Genome-Wide Semi-Automated Annotation of Transporter Systems

Oscar Dias, Daniel Gomes, Paulo Vilaça, João Cardoso, Miguel Rocha, Eugénio C. Ferreira, and Isabel Rocha

Abstract—Usually, transport reactions are added to genome-scale metabolic models (GSMMs) based on experimental data and literature. This approach does not allow associating specific genes with transport reactions, which impairs the ability of the model to predict effects of gene deletions. Novel methods for systematic genome-wide transporter functional annotation and their integration into GSMMs are therefore necessary. In this work, an automatic system to detect and classify all potential membrane transport proteins for a given genome and integrate the related reactions into GSMMs is proposed, based on the identification and classification of genes that encode transmembrane proteins. The Transport Reactions Annotation and Generation (TRIAGE) tool identifies the metabolites transported by each transmembrane protein and its transporter family. The localization of the carriers is also predicted and, consequently, their action is confined to a given membrane. The integration of the data provided by TRIAGE with highly curated models allowed the identification of new transport reactions. TRIAGE is included in the new release of *merlin*, a software tool previously developed by the authors, which expedites the GSMM reconstruction processes.

Index Terms—Bioinformatics, systems biology, Genome-scale metabolic models, transport reactions, genome transporters inference, *merlin*, TRIAGE

1 Introduction

Genome-scale metabolic models (GSMMs) can be used to simulate *in silico* the phenotype of organisms of interest in selected genetic/environmental conditions. These models are becoming increasingly common since the number of fully sequenced organisms, as well as the available data generated by high-throughput techniques, have been growing exponentially [1]. GSMMs have been applied in strain optimization tasks within the Metabolic Engineering field, and also in guiding biological discovery, analyzing global network properties, and studying evolution [2]. Several new methods, tools and databases to aid in the development of GSMMs and their application in strain optimizations tasks have been described [3], [4].

GSMMs include diverse information, such as reaction and metabolite sets, Enzyme Commission (EC) numbers [5] and gene-protein-reaction (GPR) associations. While most GSMMs have few compartments, over the years several have been released including broader compartmentalization information. Models such as the iMH805/775 [6] (15 compartments) and the iMM904 [7] (eight compartments) for *Saccharomyces cerevisiae*, the iOD907 [8] (four compartments) for *Kluyveromyces lactis* or the iRS1563 [9] for *Zea mays* (six compartments) include reactions occurring in specific cellular organelles, such as the mitochondria, chloroplasts (in photosynthetic organisms), lysosomes, cell nucleus or the Golgi apparatus. On the other hand, despite not having intracellular organelles, models of prokaryotic cells often have up to three compartments (cytoplasm, periplasm and the extracellular space), although more could be added if microcompartments (e.g., carboxysomes) were accounted for. The presence of compartments in GSMMs implies that often compounds have to cross cell or organelle-specific membranes so that reactions can take place, depending also on the localization of the enzymes.

Transport reactions present in GSMMs are normally obtained from experimental data and literature. A transport reaction is added for every metabolite known to be taken in from the medium, excreted from the cell or transported across intracellular membranes [10]. This methodology does not allow to automatically associate genes to transporter proteins and their reactions, decreasing the quality of the model's predictions (e.g., for gene knockouts). The identification of genes encoding transport proteins and the metabolites being transported by those carriers is important so that more and accurate GSMMs can be reconstructed [11] (both for eukaryotes and prokaryotes), also allowing the elemental and charge balances to be assessed more easily [10], [12]. The reconstruction of robust models involves determining the ionization state of each compound, according to the pH of the compartments [10]. Reactions must have a net charge of 0 whilst obeying to elemental balances, thus it is of the utmost importance to account for the water molecules and protons that participate in those reactions. Though, usually, water molecules are allowed to freely diffuse into all compartments, protons only change compartments through active transport reactions. Thus, the production and consumption of protons has to be properly balanced within each compartment.

- *O. Dias, D.Gomes, M. Rocha, E.C. Ferreira, and I. Rocha are with the Centre of Biological Enginnering, University of Minho, Campus de Gualtar 4710-057, Braga, Portugal. E-mail: {odias, danielg_gomes, ecferreira, irocha}@deb.uminho.pt, mrocha@di.uminho.pt.*
- *P. Vilaça and J. Cardoso are with the SilicoLife Lda., Rua do Canastreiro 15, 4715-387, Braga, Portugal. E-mail: {pvilaca, jcardoso}@silicolife.com.*

Manuscript received 12 Aug. 2014; revised 26 Jan. 2016; accepted 1 Feb. 2016. Date of publication 11 Feb. 2016; date of current version 22 Mar. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2527647

Some efforts have been undertaken by Lee et al. [13] to infer transport systems based in the genome annotation. However, to the best of our knowledge, a systematic approach to simultaneously identify, classify and annotate membrane transporters, as well as the reactions promoted by these proteins is lacking: Indeed, some authors have recently raised concerns regarding this issue. For example, Feist et al. [14] mention that new methods for this task are required and, even more recently, Hamilton and Reed [15] considered that transporters are still often poorly annotated and the existing tools cannot add transport reactions between compartments. In our vision, such a framework should also envision the (semi)-automated integration of these transporters into GSMMs, within the model reconstruction process.

In this work, TRIAGE (Transport Proteins Annotation and Reactions Generation), a tool that detects and classifies potential transport proteins based on the identification and classification of genes that encode transmembrane proteins, is proposed. Furthermore, TRIAGE automatically generates transport reactions for selected metabolites, which can be immediately integrated into GSMMs.

For this purpose, a pipeline which performs the genome-scale annotation of membrane transport proteins, by identifying genes encoding transporter proteins and the metabolites transported by each carrier, was developed. Specialized tools were used for predicting the localization of the carriers, confining their action to a specific membrane. TRIAGE is currently available in *merlin* (www.merlin-sysbio.org) [16], a software tool developed in-house that expedites the GSMMs reconstruction process.

After identifying the transporter systems within the target organism's genome, as well as their localization in the cell, an algorithm for generating transport reactions is deployed. Although such reactions are balanced and can be directly integrated into GSMMs, these should be regarded as predictions and not as reactions confirmed with experimental evidences. The transport reactions are built taking into account the metabolites annotated in the TCDB records identified as similar to the Transporter Candidate Gene (TCG) under analysis in the target genome. Several organisms were used to validate TRIAGE, namely *Kluyveromyces lactis*, *Ashbya gossypii*, *Saccharomyces cerevisiae*, *Helicobacter pylori* and *Escherichia coli*. Almost all *S. cerevisiae* GSMMs are compartmentalized, having intracellular transport reactions. *E. coli* is the most studied microbe and, despite being a prokaryote, several GSMMs with transport reactions from the outside to the periplasm and inside of the cell are available. The other cases represent less studied organisms of interest for which the authors have expertise.

2 Methods

TRIAGE was developed in Java and the information retrieved from the different data sources is kept in a MySQL relational database. As depicted in Fig. A.1 of Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2016.2527647, there are two layers on the relational database. The transporter candidates' layer (dynamic layer) is organism specific, with an instance of these tables for each organism. This layer is connected to the shared layer of the database, the transport reactions layer (static layer), by three connections that allow TCGs to be assigned with a TC family, a range of metabolites to be transported and a direction for such transport.

Five online databases are used by TRIAGE. TCDB (www.tcdb.org) is used as the main data source, since TCGs are compared against its sequences. Also, information on the metabolites used to construct transport reactions are retrieved from TCDB records. Kyoto Encyclopedia of Genes and Genomes (KEGG – www.kegg.jp) [17], Chemical Entities of Biological Interest database (ChEBI – www.ebi.ac.uk) [18] and semantics SBML 2.0 (semanticsbml.org/) [19] are used for collecting additional data for metabolite identification and characterization. UniProtJAPI [20] (www.uniprot.org) is used to retrieve the phylogenetic data of each of the TCDB transport systems that are essential for the assignment of transport reactions to the candidate genes, as it is described later. Biojava [21] was used to implement the Smith-Waterman (SW) [22] algorithm.

2.1 Development of an Internal Database of Transport Reactions

A database of transport reactions, based on information retrieved from TCDB, was compiled for TRIAGE. A concise description of that process is provided next.

Manually annotated cellular transport systems are described and stored in databases such as the Transporter Classification Database (TCDB) [23] or the TransportDB [24]. Since it is the most comprehensive, TCDB is used at the core of TRIAGE. It proposes a classification system for transport proteins, the transporter classification (TC) numbers, analogous to the EC system, but including also phylogenetic information. The system encompasses five components separated by a dot: $\#.*.\#.\#.\#$, in which $\#$ represent numbers and $*$ a letter. The first number is the class, while the letter corresponds to the subclass. TCDB classifies transport proteins in seven classes: Channels/Pores ($1.*.\#.\#.\#$); Electrochemical Potential-driven transporters ($2.*.\#.\#.\#$); Primary Active Transporters ($3.*.\#.\#.\#$); Group Translocators ($4.*.\#.\#.\#$); Transport Electron Carriers ($5.*.\#.\#.\#$); Accessory Factors Involved in Transport ($8.*.\#.\#.\#$); and Incompletely Characterized Transport Systems ($9.*.\#.\#.\#$). More information on TCDB classes can be found in [25]. The numbers after the letter indicate, respectively, the family (or superfamily), the subfamily (or the family within a superfamily) and the specific transporter system associated to a particular range of carried substrates and the polarity of transport (in or out) [23], [25]. For example, the TC number $2.A.1.1.1$ identifies a galactose:proton symport carrier of the Sugar Porter Family ($2.A.1.1$). This record is, currently, associated to a single *E. coli* gene (b2943). Enzymes are associated with EC numbers that classify the catalyzed reactions and a gene can be annotated with several EC numbers. On the other hand, TC numbers are associated to proteins that transport a specific substrate or range of substrates on a specific direction (in, out or both) using a given mechanism (uniport, symport, etc.), and are normally associated to a single organism [25]. Thus, unlike enzymes, transporter proteins should not be directly classified with TC numbers from homology, given the various criteria that must be met to match a specific TC

number. Moreover, the allocation of new TC numbers is exclusively performed by TCDB's expert curators as the deployment of TC numbers must be prudently controlled.

