The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks

Neil Swainston[1*], Kieran Smallbone[1], Pedro Mendes[1,2], Douglas B Kell[1], Norman W Paton[1]

[1] Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, M1 7DN, United Kingdom

[2] Virginia Bioinformatics Institute, Virginia Tech, Washington St. 0477, Blacksburg, VA 24061, USA

Summary

The generation and use of metabolic network reconstructions has increased over recent years. The development of such reconstructions has typically involved a time-consuming, manual process. Recent work has shown that steps undertaken in reconstructing such metabolic networks are amenable to automation.

The SuBliMinaL Toolbox (http://www.mcisb.org/subliminal/) facilitates the reconstruction process by providing a number of independent modules to perform common tasks, such as generating draft reconstructions, determining metabolite protonation state, mass and charge balancing reactions, suggesting intracellular compartmentalisation, adding transport reactions and a biomass function, and formatting the reconstruction to be used in third-party analysis packages. The individual modules manipulate reconstructions encoded in Systems Biology Markup Language (SBML), and can be chained to generate a reconstruction pipeline, or used individually during a manual curation process.

This work describes the individual modules themselves, and a study in which the modules were used to develop a metabolic reconstruction of *Saccharomyces cerevisiae* from the existing data resources KEGG and MetaCyc. The automatically generated reconstruction is analysed for blocked reactions, and suggestions for future improvements to the toolbox are discussed.

1 Introduction

The development of metabolic network reconstructions has increased over the last ten years. Such reconstructions are now available for a range of taxonomically diverse organisms, and they have been applied to a number of research topics including metabolic engineering, genome-annotation, evolutionary studies, network analysis, and interpretation of omics datasets [1].

A genome-scale metabolic reconstruction is a computational and mathematical model of the metabolic capabilities of a given organism [2]. It consists of all known metabolic reactions that can take place in a cell and the gene-protein-reaction relationships that connect the genome to the metabolome via the specification of enzymes and isoenzymes that catalyse each reaction. Specifying such gene-protein-reaction relationships will allow metabolic modelling to become increasingly integrated with transcription and signalling networks

* To whom correspondence should be addressed. Email: neil.swainston@manchester.ac.uk

through consideration of the action of metabolites on promoters and transcription factors. In addition, intra- and extra-cellular compartments can be considered, along with transport reactions and transport proteins that provide for metabolic transport across compartmental membranes. Furthermore, in order to analyse the phenotypic behaviour of the organism under a given condition, it is common to specify an objective function that is assumed to be optimised by the cell [3]. This can take a number of forms, including the maximisation or minimisation of usage of ATP, but commonly assumes that a cell attempts to maximise growth rate. In this case, a biomass function is included, which is a hypothetical reaction that uses metabolites necessary for cell growth, such as amino acids, nucleotides, lipids and cell-wall components, and required cofactors.

This work concerns itself with the automation of steps that are necessary in the development and analysis of genome-scale metabolic models. The process of completing such steps to develop reconstructions is now well defined and is recognised as being time-consuming [4]. While many of the steps associated with generating a high-quality reconstruction require manual curation, some of these are amenable to automation, providing the possibility of automating the process of generating a draft reconstruction to be used in subsequent manual curation. While a fully automated approach has been shown itself capable of the rapid generation of candidate reconstructions in a number of cases [5], it is recognised that such reconstructions still require manual validation and editing. As such, there remains a middle ground between the fully automated and fully manual approaches, where the draft reconstruction and curation process stands to benefit from dedicated software support. Such a semi-automated approach was followed in the development of recent genome-scale metabolic reconstructions for *Saccharomyces cerevisiae* and *Homo sapiens*, in which draft reconstructions were checked and enhanced by utilising SuBliMinaL Toolbox modules during an iterative development process.

Drawing upon previous experience of generating such reconstructions [6,7,8], this paper considers the development of reconstructions from the existing curated data resources KEGG [9] and MetaCyc [10]. Although both resources provide the facility for exporting metabolic models, neither of these exported models is of sufficient accuracy nor is suitably formatted for performing genome-scale, constraint-based analyses. Nevertheless, both resources provide initial pre-draft prototypes that can be developed further [11].

The SuBliMinaL Toolbox consists of a number of independent modules that can be used independently or chained together to form a reconstruction workflow allowing the generation of an initial draft of a metabolic reconstruction (see Figure 1). The importance of using community-developed standards to represent models in systems biology is well established [12]. As such, reconstructions are generated in Systems Biology Markup Language (SBML) [13] and are semantically annotated according to the MIRIAM standard [14]. They can be formatted in such a way that they can be loaded into the COBRA Toolbox [15], allowing constraint based analyses to be performed on the model, using techniques such as Flux Balance Analysis (FBA) [16].

2 Methods

SuBliMinaL Toolbox modules typically have a simple SBML-in / SBML-out interface, which take in a model or models, perform a given task and produce an updated model. Some modules should be used sequentially (for example, a reaction should only be elementally and charge balanced once the protonation states of its reactants and products have been determined). The SuBliMinaL Toolbox utilises the programming library libAnnotationSBML

[17], and web service interfaces to ChEBI [18] and KEGG to automatically retrieve required chemical data.

The SuBliMinaL Toolbox is written in Java, and is dependent upon third-party tools, which must be installed independently. Each of the modules can either be run from the command line or incorporated into custom software via a Java API. The SuBliMinaL Toolbox has been tested on Mac OS X 10.6 and 64-bit Windows 7. Instructions on the installation and use of the toolbox are available at http://www.mcisb.org/subliminal/. A description of each module of the toolbox is given below.

