ON THE RECONSTRUCTION OF THE MUS MUSCULUS GENOME-SCALE METABOLIC NETWORK MODEL

LAKE-EE QUEK1 LARS K. NIELSEN1
l.quek@uq.edu.au lars.nielsen@uq.edu.au

1 Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, St Lucia Campus, Brisbane QLD 4072, Australia

Genome-scale metabolic modeling is a systems-based approach that attempts to capture the metabolic complexity of the whole cell, for the purpose of gaining insight into metabolic function and regulation. This is achieved by organizing the metabolic components and their corresponding interactions into a single context. The reconstruction process is a challenging and laborious task, especially during the stage of manual curation. For the mouse genome-scale metabolic model, however, we were able to rapidly reconstruct a compartmentalized model from well-curated metabolic databases online. The prototype model was comprehensive. Apart from minor compound naming and compartmentalization issues, only nine additional reactions without gene associations were added during model curation before the model was able to simulate growth in silico. Further curation led to a metabolic model that consists of 1399 genes mapped to 1757 reactions, with a total of 2037 reactions compartmentalized into the cytoplasm and mitochondria, capable of reproducing metabolic functions inferred from literatures. The reconstruction is made more tractable by developing a formal system to update the model against online databases. Effectively, we can focus our curation efforts into establishing better model annotations and gene–protein–reaction associations within the core metabolism, while relying on genome and proteome databases to build new annotations for peripheral pathways, which may bear less relevance to our modeling interest.

Keywords: systems biology; metabolism; computational model; mouse

1. Introduction

Genome-scale metabolic network models (GSMs) are useful tools to represent and analyze the metabolism of an organism. They are information infrastructures containing chemically accurate descriptions of the cellular reactions and known gene–protein–reaction associations [11]. GSM provides a context to study cellular metabolism, not only to derive insights into the metabolic phenotypes that emerge from the system as a whole, but also to integrate heterogeneous datasets within a single modeling framework [1-3, 13]. Many organism-specific GSMs have been generated to date, ranging from microbial to multicellular organisms [5, 6, 11, 15].

Reconstruction of a metabolic network is a challenging task. For well-annotated genomes, a preliminary model can be assembled from online gene and protein databases; all that is required is an appropriate system for information storage and a consistent naming of network components. This is followed by an immense effort taken to curate the GSM such that the model reflects well-demonstrated and current knowledge of the organism’s metabolism. The effort increases with the degree of content fidelity required – validating network components and their interactions using direct physical evidence in
the *H. sapiens* Recon 1 model illustrate the potential challenges posed [5]. Without specialized software tools or formalized procedures, the reconstruction process is a daunting task not readily accomplished by small research groups with limited resources.

In this paper, we describe our experience with the reconstruction of the *M. musculus* GSM. We established a simple but formal approach to compile and curate a new GSM using basic software tools, namely JAVA (Sun Microsystems, Inc), Excel (Microsoft Corporation) and MATLAB (The MathWorks), which are used for information extraction, for storage and editing of the reconstructed model, and for flux simulation, respectively (Fig. 1). A new GSM is rapidly prototyped by large-scale extraction of gene, protein and reaction information from genome and proteome databases. This rudimentary GSM is then curated such that that known metabolic functions are reproduced *in silico*.

Figure 1. Workflow for the reconstruction of the *M. musculus* GSM. The GSM is compiled from online KEGG and UniProtKB databases using JAVA (grey boxes), and is stored in Excel (top-left). Gene-centric contents are parsed into reaction-centric SBML, which is a convenient intermediate for extraction of various ‘omics sub-models. One instance is the fluxomic (stoichiometric) model, which is used to curate the GSM by flux balance analysis. Flux results are visualized on a flux map drawn in Excel (bottom-left). Model curation is an iterative process, whereby the GSM is consistently reconciled against biochemical literatures and new annotation data.

