Briefings in Bioinformatics, 16(6), 2015, 1057–1068

doi: 10.1093/bib/bbv003
Advance Access Publication Date: 28 February 2015
Paper

Critical assessment of genome-scale metabolic networks: the need for a unified standard

Aarthi Ravikrishnan and Karthik Raman

Corresponding author. Karthik Raman, BT 104, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India. Tel.: +91-44-2257-4139; Fax: +91-44-2257-4102; E-mail: kraman@iitm.ac.in

Abstract

Genome-scale metabolic networks have been reconstructed for several organisms. These metabolic networks provide detailed information about the metabolism inside the cells, coupled with the genomic, proteomic and thermodynamic information. These networks are widely simulated using ‘constraint-based’ modelling techniques and find applications ranging from strain improvement for metabolic engineering to prediction of drug targets in pathogenic organisms. Components of these metabolic networks are represented in multiple file formats and also using different markup languages, with varying levels of annotations; this leads to inconsistencies and increases the complexities in comparing and analysing reconstructions on multiple platforms. In this work, we critically examine nearly 100 published genome-scale metabolic networks and their corresponding constraint-based models and discuss various issues with respect to model quality. One of the major concerns is the lack of annotations using standard identifiers that can uniquely describe several components such as metabolites, genes, proteins and reactions. We also find that many models do not have complete information regarding constraints on reactions fluxes and objective functions for carrying out simulations. Overall, our analysis highlights the need for a widely acceptable standard for representing constraint-based models. A rigorous standard can help in streamlining the process of reconstruction and improve the quality of reconstructed metabolic models.

Key words: genome-scale metabolic networks; reconstruction; blocked reactions; standards; SBML

Introduction

Genome-scale metabolic networks (GSMNs) provide an integrated view of cellular metabolism by incorporating the information obtained from literature, as well as genomic, proteomic, metabolomic and thermodynamic data [1]. Over the recent years, there has been a growing interest in the reconstruction of these metabolic networks, which can be used to infer metabolic capabilities of the cells, with applications ranging from the prediction of drug targets in pathogenic organisms [2–4] to strain design for overproduction of commercially important metabolites [5–7]. Further, the reconstructions permit the integration of high-throughput data such as ^{13}C flux measurements [8,9] and microarray data [10]. They also aid in analysing the relationships between the organisms in a community [11, 12] and guide the discovery of network properties [13, 14]. These widespread applications have led to the development of multiple genome-scale metabolic models for various organisms that are represented in different formats catering to the needs of respective toolboxes. Despite such extensive use of genome-scale reconstructions, a standardized form for their representation is still lacking. In the recent years, some formats have become popular for the exchange of these GSMNs, notably the constraint-based reconstruction and analysis (COBRA) format built on Systems Biology Markup Language (SBML) [15], but despite this, many issues such as the lack of annotations to uniquely identify metabolites remain.

GSMNs are constructed from sequenced genome information, biochemical evidence and available literature [1]. As an initial step, automated genome annotation that provides

Aarthi Ravikrishnan is a graduate student at the Department of Biotechnology, Indian Institute of Technology Madras. She works in the area of metabolic engineering, with a mix of *in silico* simulations and wet lab experiments.
Karthik Raman is an assistant professor at the Department of Biotechnology, Indian Institute of Technology Madras. His group works on the modelling of biological networks, with applications in metabolic engineering and drug target identification.

Submitted: 5 November 2014; **Received (in revised form):** 9 January 2015

information about the genes is generated from different databases. These databases could either be organism-specific like EcoCyc [16], *Saccharomyces* Genome Database [17] and *Ashbya* Genome Database [18] or a collection of genome annotations such as GenBank [19], The Institute of Genome Research TIGR, (http://www.tigr.org/) and Comprehensive Microbial Resource [20]. The information is then mapped to the biochemical reactions that take place inside the cell by extracting the data from different metabolic databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [21], MetaCyc [22] and TransportDB [23]. This step is followed by manual curation of the reconstructed network, which involves inclusion of compartments, organism-specific transporters, cofactors and directionality of the reactions. The complete information about different reconstructions can be found in Supplementary Table S1.

The assembled GSMNs are simulated by applying constraints that may often be experimentally defined [24, 25]. These constraints define the 'phenotypic space' that can be explored to gather information about different phenotypes and infer the metabolic capabilities of the cell [26]. GSMNs are often referred to as (genome-scale metabolic) models after the constraints are applied. Several 'constraint-based modelling' techniques have emerged, which make use of these metabolic models. One such technique is flux balance analysis (FBA), which uses the stoichiometric information in the GSMNs to predict the growth rate of an organism as well as the flow of metabolites through different pathways [27, 28]. Several enhancements to FBA, such as minimization of metabolic adjustment [29] and regulatory on/off minimization [30], also rely on the information provided by the reconstructions. A broad overview of constraint-based modelling techniques has been described in [31]. The increasing number of such techniques has led to a concurrent increase in the number of software and toolboxes tailored for the analysis of metabolic networks, e.g. COBRA Toolbox [15], RAVEN Toolbox [32] and OptFlux [33]. A more detailed review on the software used for constraint-based modelling can be found elsewhere [34, 35].

The reconstruction process is facilitated by several platforms, such as KEGGtranslator [36], rBioNet [37], Pathway Tools [38], ModelSEED [39], Flux Analysis and Modelling Environment [40] and MicrobesFlux [41]. Each of these tools retrieve pathway information from different databases and provide several modes of representation and simulations. The availability of alternate representations introduces non-uniformities in several components of the GSMNs, such as reaction identifiers, metabolite identifiers and representation of compartments. Further, these details are not well-catalogued in all formats, which greatly reduce the scope of reproducibility and flexibility in utilizing them on various software tools. It also hampers the automated integration of high-throughput data and greatly reduces the chances of comparison between different GSMNs [42]. The increasing number of reconstructions in varied formats and with different sets of annotations makes it difficult to utilize and analyse these networks. A recent review also focuses on the importance of optimizing and generating high-quality metabolic network reconstructions for improving the overall progress of the field [42].

With the growing number of these reconstructions and their applications, uniformity in representation of different components becomes crucial. In this work, we critically assess nearly a hundred published network reconstructions in aspects such as specification of metabolite identifiers, extent of blocked reactions and presence of gene-protein-reaction (GPR) associations. Specifically, we seek to answer the following questions: are the genome-scale metabolic networks completely specified in terms of various components such as metabolite identifiers and reaction identifiers? What is the extent of blocked reactions in each of these models? Do these models have sufficient information such as objective function and experimental bounds to enable their simulation? We also outline common issues with these metabolic reconstructions and consequently, a set of recommendations, which may aid in establishing a common denominator for all GSMNs.

