example, ATP, ADP and water for ATP reactions) is added to the appropriate side of the reaction.

If the coupling mechanism is PEP, the primary substrate is phosphorylated during the transport. To reflect this, MetaCyc is searched for a phosphorylated variant of the primary substrate. If found, it replaces the primary substrate as a product of the transport reaction. In the atypical case in which it is not found, the unmodified primary substrate remains both a reactant and a product in the reaction.

To maximize the notion of MetaCyc reactions as a controlled vocabulary, we prefer to maintain a one-to-one mapping between reaction identifiers and reaction equations. That is, we do not want MetaCyc or other PGDBs to define the same reaction under different identifiers. Thus, if a reaction with the same substrates as those in the transport reaction just inferred are found in MetaCyc, the MetaCyc reaction object is imported into the PGDB where TIP is being run. Otherwise, a new reaction object is created.

3.7 Examples

We present an extended example of the operation of the TIP algorithm for a representative protein, followed by examples of erroneous predictions and successive refinement to TIP’s behavior.

3.7.1 Extended example Consider a protein with the annotation *sodium/proline symporter*. First, the protein is identified as a transporter by the presence of the keyword ‘symporter’. Next, the annotation is scanned for substrates; TIP has a rule that considers words separated by slashes as possible multiple substrates. TIP queries MetaCyc with the word ‘sodium’, that matches the compound object Na^{+}, which has a synonym ‘sodium’. Similarly, MetaCyc is queried with ‘proline’, that matches the compound object L-proline.

The energy coupling mechanism is determined to be SECONDARY by the presence of the keyword ‘symporter’. A secondary transporter has a carrier substrate; TIP applies the rule that if either a proton or a sodium ion is present, it is the carrier. So Na^{+} is made the carrier substrate, and L-proline remains the primary substrate.

The presence of ‘symporter’ indicates that the primary and carrier substrates start and end in the same compartments. Absent other indicators, TIP assumes transport is into the cytosol. This leads to the compartment assignments of [*extracellular*] to both substrates on the left side of the reaction, and of [*cytosol*] to both substrates on the right side.

Since the coupling is not ATP or PTS, no attempt to identify transport complexes is made. Finally, the full transport reaction is constructed: $Na^{+}_{[extracellular]} + \textit{L-proline}_{[extracellular]} = Na^{+} + \textit{L-proline}$. Provenance data is attached indicating that the reaction was predicted by TIP, the date and time of the creation of the reaction, the name of the user operating TIP and an evidence code indicating that the supporting evidence for the prediction is computational in nature.

3.7.2 Examples of errors TIP fails to identify a protein with the annotation ‘arsenical pump-driving ATPase’ as a transporter because neither ‘pump-driving’ nor ‘ATPase’ had been identified as transport indicator keywords.