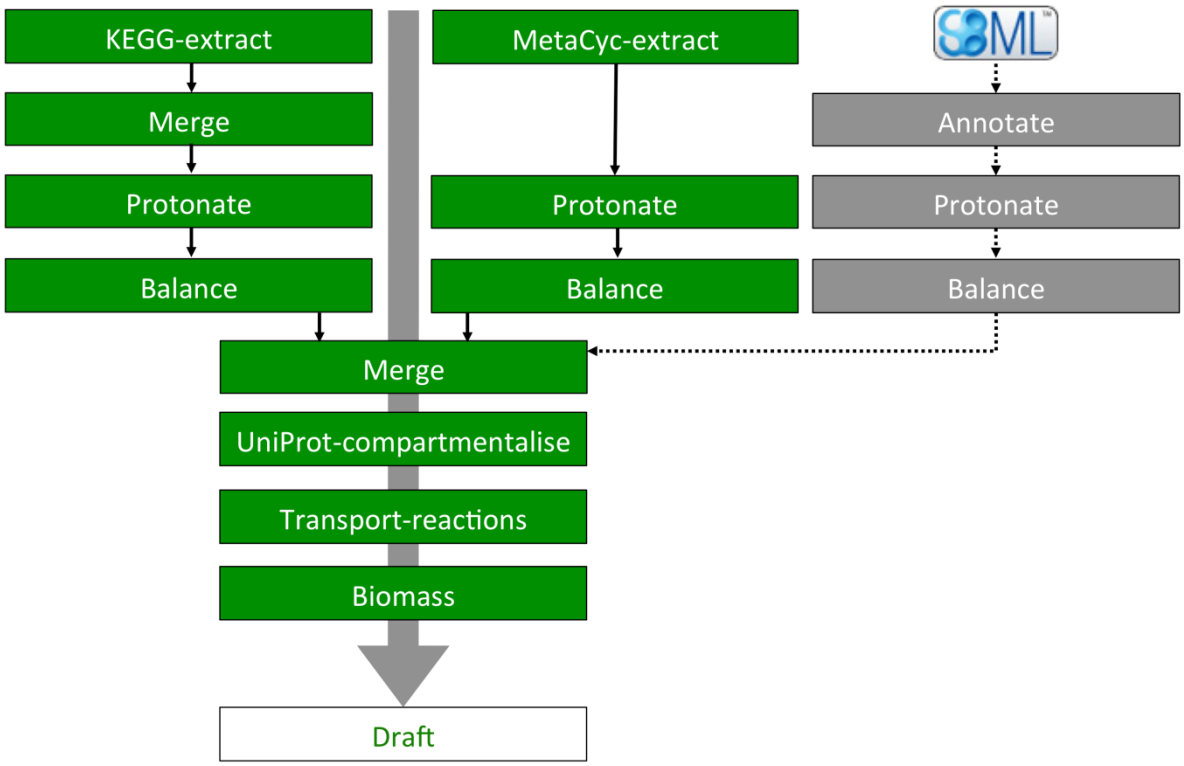

Figure 1: Flow diagram illustrating how SuBliMinaL Toolbox modules may be chained together to generate a draft metabolic network reconstruction that can be analysed in the COBRA Toolbox. The names of the boxes refer to individual SuBliMinaL Toolbox modules. The main branches with solid arrows from KEGG-extract and MetaCyc-extract indicate the pipeline that was utilised in this study. The right-hand branch with dotted arrows indicates a hypothetical addition to the pipeline, which could be used to include existing reconstructions or individual pathways marked up in SBML format.

2.1 Pre-draft reconstruction

Initial pre-draft pathways for a given organism can be generated from both KEGG and MetaCyc, using the **KEGG-extract** and **MetaCyc-extract** modules respectively.

KEGG does not allow export of pathways data in SBML format. The **KEGG-extract** module has been developed to provide this functionality. The module downloads the organism-specific KEGG KGML flat files for each represented pathway, and parses these to extract the individual metabolic reactions, in terms of metabolites and enzymes, that constitute the pathway. Where specified, reaction directionality is also considered. KEGG does not specify intracellular compartmentalisation, and as such, all metabolites are assumed to be cytoplasmic. An SBML model is then generated for each defined pathway, and each of these

is then annotated according to the MIRIAM standard, such that each metabolite and enzyme is assigned an unambiguous identifier.

It was found that both the existing tools for converting KEGG data into SBML format, KEGG2SBML (http://sbml.org/Software/KEGG2SBML) and KEGGConverter [19], were unsuitable for use in the context of generating genome-scale reconstructions, due to the presence of missing reactions or of reaction participants in their generated pathways. These shortcomings result in gaps and stoichiometric inconsistencies in the final draft reconstruction. Furthermore, both tools are reliant upon the downloading of KEGG flat files that are no longer freely available to academic users. Along with this work, the recently introduced KEGGtranslator [20] overcomes these limitations.

MetaCyc-extract downloads the appropriate organism-specific flat files and annotates the supplied SBML file to ensure consistency with the equivalent KEGG model. Again, the resulting model is updated to ensure appropriate metabolite charge state and balanced reactions. An advantage of MetaCyc over KEGG is in its definition of intracellular compartmentalisation. Where present, this intracellular compartmentalisation is extracted and added to the model. If specified, unambiguous metabolite and enzyme identifiers are extracted from the MetaCyc flat files and assigned to chemical species in the generated SBML file. If no identifiers are present, metabolite names are automatically searched against the ChEBI database in order to determine ChEBI identifiers to be assigned to metabolites. Checking against supplied chemical formulae validates the assignment of such identifiers.

Both modules generate consistently formatted models representing the union of all metabolic pathways described in each resource. These individual models can then be merged and their annotations exploited in subsequent modules.

2.2 Annotation

The modules of the toolbox are dependent upon the initial draft reconstruction being annotated with unambiguous identifiers according to the MIRIAM standard. In order to support the use of existing reconstructions and pathways in the toolbox, the **Annotate** module has been developed to automate the process of adding annotations to existing models. The **Annotate** model launches the SuBliMinaL Annotator, a graphical wizard that facilitates the annotation process. The SuBliMinaL Annotator allows the user to select a model in SBML format, which is then parsed to extract names of the model components compartments, metabolites and enzymes. Each of these terms is then searched against the databases Gene Ontology (GO) [21], ChEBI and UniProt [22] respectively. The results of these searches are presented to the user, allowing the selection of the appropriate database term with which to annotate the model component (see Figure 2). Upon completion of the annotation process, the updated model is saved and can be used with subsequent SuBliMinaL modules.