Our manual curation was focused on improving connectivity and annotation of the metabolic components in core metabolism, i.e., energy metabolism and anabolic reactions required for biosynthesis of major cell components (protein, DNA, RNA, lipids and carbohydrates). Main task of curation is to identify inconsistent compound name and to fill reaction gaps in the network. During metabolic simulation, the presence of an incorrect reaction is flagged by the failure to synthesize the required biomass precursors.
or producing unbalanced inputs and outputs. In contrast, the connectivity of peripheral pathways is progressively improved by automatically deriving new annotations from well-curated metabolic databases. All manual modifications made in core metabolism are recorded to ensure traceability, which enables the successive automated update of peripheral pathway using online databases. As a whole, this approach enabled us to make a functional \textit{M. musculus} GSM available without significant upfront investment of resources, while still supporting continued improvements.

2. **Large-scale Metabolic Reconstruction**

2.1. **Gene-Reaction Assignment**

We adopted a gene-centric organization of metabolic information, in which each of the known metabolic genes is be mapped to one or many reactions. The core of the GSM was generated using the KEGG (Kyoto Encyclopedia of Genes and Genomes) genes database for \textit{M. musculus} (Release 46) \cite{8}. The gene–reaction mappings were derived from the four different flat files available for each pathway map: the GENE and RN (reaction) files, and their corresponding COORD (coordinate) files. These files can be downloaded from KEGG’s FTP site. Each \textit{mmu} (indicating \textit{M. musculus}) gene entry is mapped to a reaction entry using their positional coordinate on the pathway map, which is contained in the COORD files. This is likened to clicking the active links in KEGG’s pathway maps to download the corresponding gene and reaction documents. The process is repeated for all available pathway maps. Redundant gene–reaction entries are subsequently removed. Metabolic reconstruction from KEGG’s reaction database is readily performed \cite{9}. Here, a simple JAVA script is used. The accompanying annotation attributes were included as well, namely the gene name, enzyme name, EC (Enzyme Commission) number, UniProtKB accession number, KO (KEGG Orthology) accession number and the name of the metabolic pathways where the gene entry was found.

A weakness of the above approach is that the coverage of the gene-reaction associations is limited to reactions presented on these maps. Half of the gene entries in the global gene list (in “mmu_genome” LST file) could not be mapped to a reaction entry because their corresponding coordinates do not exist. To overcome this, the EC number associated with the gene (in “mmu_enzyme” LST file) was used instead to link to one or many reaction entries using KEGG LIGAND’s “reaction” file. We chose to use pathway map coordinates as the primary mapping mechanism, because genes in the maps are linked to specific reactions, whereas the use of EC number leads to mapping of genes to a broader reaction categories.

2.2. **Reaction Attributes and Compartamentalization**

The reaction attributes attached to each gene-reaction association are the reaction equation and reversibility. Reaction equations are derived from the “reaction” and “reaction_name” LST files contained in KEGG LIGAND. The original reaction formula
is retained, with the compounds expressed using the full chemical name and the ID (unique entry number). As each compound ID may be associated with multiple full name aliases, it is more reliable to use the compound ID as the basis to generate the stoichiometric matrix. A full name version is kept for display purposes. The reaction reversibility (and direction) is derived from the “reaction_mapformula” LST file. Where conflicting information is encountered for the same reaction from different maps, the reaction is assumed reversible. Non-mapped reactions are automatically assumed reversible by default in absence of further information.

For reaction compartmentalization, we currently only distinguish two sub-cellular localizations: cytoplasm and mitochondria. Using the UniProtKB accession number(s) gathered for each gene entry (in “mmu_uniprot” LST file), we can interrogate the UniProtKB database for the sub-cellular localization of the corresponding protein. By default, all reactions are assigned to the cytoplasm, unless there is specific information to suggest that the reaction is localized either solely in the mitochondria or in both the cytoplasm and mitochondria.

3. Manual Curation

3.1. Data storage

The main objectives of curation are (a) to reproduce the known metabolic functions in silico by filling in network gaps and (b) to remove inconsistent naming of compounds. Metabolic modeling is performed in parallel with the curation and it is important that an appropriate data storage model is chosen that supports both curation efforts and extraction of the GSM content into structured models (i.e., stoichiometric matrix).

We used Excel as a convenient interface to curate the GSM. Contents of the GSM are easily visualized and modified. From the large-scale metabolic network reconstruction, it is relatively easy to produce a tab-delimited text file that contains a unique list of gene-reaction associations, with the accompanying gene and reaction attributes, which can be imported into Excel.