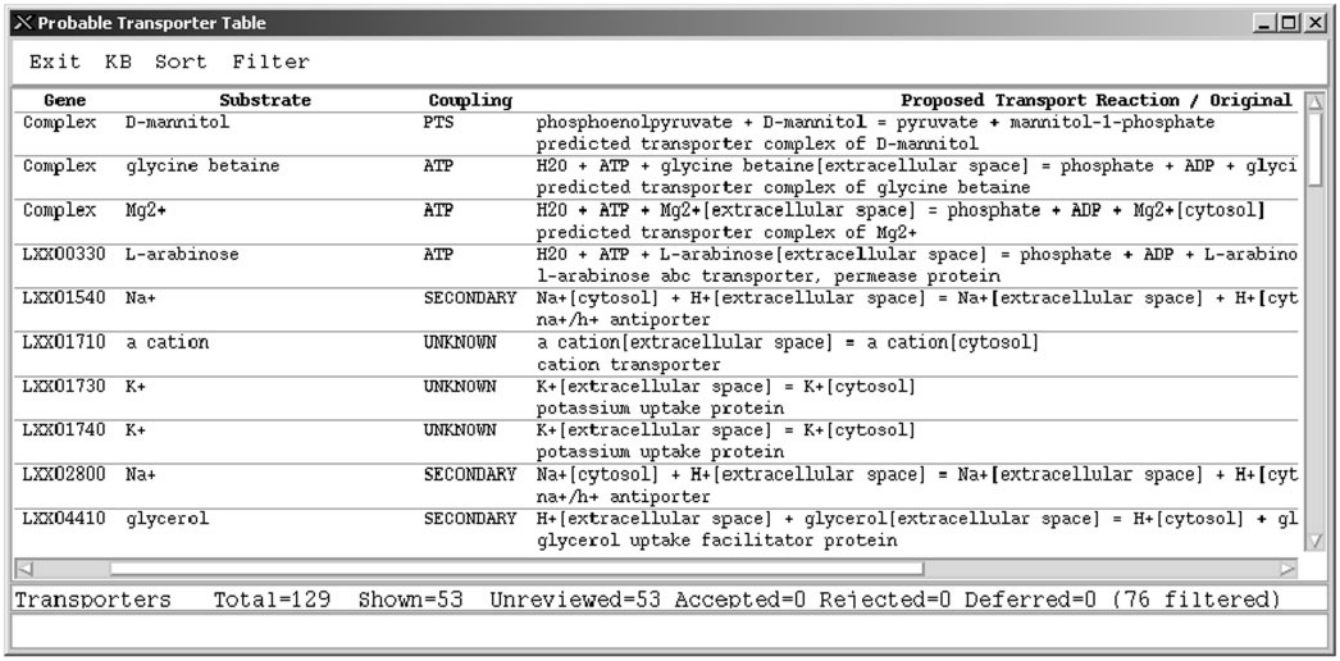
For a protein with the annotation ‘magnesium-exporting atpase’, TIP detects the substrate correctly, as it recognizes that ‘exporting’ is an acceptable suffix for a substrate. However, TIP did not, at the time this protein was encountered, include this suffix as a clue for compartment assignments; this resulted in the incorrect assignment of the cytosol as the ending compartment, rather than the starting compartment, of magnesium. Upon encountering this example, TIP is easily modified to recognize that this suffix implies transport out of the cytoplasm.

For a protein with the annotation ‘high-affinity iron permease’, TIP fails to detect a substrate because, although iron is clearly mentioned, iron has multiple valences and there is no mention of which form of iron is transported. TIP currently has no rules to disambiguate such references.

3.8 TIP Implementation

TIP can be run in both an interactive and a batch mode. Interactive mode permits review and modification of all inferences made by TIP. Interactive TIP is available as part of the PathoLogic program, as a step in construction of a PGDB.

Results are presented in a columnar GUI (Fig. 1). Transporters are sortable by gene name, substrate or coupling, facilitating systematic review and comparison of predictions. The GUI permits display of the set of



Gene	Substrate	Coupling	Proposed Transport Reaction / Original
Complex	D-mannitol	PTS	phosphoenolpyruvate + D-mannitol = pyruvate + mannitol-1-phosphate predicted transporter complex of D-mannitol
Complex	glycine betaine	ATP	H2O + ATP + glycine betaine[extracellular space] = phosphate + ADP + glyci predicted transporter complex of glycine betaine
Complex	Mg2+	ATP	H2O + ATP + Mg2+[extracellular space] = phosphate + ADP + Mg2+[cytosol] predicted transporter complex of Mg2+
LXX00330	L-arabinose	ATP	H2O + ATP + L-arabinose[extracellular space] = phosphate + ADP + L-arabino l-arabinose abc transporter, permease protein
LXX01540	Na+	SECONDARY	Na+[cytosol] + H+[extracellular space] = Na+[extracellular space] + H+[cyt na+/h+ antiporter
LXX01710	a cation	UNKNOWN	a cation[extracellular space] = a cation[cytosol] cation transporter
LXX01730	K+	UNKNOWN	K+[extracellular space] = K+[cytosol] potassium uptake protein
LXX01740	K+	UNKNOWN	K+[extracellular space] = K+[cytosol] potassium uptake protein
LXX02800	Na+	SECONDARY	Na+[cytosol] + H+[extracellular space] = Na+[extracellular space] + H+[cyt na+/h+ antiporter
LXX04410	glycerol	SECONDARY	H+[extracellular space] + glycerol[extracellular space] = H+[cytosol] + gl glycerol uptake facilitator protein

```
Transporters   Total=129  Shown=53  Unreviewed=53 Accepted=0 Rejected=0 Deferred=0 (76 filtered)
```

Fig. 1. PathoLogic GUI for TIP. Predicted transport reactions may be individually accepted or rejected (when rejected, the reaction is deleted from the PGDB). Attributes of transporters, including the compartmentalized reaction, may be edited.