TCDB records often provide direct access to specific information, namely: UniProt Accession Number, organism, Protein Name, Length, Molecular Weight, organism's species, Number of transmembrane domains and Location/ Topology/ Orientation. However, to date, they do not contain specific fields for transported metabolites and direction of transport, which have to be manually inferred from the transport system description of each record, as well as from generalized transport reactions provided by TCDB for several families and superfamilies. Thus, a data integration workflow was developed, using other databases (KEGG, ChEBI and SemanticsSBML), for extracting information from TCDB on the metabolites involved in each TC record. This information is mandatory to provide connections to GSMMs.

TRIAGE's internal transport reactions database was populated with the following information retrieved from the TCDB, when available: UniProt accession number, protein name, organism, taxonomy, TC number, TC family, transported metabolite, direction, reversibility, reacting metabolites and equation. These data were retrieved using different approaches.

In this work, 3,248 TCDB transport system records were manually examined (TCDB records having similarities with the case studies), after automatic retrieval from the HTML interface using a Java routine. The remaining records will be curated as required by the case studies. The UniProt accession number, protein name, TC number, TC family and TC number description fields were automatically extracted from these records. The taxonomy of the record is directly retrieved from the UniProt database using the accession numbers available in the TCDB records and UniProtJAPI. The direction, reversibility, reacting metabolites and generic equation were manually retrieved from the TC families description or, when not available in this subclass, superfamily descriptions. This process was complex because, in the latter case, distinct transport system families share the same equations in a many-to-one relationship. The following example illustrates this, where equation (1) represents the generalized transport reaction for the 3.A.1 ATP-binding Cassette uptake system.

$$Solute_{out} + ATP- > Solute_{in} + ADP + P_i. \quad (1)$$

However, for example, distinct families like 3.A.1.1: the Carbohydrate Uptake Transporter-1 (CUT1) Family and 3.A.1.3: the Polar Amino Acid Uptake Transporter Family will have similar ATP dependent transport systems. Therefore, although these families have the same reacting metabolites (ATP, ADP, P_i), the transported metabolites are distinct (carbohydrates and polar amino acids, respectively). The localization of the reacting metabolites is not provided as it depends on the organism in which the reaction is taking place. Thus, manual curation of the reactions in the model is of paramount importance.

The difference between transported and reacting metabolites is that the first ones are only involved in reactions across membranes, while reacting metabolites are involved in chemical transformations throughout the transport process (e.g., ATP or NADH). As mentioned above, the metabolites transported by each system have to be manually retrieved from the transport system description. Still, only metabolite names can be retrieved from the TCDB records definition. Therefore, all of the manually identified metabolites were submitted to an algorithm developed in the scope of this work for classifying them with KEGG and ChEBI identifiers. This process uses three database Application Programming Interfaces (APIs) to identify cross references for these metabolites: KEGG, ChEBI and semanticSBML. Cross-references to KEGG are extremely important since those transport reactions will be easily integrated with the GSMMs created within *merlin*, which uses KEGG's metabolic information to assemble the reaction set. Therefore, this annotation algorithm tries to assign both identifiers (KEGG and ChEBI) to each metabolite, as illustrated in Fig. 1.

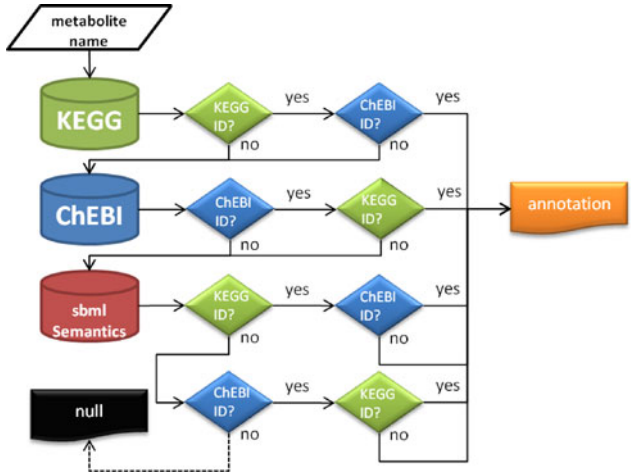


Fig. 1. Algorithm for assigning identifiers from KEGG and ChEBI to each metabolite. The algorithm stops when both identifiers are retrieved. When KEGG and ChEBI web services cannot annotate the metabolite with both identifiers, sbmlSemantics REST API is used to retrieve at least one of the identifiers. If the algorithm cannot return any identifier the metabolite is left unannotated.

Initially, the algorithm uses the KEGG REST API (http://www.kegg.jp/kegg/rest/keggapi.html) to look for a KEGG compound with a name or synonym that is a perfect match to the name manually collected from TCDB. If any KEGG entity meets the requirements, then the ChEBI cross-reference from that record is also retrieved. A valid ChEBI identifier allows the algorithm to stop and the metabolite to be annotated. On the other hand, an invalid ChEBI identifier or the lack of a match to a KEGG compound leads to the search of a match to ChEBI, performed using its Java API (www.ebi.ac.uk/chebi/webServices.do). As previously, a perfect match to a ChEBI entity name or synonym allows the algorithm to annotate the metabolite entity with a ChEBI identifier. The algorithm stops if the metabolite was previously annotated with a KEGG identifier or if ChEBI has a valid cross-reference to KEGG.

If the metabolite is not annotated with both KEGG and ChEBI identifiers after this direct search, semanticSBML is used to try to retrieve such identifiers. The semanticSBML REpresentational State Transfer (REST) API "search" method is used to search for MIRIAM annotations [26], using the metabolite name, and to get the list of matching annotation groups. The method is configured to return

only results with a precision of 1, i.e., only exact hits will be returned. If successful, the results obtained from this method allow the algorithm to annotate the metabolite with both KEGG and ChEBI identifiers. In the case none of the previous three methods assigns either a KEGG or ChEBI id, the metabolite is left unannotated.

Often, a metabolite retrieved from a TCDB record description is a generic entity (such as sugars, anions, lipids, etc.); thus, all the second-generation elements, i.e., the ones that are associated to the generic (or parent) metabolite in ChEBI by a "is a" or "has role" ontology, are also associated to the transport system of the parent metabolite. For example, α-D-glucose (CHEBI:17925) and β-D-glucose (CHEBI: 15903) are both second-generation elements of D-glucose (CHEBI:4167). For each second-generation element of a generic compound classified as substrate of a carrier encoded in the genome, a new transport reaction will be included in the GSMM and annotated with the corresponding gene. Still, not all second-generation elements retrieved from ChEBI for a given parent metabolite keep KEGG cross-references. Reactions for metabolites without KEGG identifiers will be generated, although not included in the model.

The formulae of all metabolites involved in transport reactions are inspected and the transport reaction is only accepted if the equation is balanced, i.e., if there are the same number of atoms of each element on the left and right hand sides of the equation. All previous information is kept in the transport reactions' layer of the database, according to Fig. A.1 of Appendix A, available in the online supplementary material. This database associates TCDB entries with transport reactions, including the transport type and transported metabolites, as well as other metabolites involved in the transport process.

2.2 Assignment of Transport Systems to Transporter Candidate Genes

The most important feature of TRIAGE is probably the assignment of transport systems (including the transport reactions) to genes. These reactions are usually catalysed by proteins located on membranes [23], [25], [27]. Thus, proteins with transmembrane domains may be regarded as suitable candidates to potential transport systems. There are a few tools available for the prediction of transmembrane protein topology from its sequence. The TransMembrane prediction using Hidden Markov Models (TMHMM - www.cbs.dtu.dk/services/TMHMM) tool [28] has been considered the best for this task [29]. Hence, it was therefore used to find, within the full genome, genes encoding proteins with transmembrane helices in their sequences, classifying these genes as TCGs if they have at least n transmembrane domains (n=1 was used in the scope of this work). In the future, other tools will be assessed for integration on TRIAGE, since recently Reddy *et al.* [30] claimed that other tools are able to provide better predictions than TMHMM for this task.

After identifying the transporter candidate genes, similarity searches are performed, comparing the proteins encoded in such genes with the ones available in TCDB. The similarity between sequences is calculated using the dynamic programming based algorithm SW for local alignments, guaranteeing optimality and high sensitivity when looking for homologous sequences. These alignments are performed to identify proteins with sequence similarities to known transport systems. The similarity threshold for considering the homology was of 10 percent of the maximum alignment score. However, as TCDB is a very small database (11622 records as of June 2014) a heuristic method was used to lower the similarity threshold, whenever a TCG has at least 5 transmembrane helices. In these cases, the evidence for a transporter role is stronger, justifying the special case. For each extra transmembrane helix, the similarity threshold was lowered by 0.5 percent until a minimum of half the initial similarity threshold is reached (5 percent in this case). Nonetheless, the initial similarity threshold can be beforehand set by the user.

This TMHMM/TCDB/SW coupled strategy allows identifying and annotating different types of transport proteins located in membranes. Table B.1 of Appendix B, available in the online supplementary material, shows the result of the alignment of a *K. lactis* gene (KLLA0B00264g) with TCDB. This putative membrane transport system has 10 transmembrane helices (according to TMHMM), thus the similarity threshold was set to 7.5 percent. If this heuristic was not used, this gene would only have 17 similar genes in TCDB (instead of 47). Glycerol is associated with that membrane transport systems with a final classification score (see the details below) of 0.38. However, if the initial similarity threshold of 10 percent was used, glycerol would not be included in the list of transported metabolites because the TCDB homologous gene that transports glycerol with the highest similarity to that gene has a similarity of only 8.28 percent. This gene's function is yet unknown, thus the annotation proposed by TRIAGE cannot be assessed. Nevertheless, these predictions can be used to decrease the number of experiments required to determine the gene's role.