The **Annotate** module can also be run in “Silent mode”, which does not rely on user selection of search results. In this mode, the name of the SBML element being searched is compared alphanumerically against the search results in a case-insensitive manner. Upon matching, the SBML element is annotated with the matching search term, allowing commonly named metabolites such as ATP to be quickly annotated without relying on user selection.

By default, the **Annotate** module assumes all chemical species to be metabolic. To specify that a chemical species represents a protein, the appropriate species elements in the SBML model must be annotated with the Systems Biology Ontology (SBO) Term `SBO:0000252` (denoting *polypeptide chain*) [23]. Doing so ensures that this species name is searched against UniProt. If the SBML model is annotated with an NCBI Taxonomy term [24], the UniProt

search is restricted to proteins of that organism. Figure 3 illustrates an SBML model that allows organism-specific searches of UniProt to be performed.

2.3 Model merging

The lack of consistent naming between components within existing reconstructions has been reported as an impediment to both manual and automatic comparison and construction of models [25], and was a major motivation for the use of semantic annotations that overcome this. As each of the initial pre-draft models generated by both **KEGG-extract** and **MetaCyc-extract** contain comparable identifiers, this issue is mitigated. The **Merge** module automatically merges each supplied model into a single consolidated model, in which duplicate metabolites, enzymes, and reactions are removed. If present, intracellular compartmentalisation is also considered, such that chemical species (metabolites and enzymes) are only considered to be duplicates if they share both identifier and compartment. As chemical species in different compartments are considered to be distinct, transport reactions are retained during the merge process.

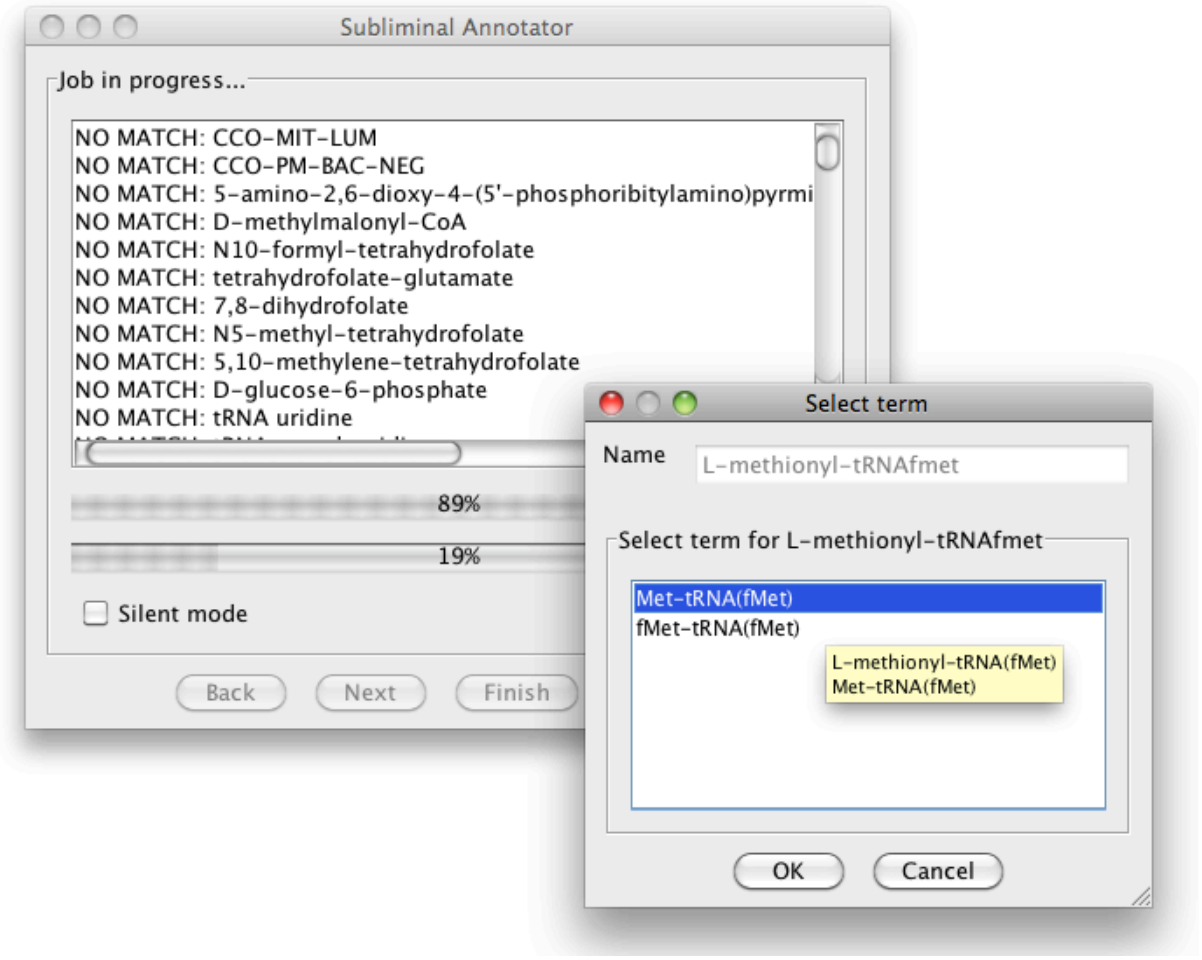

Figure 2: Screen capture of the SuBliMinaL Annotator. The main window displays progress of the annotation process, indicating terms that remain unmatched, and two progress bars displaying the percentage of terms successfully annotated and progress of the annotation process respectively. The foreground dialog box displays the results of a search for a metabolite name against ChEBI, ordered by the ChEBI Text Based Score. The user can select one of the two terms in order to annotate the metabolite. Compound synonyms are also searched, and can be viewed in a tooltip.