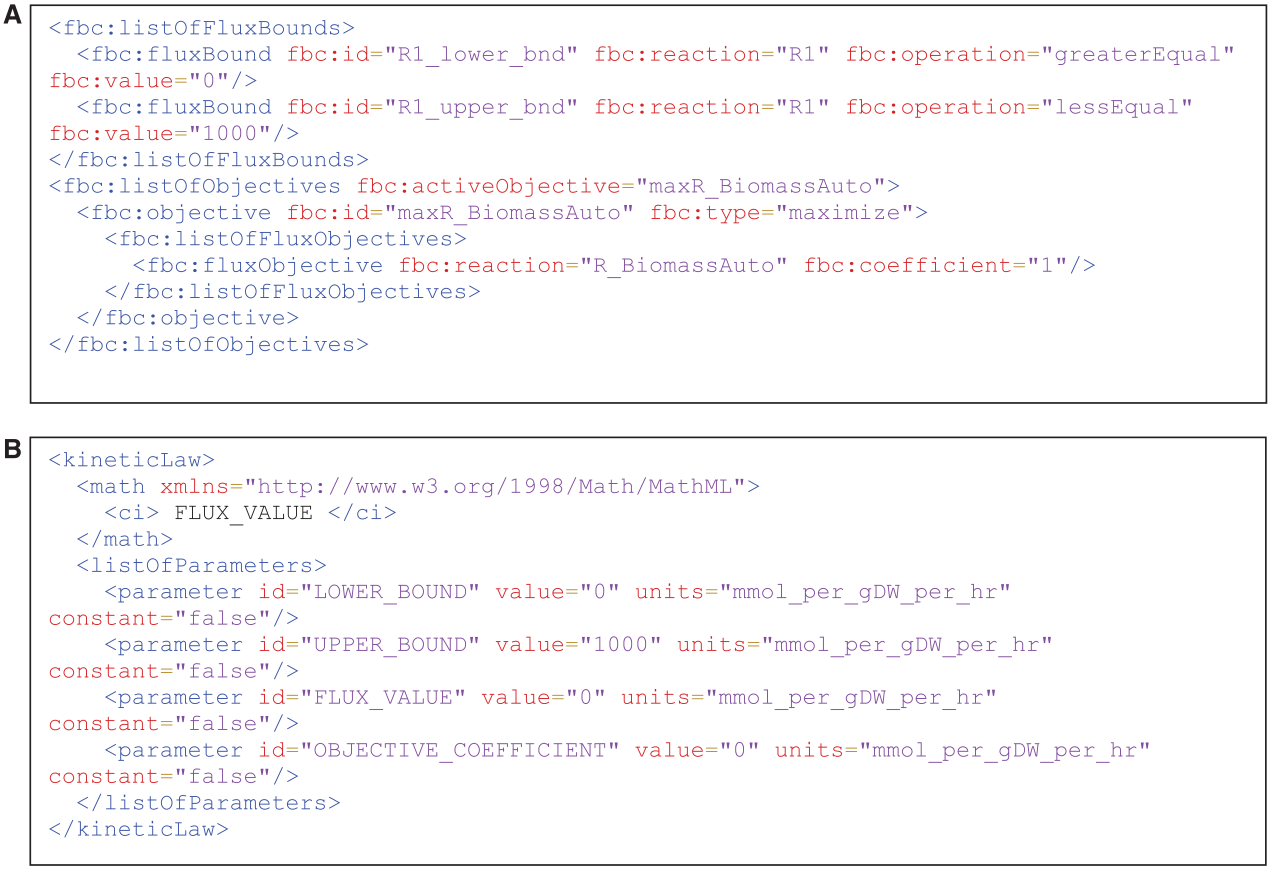
Assessment of metabolic models

Metabolic models are extensively simulated using constraint-based analysis techniques such as FBA [25, 43]. FBA considers the steady-state mass balance in a cell, and identifies the most plausible flux distribution, by optimizing a particular (linear) objective function [44], usually maximum growth, and satisfying several known biochemical and thermodynamic constraints. The FBA framework therefore necessitates the specification of the constraints, such as upper and lower bounds for the fluxes, and the objective function.

We present the results of our analyses of various GSMNs/models, on different axes, such as the representation formats, availability of annotations for genes, metabolites and proteins, information on GPR associations, availability of complete information (exact specification of constraints) for enabling simulations and the extent of blocked reactions. We carried out our analysis on the SBML models using the COBRA Toolbox v2.0 accessed via MATLAB R2012b (Mathworks Inc.) and GUROBI Solver (v5.6.3, Gurobi Inc.). We also adopted COBRApy [45] for reading and converting SBML files that could not be interfaced through COBRA in MATLAB. COBRApy, an object-oriented programming implementation, not only handles SBML models with different structures (e.g. COBRA and FBC, as described in the following section) but also models incorporating expression data. The models available as spreadsheets were analysed manually for different components of the metabolic network such as the number of reactions, genes, presence of metabolite identifiers and bounds on the reactions using built-in functions of Microsoft Excel.

Paradigms for GSMN representation

Genome-scale metabolic reconstructions are popularly exchanged in the SBML format [46], while there are other types of languages like Metabolic Flux Analysis Markup Language [47] and Web Ontology Language-based BioPAX format [48]. Although the GSMNs generated in SBML format essentially follow the basic framework of XML (eXtensible Markup Language), the syntax and semantics vary widely. A very recent review on all the relevant community modelling standards can be found in [49]. In fact, the most widely used XML-based SBML format differs in several aspects based on specifications laid down by different packages: for instance, SBML, as popularized by the COBRA toolbox [15], uses 'notes' to specify metabolite information, gene associations and named parameters that specify flux constraints, while the proposed version of Flux Balance Constraints (FBC) package for SBML (http://sbml.org/Documents/Specifications/SBML_Level_3/Packages/fbc) encourages the use of specific tags for representing the same. Figures 1 and 2 illustrate the representation of reaction constraints and gene–protein reaction associations in COBRA and FBC formats, respectively. Indeed, COBRApy extensively supports FBC-compliant SBML file.



A

```
<fbc:listOfFluxBounds>
  <fbc:fluxBound fbc:id="R1_lower_bnd" fbc:reaction="R1" fbc:operation="greaterEqual"
fbc:value="0"/>
  <fbc:fluxBound fbc:id="R1_upper_bnd" fbc:reaction="R1" fbc:operation="lessEqual"
fbc:value="1000"/>
</fbc:listOfFluxBounds>
<fbc:listOfObjectives fbc:activeObjective="maxR_BiomassAuto">
  <fbc:objective fbc:id="maxR_BiomassAuto" fbc:type="maximize">
    <fbc:listOfFluxObjectives>
      <fbc:fluxObjective fbc:reaction="R_BiomassAuto" fbc:coefficient="1"/>
    </fbc:listOfFluxObjectives>
  </fbc:objective>
</fbc:listOfObjectives>
```

B

```
<kineticLaw>
  <math xmlns="http://www.w3.org/1998/Math/MathML">
    <ci> FLUX_VALUE </ci>
  </math>
  <listOfParameters>
    <parameter id="LOWER_BOUND" value="0" units="mmol_per_gDW_per_hr"
constant="false"/>
    <parameter id="UPPER_BOUND" value="1000" units="mmol_per_gDW_per_hr"
constant="false"/>
    <parameter id="FLUX_VALUE" value="0" units="mmol_per_gDW_per_hr"
constant="false"/>
    <parameter id="OBJECTIVE_COEFFICIENT" value="0" units="mmol_per_gDW_per_hr"
constant="false"/>
  </listOfParameters>
</kineticLaw>
```

Figure 1. Different types of representations of reaction constraints in SBML format: Panel **A** illustrates how reaction constraints such as upper and lower bounds are specified using the Flux Balance Constraints package, while Panel **B** illustrates the same in the COBRA format. Note that the list of objectives in the FBC package specifies the list of reactions that serve as objective function. On the other hand, the 'objective coefficient' in the COBRA format is individually specified for every reaction.