After running the SW algorithm and retrieving the information for the most similar genes in TCDB, a routine is then used to select which metabolites will actually be assigned to each gene g, weighting the number of times each metabolite m is found within the homologous gene records (frequency), and the taxonomy of the organisms appearing in those records. This methodology builds on the assumption that related organisms will thrive in similar environments more often than dissimilar organisms, thus having closer uptake and secretion requirements. The following equation describes how this process is performed.

$$score(g)_m = \alpha.score_{frequency} + (1-\alpha).score_{taxonomy}. \quad (2)$$

The balance between the frequency score and the taxonomy score is given by a parameter α. The frequency score, which calculates the number of occurrences of a metabolite m within all TCDB similar records for that gene, is given by:

$$score_{frequency} = \frac{\sum_{i=1}^{H} s_i \times Vm_i}{\sum_{i=1}^{H} s_i}. \quad (3)$$

This score is obtained by summing up the similarities of each homologous TCDB gene that transports the metabolite m and dividing by the sum of the similarities of all homologous genes, no matter what they transport. In this notation, s_i is the similarity of the gene g with the ith record in TCDB

calculated using SW, H is the total number of hits for the gene, and Vm_i is a binary variable described as:

$$Vm_i = \begin{cases} 1, & \text{if metabolite } m \text{ is in record } i \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The frequency score considers both lateral and vertical gene transfer, whereas the taxonomy score is put forward to favour homologies of TCGs with TCDB records of closely related organisms. The latter is calculated as shown below.

$$score_{taxonomy} = \frac{\sum_{i=1}^{H} t_i \times Vm_i \times (1 - p_m \times \beta)}{M_T \times \sum_{i=1}^{H} Vm_i}. \quad (5)$$

In the numerator, the taxonomy frequency (sum of the number of common taxa between the organism being studied and the one in the TCDB record, over all hits) is multiplied by a penalty. This is used to penalize the score for metabolites that are associated to a low number of similar genes and may be the result of incorrect assignments (β is a penalty parameter, default value of 0.05). The penalty factor (pm) is calculated by subtracting the frequency of the genes that transport a given metabolite from a user defined minimal number of hits (set to 2 in the example given in Appendix B, available in the online supplementary material). If this subtraction is positive, it is multiplied by β and subtracted to 1, otherwise the penalty is zero. In the denominator, the maximum taxonomy (MT) value (number of taxa of the target organism) is multiplied by the frequency of the genes that transport the metabolite. In this notation t_i is the number of common taxa between the organism to which record i belongs and the target organism. The calculation of the metabolite penalty is described in equation (6). The default values of all parameters are shown in (Fig. A.2 of Appendix A, available in the online supplementary material).

$$p_m = \begin{cases} 0, & \text{if } \sum_{i=1}^{H} Vm_i \geq Min_{Hits} \\ Min_{Hits} - \sum_{i=1}^{H} Vm_i, & \text{otherwise.} \end{cases} \quad (6)$$

Table B.2 of Appendix B, available in the online supplementary material, describes the process of determining which metabolites will be assigned to gene KLLA0B00264g (*K. lactis*). The example shows the steps for assessing the scores for glucose, lactose, myo-inositol and glycerol. Table B.1 of Appendix B, available in the online supplementary material, shows in bold all the taxa that each TCDB hit has in common with the *K. lactis* case study gene (KLLA0B00264g). As shown in Tables B.1 and B.2 of Appendix B, available in the online supplementary material, *S. cerevisiae* has 8 taxa in common with *K. lactis*. On the other hand, the *Homo sapiens* homologue only has 1 taxon in common and Bacteria have none. The taxonomy frequency sum for D-glucose is calculated by adding all the common taxa count for the TCDB records (Table B.2 of Appendix B, available in the online supplementary material, highlighted in blue) and the final result is 119. The maximum taxonomy frequency is 10 (result obtained by counting all the *K. lactis* taxa) which will be multiplied by 26 records associated to the transport of D-glucose. This metabolite's frequency is greater than the minimum required; therefore, there will be no penalty applied. On the other hand, lactose (Table B.2 of Appendix B, available in the online supplementary material, highlighted in green) is available just one time and it will have a frequency penalty of 5 percent. The taxonomy score for D-glucose is 0.46. The final D-glucose score, for $\alpha = 0.3$, is 0.5. If α was set to 0.4 it would be 0.51. Lactose has a score of 0.67 ($\alpha = 0.3$) and for $\alpha = 0.4$ the score would be 0.58.

The same metabolite can be transported by several types of carriers, such as uniport, symport or antiport. The algorithm developed for metabolite classification was also used to classify how a metabolite is transported. In the previous example, glucose can be transported by symport (Table B.2 of Appendix B, available in the online supplementary material, emphasized in light red) or uniport (light blue). The final score for D-glucose transport by symport is 0.40 and the score for uniport is 0.49; thus, uniport will be selected by TRIAGE. If the scores were equal, both types of transport would have been selected. The user can set a list of symport currency metabolites to prevent creating excessive transport reactions (Fig. A.2 of Appendix A, available in the online supplementary material). In this work H^+ (C00080) and Na^+ (C01330) were set as currency metabolites.

It is possible that genes being classified by TRIAGE have records in TCDB and, consequently, a similarity score of 1 in the SW alignments. Nevertheless, such genes may also have similarity to other genes in the transporters classification database. Thus, all the hits are used to classify the metabolites to be transported by those genes. It is assumed that TCDB associates the transport of specific metabolites to genes according to published experiments. However, those carriers may also be able to transport other metabolites, specifically metabolites carried by similar transport systems untested in such experiments.

2.3 Prediction of the Subcellular Localization of Proteins

WoLF PSORT [31] and PSORTb 3.0 [32] were used to assign sub-cellular localizations to the identified transporters. The first was chosen because it has been reported as the best eukaryotic protein subcellular localization prediction tool in the literature [33], [34], [35], while PSORTb 3.0 is the next generation of the PSORTb tools, which continues to be the most widely employed localization prediction software for bacteria [36].

Although PSORTb 3.0 does not provide a web API for accessing the compartmentalization data, there are two approaches for retrieving these data. PSORTb 3.0 offers precomputed genome results, for genomes deposited in GenBank. These data can be retrieved from the PSORTdb database at http://db.psort.org/browse. On the other hand, if the genome in question is not available in the precomputed results, the target genome sequence files, in the FASTA format, should be submitted to the PSORTb 3.0 HTML interface. However, the maximum size allowed for submission is 100 Kb; therefore, some files may have to be split.

For WoLF PSORT (http://www.wolfpsort.org) it was possible to use a simple remote Java API from a Java archive (jar) provided by Paul Horton, in a personal communication, where it was also indicated that "intracellular organelle membranes, say mitochondrial or E.R., are lumped together with soluble proteins in their organelle".

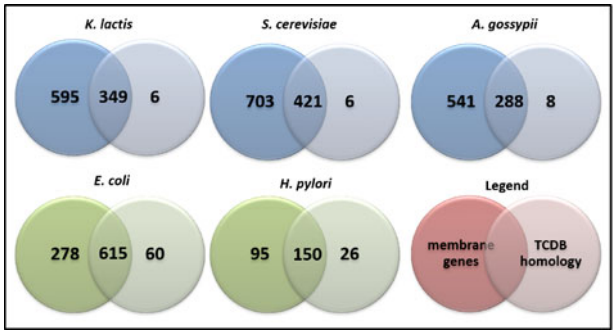


Fig. 2. Cross linking the information from protein localization and the identification of transporter candidate genes. The number of genes, classified as transporters is represented by the intersection of the genes that have similarities to TCDB records (after checking for transmembrane domains with TMHMM) and the genes with a localization prediction within an external membrane. The Fungi are represented in blue and the Bacteria are represented in green.

Secondary locations are also considered for protein subcellular localization. If any compartment(s) has(have) a score that differs less than 10 percent from the main location score, such compartment(s) is(are) also taken into account when generating transport reactions.

2.4 Automatic Assembly of Transport Reactions

After metabolite identification and transport type selection, as well as their localization in the cell, TRIAGE automatically generates the transport reactions using a few heuristic rules. If a metabolite is transported by antiport or symport by a carrier encoded in a given gene, then co-transported metabolites are used to assemble reactions. For example, if symport was selected for the previous example, a reaction including all the metabolites that are co-transported by symport with D-glucose (in this case, just the proton) would be generated and proposed to be integrated in the GSMM. This reaction is described below.

$$D\text{-}glucose_{extr} + H^{+}_{extr} <=> D\text{-}glucose_{cyto} + H^{+}_{cyto}. \quad (7)$$

For uniport the reaction is:

$$D\text{-}glucose_{extr} <=> D\text{-}glucose_{cyto}. \quad (8)$$

Moreover, the descendants of the selected metabolites are also used to generate similar reactions. Therefore, from D-glucose alone, three transport reactions will be generated and associated to this *K. lactis* gene (KLLA0B00264g). This example does not involve any source of energy to drive the transport of D-glucose. However, other types of carriers implicate energy requirements, as is the case of the P-type ATPase Superfamily transport proteins that use energy from ATP hydrolysis to transport a metabolite across a membrane. If a gene has similarities with genes of this family, that gene will be associated with ATP, ADP and P_i. The reacting metabolites are treated as currency metabolites, not being scored. A system for the assignment of partial TC numbers to these genes was also developed and is available in Appendix C, available in the online supplementary material. The generated transport reactions for several thresholds, provided by TRIAGE for the gene in the example, are displayed in Table B.3 of Appendix B, available in the online supplementary material. As expected, there is an inverse correlation between the score threshold and the number of reactions annotated to a specific gene.