The merging of models is non-trivial due to the presence of duplicate metabolites both across data resources and within a given resource [26]. Furthermore, the specification of metabolites can differ in their precision. For example, in the case of KEGG, many stereoisomers are represented, an example being D-glucose, alpha-D-glucose and beta-D-glucose. The **Merge** module can therefore be run in a "fuzzy" mode, in which these metabolites are considered to be synonymous, as it is difficult to determine whether a given reaction involving these metabolites is intended to refer to the general case or one or both of the stereospecific terms.

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level2/version4" level="2" version="4">
  <model metaid="_model">
    <!-- The following specifies that the model represents Saccharomyces cerevisiae -->
    <annotation>
      <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:bqbiol="http://biomodels.net/biology-qualifiers/">
        <rdf:Description rdf:about="#_model">
          <bqbiol:is>
            <rdf:Bag>
              <rdf:li rdf:resource="urn:miriam:taxonomy:559292"/>
            </rdf:Bag>
          </bqbiol:is>
        </rdf:Description>
      </rdf:RDF>
    </annotation>
    <listOfCompartments>
      <compartment id="c" name="cytosol" size="1"/>
    </listOfCompartments>
    <listOfSpecies>

      <!-- The following species represent metabolites (small molecules) -->
      <species id="s1" name="D-Glucose 1-phosphate" compartment="c"/>
      <species id="s2" name="D-Glucose 6-phosphate" compartment="c" sboTerm="SBO:0000247"/>

      <!-- The following species represents a protein -->
      <species id="s3" name="Phosphoglucomutase-1" compartment="c" sboTerm="SBO:0000252"/>