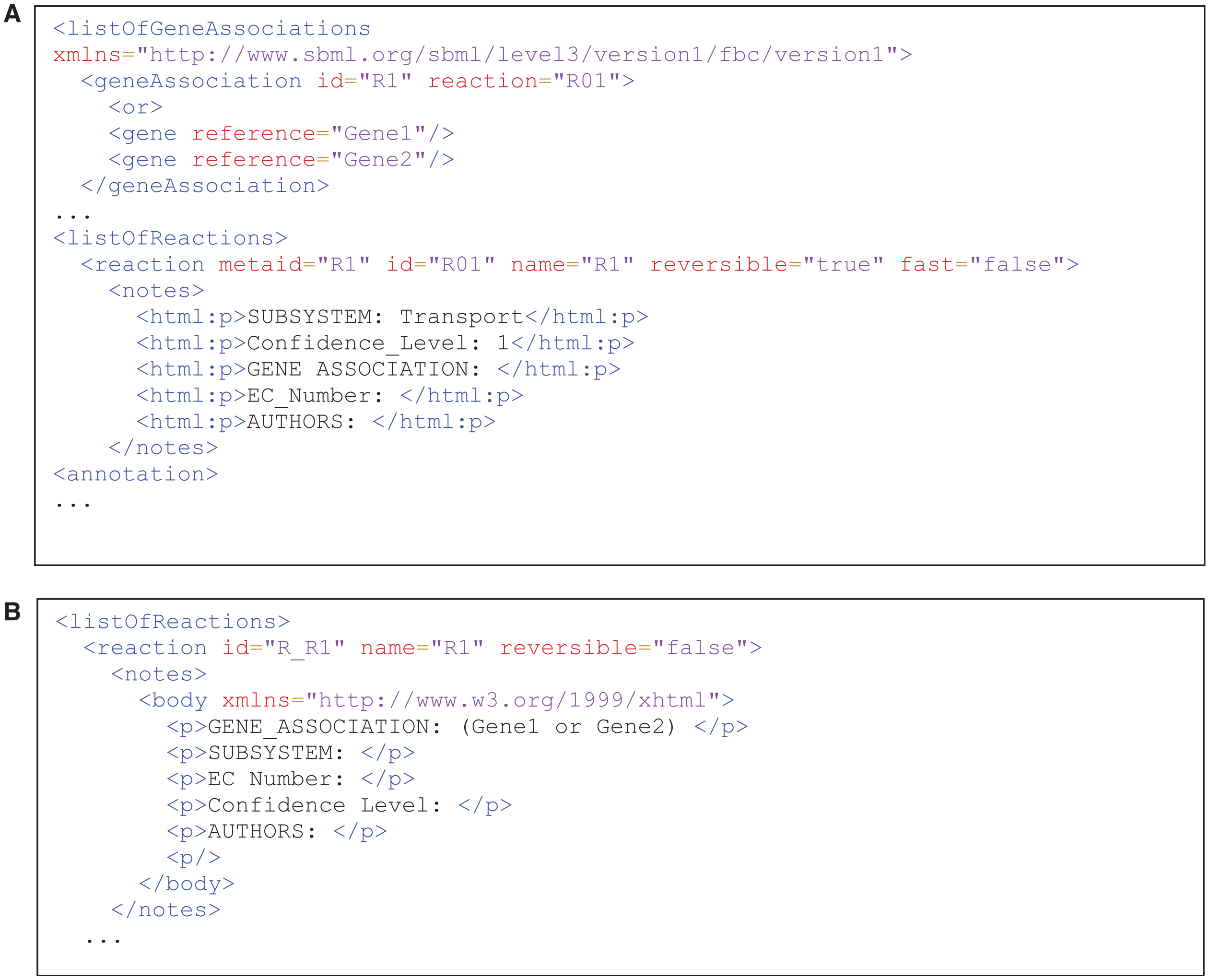
In addition to the formats mentioned above, many of the models are represented as text files and spreadsheets, thereby increasing the complexity in analysing these networks. While spreadsheets are convenient for tabulating and consolidating diverse data in metabolic networks, they do not lend themselves easily to simulation; the COBRA toolbox [50] for MATLAB does have a method to convert Excel spreadsheet files to XML, but the process may not always be straightforward and error-free. Overall, the availability of different incompatible formats hinders our ability to simulate the reconstructions on different platforms. The examples of GSMNs presented in different formats can be found in Table 1, and a detailed description of the models can be found in Supplementary Table S1. Our analysis of 99 GSMNs shows that the most widely used formats for representation are SBML and spreadsheets. We observe that 58 of the reconstructions are exchanged as SBML files and 28 are found only as Microsoft Excel spreadsheet (XLS) files; while 35 are available in both SBML format and as spreadsheet files. Besides, we note that the models are also exchanged as PDF, MATLAB mat data files and Microsoft Word document files. Supplementary Table S2 details the different formats in which the models are represented.

Annotations of metabolic networks components

The representation of different components of the GSMN like the metabolite identifiers, genes, compartments and the objective function assume a crucial role while the network is used in simulations. These components provide insights into the features of the metabolic network and help quantify different properties of the reconstructions such as growth rates and fluxes through different pathways. The inclusion of these identifiers in GSMNs is predominantly a part of the reconstruction process and involves careful manual curation from various databases. Many times during this process, several *ad hoc* notations are introduced, which lead to a wide spectrum of names and identifiers even for common reactions and metabolites in metabolic networks. In the following sections, we elucidate the importance of different components of the GSMNs and demonstrate the inconsistencies observed across the GSMNs analysed.

Metabolite identifiers

Because a GSMN provides a catalogue of the biochemical reactions taking place in an organism, it is important to uniquely identify the corresponding metabolites, genes and proteins, through annotations that uniquely map these components to standard databases. Metabolic networks are usually identified through references to different standard databases. The primary metabolite identifiers are PubChem Compound ID [62], KEGG Compound ID [21] and Chemical Entities of Biological Interest (ChEBI) ID [63], while the structure-based identifiers include IUPAC International Chemical Identifier (InChI) and the Simplified Molecular-Input Line-Entry System (SMILES) from the respective databases. The structure-based identifiers like



A

```
<listOfGeneAssociations
xmlns="http://www.sbml.org/sbml/level3/version1/fbc/version1">
  <geneAssociation id="R1" reaction="R01">
    <or>
    <gene reference="Gene1"/>
    <gene reference="Gene2"/>
  </geneAssociation>
...
<listOfReactions>
  <reaction metaid="R1" id="R01" name="R1" reversible="true" fast="false">
    <notes>
      <html:p>SUBSYSTEM: Transport</html:p>
      <html:p>Confidence_Level: 1</html:p>
      <html:p>GENE ASSOCIATION: </html:p>
      <html:p>EC_Number: </html:p>
      <html:p>AUTHORS: </html:p>
    </notes>
<annotation>
...
```

B

```
<listOfReactions>
  <reaction id="R_R1" name="R1" reversible="false">
    <notes>
      <body xmlns="http://www.w3.org/1999/xhtml">
        <p>GENE_ASSOCIATION: (Gene1 or Gene2) </p>
        <p>SUBSYSTEM: </p>
        <p>EC Number: </p>
        <p>Confidence Level: </p>
        <p>AUTHORS: </p>
        <p/>
      </body>
    </notes>
  ...
```

Figure 2. Different types of representations of Gene Protein Reaction Association in SBML format: Panel **A** illustrates how the gene-protein-reaction associations are specified in Flux Balance Constraints package, while Panel **B** illustrates the same in COBRA Format. Note that the FBC package, in addition to the `geneAssociation` field, consists of a separate notes field, which contains the details about the GPRs involved in the given reaction.

InChI also provide details on the stereochemistry of the compounds [64].

Our analysis of the reconstructions for the presence of standard metabolite identifiers illustrates that nearly 60% of the models lack any of the standard metabolite identifiers, such as KEGG LIGAND ID, PubChem Compound ID and InChI. Figure 3 illustrates an example of an SBML reconstruction where identifiers are mapped to the standard database of metabolites, while Figure 4 represents a bar graph showing the distribution of models having specific metabolite identifiers. Besides, few reconstructions possess database specific identifiers, such as SEED ID, thereby making their interpretation even more circuitous.