TABLE 1
Number of Potential Transport Systems Encoding Genes in Each of the Studied Genomes

Organism	Nr. of Genes	TCGs from TMHMM	TCGs with TCDB Hits
K. lactis	5,085	967	355
S .cerevisiae	5,882	1,144	427
A. gossypii	4,726	860	296
E. coli	4,146	1,039	675
H. pylori	1,590	330	176

Genes having transmembrane domains and similarities to TCDB records.

To conclude, the confirmation of a TCG as a membrane transport system by TRIAGE involves meeting three criteria. The first is to have transmembrane domains. The second is to have similarities with TCDB records. The third is to have a localization prediction within a membrane: inner membrane or outer membrane for prokaryotes and plasma membrane for eukaryotes. However, since intracellular membranes are lumped with the intracellular organelle predictions, it was decided that if a TCG met the first two requirements, thus having strong evidences of being a transporter, and WoLF PSORT assigned an intracellular organelle to such TCG, it would be considered that the TCG was located in the organelle membrane, encoding an intra-cellular transport system.

3 Results and Discussion

3.1 Transporter Annotation

TRIAGE was used to identify genes encoding transport systems and to generate transport reactions for the case study organisms (*K. lactis*, *A. gossypii*, *S. cerevisiae*, *H. pylori*, and *E. coli*).

3.1.1 Transporters Candidate Genes

The main results are provided in Fig. 2 and Tables 1 and 2. The former table represents the number of genes that fulfil criterion 1 (having transmembrane domains - TCGs from TMHMM) and 2 (having similarities to TCDB records - TCGs with TCDB Hits). The latter table contains the number of genes that fulfil the third criterion (location on a membrane) independently of the other 2 criteria. Fig. 2 shows the intersection of the last column of both tables, i.e., the genes that fulfill all three criteria.

TABLE 2
Number of Genes Predicted to Encode Proteins Localized in Each of the Membranes

Organism	External	Internal	Total
K. lactis	663	475	944
S. cerevisiae	805	576	1,124
A. gossypii	625	403	829
E. coli	25	869	893
H. pylori	9	237	245

The last column does not represent the sum of the preceding ones, since a gene can be associated with multiple membranes.

TABLE 3
Number of Transport Systems Encoding Genes Associated to Transport Reactions for Different Score Thresholds

Organism	Threshold				
	0.0	0.1	0.2	0.3	0.4
K. lactis	324	324	317	303	262
S. cerevisiae	392	387	375	364	323
A. gossypii	268	266	254	243	208
E. coli	559	555	539	523	490
H. pylori	137	136	127	96	61

The default values used for the scorer parameters α and β were 0.3 and 0.05, respectively.

TABLE 4
Number of Transport Reactions Provided by TRIAGE

Organism	Threshold				
	0.0	0.1	0.2	0.3	0.4
K. lactis	14,322	9,188	8,217	6,853	5,310
S. cerevisiae	18,038	12,187	9,604	8,863	6,974
A. gossypii	13,051	8,087	7,163	5,484	4,867
E. coli	18,740	15,868	11,339	10,152	8,470
H. pylori	14,667	12,700	7,622	2,789	488

This table shows the number of reactions created by TRIAGE in which all metabolites have KEGG identifiers. The default values used for the scorer parameters α and β were 0.3 and 0.05, respectively.

Table 1 and Fig. 2 clearly show that, as expected, increasing the stringency of the conditions that a given gene has to meet reduces the number of TCG's annotated as carriers. In all of the studied organisms, 18 to 20 percent of the genes were identified as TCGs by TMHMM, except for *E. coli* (with 25 percent). These results are consistent with previous studies [37], [38] that suggest that approximately $20 - 30$ percent of all proteins in the genome are predicted to encode membrane proteins. In Fungi 34 to 37 percent of the TCG's have similarities with TCDB entries, whereas *H. pylori* and *E. coli* have more than half of the TCG's with homology to TCDB records. These results suggest that TCDB has more entries for prokaryotes than for lower eukaryotes.

The number of genes predicted to encode proteins localized in the different membranes, according to PSORTb and WoLF PSORT, is indicated in Table 2. Fungi have a nearly constant ratio (13 percent) of genes predicted to encode proteins localized in the plasma membrane. Bacterial predictions are somewhat different. *H. pylori* is predicted to have less than 1 percent of the genes encoding outer membrane proteins and 15 percent encoding cytoplasmic membrane proteins. Likewise, *E. coli* is expected to have less than 1 percent of genes encoding outer membrane proteins and 21 percent encoding cytoplasmic membrane proteins.

Nevertheless, more than half of the genes predicted by TMHMM to encode membrane proteins have similarities to genes available in the TCDB, except for *A. gossypii* and *H. pylori*. The high number of putative membrane protein encoding genes without similarities to TCDB is probably due to the reduced number of entries still available in TCDB. As shown in Fig. 2, only a small number of genes not predicted to be localized in membranes have transmembrane helices and similarities to the TCDB genes.

The fact that there are about 10 times more TCGs between periplasm and the cytoplasm than between the exterior and the periplasm can be, at least in part, justified by limitations in the methodologies used. Although plasma membrane proteins consist of mostly transmembrane α-helices, outer membrane proteins are typically composed of β- strands in which barrel sizes are associated with different functions. The latter comprise active ion transporters for nutrient uptake, membrane anchors, membrane-bound enzymes and protection against pathogenic proteins [38]. However, TMHMM only predicts transmembrane helices, impairing the predictions for the outer membrane.

As shown in Fig. 2, only a small number of genes not predicted to be localized in membranes have transmembrane helices and similarities to the TCDB genes. On the other hand the number of membrane protein encoding genes without similarities in TCDB is well over a half for eukaryotes and about a third for prokaryotes.

The number of TCGs effectively associated to reactions by TRIAGE is shown in Table 3. In this table it is shown that, as expected, increasing the transport reactions score threshold decreases the number of TCGs associated to transport reactions. By comparing with the results of the intersections shown in Fig. 2, it is clear that in none of the case studies the full set of TCG's is used to generate transport reactions. This means that such genes have similarities with records poorly annotated in TCDB and hence our internal database.

3.1.2 TRIAGE's Internal Transport Reactions Database

The database of Transport Reactions, compiled throughout this work, contains information for 5,495 TCDB records, which were associated to 1,053 distinct primary metabolites. These were submitted to the metabolite names annotation pipeline to be assigned with KEGG and ChEBI identifiers. For the majority (603) both identifiers were retrieved, while 224 were not assigned with any identifier, 220 were only assigned with ChEBI identifiers and six were only assigned with KEGG identifiers. Moreover, 29,034 second-generation metabolites (of which 6,964 also had KEGG identifiers) were retrieved from the 823 primary metabolites with ChEBI identifiers. The 829 (603+6+220) metabolites that had at least one database identifier were used to generate 2,491 main (1^{st} generation) reactions associated with the respective proteins in TCDB.

3.1.3 Transport Reactions Assembly

The connection between the membrane transport systems and the metabolites to be carried was performed by using the routine described above and empirically setting the default threshold of the overall score to 0.2 (values ranging from 0.0 to 0.4 are also shown in Table 4).

Table 4 presents the total number of reactions in which all metabolites have KEGG identifiers, for each organism. The consequences of retrieving all the second-generation metabolites from the ChEBI ontology, so that all possible substrates of a given transporter could be found, are well demonstrated. Transporters are often able to transport various substrates [13], [39], and multiple transporters may exist for one specific substrate [39]. Thus, TRIAGE mimics this behaviour, proposing that a transporter can transport

similar compounds and that different transporters may carry the same compound using dissimilar mechanisms.

The high number of transport reactions is thus explained by several facts: i) for each second-generation metabolite of a given ChEBI entity, a transport reaction similar to the reaction of the parent metabolite is created; ii) moreover, each metabolite (and its second-generation elements) may be transported through several transport systems (uniport, symport, etc.), depending on the classification of each gene associated with that metabolite; and, iii) the same reaction on different membranes is regarded as distinct reactions. Yet, if the reactions in which metabolites without KEGG identifiers (some metabolites were only assigned with ChEBI identifiers) were shown in this table, the number of reactions would increase even further.

Still, regardless of the number of reactions generated, another filtering process is performed when integrating this information with the GSMMs. Only transport reactions in which all metabolites are already present in the genome-scale models are selected for integration. Moreover, when performing simulations, those reactions will only be active in the compartments that hold the reactants (or products) of the transport reaction. Worth mentioning is the fact that increasing the threshold decreases the number of reactions (Table 4) significantly more than the number of genes that encode reactions is reduced (Table 3). Thus, a threshold of 0.2 is recommended for scoring, according to Table 4 and Table B.4 of Appendix B, available in the online supplementary material, as we think it provides reasonable results. A threshold of zero would include all available reactions for each gene, while an increase from 0.2 to 0.4 can be too restrictive for less characterized organisms, as shown in the case of *H. pylori*.

3.2 Integration with Genome–Scale Models

TRIAGE was not compared to other tools as we could not find any other tools that generated transport reactions for annotating proteins in the mentioned case studies. Instead the results obtained with TRIAGE were compared with metabolic models of *E. coli* and *S. cerevisiae*, iAF1260 (406 genes for 718 transport reactions) and iMM904 (202 genes for 409 transport reactions), respectively.

3.2.1 Genes Integration

To gather the genes encoding transport systems in a model, the rule was the following: if a given reaction in the model has substrates and products in different compartments, the reaction is regarded as of transport and the genes associated with that reaction are considered genes encoding membrane transport systems. TRIAGE was executed using the default parameters for the similarity alignments and metabolite scoring.