    </listOfSpecies>
  </model>
</sbml>
```

Figure 3: Simple SBML model indicating how metabolites and enzymes are distinguished in the Annotate module by use of `sboTerm` attributes on the `species` elements. By default, all `species` are considered to be metabolic (small molecules) but this can be made explicit through use of the SBO term `SBO:0000247` (*simple chemical*). Enzymes are specified with the SBO term `SBO:0000252` (*polypeptide sequence*). Annotating the `model` element with an NCBI Taxonomy term (in this case, `559292`, representing *Saccharomyces cerevisiae*) limits the subsequent UniProt search to proteins belonging to the specified organism.

As such, the assumption can be made that a reaction applies to all synonymous metabolites, and these are then collapsed into a single metabolite in the merge process, with the intention of increasing the network connectivity of the merged reconstruction. The **Merge** module can determine whether two metabolites share the same chemical formula, and whether both have a shared ancestor in the ChEBI ontological tree. If so, these metabolites can be collapsed into a single term. For well-curated reconstructions, the **Merge** module can also be run in a more simplified mode, in which two metabolites will only be considered to be the same if they share the same semantic annotation.

2.4 Metabolite pKa prediction and determination of appropriate charge state

The **Protonate** module utilises the ChEBI web service to harvest SMILES strings [27] representing each metabolite. These are then passed to the `MajorMicrospeciesPlugin` method in the API of the cheminformatic library Marvin Beans for Java Developers (ChemAxon Kft., Budapest, Hungary; http://www.chemaxon.com), which relies on the Hammett−Taft approach [28] to estimate pKas and thus predict the dominant protonation state of the metabolites at a supplied pH. Specific pHs may also be applied to metabolites in a given intracellular compartment. Updated chemical formula and charge are then added to each metabolite, and, if appropriate, both the name and the ChEBI annotation of the molecule are updated to reflect its corrected charge state. Exploiting the ChEBI ontology specification of `isConjugateBaseOf` and `isConjugateAcidOf` predicates, which allow relationships between de/protonated molecules to be automatically determined, enables this functionality.

An example of this is KEGG compound `C00022`. Although KEGG names this metabolite pyruvate, both the molecular formula ($C_3H_4O_3$) and the cross-reference link to ChEBI (`CHEBI:32816`) for this entry indicate that the metabolite is actually the protonated form, pyruvic acid. Marvin Beans predicts that the metabolite is deprotonated at a pH of 7.0. As such, the **Protonate** module updates the metabolite in the reconstruction, setting the molecular formula to $C_3H_3O_3$, the charge to -1, and updates the annotation to that of the ChEBI term for pyruvate, `CHEBI:15361`. This illustrates the inconsistencies that are often present in biochemical resources, such that both conjugate acid and bases are sometimes collapsed into a single, ambiguous entry. Such inconsistencies can be resolved with this approach, producing unambiguous definitions of both metabolites and reactions that more accurately reflect physiological conditions.

2.5 Elemental and charge balancing

Balancing all metabolic reactions ensures that a reconstruction is free of stoichiometric inconsistencies [29]. Stoichiometric inconsistencies violate mass conservation, and can be illustrated in the example below:

$$\text{R1: A} \Leftrightarrow \text{B}$$

$$\text{R2: A} \Leftrightarrow \text{B + C}$$

It is intuitively clear that a network containing these two reactions contains an inconsistency. That is, metabolite C could only satisfy the above two equations if it were to have a mass and charge of zero. While the above example is simple, determining such errors in genome-scale models is non-trivial but can be performed algorithmically by the ScrumPy package [30]. While the ScrumPy package can detect such inconsistencies, their correction in the reconstruction relies upon manual curation. As such, it is preferable to reduce such inconsistencies by performing elemental and charge balancing where possible.

The **Balance** module automates this process of mass and charge balancing of reactions. Consider the following reaction:

$$\text{2-Acetolactate + carbon dioxide} \Leftrightarrow \text{Pyruvate}$$

$$C_5H_7O_4^- + CO_2 \Leftrightarrow C_3H_3O_3^-$$

Manual inspection can quickly determine that, in terms of elemental and charge balancing, two pyruvates must be produced, and the list of products is also deficient in a proton.

The **Balance** module attempts to detect (and fix) such issues automatically through mixed integer linear programming (MILP). Each reaction is represented as a matrix, A, containing elemental counts and charges for each reactant and product. Metabolites that are commonly absent from reaction definitions [31], such as water, protons and carbon dioxide, are also considered, and are added as both potential reactants and products. Reactant elemental and charge counts are specified as positive, those of products negative. Optional cofactor metabolites are only added to the matrix if they are not present in the original reaction. It is for this reason that carbon dioxide is absent from the specified optional reactants (see Figure 4).

The stoichiometric coefficients of each reactant are represented by the vector, b. Mixed integer linear programming is applied to solve $Ab = 0$, satisfying the constraint $b_j >= b_{j,min}$, where $b_{j,min}$ represents the minimum allowed stoichiometric coefficient for a given metabolite (1 for specified metabolites, 0 for optionally considered metabolites). This produces the vector of stoichiometric coefficients, b, to be applied to each reactant and product to balance the equation. (The vector, b, is minimised to return the minimum collection of stoichiometric coefficients that are required to balance the equation, thus preventing mathematically correct but non-optimal solutions, such as the spurious addition of water to both sides of the equation).

	Reactants		Products	Optional reactants		Optional products		
	CO2	C5H7O4	C3H3O3	H+	H20	H+	H20	CO2
C	**1**	**5**	**-3**	*0*	*0*	*0*	*0*	*-1*
O	**2**	**4**	**-3**	*0*	*1*	*0*	*-1*	*-2*
H	**0**	**7**	**-3**	*1*	*2*	*-1*	*-2*	*0*
charge	**0**	**-1**	**1**	*1*	*0*	*-1*	*0*	*0*
b_{min}^T	**1**	**1**	**1**	*0*	*0*	*0*	*0*	*0*

Figure 4: A matrix representing elemental count and charge of reactants and products in the reaction 2-Acetolactate + carbon dioxide ⇔ Pyruvate. Required reactant and product elemental and charge counts are specified in bold; those of optional reactants and products in italics. The vector b_{min}, representing the minimum permitted stoichiometric coefficients for each reaction participant, is also shown.

Consequently, the linear solver returns the solution, $b^T = (1\ 1\ 2\ 0\ 0\ 1\ 0\ 0)$, indicating that in order to balance the reaction, two pyruvates should be produced and one proton should be added as a product. In the case of a solution being found, the reaction is updated in the reconstruction to reflect this:

2-Acetolactate + carbon dioxide ⇔ 2 Pyruvate + proton

$$C_5H_7O_4^- + CO_2 \Leftrightarrow 2\ C_3H_3O_3^- + H^+$$

In many cases, however, reactions cannot be balanced with the above approach. This could be due to a number of reasons. An obvious limitation occurs when attempting to balance reactions in which the chemical formula of one or more participants is unknown, which is a result of missing information in the data resources. A further trivial problem is the specification of incorrect reactions, in which key reactants or products, over and above commonly absent metabolites such as water, are missing. In both cases, manual curation will be necessary to correct the errors, and calculating an elemental difference between the

reactants and products, which could suggest the chemical formula of a missing participant, may drive this process.

The **Balance** module uses the linear solver glpk (http://www.gnu.org/s/glpk/) and the java interface GLPK for Java (http://glpk-java.sourceforge.net).