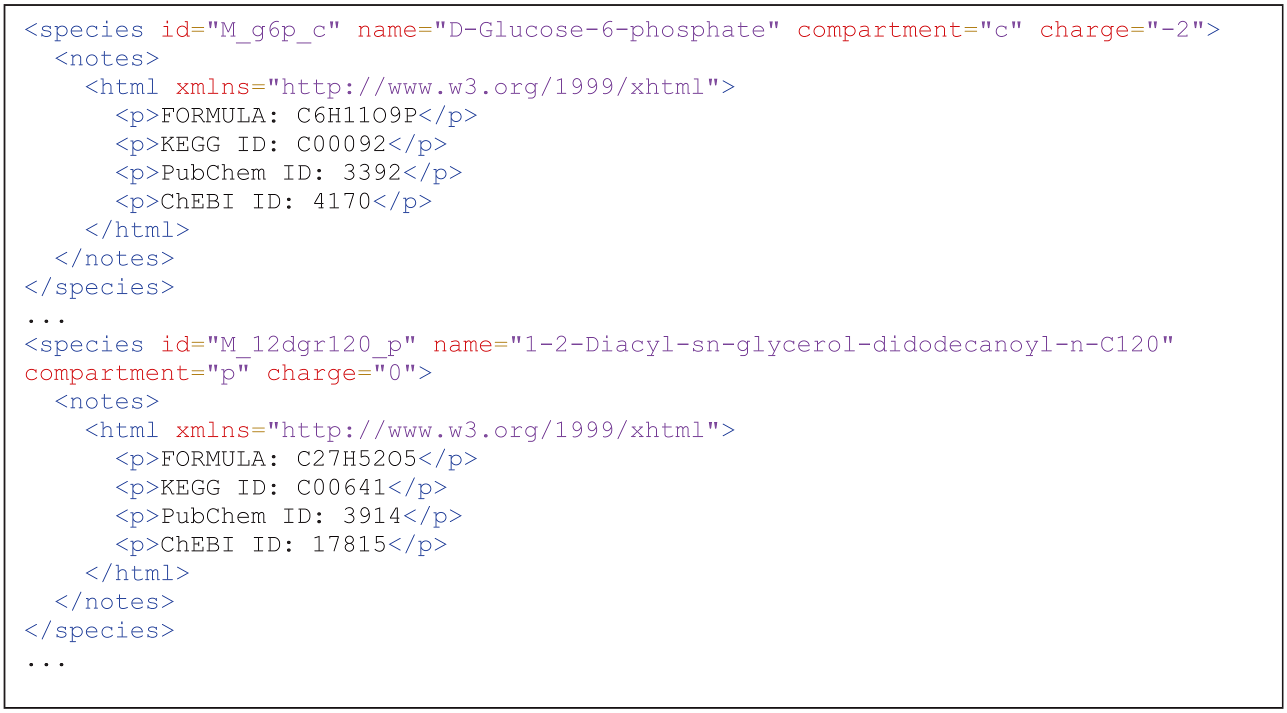
Mass and charge balance of reactions

We also examined the reactions in the GSMNs for mass and charge balances, using the molecular formulae of the metabolites involved. The stoichiometric inconsistencies arising out of the unbalanced reactions in the metabolic network not only affect the accuracy of predictions but also violate the fundamental law of mass conservation [65], which is the key

Table 1: Examples of reconstruction formats

Serial number	Format	Example GSMNs in this format
1	SBML:COBRA	*Aspergillus terreus* iJL1454 [51], *Kluyveromyces lactis* iOD907 [52], *Pseudomonas putida* iJP815 [53], *Salmonella enterica* subsp. enterica serovar Typhimurium str. LT2 iIT1271 [2], *Spirulina platensis* C1 iAK692 [54], *Geobacter metallireducens* iAF987 [55]
2	SBML:RAVEN	*Saccharomyces cerevisiae* iTO977 [56], *Penicillium chrysogenum* Wisconsin 54-1255 iAL1006 [32], *Pseudomonas fluorescens* SBW25 iSB1139 [57]
3	SBML: FBC	*Synechocystis* sp. PCC 6803 iTM686 [58], *Geobacter metallireducens* iAF987 [55]
4	Simpheny	*Saccharomyces cerevisiae* iMM904 [59]
5	Spreadsheets	*Pichia stipitis* iBB814 [60], *Pichia pastoris* iLC915 [61]

Note. The table lists example of formats adhered to by different GSMNs.



```
<species id="M_g6p_c" name="D-Glucose-6-phosphate" compartment="c" charge="-2">
  <notes>
    <html xmlns="http://www.w3.org/1999/xhtml">
      <p>FORMULA: C6H11O9P</p>
      <p>KEGG ID: C00092</p>
      <p>PubChem ID: 3392</p>
      <p>ChEBI ID: 4170</p>
    </html>
  </notes>
</species>
...
<species id="M_12dgr120_p" name="1-2-Diacyl-sn-glycerol-didodecanoyl-n-C120"
compartment="p" charge="0">
  <notes>
    <html xmlns="http://www.w3.org/1999/xhtml">
      <p>FORMULA: C27H52O5</p>
      <p>KEGG ID: C00641</p>
      <p>PubChem ID: 3914</p>
      <p>ChEBI ID: 17815</p>
    </html>
  </notes>
</species>
...
```

Figure 3. Example representation of metabolite identifiers in models: figure illustrates the presence of metabolite identifiers in an SBML reconstruction following COBRA format. Note the presence of unique identifiers pointing to standard databases such as KEGG, PubChem and ChEBI.

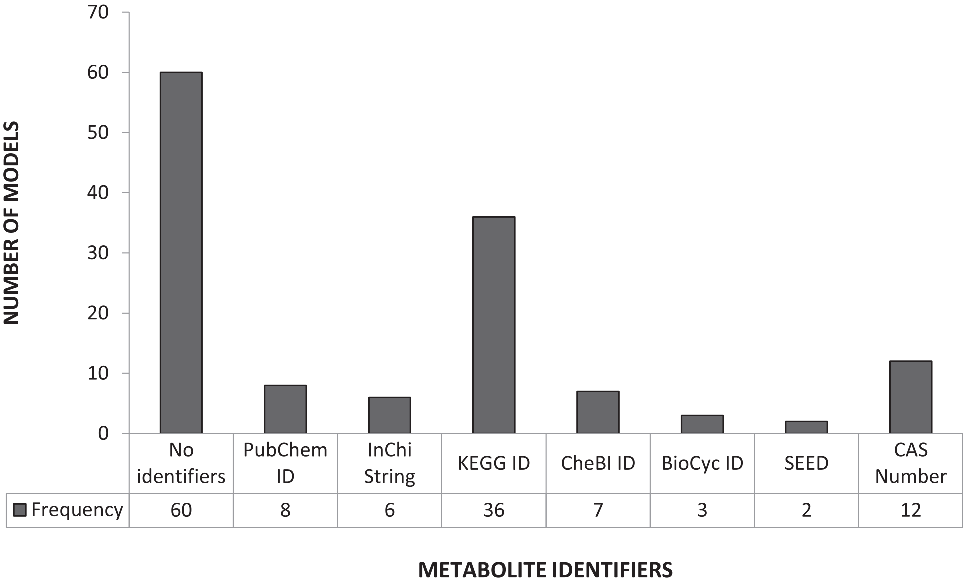


Figure 4. Presence of metabolite identifiers in the GSMNs: figure shows a bar graph for the number of models having different metabolite identifiers, which can be used to uniquely identify the metabolite and also retrieve the appropriate information from various databases.

principle on which methods such as FBA are based. It has been previously shown that rebalancing all the unbalanced reactions of a GSMN had significant impact on the predictions of FBA [66].

Although the reconstructions exchanged as spreadsheets contained the metabolite formulae, we limited our analysis to GSMNs exchanged as SBML files due to the difficulties involved in parsing spreadsheets in different formats. Of the 59 reconstructions in SBML format, we were able to perform this analysis only on 21 models. The other SBML reconstructions either did not follow the structure specified by COBRA Toolbox or did not contain metabolite formulae. We observe that most of the unbalanced reactions result owing to the differences in hydrogen atom numbers (Supplementary Table S3). For a few reactions in any given reconstruction, we were unable to determine the mass balance owing to the absence of corresponding metabolite formulae (Figure 5). It is interesting to note that in few cases such as