However, the GSM stored in Excel is gene-centric. For metabolic (stoichiometric) modeling, the contents must be organized into a reaction-centric form. A solution is to convert the GSM (in Excel) into SBML (System Biology Markup Language) data format as an intermediate storage medium (www.sbml.org), from which the stoichiometric model is generated. The key advantages are that (1) the gene–protein–reaction–metabolite association can be described in a hierarchical format, (2) the reaction–metabolite elements are easily transformed into a stoichiometric matrix and (3) the approach is consistent with the community’s practice for storing biochemical network models. There is no specific element in SMBL allocated to store the gene–protein–reaction associations (e.g., splice-variants, isozymes, protein complex). Accordingly, additional sub-elements under the “reaction” element were created to accommodate these associations. The storage of the GSM in a hierarchical data format supports efficient
interrogation and clustering of the GSM’s content, especially for processing of omics datasets.

3.2. Checking consistency of reaction equation

Inconsistent labels used to describe the same compound are manifested as network gaps when performing stoichiometric modeling. To avoid this problem, a single candidate must be chosen to represent all other alternatives, and this chosen candidate is consistently applied throughout the GSM. The most common problem is the non-specific and specific reference to sugar stereoisomers (e.g., D-Glucose versus α-D-Glucose). Table 1 contains the modifications that were made to maintain consistent usage of compound name. Where KEGG had two identical reactions using different compound ID, only one the reaction with the desired compound name was retained.

<table>
<thead>
<tr>
<th>Compound name</th>
<th>Compound entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-Glucose → alpha-D-Glucose</td>
<td>C00031 → C00267</td>
</tr>
<tr>
<td>D-Glucose 1-phosphate → alpha-D-Glucose 1-phosphate</td>
<td>C00103a</td>
</tr>
<tr>
<td>D-Fructose → beta-D-Fructose</td>
<td>C00095 → C02336</td>
</tr>
<tr>
<td>D-Fructose 6-phosphate → beta-D-Fructose 6-phosphate</td>
<td>C00085 → C05345</td>
</tr>
<tr>
<td>D-Fructose 1,6-bisphosphate</td>
<td>C00354 → C05378</td>
</tr>
<tr>
<td>beta-D-Fructose 1,6-bisphosphate</td>
<td></td>
</tr>
<tr>
<td>N-Acetyl-alpha-D-glucosamine 1-phosphate</td>
<td>C04501 → C04256</td>
</tr>
<tr>
<td>N-Acetyl-D-glucosamine 1-phosphate</td>
<td></td>
</tr>
<tr>
<td>Electron-transferring flavoprotein → FAD</td>
<td>C04253 → C00016</td>
</tr>
<tr>
<td>Reduced electron-transferring flavoprotein → FADH2</td>
<td>C04570 → C01352</td>
</tr>
<tr>
<td>Inositol 1-phosphate → 1L-myo-Inositol 1-phosphate</td>
<td>C01177a</td>
</tr>
<tr>
<td>CMPa</td>
<td>G10621 → C00055</td>
</tr>
<tr>
<td>UDPa</td>
<td>G10619 → C00015</td>
</tr>
<tr>
<td>GDPa</td>
<td>G10620 → C00035</td>
</tr>
<tr>
<td>Dolichyl phosphate</td>
<td>G10622 → C00110</td>
</tr>
<tr>
<td>GDP-L-fucose</td>
<td>C00325 → G10615</td>
</tr>
</tbody>
</table>

Some reaction formulas in KEGG’s were found to violate atom conservation. They are typically encountered in reactions that involve (1) the synthesis and breakdown of polymers, (2) the use of a generic atom “R”, and (3) the consumption or production of H2O, H+, and redox equivalents (e.g., NAD(P)H, FADH2). In this GSM, polymers were described in the form of their corresponding monomers, and the use of the generic atom “R” was avoided. The active reaction set is checked when inconsistent atom balance is detected at the cellular input/output level during flux simulation. It was more difficult to close the balance for hydrogen and oxygen because metabolites like H2O and redox units are highly connected. In recent work, automated atom mapping algorithms have been generated to validate reaction equation for these inconsistencies [7].