2.6 Compartmentalisation

Thus far, the SuBliMinaL Toolbox generates largely uncompartmentalised reconstructions. Some intracellular compartmentalisation is provided by MetaCyc, but given the dependency of the pipeline described in figure 1 on KEGG, which does not consider compartmentalisation, most metabolites are considered to be cytoplasmic by default.

The Compartmentalise modules provide the facility for extending reconstructions generated from KEGG or MetaCyc alone to generate semi-compartmentalised models. Two compartmentalise modules exist. **UniProt-compartmentalise** extracts protein localisation information directly from UniProt annotation. Only Swiss-Prot entries are considered, and as the metadata associated to these entries is manually curated, the localisation specified for such entries is therefore likely to be accurate. For cases where no such curated data exist, the **PSORT-compartmentalise** module can be used. This module harvests protein sequences from the UniProt web services for each enzyme and passes these to the protein localisation service WoLFPSORT [32], a web interface to the PSORT algorithm [33]. From the curated or predicted intracellular compartmentalisation of a given enzyme, the localisation of metabolic reactions catalysed by this enzyme is inferred. As such, reactions, enzymes and metabolites are localised, and can be inferred to be present in multiple compartments, depending on the UniProt annotation or prediction of WoLFPSORT. In the case of a reaction being catalyzed by isoenyzmes that are present in different intracellular compartments, the reaction is duplicated such that an instance appears in each compartment, with the appropriate isoenzyme specified as the reaction modifier. Where predictions suggest that metabolites are found in multiple compartments, putative intracellular transport reactions are added to the reconstruction to allow for their transport between compartments.

2.7 Transport

Transport reactions are important for both natural metabolites and xenobiotics [34,35]. The **Transport-reaction** module adds a generic set of import reactions to the reconstruction in order to allow for uptake of metabolites from the growth medium. The set of generic import reactions are taken from the BIGG database [36], which contains 9 published and well-curated reconstructions from a range of taxonomically diverse organisms[1]. Import reactions across the cell membrane are added if the extracellular metabolite is also present in the reconstruction's cytoplasm. The addition of this generic set of import reactions is essential if subsequent analysis by the COBRA Toolbox is to be performed, as neither KEGG nor MetaCyc provide such cell-membrane transport reactions, which effectively means that reconstructions generated from these resources would be "starved" of growth media metabolites.

[1]These reconstructions are *S. cerevisiae* iND750, *E. coli* iAF1260, *E. coli* iJR904, *E. coli* textbook, *H. pylori* iIT341, *H. sapiens* Recon_1, *M. barkeri* iAF692, *M. tuberculosis* iNJ661, and *S. aureus* iSB619. Uptake reactions specific to the *H. sapiens* reconstruction were excluded from the set of selected uptake reactions, as they accounted for a number of metabolites for which transporters would be unlikely to be present in the majority of organisms.

Irreversible export reactions are added for all cytoplasmic metabolites, providing by default the possibility of excreting all cytoplasmic metabolites from the cell. This approach relies upon no *a priori* knowledge of the transport capabilities of the cell, and follows the philosophy of Fell *et al.* [37], which states that a simple solution to the problem of adding reactions to a reconstruction is to "add more than is likely to be necessary and to remove at a later date the ones that are not functional". It is envisaged that subsequent flux balance analysis of the completed draft model will provide an indication of which intracellular metabolites will need to be excreted in order for the model to fulfill the objective function. Superfluous export reactions can then be purged from the model, leaving a subset that can be manually validated according to the known capabilities of the organism, which may have been tested experimentally by metabolic footprinting [38]. The approach of adding more transport reactions than may be biologically feasible mirrors that of the **compartmentalise** modules, in which compartments are added with the intention of removing or reconciling these later as the reconstruction is manually validated.

2.8 Biomass function

The **Biomass** module assigns a generic biomass reaction to the reconstruction, and performs reformatting that allows FBA simulations to be performed by the COBRA Toolbox. The generic biomass reaction consists of the 20 most common amino acids, the four nucleotide precursors of DNA, ATP and lipids. In addition, the biomass reaction contains ADP, phosphate and protons as products. These "by-products" of biomass formation are then subsequently available to the model.

While the first biomass components are static and are applied to all reconstructions, the lipid term is built dynamically, and is dependent upon the constituents of the reconstruction. The generic metabolite "lipid" is first added to the list of biomass components. A number of modelling reactions are then added to the reconstruction, in which any metabolites in the model that exhibit an "is a" lipid relationship in the ChEBI database are added as reactants, with lipid as product.

Each of the reactants and products in the biomass reaction are given a stoichiometry of 1. This simple approach allows the reconstruction to be analysed to determine network connectivity, i.e., testing if the reconstruction allows for growth of the organism under given conditions. However, by not quantifying the components in biomass relative to one another, the reconstruction is unable to predict growth rate. This limitation can be rectified by manual curation.

3 Results

From applying the pipeline illustrated in Figure 1, a draft version of a metabolic reconstruction for *Saccharomyces cerevisiae* was generated for comparison against a manually generated version [7], which has been updated iteratively over a number of years. A comparison of both models is given in Table 1.

While SuBliMinaL generates a model with an increased number of metabolites and metabolic reactions with respect to the manually generated version (an increase of 92% and 90% respectively), it remains unclear whether this increase is due to the combined coverage of the original resources, KEGG and MetaCyc, or an indication of incomplete merging of data from each source. While the **Merge** module attempts to ensure that duplicate metabolites and reactions are not added to the consensus, metabolites that are lacking in comparable identifiers across the two sources may be duplicated. A limitation of both KEGG and

MetaCyc (and hence also of reconstructions generated from these resources) is the lack of defined multimeric enzymatic complexes.

While reactions are associated with genes and proteins where possible, specification of multimeric complexes in reconstructions developed by SuBliMinaL remains a task for manual curation, as it appears that no data resource describing such complexes currently exists, preventing the automation of this step.

Table 1: Comparison of SuBliMinaL- and manually-generated *S. cerevisiae* metabolic reconstructions. Values for unique metabolites, enzymes and metabolic reactions refer to "flattened" versions of the reconstructions, in which metabolites and enzymes in different intracellular compartments are considered one. In the case of SuBliMinaL, unblocked reactions were calculated on a minimal growth medium as described below.

Components	SuBliMinaL[2]	Manual
Compartments	8	17
Unique metabolites	1397	728
Unique enzymes	936	939
Unique metabolic reactions	1803	947
Unblocked reactions	1428/1803 (79%)	759/947 (80%)