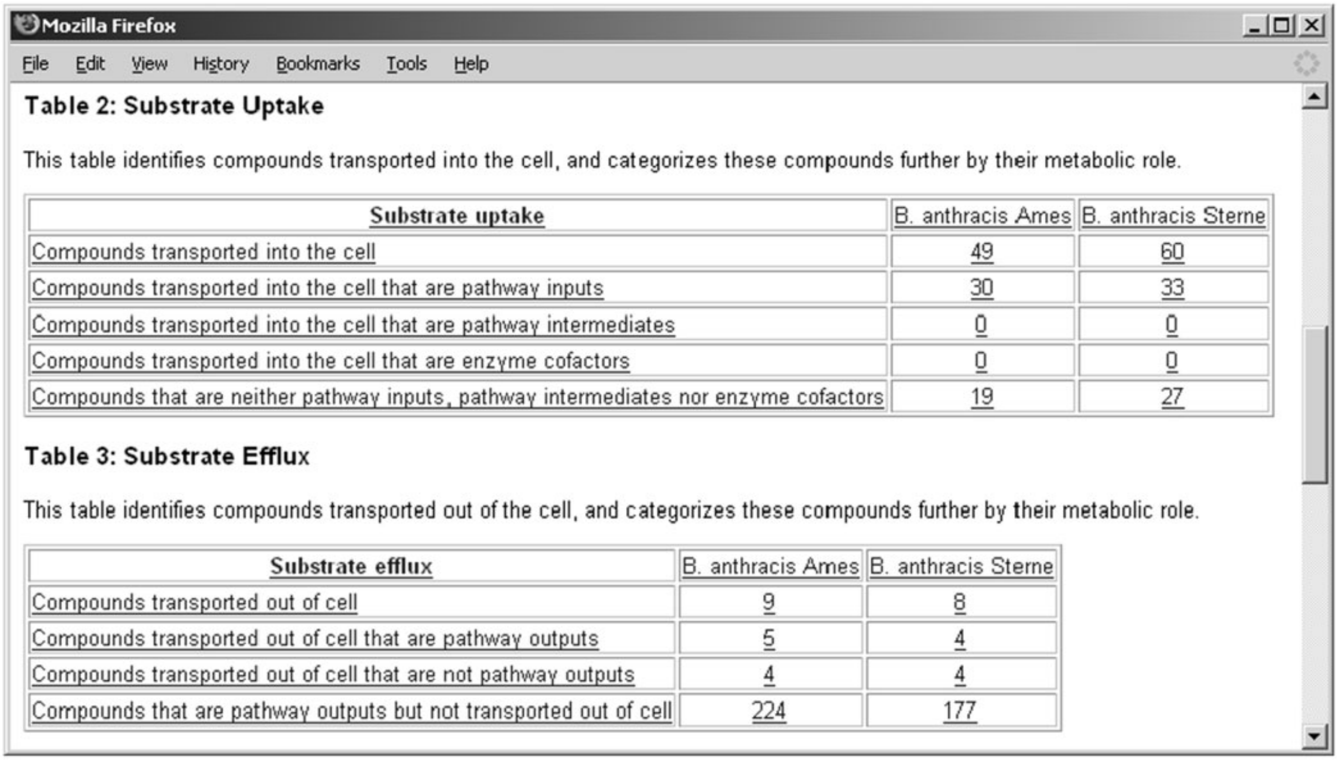


Table 2: Substrate Uptake

This table identifies compounds transported into the cell, and categorizes these compounds further by their metabolic role.

Substrate uptake	B. anthracis Ames	B. anthracis Sterne
Compounds transported into the cell	49	60
Compounds transported into the cell that are pathway inputs	30	33
Compounds transported into the cell that are pathway intermediates	0	0
Compounds transported into the cell that are enzyme cofactors	0	0
Compounds that are neither pathway inputs, pathway intermediates nor enzyme cofactors	19	27

Table 3: Substrate Efflux

This table identifies compounds transported out of the cell, and categorizes these compounds further by their metabolic role.