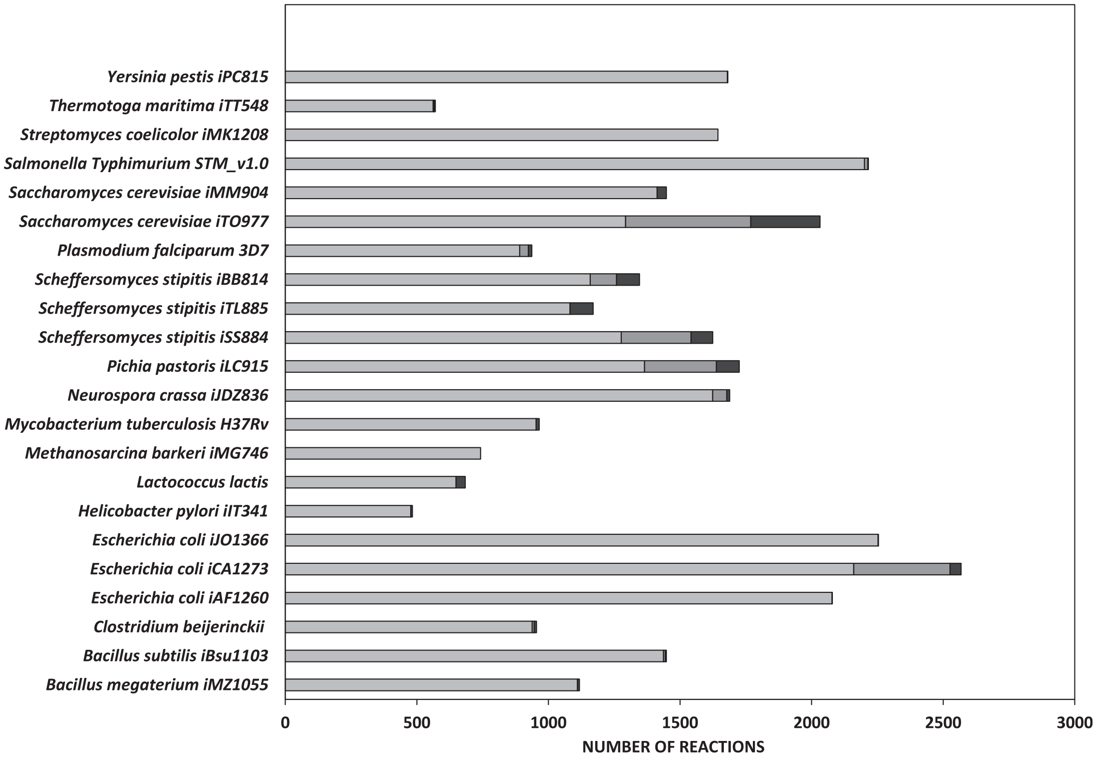


Figure 5. Mass and charge balance analysis: figure shows a stacked graph representing the fraction of reactions that are balanced (light grey), fraction of reactions for which the mass balance cannot be determined (dark grey) owing to absence of metabolite formulae corresponding to that reaction and the fraction of reactions that are unbalanced (black). Note that few reconstructions such as *Methanosarcina barkeri* iMG746 and *Streptomyces coelicolor* iMK1208 do not have any unbalanced reactions, while some reconstructions such as *Escherichia coli* iJO1366 and *Escherichia coli* iAF1260 have just one unbalanced reaction.

Streptomyces coelicolor iMK1208 [67], all the reactions were balanced, while in some others such as *Escherichia coli* iAF1260 [68], only the biomass reaction was unbalanced. However, to reduce the model complexity, reactions such as biomass production and exchange reactions are often left unbalanced.

Genes, proteins and GPR associations

We also checked for the presence of the enzymes participating in different reactions of the GSMNs. The enzymes are represented with their Enzyme Commission numbers, obtained from databases such as BRaunschweig ENzyme DAtabase (BRENDA) [69] and ExPaSy [70]. Apart from these, the GSMNs also harbour unique protein identifiers retrieved from databases such as Universal Protein Resource (UniProt) [71] and SWISS-PROT [72]. We analysed the reconstructions for the presence of the protein identifiers, and find nearly 25% of them lack information on the protein identifiers.

Another important component of GSMNs is the presence of information about genes/proteins that are involved in the catalysis of a reaction, indicated as GPRs. The gene and protein identifiers are crucial, as they help in deriving useful hypotheses about gene essentiality and understanding the regulatory mechanisms of the biological systems. The gene IDs, as mentioned earlier, are usually retrieved from GenBank [19] or organism-specific databases, and the GPR associations are formulated as Boolean rules based on the participation of each gene and its associated product in every reaction. These rules also take into account the presence of isozymes and multifunctional enzymes in the metabolic network.

The presence of information on genes is critical when the model is applied for strain optimization using constraint-based modelling. They also help in deriving useful hypotheses about gene essentiality and understanding the regulatory mechanisms of the biological systems. Several previous studies rely on the gene information provided by the GSMNs to discover and formulate hypothesis about gene essentiality [73, 74]. We find that nearly all GSMNs (90%) provide details of the GPR associations, in either SBML or XLS files. However, nearly 35% of the models available in both XLS and SBML formats fail to incorporate the GPRs in their respective SBML files, making it difficult to generate predictions on gene knockouts or overexpression. Supplementary Table S4 lists the information on genes and GPRs present in various reconstructions. It is often the case that the same reconstruction, when exchanged in different formats, say as SBML and XLSs, fails to incorporate all the information in both formats.

Analysis of simulation capabilities of the models

Metabolic models are often simulated to capture vital information like growth rate, essential genes, active pathways and essential reactions in a given medium. The metabolic models that underlie these metabolic networks are highly underdetermined; solving them requires the specification of additional constraints, to narrow down the space of solutions that are more likely to be biologically feasible. The constraints could be physico-chemical, environmental or thermodynamic, which are defined to analyse different functions of a biological system. Mathematically, these are indicated as the bounds and the balances on the reaction fluxes. The former are often derived from enzyme capacity measurements and environmental conditions, whereas the latter are derived from fundamental mass, energy and charge conservation [26]. The integration of the objective function and environment/medium conditions into the metabolic network is a part of the reconstruction step itself and plays a crucial role in determining the usability of the model for simulations. Often in FBA, biomass maximization is chosen as the objective function; the biomass equation is formulated based on cellular composition and energy requirements of the cell [75]. In models such as human tissues, where maximizing growth rate has little meaning, an alternate objective function is chosen [76]. Moreover, in alternate FBA formulations such as flux minimization [77] and loopless FBA [78], biologically meaningful solutions can be obtained by constraining the uptake and secretion bounds.

In some cases, transcriptomic data such as those obtained from gene expression and protein expression are incorporated in metabolic models to further constrain the flux space of possible solutions [79]. To this end, several methods such as integration of Boolean rules defining gene expression [79], E-Flux [10], PROM [80] and MADE [81] have been developed. Although such regulatory constraints improve metabolic predictions, their applications are still restricted owing to the lack of understanding about the transcriptional mechanisms, absence of a significant correlation between gene and protein expressions and experimental variations in detecting different levels of transcripts [82].

We note that around 88% of the models use 'default bounds' for simulations, i.e. $0\,\mathrm{mmol\,gDCW^{-1}h^{-1}}$ to $1000\,\mathrm{mmol\,gDCW^{-1}h^{-1}}$ for irreversible reactions and $-1000\,\mathrm{mmol\,gDCW^{-1}h^{-1}}$ to $1000\,\mathrm{mmol\,gDCW^{-1}h^{-1}}$ for reversible reactions. More details on the bounds of the reactions in the models can be found in the Supplementary Table S5. While these sets of constraints do lead to the generation of growth rates, to obtain a biologically feasible solution, the inclusion of the experimental constraints based on experimental measurements is necessary. These constraints ensure that the predictions of models do not diverge too far from the reality. We evaluated the biomass formation in each of these SBML models by applying environmental conditions specified in the corresponding publication. We find that some models do not explicitly provide the constraints and objective function required for simulation. In such cases, we tried to simulate the models for growth by unrestricted exchange of all possible metabolites. Despite this, we were able to precisely correlate the growth rates of only 22 SBML models (Supplementary Table S6). We also note that the regulatory data have been incorporated in few genome-scale models such as *Saccharomyces cerevisiae* iFF708 [83], *Chlamydomonas reinhardtii* iRC1080 [84] and *Pseudomonas aeruginosa* iMO1056 [85].

Blocked reactions

Blocked reactions in metabolic networks are those reactions that cannot carry a flux at steady state in a given growth medium. These reactions may arise owing to multiple reasons: (i) the metabolic network model is incomplete, with gaps (unknown connecting reactions) between the metabolites, (ii) medium conditions used to simulate the model and (iii) incorrect bounds on the reactions. Although these reactions are superfluous for simulations, it is often necessary to specify them in the model, as they highlight the knowledge gaps in the reconstructed metabolic network. It has been previously reported that by removing and connecting a few dead-end metabolites in *Saccharomyces cerevisiae* iFF708, the predictions of the FBA algorithm improved from 40–53% to 68–80% for single lethal deletions [77].