It is not surprising that TRIAGE annotates a number of genes with transport functions larger than published models. According to Fig. 2, *E. coli* and *S. cerevisiae* have 618 and 407 transporter genes, respectively. Nevertheless, as shown in Table B.5 of Appendix B, available in the online supplementary material, there are 65 genes for *S. cerevisiae* and, surprisingly, 190 for *E. coli* assigned with transport reactions in the models not identified by TRIAGE.

Analyzing the results further, about 22 percent (14) of the unidentified yeast model's genes are not identified as transporters because WoLF PSORT predicted such genes to be expressed in the cytoplasm, thus not complying with one of the criteria of this approach. Moreover, 68 percent (44) of such genes were not annotated as carriers because TMHMM did not predict any transmembrane region and 7 percent (5) of those genes did not have similarities with TCDB. The remaining 3 percent (two) genes were associated to transport reactions with scores below the threshold.

For *E. coli*, as shown in Table B.5 of Appendix B, available in the online supplementary material, PSORTb 3.0 indicated approximately 32 percent (61) of the unidentified carriers as periplasmic, cytoplasmic or in one unknown location. Furthermore, 51 percent (97) of the genes classified as carriers in the model and assigned to the cytoplasmic membrane or outer membrane by PSORTb 3.0, could not be identified as membrane transporters, as TMHMM did not predict any transmembrane region in these genes. Finally, 7 percent (13) of the remaining genes did not have any similarity with TCDB records. The remaining 10 percent of the genes (19) could not be associated to transport reactions with scores above the threshold.

On the other hand, as shown in Table B.5 of Appendix B, available in the online supplementary material, TRIAGE proposes 228 new carrier genes for *E. coli* and 148 for yeast. These predictions should be confirmed with wet-lab experiments and their impact verified through phenotype simulations using GSMMs.

3.2.2 Transport Reactions Integration

The reactions generated by TRIAGE were also integrated in the GSMMs. For that purpose, the pipeline used for performing metabolites identification (Fig. 1) was also used to assign KEGG identifiers to the metabolites in both models. For the yeast model 674 out of 713 metabolites (ignoring location) were assigned with identifiers. For *E. coli*, 883 within 1039 metabolites (ignoring location) were identified. The remaining metabolites are either absent from the databases or, although present, their labels in the models are not similar to the names available in the selected databases. For instance, 4-methylzymosterol, (identified in iMM904 as M_4mzym_c) does not exist in KEGG, thus it cannot be assigned with a KEGG identifier. Yet, KEGG has a variation of this metabolite (4alpha-methylzymosterol), which was not identified by the algorithm developed in this work.

TRIAGE's reactions were then filtered to keep only reactions that included metabolites present in the models. In these models, metabolites in different compartments are regarded as different entities, called species. The same metabolite in different compartments are two distinct metabolic species. For instance, glucose in the cytoplasm is a species and extracellular glucose is another species. Hence, if glucose is available in the cytoplasm (or periplasm for some prokaryotes) but not in the external compartment of the model, relaxing the integration criteria will allow adding extracellular glucose to the model, thus adding a new transport reaction to the model. These reactions can be useful for extending the model, in order to enable predictions in different environmental conditions.

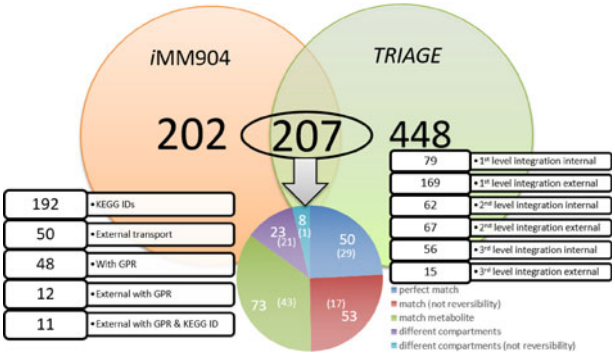


Fig. 3. Comparison of the results for transport reactions obtained with TRIAGE and from the iMM904 GSMM for *S. cerevisiae*. The above figure represents the intersection of the results of TRIAGE and the iMM904 model. The pie chart classifies the intersection results within three classes. The numbers between brackets in the pie chart represent the number of reactions that had no gene assigned in the model but were assigned to a gene by TRIAGE. The block list displays some properties of the reaction set, to which it is connected to, that help understanding the mispredictions.

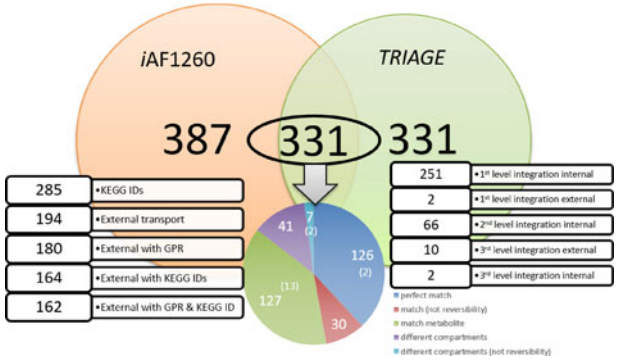


Fig. 4. Comparison of the results for transport reactions obtained with TRIAGE and from the iAF1260 GSMM model for *E. coli*. The above figure represents the intersection of the results of TRIAGE and the iAF1260 model. The pie chart classifies the intersection results within three classes. The numbers between brackets in the pie chart represent the number of reactions that had no gene assigned in the model but were assigned to a gene by TRIAGE. The block list displays some properties of the reaction set, to which it is connected to, that help understanding the mispredictions.

We considered three different types of transport reactions regarding their integration with the model. If all metabolic species for a given reaction are available in a model, it is considered a 1^{st} level reaction. These reactions can be directly added to the model not demanding any changes in the metabolites. When the reaction contains at least one metabolic species unavailable in the model (i.e., all metabolites are present in the model, but at least one is not present in the same compartment), we call it a 2^{nd} level reaction. If all metabolic species are absent from the model (i.e., all metabolites are present in the model, but none is in the same compartment), the reaction is considered of the 3^{rd} level. Thus, 2^{nd} and 3^{rd} level reactions can only be integrated in the model, if the integration criterion is relaxed, allowing the addition of new metabolic species to the model.

Score threshold values ranging from 0.0 to 0.4 and similarity threshold values ranging from 0.05 to 0.3 were analyzed for *S. cerevisiae* and *E. coli*, and results are shown in Table B.4 of Appendix B, available in the online supplementary material. In this table, it is shown that lowering the overall score threshold increases the number of perfect matches between TRIAGE's reactions and the models reactions. Likewise, decreasing the similarity score threshold also rises the number of perfect matches between the tool and the model, except for a few cases in which the taxonomy of the TCDB hits lowers the overall score as shown in Table B.4.1 of Appendix B, available in the online supplementary material. The number of new integratable reactions also increases for lower thresholds. Results for a score threshold of 0.2 and a similarity threshold of 0.1 are further discussed.

As depicted in Fig. 3 (and Table B.4 of Appendix B, available in the online supplementary material), 655 reactions were selected from TRIAGE to be integrated in the yeast GSMM. Most of these transport reactions (448) are not available in the model, thus being new reactions. For the iMM904 model, loosening the criteria allows adding 200 (2^{nd} and 3^{rd} level) new transport reactions to the model. Moreover, 56 percent of the new reactions are associated to the transport of metabolites between the external compartment and the cytoplasm. This model has originally 409 transport reactions, 155 of which carry metabolites from the exterior to the cytoplasm (data not shown). TRIAGE was able to obtain a recall of 51 percent (207) over all model reactions. From those, 50 reactions fully matched the ones in the model, including 29 that were not assigned to any genes in the GSMM and, therefore, can now be assigned with a gene-reaction rule. Also, 52 model reactions were matched by TRIAGE (assigning new genes to 17), but in this case the reversibility was different. Moreover, 73 reactions corresponding to metabolites for which TRIAGE predicted different transport mechanisms (43 of which had no gene assigned in the model) were found. For example, the model has a transport reaction for L-Asparagine by proton symport, but TRIAGE chose uniport. Overall, TRIAGE was able to assign GPRs to a total of 111 non-gene associated transport reactions in the iMM904 model.

Due to several reasons, 202 transport reactions in the model were not matched in this work. Most of those reactions (over 75 percent) are associated to internal transport reactions. Although TRIAGE can predict intracellular reactions, only four internal compartments were selected for yeast, because transport reactions to other compartments were associated to metabolites not existing in such compartments; thus, these reactions were ignored. Also, 76 percent (154) of the 202 reactions not matched by TRIAGE were not associated to any gene and thus these reactions were probably added without genomic evidences. Only 11 external transport reactions with gene associations, for metabolites identified with KEGG identifiers in the model were not matched, giving an accuracy of 86 percent for external exchange reactions with gene associations and metabolites with KEGG identifiers.

E. coli's model contains 718 transport reactions, 299 of which carry metabolites from the exterior to the periplasm. As shown in Fig. 4 (and Table B.4 of Appendix B, available in the online supplementary material), 662 reactions were selected to be integrated in the model.

Half of these transport reactions (331) are not available in the model, thus being new reactions. Loosening the integration criteria allows adding 78 2^{nd} and 3^{rd} reactions most of which (75) internal as, again, there are very few genes

assigned to the outer membranes. Indeed, 96 percent of the new reactions (1^{st}, 2^{nd} and 3^{rd}) are associated to the transport of metabolites between the periplasm and the cytosol (319) and only four reactions between the external compartment and the periplasm.