a No change
3.3. Adding Membrane Transporters

Exchange equations are used to describe the inter-compartmental exchange of metabolites: cytoplasm–extracellular and cytoplasm–mitochondria. Predominantly, the exchange equations are added to the GSM on the basis that these transporters are necessary components of normal metabolic functions. For example, the uptake of macro nutrients (e.g., amino acids, glucose), the secretion of by-products (e.g., alanine, lactate, ammonia) and the exchange of free compounds (H₂O, CO₂, O₂) are added because they represent essential cellular inputs and outputs. Similarly, the intracellular exchange of compounds between the cytoplasm and mitochondria are inferred from known mitochondrial functions, such as cellular respiration, synthesis of biomass precursors (e.g., acetyl-CoA, oxaloacetate), and oxidation of aliphatic compounds (e.g., fatty acids, branched-chain amino acids). The final GSM consists of a total of 33 and 31 inter-compartmental exchange equations added for cytoplasm–extracellular and cytoplasm–mitochondria transporters, respectively.

3.4. Lumping Reactions

A series of elementary reactions catalyzed by an enzyme complex should be lumped into a single overall reaction. The importance of lumping reactions is that a single flux parameter is used to describe the activity of an enzyme complex. Physiologically, this approach is used to represent the channeling of substrate to product. In KEGG for example, the pyruvate dehydrogenase complex catalyze four separate reactions: pyruvate decarboxylation (2 steps, via an intermediate thiamine pyrophosphate cofactor), dihydrolipoyl transacetylase and dihydrolipoyl dehydrogenase (Table 2). These four reactions are summed into an overall reaction by removing the intermediate metabolites. The lumping of these reactions reinforces the fact that dihydrolipoyl dehydrogenase is not shared between pyruvate dehydrogenase, oxoglutarate dehydrogenase and branched-chain oxo-acid dehydrogenase, which adopts similar reaction mechanism.

Table 2. List of elementary reactions catalyzed by the pyruvate dehydrogenase complex. These reactions are summarized into an overall reaction equation.

<table>
<thead>
<tr>
<th>Reaction entry</th>
<th>Reactant side</th>
<th>Product side</th>
</tr>
</thead>
<tbody>
<tr>
<td>R00014</td>
<td>Pyruvate + ThPP = 2-Hydroxyethyl-ThPP + CO₂</td>
<td></td>
</tr>
<tr>
<td>R03270</td>
<td>2-Hydroxyethyl-ThPP + Lipoamide-E = S-Acetyldihydrolipoamide-E + ThPP</td>
<td></td>
</tr>
<tr>
<td>R02569</td>
<td>CoA + S-Acetyldihydrolipoamide-E = Acetyl-CoA + Dihydrolipoamide-E</td>
<td></td>
</tr>
<tr>
<td>R07618</td>
<td>Dihydrolipoamide-E + NAD⁺ = Lipoamide-E + NADH + H⁺</td>
<td></td>
</tr>
<tr>
<td>R00209</td>
<td>Pyruvate + CoA + NAD⁺ = Acetyl-CoA + CO₂ + NADH + H⁺</td>
<td></td>
</tr>
</tbody>
</table>

Lumping is also introduced to define how NADH and FADH₂ contribute their redox equivalent to the electron transport chain. Without a clear description of the mechanism of the electron transport chain, and a satisfactory proton balance in both the inner
membrane space (i.e., cristae) and the mitochondria matrix, it is more efficient and simpler to describe the electron transport chain with a generic oxidative phosphorylation reaction, using P/O ratio of 2.5 and 1.5 for NADH and FADH$_2$ respectively. Following this modification, dehydrogenase reactions that contain ubiquinone–ubiquinol cofactor pair are replaced by FAD–FADH$_2$ cofactor pair (e.g., succinate dehydrogenase). This is necessary to define the entry point of the redox equivalent generated.