Substrate efflux	B. anthracis Ames	B. anthracis Sterne
Compounds transported out of cell	9	8
Compounds transported out of cell that are pathway outputs	5	4
Compounds transported out of cell that are not pathway outputs	4	4
Compounds that are pathway outputs but not transported out of cell	224	177

Fig. 2. A portion of the Comparative Genomics display available at www.biocyc.org. Supplementing BioCyc PGDBs with transport reactions enhances metabolic analysis of a single organism and of multiple organisms.

either high-confidence, low-confidence or all transporters predicted. Each transport reaction is individually selectable, and may be rejected, accepted as is, or modified. Permitted modifications consist of changing the energy coupling, editing the protein annotation and editing the reaction (including the primary substrate and compartment assignments). Typical modifications include changing a predicted coupling of UNKNOWN to a more specific coupling, and providing a reaction for a low-confidence transporter in which the primary substrate was not detected.

In batch mode, TIP is applied to a set of organisms. The batch mode of TIP performs the same predictions as interactive mode except that low-confidence predictions are discarded. Results are not presented for review, but are incorporated immediately into the PGDB. Batch mode permits automation of transporter predictions, such as for the large-scale PGDB creation effort within the BioCyc project (Karp *et al.*, 2005). In BioCyc version 11.5, TIP was used to predict transport functions for its 349 Tier 3 PGDBs.

In both interactive and batch modes, provenance information is attached to predictions. This information includes a timestamp, identification of the software making the prediction (TIP), the user that is operating TIP and an indication that the evidence supporting the prediction is computational in nature.

4 EVALUATION

We evaluated TIP on three genomes chosen randomly from the set of BioCyc Tier 3 (version 11.5) organisms (Caspi *et al.*, 2008): *Helicobacter hepaticus ATCC 51449* (sequenced by MWG Biotech, University of Wuerzburg, Massachusetts Institute of Technology, and GeneData), *Fusobacterium nucleatum ATCC 25586* (sequenced by Integrated Genomics Inc., Chicago, Illinois, USA) and *Leifsonia xyli CTCB07* (sequenced by University of Campinas, Campinas, Brazil). Neither these genomes nor genomes from these centers were used to formulate or to tune the TIP heuristics.

We partitioned the evaluation into two parts: identification of transporters and prediction of attributes of transport reactions. We do not present an evaluation of transport complex prediction. Since no gold standard exists for evaluation of transporter function predictions, we generated a gold standard for this work. We used TransportDB (http://www.membranetransport.org/, (Ren *et al.*, 2007)) as a basis for this standard.

TransportDB is an extensive DB of transporters, supported by evidence of varying type and strength. Like most gold standards in bioinformatics, TransportDB has high-quality data, but it is probably not perfect. We did not take into account the type or strength of evidence in our scoring. We adapted the data from TransportDB in cases of clear annotation discrepancies between it and the PGDB being evaluated. That is, when a PGDB gene function differed from a TransportDB gene function, we used the PGDB gene function in our gold standard since TIP would be run on the gene function present in the PGDB.

4.1 Evaluation of transporter identification

Our standard for evaluation of TIP transporter identification was formed by comparing the set of predictions made by TIP with the set of transporters in TransportDB, and reviewing each discrepancy. If there is a transporter in TransportDB that is not predicted by TIP, it is excluded from the standard if the protein function in the PGDB was different for that protein in TransportDB and was non-specific with respect to transport function (for example, a protein with no annotation at all in PGDB would be excluded). Recall that TIP does not perform functional annotation; its goal is to interpret a functional prediction that is in natural-language form. If there is no transporter in TransportDB corresponding to a TIP prediction, the TIP prediction is added to the gold standard if there is a clear indication of transport function in the PGDB annotation.

Transporter identification results are shown in Table 2. `TransportDB size` is the total number of monomeric transporters in TransportDB. `Standard size` is the number of monomeric transporters in the evaluation standard. `False negatives` is the number of transporters in the standard that are not predicted to be transporters. `True positives` is the number of transporters in the standard that are correctly predicted to be transporters. `False positives` is the number of transporters not in the standard that are predicted to be transporters. True negatives—proteins that are not transporters and not identified as such—are not shown, and typically number in the thousands for an organism. `Precision` is the ratio of true positives to all positives predicted. `Recall` is the ratio of true positives to the size of the standard. Precision and recall are shown for each organism and in aggregate.