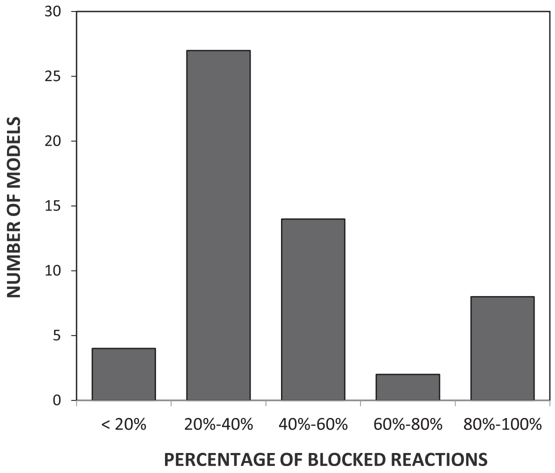


Figure 6. Blocked Reaction Analysis: histogram represents the number of models having different degrees of blocked reactions. Blocked reactions refer to those reactions that cannot carry a flux in either direction at steady state in the given growth medium.

We identified blocked reactions by solving a set of linear programming problems that involve maximizing and minimizing the flux through every reaction, to pinpoint the reactions that cannot carry fluxes in either direction [86]. For this purpose, we simulated the metabolic model with default bounds as given in the reconstructions. It is striking to note that 50% of the reconstructions have nearly 20–40% of the reactions blocked, while only 7% of all the reconstructions have <20% blocked reactions (Figure 6). We also observe that a minor fraction of the models have all their reactions to be blocked. This owes to the absence of bounds for reactions in the GSMNs, owing to which all the reactions assume a zero flux when simulated.

Compartments

All higher organisms have complex compartmentalization of the cell into multiple organelles; in prokaryotes, there are no organelles. For the purpose of modelling, prokaryotes are usually depicted as two-compartment systems with the internal compartments being 'cytoplasm' and 'periplasm', along with an 'extracellular' compartment. Eukaryotes have multiple compartments corresponding to their complex organellar organization. Compartmentalization in models helps in distinguishing the metabolic machinery from the surrounding and also helps demarcate between different organelles, which is particularly important when the models of eukaryotic organisms are reconstructed and simulated. Lack of compartment definitions makes it difficult to discern reactions taking place in different parts of the cell and accurately determine flux distributions. We find that 36% of the prokaryotic models lack 'periplasm' as one of the compartments and consist of only two compartments, i.e. 'cytoplasm' and 'extracellular', while 19% of all the models do not contain compartments at all. Besides, we also encounter several inconsistencies associated with the compartments IDs such as representation of endoplasmic reticulum with 'e', 'er' or 'r' and peroxisome with 'p' or 'x'. Supplementary Table S7 represents different compartments in the metabolic network.

Discussion

GSMNs have been reconstructed for several organisms in the past years and are being simulated for understanding the metabolic capabilities of biological systems. The reconstructions consist of components such as metabolite and reaction identifiers, genes, GPRs and protein identifiers, many of which lack a standard form of representation. Further, simulation of the metabolic models demand specification of constraints such as bounds on reactions and objective function to predict the metabolite flux through different pathways.

In this study, we assess nearly 100 reconstructions on various axes to determine the suitability of simulations on multiple toolboxes. We find multiple formats for model representation and lack of standard annotations for different components of the metabolic models such as metabolite and protein identifiers. The absence of such annotations that provide references to standard databases greatly limit our ability to compare and analyse multiple metabolic models, especially while interrogating metabolic interactions between organisms in a community.

Broadly, we make four observations that illustrate the need for a unifying standard, across reconstructions. First, we see that there are many different formats for the exchange of these models, with each format having a different set of specifications, to denote various components of the genome-scale reconstruction. Also, XLS files are most unsuitable for representation, as they are unstructured and are not easy to validate. The diversity in the formats for representation increases the complexity in analysing and simulating the GSMNs. These challenges can be surmounted if the reconstructions adhere to one standard format of representation and also harbour appropriate annotations for all the metabolites, genes, etc., which will also greatly facilitate interconversion between different formats and enhance interoperability and the ability to simulate on multiple software platforms. Also, such a standard would help overcome the customized instructions and dedicated software that one needs to adopt while simulating the models.

Second, we point out that nearly 60% of the models lack annotations for metabolites, with identifiers connecting to standard databases. Our results clearly suggest the need to include these standard metabolite identifiers to improve the quality of the reconstruction, which is often overlooked. Nevertheless, several online resources such as MetRxn [66], MetaNetX [87] have been developed; these databases strive to enable standardization of the metabolic models by assigning a common namespace for metabolites. However, these resources have been largely underutilized.

Third, we note that the GPR information is not often clearly specified in the models—even for the same model, the XLS file may have more (or less) information on the GPRs, compared with the SBML file. Nearly 35% of the models do not have well-defined gene annotations at all. These aspects limit our ability to decipher the underlying associations between genes and reactions and also predict gene targets for amplification or knockout studies.

Fourth and most important, we find that many models lack clear information on the constraints and objective functions used for the simulations. Also, we observe a large fraction of blocked reactions in many of the models. Overall, many models do not lay out a clear specification of the constraints and the environment conditions that were used to simulate the model, which severely limits our ability to simulate for models and predict growth rates.

Further, it is also desirable that reconstructions accommodate the changes in reaction bounds when simulated under different environment conditions. It is common to simulate a model under different environmental conditions (growth media) or with different objective functions (that model plausible alternate behaviours). The ability to 'compose' multiple environmental conditions and objective functions into reconstructed models will be a very valuable feature in any standard for representing these models. Currently, only a very few models such as *i*RC1080 of *Chlamydomonas reinhardtii* [84] and *i*TM560 of *Neisseria meningitis* [88] incorporate multiple objective functions in the same reconstruction. Moreover, we observe that models do not explicitly specify the choice of medium or the objective function, which makes it difficult to faithfully reproduce the original observations from other models. We also note that many of the reconstructed networks are constantly revised, based on newly acquired biochemical knowledge. In such cases, it would be instructive to adopt versioning, although incremental changes can be tracked by examining the model history. Version control becomes particularly important when the biochemical models are deposited in standard repositories. Most of the available model version control systems (VCS) operate on software codes; only few of them, such as XyDiff-algorithm-based *BiVeS* library [89] (Biochemical Model Version Control System), are suited for XML-based models. Currently, versioning is in practice for few reconstructions such as *Saccharomyces cerevisiae* [90] and *Salmonella enterica subsp. enterica serovar* Typhimurium LT2 [2].

In sum, while we find that most of the reconstructions strive to provide adequate information for analysis and simulation, the establishment of a clear standard, which will entail only minor modifications will greatly improve the quality of the reconstructions, particularly their reusability. Further, it will help us garner maximum benefits from the enormous efforts that go into the creation of these networks. Various community efforts like organization of 'jamborees' for annotating the genome-scale reconstructions of different organisms have been initiated [91]. The jamborees are aimed are reconciling and consolidating the existing reconstructions of the target organisms. Besides, there have also been considerable efforts in developing online resources that facilitate the generation and reconciliation of draft genome-scale reconstructions [66, 92, 93]. Notably, Path2Models project [93] consolidates the biochemical pathways catalogued in well-curated online resources such as KEGG and pathway tools and generates draft SBML models that comply with MIRIAM (Minimum Information Required in Annotated Models) specifications.

Although unique annotations for different model components such as reactions, metabolites, genes and compartments are a part of MIRIAM standards [94], many of these guidelines are rarely followed. Availability of different forms of encoding constraint-based models such as COBRA-compliant SBML, FBC-compliant SBML and spreadsheets have resulted in multiple ways of representing the same information. Not only does this compromise the ability to simulate the reconstructions, it also hampers the automated integration of high-throughput data. Although MIRIAM guidelines address most of these concerns, a unified standard for representing GSMNs is still lacking. We reiterate that the reconstructions must provide the following:

i. sufficient information about the reactions and the reaction IDs from the respective databases based on which the model was constructed;
ii. standard metabolite annotations like KEGG ID, InChI String, PubChem ID and metabolite formula, charge or references to the corresponding database;

Table 2: List of recommendations to FBC Package for SBML

Components	Suggestions
Reactions	EC Numbers, annotations from databases such as KEGG Reaction (in accordance with MIRIAM)
Metabolite formula	Annotations from databases such as KEGG, PubChem and ChEBI (in accordance with MIRIAM)
Compartments	Annotations that are either MIRIAM compliant or follow SBO conventions
Subsystems	An additional attribute to reaction class to specify subsystems derived from standard database such as SEED [95] (optional)
Genes and GPRs	References pointing to standard databases like KEGG (planned for FBC Version 2)
Constraints and Objective functions	Additional object under FBC FluxBound class to specify regulatory constraints, Class to specify multiple objective functions (planned for FBC Version 2)
Miscellaneous	a. Special attribute or a separate class to describe models that do not have experimental validations, such as in uncultivable bacteria where it is often challenging to determine the constraints such as uptake or the secretion rates. b. An optional attribute to reactions to specify its addition to the currently existing models (to incrementally track changes)

Note. The table provides the list of components in the metabolic network and the corresponding suggestions.

iii. information about the genes associated with the network;
iv. GPR association in Boolean format;
v. the bounds and the experimental conditions used to simulate the network;
vi. details on biomass formulation;
vii. validation about the growth rate and the values observed when simulated on different media;
viii. compartments and subsystems where the reactions occur;
ix. conventions for versioning such as model VCS.