In this case, TRIAGE only obtained a recall of 46 percent (331) of the reactions in the GSMM. TRIAGE perfectly matched 126 reactions, matched the reaction except for the reversibility on 30 occasions, and for 129 reactions the metabolite to be transported was matched in a reaction different than the available in the model. This model has a smaller number of reactions without gene associations, but still a total of 17 transport reactions without GPRs in the iAF1260 model were associated to genes by TRIAGE. There is a large number of reactions in the model not matched by TRIAGE, due to several reasons. For instance, 102 reactions were associated to metabolites without KEGG identifiers, thus such reactions could not be matched. Moreover, 166 of the 285 reactions in which metabolites had KEGG identifiers, were associated to reactions in the outer membrane. However, as shown in Table2, PSORTb 3.0 predicted that only 85 genes in the entire genome encode proteins in that location, while for the cytoplasmic membrane the number of genes is 12 fold higher. Finally, only 103 reactions for transporting metabolites with KEGG identifiers, in the cytoplasmic membrane with gene associations (out of 315 –) were not matched (data not shown), including several chemically complex transport reactions that could not be directly compared to the generated reactions. Therefore, the accuracy was about 68 percent for transport reactions through the cytoplasmic membrane, with gene associations and metabolites with KEGG identifiers.

3.2.3 Integration Outcome

Although thousands of reactions were made available by TRIAGE (as shown in Table 4), only a few hundred were integrated in the model. As stated earlier, several transport reactions may exist for the same metabolite, like several antiport reactions with various counter-transported substrates, etc. Moreover, specific variations of the same metabolite (for instance, α and β conformations) are also included as different reactions in this set, though often only one of the variations is available in the model. Therefore, it was expected that the number of reactions that can be integrated in the model is in the order of the hundreds.

Nevertheless, TRIAGE allows identifying and assigning functions to genes previously unannotated or new functions to annotated genes. For example *E. coli*'s b3876 gene has a putative annotation in UniProt and is unavailable in TCDB. TRIAGE associated this gene to the symport of melibiose as shown in equations (9) and (10):

$$melib_{perip} + h_{perip} => melib_{cyto} + h_{cyto}, \quad (9)$$

$$melib_{perip} + na1_{perip} => melib_{cyto} + na1_{cyto}. \quad (10)$$

For *S. cerevisiae*, the YBR220C gene was annotated as a peptide-Acetyl-Coenzyme A antiporter as in equation (11):

$$CoA_{cyto} + AcCoA_{extr} <=> CoA_{extr} + AcCoA_{cyto}. \quad (11)$$

using TRIAGE, and such gene was previously annotated as a protein of unknown function. Furthermore, TRIAGE allows assigning new genes to existing transport reactions in both *E. coli* and *S. cerevisiae* models, thus proving to be valuable when reconstructing metabolic models. Also, TRIAGE generated alternative reactions for transporting metabolites already available in the model. Finally, relaxing the integration criteria allows adding several new transport reactions, as shown in Table B.6 of Appendix B, available in the online supplementary material. These reactions should be carefully reviewed, as adding new species to the models may impair predictions.

TABLE 5
Association between the Presence of Transport Reactions and Growth in Different Carbon Sources

Phenotype	*S. cerevisiae*		*E. coli*	
	iMM904	TRIAGE	iAF1260	TRIAGE
Growth	53%	53%	87%	26% (83%)
No Growth	33%	44%	26%	11% (52%)

This table represents the comparison of the percentage of existing transport reactions for carbon sources in which the organism is known to exhibit or lack growth. In brackets are the results obtained for the metabolites transported between the periplasm and the cytoplasm.

3.3 Carbon Sources Assessment

According to CBS-KNAW Fungal Biodiversity Centre (http://www.cbs.knaw.nl/) and EcoCyc (http://ecocyc.org), *S. cerevisiae* and *E. coli* are able to grow in a broad number of carbon sources. As shown in Table B.7 of Appendix B, available in the online supplementary material, considering that the presence of a transport reaction promotes growth on a given carbon source, the following results were generated. Table 5 shows that TRIAGE provides reactions for carrying 53 percent (10 out of 19) of the carbon sources in which this yeast is known to thrive.

The iMM904 model also provides these reactions for 53 percent (10 out of 19) of the carbon sources. Yet, only half of these reactions are associated to genes in the model. TRIAGE proposes more reactions for the carbon sources in which the *in vivo* growth was not verified: 44 percent (16 in 36) against 33 percent - 12 out of 36, of which gene associations are only provided for 2 reactions. As previously stated, all reactions provided by TRIAGE are associated to genes. The high number of false positives provided both by TRIAGE and the model may be associated with limitations in the experimental conditions used. Although this yeast cannot grow using some metabolites as sole carbon sources in the conditions of the experiments, this does not mean that the yeast cannot metabolize these compounds in other experimental conditions. Moreover, the organisms can possess the transport reactions for a given metabolite and still not be able to metabolize it due to the lack of specific enzymes. This may be due to the non-specificity of the transporters, as well as to the loss of metabolic functions along evolution.

For *E. coli*, TRIAGE predictions are, once again, impaired by the number of proteins predicted to be located in the outer membrane. Only carriers for 6 (out of 23) carbon sources are identified by TRIAGE in this membrane. However, searching for transporters of the same metabolites, but from the periplasm to the cytoplasm, the numbers are quite

different and the percentage of carbon sources transported by *E. coli* increases over three fold to 83 percent. The model has transport reactions for 87 percent (20 in 23) of the carbon sources in which growth is confirmed. The number of carriers for carbon sources, in which this organism cannot attain *in vivo* growth, are 26 percent (seven in 27) for the model and 11 percent (or 52 percent if transport from the periplasm is accounted for) in TRIAGE. All reactions available in the model for the carbon sources listed in Table B.7 of Appendix B, available in the online supplementary material, (20+7) are associated to genes. However, these reactions are associated to four genes (b0929, b2215, b1377, b0241) except for reactions that transport glucose and maltose, which are associated to gene b4036. None of these genes has transmembrane helices (all of them have β-strands), thus these genes were discarded by TRIAGE.

4 Conclusions

TRIAGE, a new methodology for identifying membrane transport systems and automatically generating transport reactions for every metabolite transported by those carriers, is proposed in this work. These reactions can be directly integrated with GSMMs since all metabolites involved have KEGG and/or ChEBI identifiers. TRIAGE combines several tools to obtain more reliable results, minimizing the possibility of adding incorrect transporters as the pipeline for TCG identification is very stringent. This implies that a negative prediction in one of the modules will exclude the gene of the membrane transport systems encoding genes set.

TRIAGE is the only tool able to generate transport reactions associated to genes based in the organisms genome sequence and to sort such reactions in a way that these can be immediately integrated into metabolic models, both for internal and external transporters. Toolboxes like RAVEN [40] (which allows adding transport reactions manually) or SuBliMinaL [41] (which includes a tool that allows adding a default standard set of transporters) also try to approach this problematic, yet without relying on genomic information. ModelSEED [4], KBASE (http://kbase.us/) and Pathway Tools [42] all perform the annotation of transporters, the first two using the RAST annotations to develop models. These tools are targeted at prokaryotic organisms adding spontaneous reactions to fill in pathways when necessary, whilst external transport reactions can also be added based on information from the genome. However, to our knowledge, TRIAGE is currently the only tool annotating transport proteins, while simultaneously generating transport reactions for eukaryotic organisms.

The first step performed by TRIAGE, transmembrane domain identification is crucial, because if TMHMM does not predict transmembrane helices, the gene is excluded. In the future, the authors intend to consider more elaborate methods and tools for this first step, taking more information into account, such as the inclusion of methods that can predict β-strands in protein sequences since the identification of these structures might improve the annotation of outer membrane transporters encoding genes from gram negative bacteria, and considering the use of machine learning approaches, similar to those already used for protein localization methods.

WoLF PSORT and PSORTb 3.0 also have a prominent role in the transport systems assignment. A wrong compartment prediction will also exclude the gene of the carriers set. Being such an important step, though the best methodologies published to date are being used, the study of alternative databases and tools will be considered.

Overall, further tests will be performed to try to relax some of the strict rules that TRIAGE uses, if it is proven that false negative rates are high.

Furthermore, although all TCDB records are linked to literature references and thus experimental data are directly used, these data could not be used to validate TRIAGE. When studying a genome, the transport protein encoding genes available in TCDB are annotated to that genome. As such, a direct comparison between the experimentally validated TCGs and the ones predicted by TRIAGE would be too biased.

The results for the integration of the data provided by TRIAGE with curated models (continuously curated for over 10 years) are quite acceptable, including the association of transport reactions to genes that were previously not annotated, providing gene-reaction associations required for several simulations, and the identification of new reactions that can be added to the existing models. TRIAGE was able to provide uptake transport reactions for metabolites that the yeast can use as sole carbon sources, not previously identified in the model. For *E. coli* the results concerning carbon sources were impaired by the lack of a tool that could predict genes with β-strands. Nonetheless, the results (considering transport between the periplasm and cytosol) are fairly satisfactory.

Also, when comparing TRIAGE's results with existing models and data for growth on carbon sources the number of false positives seems more relevant than the number of false negatives, except for the particular case of *E. coli* due to the limitations of TRIAGE in predicting transmembrane proteins in the outer membrane. Nevertheless, when analysing GSMMs, we cannot assume that the coverage of transporters is high, as in most cases a systematic analysis has not been performed to detect transporters as it is done for enzymatic activities. Also, the amount of data available for the carbon sources growth is scarce and, thus, the conclusions should be considered preliminary. Unfortunately, there are not sufficiently large experimental datasets on transporters that can be used to validate TRIAGE and many of the predictions made would need to be experimentally validated to verify TRIAGE's accuracy.

As a final remark, we believe that TRIAGE can be used as a first step of a semi-automated methodology to identify genome-associated transport reactions, which after manual curation can be integrated with existing and under development models, offering reliable results.

Competing Interests

The authors declare no competing interests.