4.2 Evaluation of transporter attribute prediction

The standard for evaluation of transport reaction attribute prediction was formed by using the attributes identified in TransportDB, adjusted for any clear discrepancies with the annotation of the PGDB protein. The attributes included are substrates, substrate compartments and energy coupling. TransportDB does not include reactions, but lists substrates and couplings, as well as characterizing transporters as symporters, antiporters, exporters and so on, implying a compartment. In the absence of an explicit compartment, we assume that the primary substrate is imported into the cytoplasm.

If either TransportDB or a PGDB annotation is non-specific with respect to substrates, the transporter is excluded from the standard. For example, the annotation ‘Probable transporter’ contains no clue as to the substrate because the annotator could not confidently infer the substrate from the protein sequence.

If there is a conflict between the substrates in TransportDB and those apparent in the PGDB annotation, the annotation substrates are

Table 2. Transport identification results for transport reactions

Organism	TransportDB size	Standard size	False negative	True positive	False positive	All positive	Precision	Recall
H.hepaticus	117	48	1	47	0	47	1.0	0.98
F.nucleatum	261	263	32	231	32	263	0.88	0.88
L.xyli	179	165	15	150	0	150	1.0	0.91
Total	557	476	48	428	32	460	0.93	0.90

used, thereby overriding TransportDB. For example, the protein for gene FN1747 in *F.nucleatum* is annotated as 'cysteine permease' but its primary substrate is identified as *alanine* in TransportDB; *cysteine* is used in the standard.

In the absence of conflicting or non-specific substrates, the TransportDB reaction attributes are included in the standard. Conflicts between TransportDB compartments or couplings and PGDB annotations are possible, but did not occur for the PGDBs evaluated.

Attribute prediction results are shown in Table 3. To evaluate prediction of transport reaction attributes, we score all correctly identified transporters (that is, all true positives) in comparison with the standard. A substrate is scored as `Too general` if the prediction is a superclass of the actual substrate (for example, predicting a substrate 'amino acid' when the actual substrate is 'a branched-chain amino acid'). Any other substrate that is not an exact match with the standard is scored as an `Error`. A prediction is scored as a `Compartment Error` if the transport direction of the primary substrate does not match the standard (for example, if the ending compartment of the primary substrate is the cytosol in an export reaction). The predicted energy coupling is scored as an `Error` if it does not match the standard, except in the case where the prediction is UNKNOWN; in this case it is scored as `Too general`.

Each reaction is attributed to exactly one category in the results. If a reaction's substrate is not predicted correctly, then neither its compartment nor its coupling is scored; if a reaction's compartment is not predicted correctly, its coupling is not scored. The `% Perfect` column shows the percentage of reactions whose predicted attributes match the standard exactly. The `% General` column shows the percentage of reactions that match the standard except for a substrate or coupling that is more general than in the standard; that is, those reactions that are accurate but not precise. The `% Error` column shows the percentage of reactions that are neither `Perfect` nor `General`, that is, the reactions containing an attribute that is inconsistent with the standard.

It is possible to supplement our method with rules that increase the prediction accuracy for various attributes. However, depending on how the predictions are used, this may not be desirable. For example, many energy couplings that are predicted as being too general (typically UNKNOWN) are in fact channel transporters. Predicting CHANNEL in these cases would increase accuracy, but would also increase errors. In addition, if the prediction results are being reviewed by a curator, it is a natural workflow to examine generic predictions for possible refinements; hence the less specific predictions may lead to a higher quality final result.

5 APPLICATIONS

Once transporter activities are encoded in a declarative (computable) form, they can be exploited in several ways.

5.1 Database queries

Transport reactions can be searched through DB queries such as 'Enumerate the set of all influx substrates for this organism' (Keseler *et al.*, 2007), or 'Find all transporters of organic anions for this organism.' Simple queries such as these require painstaking manual analyses for most genome DBs, but are trivial queries within the Pathway Tools ontology.