3.5. Adding biomass drain equations

Similar to the concept of adding membrane transporters to describe the efflux of by-products, the biomass drain equations are incorporated into the GSM as the accumulation terms of the biomass precursors. It is useful to describe these accumulation terms individually (e.g., “Cholesterol = Cholesterol_biomass”), in order to simplify the task of uncovering the pathway gaps in each of the biosynthetic routes separately. For example, the pathway for cholesterol synthesis can be visualized independent of other biomass components by allowing only the production of cholesterol. A zero drain value indicates the presence of reaction gaps in the cholesterol pathway. This process is iterated for all biomass precursors. Once the network gaps are filled, all biomass accumulation terms are then combined into an overall biomass synthesis equation, with the appropriate coefficients assigned to each precursor to define the composition of biomass.

For the GSM, a total of 17 biomass drain equations were added. They were used to describe the accumulation of phospholipids (7), nucleotides (8), glycogen (1) and cholesterol (1). The drains of amino acid are described via their respective amino-acyl-tRNA synthetase reactions.

3.6. Finding Network Gaps

Finding breaks in metabolic pathways is not an intuitive task. PathoLogic (Pathway Tools, SRI International) is an elegant program that can infer pathway gaps, from the genome annotations, using reference pathways (e.g., MetaCyc). However, Pathway Tools does not support compartmentalization and its content is not readily transformed into a stoichiometric model for flux simulations. Instead, we adopted a few elementary approaches to find the network gaps. The priority of finding network gaps is to enable the synthesis of biomass precursors in silico. The secondary objective is to reproduce known metabolic functions deduced from biochemical literature [10].

Visual inspection of metabolic maps from KEGG PATHWAY is a quick technique to find network gaps. Operating on a similar concept as PathoLogic, one can browse through the organismal pathway maps to deduce missing reactions using visual evidence that most of the reactions in the given pathway exist. This approach is particularly effective for tracing synthetic pathways for biomass components. These pathways are generally linear, and can also be checked against biochemical literature. Overall, this approach led to the identification of six missing reactions essential for biosynthesis (Table 3). These reactions were present in the human GSM [5].
Table 3. List of new reactions identified by visual inspection of KEGG PATHWAY.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Reaction equation</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>R01321</td>
<td>CDP-choline + 1,2-Diacyl-sn-glycerol = CMP + Phosphatidycholine</td>
<td>Phospholipid</td>
</tr>
<tr>
<td>R01514</td>
<td>ATP + D-Glycerate = ADP + 3-Phospho-D-glycerate</td>
<td></td>
</tr>
<tr>
<td>R01801</td>
<td>CDP-diacylglycerol + sn-Glycerol 3-phosphate = CMP + Phosphatidylglycerophosphate</td>
<td>Phospholipid</td>
</tr>
<tr>
<td>R02029</td>
<td>Phosphatidylycerophosphate + H2O = Phosphatidylglycerol + Orthophosphate</td>
<td>Phospholipid</td>
</tr>
<tr>
<td>R02057</td>
<td>CDP-ethanolamine + 1,2-Diacyl-sn-glycerol = CMP + Phosphatidylenethanolamine</td>
<td>Phospholipid</td>
</tr>
<tr>
<td>R07496</td>
<td>alpha-Methylzymosterol = Zymosterol</td>
<td></td>
</tr>
</tbody>
</table>

The alternative approach is to use flux balance analysis (FBA) (i.e., linear optimization [14]) to check whether known metabolic functions can be reproduced in silico. To set the problem up for identifying network gaps in the biosynthetic pathways, only the uptake fluxes of nutrients that the organism is auxotrophic for are set free, while all other input nutrient fluxes are constrained to zero. An infeasible biosynthetic pathway is manifested as a zero flux value calculated for the drain of the biomass component, despite the flux being maximized. Underlying the problem may be either gap(s) in the pathway’s connectivity or reversibility constraints that prevents the use of the pathway. Troubleshooting the problem, one must not only progressively trace from the end-point to the start-point to inspect potential breakage in the network connectivity, but also check whether the reversibility setting for each of the encountered reactions is realistic when compared against a given set of guidelines, such as the irreversible hydrolysis of high-energy phosphate bond [9]. The appropriate corrections are made, either by adding new reactions or by relaxing the reversibility constraint.