We observe that even in the widely used *de facto* COBRA standard for SBML files, important aspects such as the standard metabolite annotations are a part of additional information that are often left empty. Further, it encourages the introduction of several *ad hoc* notations for representing different components such as genes, compartments and reaction names. This has resulted in diverse annotations representing same information. Also, objective functions and the constraints used to simulate these models are less well-defined in the COBRA format. Moreover, several functions within the commonly used COBRA toolbox are restricted to analyse SBML Level 2 files. To address these issues and capture more information, community efforts to develop FBC package for SBML have been initiated. A more recent proposal of FBC seeks to address some of these issues, including specification of genes, however does not yet have all the features. Additionally, to enhance the interoperability and guarantee a correct evaluation of the models, it would be highly desirable if information provided in the 'notes' section of COBRA format comply to MIRIAM standards. Further, a separate attribute may be provided in the model class of FBC to describe draft models of uncultivable non-model organisms for which it is difficult to determine the uptake or secretion constraints. A list of suggestions to FBC package can be found in Table 2.

Given that the quantity of genomic data being generated is steadily increasing, the time is really ripe for the community to get together and embrace a versatile, strict and widely adoptable 'gold standard' for specifying/annotating genome-scale metabolic reconstructions. Over the past 3 years, we find that reconstructions incorporate most of the information required for simulation and analysis. Nearly 85% of the reconstructions are exchanged as SBML files, which are either COBRA- or FBC-compliant, 60% of the reconstructions have at least one metabolite identifier, 65% of the reconstructions specify GPRs and the blocked reactions in these genome-scale models have reduced to an average of <33%. However, a unified standard to represent these reconstructions is still lacking. We believe that most of these inconsistencies can be sorted with the improved version of FBC package for SBML. These standards would serve as a common denominator for the model reconstruction process. Well-crafted reconstructions will provide a platform for more reliable simulations and pave the way for a better understanding of the underlying biology to help propel research forward.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors wish to thank Aditya Pratapa and Priyanka Barman for their comments and suggestions. The authors also thank three anonymous reviewers for a critical reading of the manuscript and valuable suggestions.

Funding

This work was supported by the Indian Institute of Technology Madras, the grant from the Department of Biotechnology, Government of India [BT/PR4949/BRB/10/1048/2012] and the INSPIRE fellowship, Department of Science and Technology, Government of India (to A.R.).

Key Points

- The recent years have seen a tremendous increase in the number of GSMNs, owing to the growing number of sequenced genomes and applications in multiple areas such as metabolic engineering and drug target identification.
- Many of these reconstructions fail to incorporate the standard identifiers in annotations and other features, which assume a crucial importance when the network is used for simulations and predictions.
- A number of models lack clear information on the constraints and objective functions used for the simulations and also contain a large fraction of blocked reactions: overall, many models do not lay out a clear specification of the constraints and the environment

conditions that were used to simulate the model, severely limiting our ability to perform simulations.

- As the quantity of the data generated through high-throughput techniques is steadily increasing, we need a versatile and widely adaptable 'gold standard' for specifying/annotating genome-scale metabolic reconstructions; this would aid in streamlining the standards of the growing number of reconstructions.

References

1. Feist AM, Herrgård MJ, Thiele I, *et al*. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* 2009; **7**:129–43.
2. Thiele I, Hyduke DR, Steeb B, *et al*. A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella* Typhimurium LT2. *BMC Syst Biol* 2011;**5**:8.
3. Sigurdsson G, Fleming RMT, Heinken A, *et al*. A systems biology approach to drug targets in *Pseudomonas aeruginosa* biofilm. *PLoS One* 2012;**7**:e34337.
4. Kim HU, Kim SY, Jeong H, *et al*. Integrative genome-scale metabolic analysis of *Vibrio vulnificus* for drug targeting and discovery. *Mol Syst Biol* 2011;**7**:460.
5. Li S, Gao X, Xu N, *et al*. Enhancement of acetoin production in *Candida glabrata* by *in silico*-aided metabolic engineering. *Microb Cell Fact* 2014;**13**:55.
6. Alper H, Jin Y-S, Moxley JF, *et al*. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng* 2005;**7**:155–64.
7. Choi HS, Lee SY, Kim TY, *et al*. *In silico* identification of gene amplification targets for improvement of lycopene production. *Appl Environ Microbiol* 2010;**76**:3097–105.
8. Kjeldsen KR, Nielsen J. *In silico* genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol Bioeng* 2009;**102**:583–97.
9. Widiastuti H, Kim JY, Selvarasu S, *et al*. Genome-scale modeling and *in silico* analysis of ethanologenic bacteria Zymomonas mobilis. *Biotechnol Bioeng* 2011;**108**:655–65.
10. Colijn C, Brandes A, Zucker J, *et al*. Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Comput Biol* 2009;**5**: e1000489.
11. Zhuang K, Izallalen M, Mouser P, *et al*. Genome-scale dynamic modeling of the competition between *Rhodoferax* and *Geobacter* in anoxic subsurface environments. *ISME J* 2011;**5**: 305–16.
12. Ye C, Zou W, Xu N, *et al*. Metabolic model reconstruction and analysis of an artificial microbial ecosystem for vitamin C production. *J Biotechnol* 2014;**182-183**:61–7.
13. Larhlimi A, Basler G, Grimbs S, *et al*. Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. *Bioinformatics* 2012;**28**:i502–8.
14. Kim TY, Kim HU, Lee SY. Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metab Eng* 2010;**12**:105–11.
15. Schellenberger J, Que R, Fleming RMT, *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 2011;**6**:1290–307.
16. Keseler IM, Collado-Vides J, Gama-Castro S, *et al*. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res* 2005;**33**:D334–7.
17. Christie KR, Weng S, Balakrishnan R, *et al*. *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* 2004;**32**: D311–14.
18. Hermida L, Brachat S, Voegeli S, *et al*. The *Ashbya* Genome Database (AGD)–a tool for the yeast community and genome biologists. *Nucleic Acids Res* 2005;**33**:D348–52.
19. Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2007;**35**:D21–5.
20. Peterson JD, Umayam LA, Dickinson T, *et al*. The comprehensive microbial resource. *Nucleic Acids Res* 2001;**29**: 123–5.
21. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;**28**:27–30.
22. Caspi R, Altman T, Dreher K, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2012;**40**: D742–53.
23. Ren Q, Kang KH, Paulsen IT. TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* 2004;**32**:D284–8.
24. Wiback SJ, Mahadevan R, Palsson BØ. Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the *Escherichia coli* spectrum. *Biotechnol Bioeng* 2004;**86**:317–31.
25. Chen X, Alonso AP, Allen DK, *et al*. Synergy between (13)C-metabolic flux analysis and flux balance analysis for understanding metabolic adaptation to anaerobiosis in *E. coli*. *Metab Eng* 2011;**13**:38–48.
26. Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;**2**:886–97.
27. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* 2010;**28**:245–8.
28. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform* 2009;**10**: 435–49.
29. Segrè, D., Vitkup, D., Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 2002;**99**:15112–17.
30. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 2005;**102**:7695–700.
31. Lewis NE, Nagarajan H, Palsson BØ. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nat Rev Microbiol* 2012;**10**: 291–305.
32. Agren R, Liu L, Shoaie S, *et al*. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Comput Biol* 2013;**9**:e1002980.
33. Rocha I, Maia P, Evangelista P, *et al*. OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC Syst Biol* 2010;**4**:45.
34. Lakshmanan M, Koh G, Chung BKS, *et al*. Software applications for flux balance analysis. *Brief Bioinform* 2014;**15**: 108–22.
35. Dandekar T, Fieselmann A, Majeed S, *et al*. Software applications toward quantitative metabolic flux analysis and modeling. *Brief Bioinform* 2014;**15**:91–107.
36. Wrzodek C, Dräger A, Zell A. KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics* 2011;**27**:2314–15.
37. Thorleifsson SG, Thiele I. rBioNet: A COBRA toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics* 2011;**27**:2009–10.