5.2 Cellular Overview and Cellular Omics Viewer

Once transporters have corresponding transport reactions, transporters can be automatically added to the Pathway Tools Cellular Overview diagram (see http://biocyc.org/ECOLI/NEW-IMAGE?type=OVERVIEW. This diagram is automatically generated by Pathway Tools, but generation of transporters in the diagram requires that the transporters have transport reactions. Once the diagram is generated, it can be used for visual interrogation of cellular networks (e.g., the software will draw connections between a transported metabolite and all metabolic pathways that involve the metabolite). A large version of the diagram can be generated and printed to provide a metabolic/transport wall chart for the organism. Furthermore, the diagram can be painted with omics data (e.g. gene expression, proteomics or metabolomics data), allowing visual analysis of the data in that experiment, e.g. what transporter genes are upregulated in a given experiment (Paley and Karp, 2006)?

5.3 Anomaly detection in metabolic/transport networks

Existence of transport reactions enables detection of anomalies within the metabolic/transport networks. One type of anomaly is a dead-end metabolite (Keseler *et al.*, 2007), which is a compound that is only produced by or consumed by the metabolic network, and has no associated transporter. Most dead-end metabolites are due to errors or incompleteness in the metabolic network. Even the *Escherichia coli* metabolic network, which is probably the best studied and curated network of any free-living organism (Feist *et al.*, 2007; Keseler *et al.*, 2007) contains 169 dead ends (Keseler *et al.*, 2007).

Another type of potential anomaly between metabolism and transport is the case of metabolites other than inorganic ions for which a transporter exists, but no metabolic pathway or reaction exists that consumes the metabolite. *Escherichia coli* has a number

Table 3. Attribute prediction results

		Substrates				Coupling				
Organism	Standard size	Error	Too general	Correct	Compartment error	Error	Too general	% Perfect	% General	% Error
H.hepaticus	39	2	0	37	0	2	5	76.9	12.8	10.3
F.nucleatum	147	29	1	117	0	8	20	60.5	14.3	25.2
L.xyli	88	6	1	81	0	1	14	75.0	17.0	8.0
Total	274	37	2	235	0	11	39	67.5	15.0	17.5

of such metabolites, which could be due to genome annotation errors in transporters or enzymes, or unknown metabolic transformations.

5.4 Comparative genomics

Declarative representation of transport activities also enables comparative analyses of the transport capabilities of an organism. Pathway Tools implements several comparative analyses, such as those shown at `http://biocyc.org/comp-genomics`. These comparative analyses are based on functional capabilities, not on sequence. They provide the user with the following comparisons:

- Number of genes coding for transporter proteins.
- All transporters; and those that catalyze efflux versus influx transport.
- All compounds transported into the cell; and then a breakdown of those influx substrates that are also pathway inputs, pathway intermediates or enzyme cofactors, and those that fall into none of the preceding categories.
- All compounds transported out of the cell; and then a breakdown of those efflux substrates that are pathway inputs and those that are not pathway outputs
- Lists of transporters with multiple substrates; and of substrates with multiple transporters
- Operon analysis of transporters, listing transporters that are in the same operon as an enzyme operating on the same substrate; and transporters of unknown function in the same operon as an enzyme, which may yield clues to the transported substrate

6 RELATED WORK

Many genome DBs are unable to compute with transporter functions, such as answering the questions given in Section 1, because they lack an ontology-based representation of transporter function. Even genome DBs based on Gene Ontology (GO; Consortium, 2008) would find such queries difficult to answer. GO does provide controlled vocabulary terms for specific transporter functions. However, at best those GO terms contain transport reactions within their comments that would have to be parsed (e.g. GO:0005330, 'dopamine:sodium symporter activity'), but those reactions do not refer to a controlled vocabulary of chemical compound names such as that present in MetaCyc (Caspi *et al.*, 2008). At worst, many GO terms do not contain transport reactions.

The Transporter Classification (TC) system and associated Transporter Classification DB (Saier *et al.*, 2006) (http://www.tcdb.org) provides a different taxonomic classification of transporter than does GO. TCDB does not provide transport reactions, nor does it employ a controlled vocabulary of chemical compounds; therefore, it could not provide an ontological foundation for answering the questions in Section 1.