Authors' Contributions

OD conceived the study, created TRIAGE, carried out the transporters annotation, curated the transporters database, analysed integration of TRIAGE outputs with the *E. coli* and *S. cerevisiae* models and drafted the manuscript. DG carried

out the transporters annotation for *Ashbya gossypii*, curated the transporters database and helped to draft the manuscript. MR, EF and IR participated in the TRIAGE's and studies' design and reviewed the manuscript. PV and JC participated in software implementation and in the analysis of the results for the different models. All authors read and approved the manuscript.

Appendix A

Fig. 1.A File with Additional figure in PDF format. Relational database schema of the TRIAGE tool. Fig. 2.A TRIAGE's GUI screenshot.

Appendix B

File with Additional tables in Excel format. Table B.1 – Result of the SW similarity alignment between a single Kluyveromyces lactis gene (KLLA0B00264g) and TCDB. Table B.2 – Transport information retrieved for each of the TCDB homologue genes. Table B.3 – Reactions generated for three different score thresholds (ranging from 0.0 to 0.4). Table B.4 – Assessment of the reactions created by TRIAGE against the iAF1260 and iMM904 for different score and similarity thresholds. Table B.5 – Genes associated with transport reactions in the model not identified in TRIAGE (and vice-versa). Table B.6 – New reactions available for integration in the model. Table B.7 – Comparison of the predictions of the GSMMs and TRIAGE regarding transport reactions for known carbon sources (extended version).

Appendix C

File in PDF format. Description of the methodology for the assignment of partial TC numbers in TRIAGE.

Acknowledgments

Funding: This work was partially supported by a PhD grant (SFRH /BD/47307/2008) and by the ERDF—European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), and National Funds through the FCT within the projects FCOMP-01-0124-FEDER-009707 (HeliSysBio—molecular Systems Biology in Helicobacter pylori) and PTDC/EIA-EIA/115176/2009. The authors would also like to thank the FCT Strategic Project PEst-OE/EQB/LA0023/2013 and the Projects "BioInd - Biotechnology and Bioengineering for improved Industrial and Agro-Food processes", REF. NORTE-07-0124-FEDER-000028 and "PEM – Metabolic Engineering Platform", project number 23060 , both co-funded by the Programa Operacional Regional do Norte (ON.2 – O Novo Norte), QREN, FEDER.

References

[1] I. Rocha, J. Förster, and J. Nielsen. (2008, Jan.). Design and application of genome-scale reconstructed metabolic models, *Methods in Molecular Biology (Clifton, N.J.)* [Online]. *416*, pp. 409–31. Available: http://www.ncbi.nlm.nih.gov/pubmed/18392985

[2] A. M. Feist and B. Ø. Palsson, "The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli," *Nature Biotechnol.*, vol. 26, no. 6, pp. 659–667, Jun. 2008.

[3] I. Rocha, P. Maia, P. Evangelista, P. Vilaça, S. Soares, J. P. Pinto, J. Nielsen, K. R. Patil, E. C. Ferreira, and M. Rocha. (2010, Jan.) OptFlux: An open-source software platform for in silico metabolic engineering, *BMC Syst. Biol.* [Online]. *4(1)*, p. 45. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2864236&tool=pmcentrez&rendertype=abstract.

[4] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. (2010, Sep.). High-throughput generation, optimization and analysis of genome-scale metabolic models," *Nature Biotechnol.*, vol. 28, no. 9 [Online]. *28(9)*, pp. 977–982. Available: http://www.ncbi.nlm.nih.gov/pubmed/20802497.

[5] A. J. Barrett, C. R. Canter, C. Liebecq, G. P. Moss, W. Saenger, N. Sharon, K. F. Tipton, P. Vnetianer, and V. F. G. Vliegenthart, *Enzyme Nomenclature*, NC-ICBMB and E. C. Webb, Eds. San Diego, CA, USA: Academic, 1992.

[6] M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novère, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasić, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttilä, E. Klipp, B. Ø. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen, and D. B. Kell. (2008, Oct.). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology, *Nature Biotechnology* [Online]. *26(10)*, pp. 1155–1160. Available: http://eprints.ma.man.ac.uk/1152/, http://www.ncbi.nlm.nih.gov/pubmed/18846089.

[7] M. L. Mo, B. O. Palsson, and M. J. Herrgård. (2009, Jan.). Connecting extracellular metabolomic measurements to intracellular flux states in yeast, *BMC Syst. Biol.* [Online]. *3(1)*, p. 37. Available: http://www.biomedcentral.com/1752-0509/3/37, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2679711&tool=pm centrez&rendertype=abstract.

[8] O. Dias, R. Pereira, A. K. Gombert, E. C. Ferreira, and I. Rocha, "iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis," *Biotechnol. J.*, vol. 9, no. 6, pp. 776–790, Apr. 2014.

[9] R. Saha, P. F. Suthers, and C. D. Maranas. (2011, Jan.). Zea mays iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism, *PloS One*, [Online]. *6(7)*, p. e21784. Available: http://dx.plos.org/10.1371/journal.pone.0021784, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3131064&tool=pm centrez&rendertype=abstract.

[10] I. Thiele and B. Ø. Palsson. (2010, Jan.). A protocol for generating a high-quality genome-scale metabolic reconstruction, *Nature Protocols*, [Online]. *5(1)*, pp. 93–121. Available: http://www.ncbi.nlm.nih.gov/pubmed/20057383, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125167&tool=pm centrez&rendertype=abstract.

[11] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. (2007.). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information, *Molecular Syst. Biol.* [Online]. *3(121)*, p. 121. Available: http://www.ncbi.nlm.nih.gov/pubmed/17593909.

[12] N. C. Duarte. (2004, Jun.). Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized Genome-scale metabolic model, *Genome Res.* [Online]. *14(7)*, pp. 1298–1309. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=442145&tool=pmcentrez&rendertype=abstract, http://www.genome.org/cgi/doi/10.1101/gr.2250904.

[13] T. J. Lee, I. Paulsen, and P. Karp. (2008, Jul.). Annotation-based inference of transporter function, *Bioinformat. (Oxford, England)* [Online]. *24(13)*, pp. i259–267. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/13/i25 9.

[14] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson. (2009, Feb.). Reconstruction of biochemical networks in microorganisms, *Nature Rev. Microbiol.* [Online]. *7(2)*, pp. 129–143. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119670&tool=pmcentrez&rendertype=abstract.

[15] J. J. Hamilton and J. L. Reed. (2014, Jan.). Software platforms to facilitate reconstructing genome-scale metabolic networks, *Environmental Microbiol.* [Online]. *16(1)*, pp. 49–59. Available: http://www.ncbi.nlm.nih.gov/pubmed/24148076.

[16] O. Dias, M. Rocha, E. C. Ferreira, and I. Rocha. (2015, Apr.). Reconstructing genome-scale metabolic models with Merlin, *Nucleic Acids Res.* [Online]. pp. 1–12. Available: http://www.ncbi.nlm.nih.gov/pubmed/25845595, http://nar.oxfordjournals.org/content/early/2015/04/06/nar.gkv294.abstract?keytype=ref&ijkey= a42xhDf2nONcn2W.

[17] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. (1999, Jan.). KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.* [Online]. 27(1), pp. 29–34. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&too l=pmcentrez&rendertype=abstract.

[18] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. (2008, Jan.). ChEBI: A database and ontology for chemical entities of biological interest, *Nucleic Acids Res.* [Online]. *36*, pp. D344–D350. Available: http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D344.

[19] W. Liebermeister, F. Krause, J. Uhlendorf, T. Lubitz, and E. Klipp. (2009, Apr.). SemanticSBML: A tool for annotating, checking, and merging of biochemical models in SBML format, *Nature Precedings* [Online]. Available: http://precedings.nature.com/documents/3093/version/1, http://precedings.nature.com/doifinder/10.1038/npre.2009.3093.1.

[20] S. Patient, D. Wieser, M. Kleen, E. Kretschmann, M. Jesus Martin, and R. Apweiler. (2008, May). UniProtJAPI: A remote API for accessing UniProt data, *Bioinformatics (Oxford, England)*, [Online]. *24(10)*, pp. 1321–1322. Available: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/10/1321, http://www.ncbi.nlm.nih.gov/pubmed/18390879.

[21] A. Prlić, A. Yates, S. E. Bliven, P. W. Rose, J. Jacobsen, P. V. Troshin, M. Chapman, J. Gao, C. H. Koh, S. Foisy, R. Holland, G. Rimsa, M. L. Heuer, H. Brandstätter-Müller, P. E. Bourne, and S. Willis. (2012, Oct.). BioJava: An open-source framework for bioinformatics in 2012, *Bioinformatics (Oxford, England)*, [Online]. *28 (20)*, pp. 2693–2695. Available: http://bioinformatics.oxfordjournals.org/content/28/20/2693.abstract.

[22] T. F. Smith and M. S. Waterman. (1981, Mar.). Identification of common molecular subsequences, *J. Molecular Biol.* [Online]. *147 (1)*, pp. 195–197. Available: http://linkinghub.elsevier.com/retrieve/pii/0022283681900875, http://www.ncbi.nlm.nih.gov/pubmed/7265238.

[23] M. H. Saier, V. S. Reddy, D. G. Tamang, and A. Västermark. (2014, Jan.). The transporter classification database, *Nucleic Acids Res.*, [Online]. *42*, pp. D251–D258. Available: http://nar.oxfordjournals.org/content/42/D1/D251.abstract.

[24] Q. Ren, K. Chen, and I. T. Paulsen. (2007, Jan.). TransportDB: A comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels, *Nucleic Acids Res.* [Online]. *35*, pp. D274–D279. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1747178&tool=pmcentrez&rendertype=abstract, http://nar.oxfordjournals.org/content/35/suppl_1/D274.full.

[25] M. H. Saier. (2000, Jun.). A functional-phylogenetic classification system for transmembrane solute transporters, *Microbiol. Molecular Biol. Rev. : MMBR* [Online]. *64(2)*, pp. 354–411. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=98997&tool =pmcentrez&rendertype=abstract.