38. Karp PD, Paley SM, Krummenacker M, *et al*. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* 2010;**11**: 40–79.

39. Henry CS, DeJongh M, Best AA, *et al*. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 2010;**28**:977–82.

40. Boele J, Olivier BG, Teusink B. FAME, the flux analysis and modeling environment. *BMC Syst Biol* 2012;**6**:8.

41. Feng X, Xu Y, Chen Y, *et al*. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC Syst Biol* 2012;**6**:94.

42. Monk J, Nogales J, Palsson BØ. Optimizing genome-scale network reconstructions. *Nat Biotechnol* 2014;**32**: 447–52.

43. Knoop H, Gründel M, Zilliges Y, *et al*. Flux balance analysis of cyanobacterial metabolism: the metabolic network of *Synechocystis* sp. PCC 6803. *PLoS Comput Biol* 2013;**9**: e1003081.

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

45. Ebrahim A, Lerman JA, Palsson BØ, *et al*. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* 2013;**7**:74.

46. Hucka M, Finney A, Sauro HM, *et al*. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003; **19**:524–531.

47. Yun H, Lee D-Y, Jeong J, *et al*. MFAML: a standard data structure for representing and exchanging metabolic flux models. *Bioinformatics* 2005;**21**:3329–30.

48. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**: 935–42.

49. Dräger A, Palsson BØ. Improving collaboration by standardization efforts in systems biology. *Front Bioeng Biotechnol* 2014; **2**:1–20.

50. Schellenberger J, Que R, Fleming RMT, *et al*. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* 2011;**6**: 1290–1307.

51. Liu J, Gao Q, Xu N, *et al*. Genome-scale reconstruction and in silico analysis of *Aspergillus terreus* metabolism. *Mol Biosyst* 2013;**9**:1939–48.

52. Dias O, Pereira R, Gombert AK, *et al*. iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. *Biotechnol J* 2014;**9**:776–90.

53. Puchałka J, Oberhardt MA, Godinho M, *et al*. Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. PLoS Comput. Biol. 2008;**4**:e1000210.

54. Klanchui A, Khannapho C, Phodee A, *et al*. iAK692: a genome-scale metabolic model of *Spirulina platensis* C1. *BMC Syst Biol* 2012;**6**:71.

55. Feist AM, Nagarajan H, Rotaru AE, *et al*. Constraint-based modeling of carbon fixation and the energetics of electron transfer in *Geobacter metallireducens*. *PLoS Comput Biol* 2014;**10**: e1003575.

56. Österlund T, Nookaew I, Bordel S, *et al*. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Syst Biol* 2013;**7**:36.

57. Borgos SEF, Bordel S, Sletta H, *et al*. Mapping global effects of the anti-sigma factor MucA in *Pseudomonas fluorescens* SBW25 through genome-scale metabolic modeling. *BMC Syst Biol* 2013;**7**:19.

58. Maarleveld TR, Boele J, Bruggeman FJ, *et al*. A data integration and visualization resource for the metabolic network of *Synechocystis* sp. PCC 6803. *Plant Physiol* 2014;**164**: 1111–21.

59. Mo ML, Palsson BØ, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* 2009;**3**:37.

60. Balagurunathan B, Jonnalagadda S, Tan L, *et al*. Reconstruction and analysis of a genome-scale metabolic model for *Scheffersomyces stipitis*. *Microb Cell Fact* 2012;**11**:27.

61. Caspeta L, Shoaie S, Agren R, *et al*. Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials. *BMC Syst Biol* 2012; **6**:24.

62. Wang Y, Xiao J, Suzek TO, *et al*. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 2009;**37**:W623–33.

63. Hastings J, de Matos P, Dekker A, *et al*. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* 2013;**41**: D456–63.

64. Haraldsdóttir HS, Thiele I, Fleming RM. Comparative evaluation of open source software for mapping between metabolite identifiers in metabolic network reconstructions: application to Recon 2. *J Cheminform* 2014;**6**:2.

65. Gevorgyan A, Poolman MG, Fell DA. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics* 2008;**24**:2245–51.

66. Kumar A, Suthers PF, Maranas CD. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics* 2012;**13**:6.

67. Kim M, Sang Yi J, Kim J, *et al*. Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in *Streptomyces coelicolor* A3(2). *Biotechnol J* 2014;**9**: 1185–94.

68. Feist AM, Henry CS, Reed JL, *et al*. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007;**3**:121.

69. Schomburg I, Chang A, Ebeling C, *et al*. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;**32**:D431–3.

70. Gasteiger E, Gattiker A, Hoogland C, *et al*. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* 2003;**31**:3784–88.

71. Wu CH, Apweiler R, Bairoch A, *et al*. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 2006;**34**:D187–91.

72. Boeckmann B, Bairoch A, Apweiler R, *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;**31**:365–370.

73. Durot M, Le Fèvre F, de Berardinis V, *et al*. Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst Biol* 2008;**2**:85.

74. Förster J, Famili I, Palsson BØ, *et al*. Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *OMICS* 2003; **7**:193–202.

75. Feist AM, Palsson BØ. The biomass objective function. *Curr Opin Microbiol* 2010;**13**:344–9.

76. Chang RL, Xie L, Xie L, *et al*. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol* 2010;**6**:e1000938.
77. Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res* 2005;**15**: 1421–30.
78. Schellenberger J, Lewis NE, Palsson BØ. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys J* 2011;**100**:544–53.
79. Covert MW, Palsson BØ. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J Theor Biol* 2003;**221**:309–25.
80. Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 2010;**107**:17845–50.
81. Jensen PA, Papin JA. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* 2011;**27**:541–47.
82. Reed JL. Shrinking the metabolic solution space using experimental datasets. *PLoS Comput Biol* 2012;**8**:e1002662.
83. Förster J, Famili I, Fu P, *et al*. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 2003;**13**:244–53.
84. Chang RL, Ghamsari L, Manichaikul A, *et al*. Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol Syst Biol* 2011;**7**:518.
85. Oberhardt MA, Puchałka J, Fryer KE, *et al*. Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *J Bacteriol* 2008;**190**: 2790–803.
86. Burgard AP, Nikolaev EV, Schilling CH, *et al*. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* 2004;**14**:301–12.
87. Ganter M, Bernard T, Moretti S, *et al*. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* 2013;**29**:815–6.
88. Mendum TA, Newcombe J, Mannan AA, *et al*. Interrogation of global mutagenesis data with a genome scale model of *Neisseria meningitidis* to assess gene fitness in vitro and in sera. *Genome Biol* 2011;**12**:R127.
89. Waltemath D, Henkel R, Hälke R, *et al*. Improving the reuse of computational models through version control. *Bioinformatics* 2013;**29**:742–8.
90. Heavner BD, Smallbone K, Barker B, *et al*. Yeast 5 - an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Syst Biol* 2012;**6**:55.
91. Thiele I, Palsson BØ. Reconstruction annotation jamborees: a community approach to systems biology. *Mol Syst Biol* 2010;**6**: 361.
92. Chindelevitch L, Stanley S, Hung D, *et al*. MetaMerge: scaling up genome-scale metabolic reconstructions with application to *Mycobacterium tuberculosis*. *Genome Biol* 2012;**13**:r6.
93. Büchel F, Rodriguez N, Swainston N, *et al*. Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Syst Biol* 2013;**7**:116.
94. Le Novère N, Finney A, Hucka M, *et al*. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 2005;**23**:1509–15.
95. Overbeek R, Olson R, Pusch GD, *et al*. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 2014;**42**:D206–14.