The manually generated reconstruction also contains 9 compartments in addition to those in the SuBliMinaL-generated version. This is due to the specification of membrane compartments in the manual version, in which transport proteins are assigned. SuBliMinaL assumes all transport proteins to be present in the cytoplasm. This is simply a design decision to reduce the complexity of the reconstruction, and has no effect on its subsequent analysis.

A goal of the pipeline was to generate a reconstruction that was capable of simulating the production of biomass from minimal growth media automatically. It was found that the reconstruction could successfully simulate biomass production from a growth medium of D-glucose, ammonium, phosphate, sulphate, oxygen, water and protons. In doing so, it was found that, of the putative extracellular transport reactions added by the **Transport-reaction** module, all but 12 could be removed for the objective to be realised. The retention of these putative extracellular transport reactions provide sinks for product metabolites that are generated in reactions required to fulfill the biomass objective function. Of these 12 extracellular reactions that had to be retained, 3 involved metabolites involved in purine metabolism, suggesting reactions in this pathway that are incapable of carrying flux, which could act as a starting point for manual curation efforts.

The `fluxVariability` functionality of the COBRA Toolbox was used to assess the reconstruction. In order for a metabolic reaction to carry flux, all of its reactants and products must be connected to other reactions. As such, the proportion of reactions with capacity to carry flux is a measure of the connectivity of the network. The SuBliMinaL-generated reconstruction is found to be highly connected (75% unblocked), though slightly less than the manually curated version (80%).

[2]The SuBliMinaL-generated *S. cerevisiae* model was generated with KEGG release 59.0 (1 Jul 2011), MetaCyc version 15.1 (8 June 2011) and ChEBI release 83 (5 September 2011).

4 Discussion

The finding that the draft reconstruction contains suspected gaps in certain pathways illustrates the philosophy behind the development of draft reconstructions that are capable of undertaking constraint-based analysis: that is, that the results of such analyses can drive the curation process in an iterative manner through performance of cycles of analysis and refinement. The results of the analyses can be inspected, indicating potential errors, which can then be corrected manually.

The feasibility of performing such iterative cycles is made possible due to the speed at which genome-scale reconstructions can be automatically generated and checked. The pipeline described in Figure 1 generated the draft yeast reconstruction in under four hours on an Apple MacBook Pro 2.8GHz Intel Core i7. An existing protocol describing the generation of metabolic reconstructions suggests that the manual reconstruction refinement stage can take months to a year to complete [39]. This stage contains a number of steps that are covered by the SuBliMinaL Toolbox (such as charge state determination, reaction balancing, determination of metabolic identifiers), and as such, use of the toolbox should reduce the duration of both the initial stage of generating draft reconstructions and the checking of reconstructions in the following refinement phase.

The SuBliMinaL Toolbox has been used in the development of community-produced reconstructions of *Saccharomyces cerevisiae* and *Homo sapiens*. The use of the toolbox could be extended to the incremental development of such resources. As the development of reconstructions is an ongoing process, one could imagine a paradigm in which domain experts extract individual pathways from reconstructions, and then perform validation and curation on those areas of metabolism in which they have expertise. Such curated pathways could subsequently be re-collated into the reconstruction, which would then be formatted and reanalysed, following the iterative cycle described above. Such crowd-sourcing methods are already exploited in the web-based curation of individual pathways [40,41] and would prove useful in the iterative development of larger-scale networks. The recently developed software MEMOSys [42] may support such an approach, providing a secure web-enabled interface for community driven, multi-user development and refinement of reconstructions. The use of such tools, combined with automated modules for performing such tasks as checking of metabolite charge state determination and reaction balancing described here, may mitigate the need for jamborees: focused curation meetings that have become the preferred method of developing community-driven reconstructions over recent years [43].

Nevertheless, before such a more automated approach to community development could become more prevalent, there remain a number of issues within modules of the SuBliMinaL Toolbox that will need to be improved over time. While some reaction directionality is specified by KEGG, other reactions are initially specified to be reversible, which can result in thermodynamically infeasible flux patterns being predicted by model analyses. Specifying reaction directionality – either by automated or manual means - adds constraints to the model, which are likely to improve the model's predictive accuracy by preventing reactions that are thermodynamically infeasible. Due to the accessibility of InChI strings for many of the metabolites present in the reconstructions generated by the toolbox, there exists the possibility of automating the determination of reaction directionality, following the approach of Fleming *et al.* [44,45]. Integration of predictions of reaction directionality is therefore a likely future development.

The **Compartmentalise** module provides a useful first pass at automating the generation of compartmentalised reconstructions. While such an approach is preferable to a fully manual

approach to determining compartmentalisation that is currently followed, it is recognised that this approach is completely dependent upon the accuracy of the UniProt annotation or WoLFPSORT predictions. It is therefore likely that such an approach would reduce the connectivity of the network, as many pathways in a given intracellular compartment would be incomplete unless all enzymes within the pathway were correctly predicted to be present in the compartment. Applying this approach would require the addition of missing reactions in a gap-filling step, which may be performed by inference. For example, if an almost complete mitochondrial TCA cycle was predicted to be missing an enzyme that is present in the cytoplasm, it may be inferred that the enzyme (and the reaction that it catalyses) is indeed present in the mitochondria, despite the UniProt annotation or WoLFPSORT algorithm predicting otherwise. Such inferred reactions would be added as "modeling" reactions, and associated with an appropriate evidence code, indicating that they should be subject to subsequent manual curation. The inference of enzyme and reaction localisation, based upon the network topology of partially compartmentalised metabolic models, has been reported [46], and the approach followed by the toolbox – to generate a partially compartmentalised model for subsequent refinement – supports this inference method.

Limitations of the current **Biomass** module implementation are its assumption that *all* lipids may be constituents of the biomass objective function, and that other cell wall constituents and storage carbohydrates are not considered. Determining more specific biomass objective functions, perhaps tailored towards the taxonomy of the organism under reconstruction, would be a useful improvement for future work.

The possibility exists to extend the toolbox to consider transport proteins. Transport proteins for a given organism can be automatically extracted from the TransportDB database [47] and the potential exists to add these to the reconstruction. However, while the transport proteins can be extracted, TransportDB does not yet fully characterise its transport proteins in such a way that the corresponding transported metabolites can be retrieved in an automated fashion. It is hoped that, as such resources that describe transport proteins develop, the task of assigning such proteins to individual reactions will be able to be automated.