We are not aware of previous approaches to converting natural language descriptions of transporter functions. The closest work would probably be programs that automatically assign GO terms to transporters through sequence analysis. However, the outputs of those programs would be subject to the limitations discussed in the first paragraph. Similarly, Lin *et al.* (2006) developed a method for assigning a transporter to a TC class based on its amino acid sequence.

7 FUTURE WORK

TIP is the result of an ongoing process of successive refinement. As new organisms annotated by different genome centers are studied, TIP's rules are enhanced. To prevent regressions and to evaluate cases in which there are tradeoffs in prediction accuracy among various rules, we maintain a testbed of representative proteins.

TIP is currently oriented toward bacteria. We plan to enhance TIP to cover eukaryotes. We expect this will include parsing for eukaryotic compartments, associating particular substrates with their most likely origin or destination compartment, and exploiting other knowledge sources besides annotation. Furthermore, we plan to extend the rules for transporter complex prediction for cases in which operons are absent or unknown.

8 CONCLUSIONS

We have presented an approach to prediction of transporters and transport reactions based primarily on the textual analysis of the functional annotations of the proteins of an organism. We have discussed its implementation in the Pathway Tools software, which enhances its PGDB construction capabilities by supporting predictions for both curated and non-curated DBs.

We have evaluated the performance of TIP on several randomly selected organisms versus TransportDB, a high-quality standard for transporter knowledge. TIP achieves precision and recall rates of 0.93 and 0.90 respectively in identifying transporter proteins, and 67.5% accuracy in predicting complete transport reactions; if allowance is made for predictions that are overly general yet not incorrect, reaction prediction accuracy is 82.5%.

Once transporter activities are encoded as transport reactions, a number of computational analyses are possible including DB queries by transporter activity: inclusion of transporters into an automatically generated metabolic-map diagram that can be painted with omics data to aid in their interpretation, detection of anomalies in the metabolic and transport networks, and comparative analyses of the transport capabilities of different organisms.

ACKNOWLEDGEMENTS

We thank Pallavi Kaipa and Ron Caspi for their assistance in testing TIP.

Funding: This work was supported by grant GM75742 from the National Institutes of Health. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Bult,C.J. *et al.* (2008) The mouse genome database (MGD): mouse biology and model systems. *Nucleic Acids Res.*, **36**, D724–D728.

Caspi,R. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.

Caspi,R. *et al.* (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **36**, D623–D631.

Chisholm,R.L. *et al.* (2006) Dictybase, the model organism database for dictyostelium discoideum. *Nucleic Acids Res.*, **34**, D423–D427.

Consortium,G.O. (2008) The gene ontology project in 2008. *Nucleic Acids Res.*, **36**, D440–D444.

Feist,A. *et al*. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* k-12 mg1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, Article121.

Karp,P. (2001) Pathway databases: a case study in computational symbolic theories. *Science*, **293**, 2040–2044.

Karp,P. *et al*. (2002) The Pathway Tools Software. *Bioinformatics*, **18**, S225–S232.

Karp,P. *et al*. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.

Keseler,I. *et al*. (2007) Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res.*, **35**, 7577–7590.

Krummenacker,M. *et al*. (2005) Querying and computing with BioCyc databases. *Bioinformatics*, **21**, 3454–3455.

Lin,H.H. *et al*. (2006) Prediction of transporter family from protein sequence by support vector machine approach. *Proteins*, **62**, 218–231.

Mo,M.L. *et al*. (2007) A genome-scale, constraint-based approach to systems biology of human metabolism. *Mol. Biosyst.*, **3**, 598–603.

Nash,R. *et al*. (2007) Expanded protein information at sgd: new pages and proteome browser. *Nucleic Acids Res.*, **35**, D468–D471.

Paley,S. and Karp,P. (2006) The Pathway Tools cellular overview diagram and omics viewer. *Nucleic Acids Res.*, **34**, 3771–3778.

Ren,Q. *et al*. (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res.*, **35**, D274–D279.

Romero,P. and Karp,P. (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway/genome databases. *Bioinformatics*, **20**, 709–717.

Rzhetsky,A. *et al*. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.*, **37**, 43–53.

Saier,M.H. *et al*. (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.*, **34**, D181–D186.