[26] N. Le Novère, A. Finney, M. Hucka, U. S. Bhalla, F. Campagne, J. Collado-Vides, E. J. Crampin, M. Halstead, E. Klipp, P. Mendes, P. Nielsen, H. Sauro, B. Shapiro, J. L. Snoep, H. D. Spence, and B. L. Wanner. (2005, Dec.). Minimum information requested in the annotation of biochemical models (MIRIAM) *Nature Biotechnol.*, vol. 23, no. 12 [Online]. *23(12)*, pp. 1509–1515. Available: http://www.ncbi.nlm.nih.gov/pubmed/16333295.

[27] M. H. Saier, M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan. (2009, Jan.). The transporter classification database: Recent advances *Nucleic Acids Res.* [Online]. *37*, pp. D274–D278. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686586&tool=pmcentrez&rendertype=abstract.

[28] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. (2001, Jan.). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J. Molecular Biol.* [Online]. *305(3)*, pp. 567–580. Available: http://www.ncbi.nlm.nih.gov/pubmed/11152613.

[29] S. Moller, M. D. R. Croning, R. Apweiler, and S. Möller. (2001, Jul.). Evaluation of methods for the prediction of membrane spanning regions, *Bioinformatics* [Online]. *17(7)*, pp. 646–653. Available: http://www.ncbi.nlm.nih.gov/pubmed/11448883, http://bioinformatics.oxfordjournals.org/content/17/7/646.short, http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/17.7.646.

[30] A. Reddy, J. Cho, S. Ling, V. Reddy, M. Shlykov, and M. H. Saier. (2014, Jan.). Reliability of nine programs of topological predictions and their application to integral membrane channel and carrier proteins, *J. Molecular Microbiol. Biotechnol.* [Online]. *24(3)*, pp. 161–190. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4125430&tool=pmcentrez&rendertype=abstract.

[31] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, and K. Nakai. (2007, Jul.). WoLF PSORT: Protein localization predictor *Nucleic Acids Res.* [Online]. *35*, pp. W585–W587. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1933216&tool=pmcentrez&rendertype=abstract.

[32] N. Y. Yu, J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester, L. J. Foster, and F. S. L. Brinkman. (2010, Jul.). PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, *Bioinformatics (Oxford, England)* [Online]. *26(13)*, pp. 1608–1615. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2887053&tool=pmcentrez&rendertype=abstract.

[33] J. Liu, S. Kang, C. Tang, L. B. M. Ellis, and T. Li. (2007, Jan.). Meta-prediction of protein subcellular localization with reduced voting, *Nucleic Acids Res.* [Online]. *35(15)*, p. e96. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1976432&tool=pmcentrez&rendertype=abstract.

[34] E. W. Klee and C. P. Sosa. (2007, Mar.). Computational classification of classically secreted proteins, *Drug Discovery Today* [Online]. *12(5–6)*, pp. 234–240. Available: http://www.ncbi.nlm.nih.gov/pubmed/17331888.

[35] W. Qian and J. Zhang. (2009, Jan.). Protein subcellular relocalization in the evolution of yeast singleton and duplicate genes *Genome Biol. Evolution* [Online]. *1*, pp. 198–204. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817416&tool=pmcentrez&rendertype=abstract.

[36] J. L. Gardy and F. S. L. Brinkman, "Methods for predicting bacterial protein subcellular localization," *Nature Rev. Microbiol.*, vol. 4, no. 10, pp. 741–51, Oct. 2006.

[37] E. Wallin and G. von Heijne. (1998, Apr.). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms, *Protein Science : A Publication of the Protein Society* [Online]. *7(4)*, pp. 1029–1038. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2143985&tool=pmcentrez&rendertype=abstract.

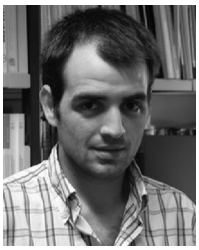
[38] N. K. Natt, H. Kaur, and G. P. S. Raghava. (2004, Jul.). Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods, *Proteins* [Online]. *56(1)*, pp. 11–18. Available: http://www.ncbi.nlm.nih.gov/pubmed/15162482.

[39] N. S. Schaadt, J. Christoph, and V. Helms Classifying substrate specificities of membrane transporters from Arabidopsis thaliana," *J. Chemical Inform. Model.*, vol. 50, no. 10, pp. 1899–1905, Oct. 2010.

[40] R. Agren, L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen. (2013, Jan.). The RAVEN toolbox and its use for generating a Genome-scale metabolic model for Penicillium chrysogenum, *PLoS Comput. Biol.* [Online]. *9(3)*, p. e1002980. Available: http://dx.plos.org/10.1371/journal.pcbi.1002980, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605104&tool=pm centrez&rendertype=abstract.

[41] N. Swainston. (2012, Mar.). Systems biology informatics for the development and use of genome-scale metabolic models Ph.D. dissertation [Online]. Available: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:157795.

[42] P. D. Karp. (2001, Sep.). Pathway databases: A case study in computational symbolic theories, *Science (New York, N.Y.)* [Online]. *293 (5537)*, pp. 2040–2044. Available: http://www.ncbi.nlm.nih.gov/pubmed/11557880.



Oscar Dias studied at the University of Minho where he received the BSc degree in biological engineering, in 2005, and the MSc dgree in informatics, in 2008, and the PhD dgree in chemical and biological engineering, in 2013. He published eight articles in specialized journals, 17 papers in conference proceedings, and four chapters of published books. Between 2010 and 2015, he participated in four research projects and is currently participating in two research projects. He currently holds a post-doc position at the University of Minho and his current research interests are industrial biotechnology and metabolic engineering with emphasis in systems biology and bioinformatics, namely, reconstruction of genome-scale metabolic models.

Daniel Gomes received the MSc degree in biological engineering from the University of Minho, in 2009. In 2010, he initiated his research activities at the Centre of Biological Engineering, University of Minho, in the field of bio-ethanol production and afterwards the reconstruction of a genome scale metabolic model for Ashbya gossypii. His current research interests include systems biology, industrial biotechnology, microbiology, and bio-ethanol production. In 2013, he initiated the PhD degree in the field of second generation bio-ethanol production with recycling of cellulases.

Paulo Vilaça graduated in informatics engineering and received the master's degree in bioinformatics from the University of Minho. He is also a member of the bioinformatics and systems biology group at the same university. He is currently working toward the industrial PhD degree at SilicoLife, a company creating computational solutions for the fast growing industrial biotechnology applications.

João Cardoso received the BSc degree in biosciences from the Catholic University of Portugal, in 2009. He received the MSc degree in bioinformatics from the University of Minho, in 2013. He was a guest researcher at the Helmholtz-Zentrum für Infektionsforschung from 2009 to 2010, in the Systems and Synthetic Biology research group. He also worked as a grant researcher at the same institution. From 2011 until 2014, he worked at the SilicoLife, Lda, doing R&D in bioinformatics. Since October 2014, he has been working toward the PhD degree at the Novo Nordisk Foundation Center for Biosustainability in the sequencing, informatics, and modeling group.

Miguel Rocha studied informatics engineering at the University of Minho, where he received the BSc degree, in 1995, the MSc degree, in 1998 and, finally the PhD degree, in 2004. He is currently an associated professor in the School of Engineering, University of Minho, and also the director of the master course Bioinformatics and coleads the Bioinformatics and Systems Biology team of the Centre of Biological Engineering. He has published over 100 papers in international peer-reviewed journals and conferences, has been the PI of several funded projects, and supervised seven PhD students. The main research interests are related to bioinformatics, mainly in the topics of machine learning, evolutionary computation, computational systems biology, reconstruction and optimization of metabolic models, strain optimization, biomedical text mining, and omics data analysis and mining.

Eugénio C. Ferreira graduated in chemical engineering in 1986 and received the PhD degree in 1995, both from the University of Porto. During the PhD preparation, he was a visiting research scholar at the Université Catholique de Louvain, Belgium. He is a full professor in the Biological Engineering Department of the University of Minho, Braga, Portugal, where he leads a research group on bioprocess engineering and computational biosystems. He was associate dean for research of the School of Engineering, University of Minho, from 2010 to 2013 and is director of the PhD program on bioengineering (MIT-Portugal Program). From 2004 to 2010, he served as vice-chair of the Department of Biological Engineering. He was the chair of the Chemical Engineering Section of the Portuguese Engineers Institution and the editor-in-chief of *Engenharia Química* from 2006 to 2008. Also, he is an associate editor of *Frontiers in Bioengineering and Biotechnology* and a member of the editorial boards of *Biomedical Research International*, *Chemical Product and Process Modeling*, the *Brazilian Journal of Chemical Engineering*, and the *Environmental Engineering and Management Journal*. He published nine books and is the author of more than 150 international refereed papers and book chapters. In addition, at least 280 papers and communications in conferences have been published by him.

Isabel Rocha received the PhD degree from Minho University, in chemical and biological engineering in 2003. She is an assistant professor and principal investigator in the Biological Engineering Department at the University of Minho, Portugal. In the last eight years, she has established a research group within the fields of metabolic engineering, systems, and computational biology. Her main research interests are the development and application of novel algorithms for metabolic model reconstruction, strain optimization, and data analysis. She is the PI of several projects in systems biology applied to industrial biotechnology and life sciences and has published more than 100 peer-reviewed papers. She is also one of the founders of two spin-off companies and was the president of APBio – the Portuguese Bioindustries Association from 2007 to 2009. In 2004, she was a post-doctoral research fellow at the Centre of Microbial Biotechnology – Technical University of Denmark in metabolic engineering and systems biology. In 2007, she was a short-term visiting scholar at the Massachusetts Institute of Technology, as part of the MIT-Portugal Program.