It is again emphasised that manual curation and validation are essential steps in generating a high-quality reconstruction. Referring to literature commonly drives this validation process, and recently developed reconstructions have illustrated the importance of applying literature references and confidence scores to components within the model. Doing so increases the confidence that users apply to reconstructions (or at least, individual pathways or reactions within reconstructions), and also can be used to prioritise refinement efforts. The determination of literature references may be aided through tighter integration with text-mining tools such as PathText [48] in order to simplify the arduous, but necessary, task of finding evidence for present (and missing) reactions in the literature [49].

Acknowledgements

The authors thank the BBSRC and EPSRC for their funding of the Manchester Centre for Integrative Systems Biology (http://www.mcisb.org), BBSRC/EPSRC Grant BB/C008219/1. The authors also thank Michael Howard and Daniel Jameson for help with the preparation of the manuscript.

References

[1] M. A. Oberhardt, B. Ø. Palsson, J. A. Papin. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:320, 2009.

[2] M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov, B. Ø. Palsson. Metabolic modeling of microbial strains in silico. *Trends Biochem Sci*, 26:179-86, 2001.

[3] R. Schuetz, L. Kuepfer, U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol Syst Biol*, 3:119, 2007.

[4] I. Thiele, B.Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5:93-121, 2010.

[5] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, R. L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol*, 28:977-982, 2010.

[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, D. B. Kell. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol*, 26:1155-1160, 2008.

[7] P. D. Dobson, K. Smallbone, D. Jameson, E. Simeonidis, K. Lanthaler, P. Pir, C. Lu, N. Swainston, W. B. Dunn, P. Fisher, D. Hull, M. Brown, O. Oshota, N. J. Stanford, D. B. Kell, R. D. King, S. G. Oliver, R. D. Stevens, P. Mendes. Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol*, 4:145, 2010.

[8] I. Thiele I, D. R. Hyduke, B. Steeb, G. Fankam, D. K. Allen, S. Bazzani, P. Charusanti, F. C. Chen, R. M. Fleming, C. A. Hsiung, S. C. De Keersmaecker, Y. C. Liao, K. Marchal, M. L. Mo, E. Özdemir, A. Raghunathan, J. L. Reed, S. I. Shin, S. Sigurbjörnsdóttir, J. Steinmann, S. Sudarsan, N. Swainston, I. M. Thijs, K. Zengler, B. O. Palsson, J. N. Adkins, D. Bumann. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Syst Biol*, 5:8, 2011.

[9] M. Kanehisa, S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28:27-30, 2000.

[10] P. D. Karp, M. Riley, M. Saier, I. T. Paulsen, S. M. Paley, A. Pellegrini-Toole. The EcoCyc and MetaCyc databases. *Nucleic Acids Res*, 28:56-59, 2000.

[11] K. Radrich, Y. Tsuruoka, P. Dobson, A. Gevorgyan, N. Swainston, G. Baart, J. M. Schwartz. Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst Biol*, 4:114, 2010.

[12] D. B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov Today*, 11:1085-92, 2006.

[13] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novère, L. M.

Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang J; SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19:524-531, 2003.

[14] 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, B. L. Wanner. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol*, 23:1509-15, 2005.

[15] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B.Ø. Palsson, M. J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*, 2:727-38, 2007.

[16] J. D. Orth, I. Thiele, B. Ø. Palsson. What is flux balance analysis? *Nat Biotechnol*, 28:245-8, 2010.

[17] N. Swainston, P. Mendes. libAnnotationSBML: a library for exploiting SBML annotations. *Bioinformatics*, 25:2292-2293, 2009.

[18] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, 36:D344-350, 2008.

[19] K. Moutselos, I. Kanaris, A. Chatziioannou, I. Maglogiannis, F. N. Kolisis. KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. *BMC Bioinformatics*, 10:324, 2009.

[20] C. Wrzodek, A. Dräger, A. Zell. KEGGtranslator: visualizing and translating the KEGG PATHWAY database. *Bioinformatics*, 27:2314-2315, 2011.

[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25:25-9, 2000.

[22] UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, 38:D142-148, 2010.

[23] N. Le Novère. Model storage, exchange and integration. *BMC Neuroscience*, 7:S11, 2006.

[24] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y . Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, J. Ye J. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 37:D5-15, 2009.

[25] L. Kuepfer. Towards whole-body systems physiology. *Mol Syst Biol*, 6:409, 2010.

[26] M. G. Poolman, B. K. Bonde, A. Gevorgyan, H. H. Patel, D. A. Fell. Challenges to be faced in the reconstruction of metabolic networks from public databases. *IEE Proc Syst Biol*, 153:379-84, 2006.

[27] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 28:31-36, 1988.

[28] F. Csizmadia, A. Tsantili-Kaoulidou, I. Paderi, F. Darvas. Prediction of distribution coefficient from structure. 1. Estimation method. *J Pharm Sci*, 86:865– 871, 1997.

[29] A. Gevorgyan, M. G. Poolman, D. A. Fell. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, 24:2245-2251, 2008.

[30] M. G. Poolman. ScrumPy: metabolic modelling with Python. *IEE Proc Syst Biol*, 153:375-378, 2006.

[31] M. A. Ott, G. Vriend. Correcting ligands, metabolites, and pathways. *BMC Bioinformatics*, 7:517, 2006.

[32] P. Horton, K. J. Park, T. Obayashi, N. Fujita, H. Harada, C. J. Adams-Collier, K. Nakai. WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, 35:W585-7, 2007.

[33] K. Nakai, P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24:34-36, 1999.

[34] P. D. Dobson, D. B. Kell. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? *Nat Rev Drug Discov*, 7:205-20, 2008.

[35] P. D. Dobson, K. Lanthaler, S. G. Oliver, D. B. Kell DB. Implications of the dominant role of transporters in drug uptake by cells. *Curr Top Med Chem*, 9:163-81, 2009.

[36] J. Schellenberger, J. O. Park, T. M. Conrad, B.Ø. Palsson. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11:213, 2010.

[37] D. A. Fell, M. G. Poolman, A. Gevorgyan. Building and analysing genome-scale metabolic models. *Biochem Soc Trans*, 38:1197-201, 2010.

[38] J. Allen, H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver, D. B. Kell. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol*, 21:692-6, 2003.

[39] I. Thiele, B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5:93-121, 2010.

[40] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, C. Evelo. WikiPathways: pathway editing for the people. *PLoS Biol*, 6:e184, 2008.

[41] Y. Matsuoka, S. Ghosh, N. Kikuchi, H. Kitano. Payao: a community platform for SBML pathway model curation. *Bioinformatics*, 26:1381-3, 2010.

[42] S. Pabinger, R. Rader, R. Agren, J. Nielsen, Z. Trajanoski. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *BMC Syst Biol*, 5:20, 2011.

[43] I. Thiele, B. Ø. Palsson. Reconstruction annotation jamborees: a community approach to systems biology. *Mol Syst Biol*, 6:361, 2010.

[44] R. M. Fleming, I. Thiele, H. P. Nasheuer. Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. *Biophys Chem*, 145:47-56, 2009.

[45] R. M. Fleming, I. Thiele. von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. *Bioinformatics*, 27:142-3, 2011.

[46] S. Mintz-Oron, A. Aharoni, E. Ruppin, T. Shlomi. Network-based prediction of metabolic enzymes' subcellular localization. *Bioinformatics*, 25:i247-52, 2009.

[47] Q. Ren, K. Chen, I. T. Paulsen. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res*, 35:D274-279, 2007.

[48] B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, J. Tsujii. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26:i374-381, 2010.

[49] C. Nobata, P. D. Dobson, S. A. Iqbal, P. Mendes, J. Tsujii, D. B. Kell, S. Ananiadou. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, 7:94-101, 2001.