FBA led to the discovery of eight additional reactions essential for biosynthesis. Three of these reactions have no known gene associations, but were required to catalyze the reverse of pre-existing reactions in the GSM (Table 4). The remaining five reactions were originally compartmentalized into the mitochondria, but existing biosynthetic pathways dictate their placement in the cytoplasm (Table 5), three of which could be found in the cytoplasmic compartment of the human GSM [5].

Table 4. List of new reactions identified by FBA and their corresponding irreversible reaction that catalyze a similar reaction but in the reverse direction.

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Reaction equation</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>R00841b</td>
<td>sn-Glycerol 3-phosphate + H2O = Glycerol + Orthophosphate</td>
<td>Glycerol</td>
</tr>
<tr>
<td>R00847</td>
<td>ATP + Glycerol = ADP + sn-Glycerol 3-phosphate</td>
<td></td>
</tr>
<tr>
<td>R01131b</td>
<td>ATP + Inosine = ADP + IMP</td>
<td></td>
</tr>
<tr>
<td>R01126</td>
<td>IMP + H2O = Inosine + Orthophosphate</td>
<td></td>
</tr>
<tr>
<td>R06517n</td>
<td>Acyl-CoA + Sphinganine = CoA + Dihydroceramide</td>
<td></td>
</tr>
<tr>
<td>R06518</td>
<td>Dihydroceramide + H2O = Fatty acid + Sphinganine</td>
<td></td>
</tr>
</tbody>
</table>

b Reaction with no gene association
### Table 5. List of pre-existing reactions added to the cytoplasmic compartment

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Reaction equation</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>R00848c</td>
<td>sn-Glycerol 3-phosphate + FAD = Glycerone phosphate + FADH2</td>
<td>Glycerol</td>
</tr>
<tr>
<td>R01221</td>
<td>Glycine + THF + NAD+ = 5,10-MethyleneTHF + NH3 + CO2 + NADH + H'</td>
<td>Glycine</td>
</tr>
<tr>
<td>R01867c</td>
<td>(S)-Dihydroorotate + Oxygen = Orotate + H2O2</td>
<td>Pyrimidine</td>
</tr>
<tr>
<td>R02030c</td>
<td>Phosphatidylglycerol + CDP-diacylglycerol = Cardiolipin + CMP</td>
<td>Phospholipid</td>
</tr>
<tr>
<td>R03657</td>
<td>ATP + L-Leucine + tRNA(Leu) = AMP + Pyrophosphate + L-Leucyl-tRNA</td>
<td>Biomass drain</td>
</tr>
</tbody>
</table>

Apart from the reactions that were essential for the synthesis of biomass components, an additional 43 reactions with no gene associations were added to the GSM based on literature data. Overall, these modifications involve network gaps that were found in nucleotide salvage and degradation pathways, as well as in essential amino acid degradation pathways. Also, the sub-cellular localization of a further 21 reactions were corrected. From the curated GSM, there was a general sense that the degradative pathways were mostly compartmentalized into the mitochondria.

### 4. Metabolic Network Properties

The final version of the *M. musculus* GSM consists of 1399 genes mapped to 1757 different reactions (model available in SBML format in Supplementary). Altogether, this produced 4619 unique gene-reaction associations. A total of 52 reactions with unknown gene associations were added to the GSM. This list excludes membrane transporters (68), biomass drains (21) and auto-catalytic reactions (7) that were added on the basis that they were required. There are a total of 2037 reactions in the stoichiometric model, 387 of which are located in the mitochondria.

An interaction map of the metabolites reveals global features of the GSM (Fig. 2). Only a very small set of reactions are essential for biosynthesis of major biomass compounds. The number of essential reactions is approximately 270, although this number varies depending on the imposed input constraints. The input constraints dictate the availability of input nutrients, and therefore the biosynthetic pathways that must be activated for growth. Nodes from the essential reaction set tend to be clustered to the center of the interaction map (Fig. 2, right), suggesting that these metabolites tend to have a higher degree of connectivity. Despite a large number of reactions contained in the GSM, only 1050 reactions are considered to be active (i.e., have non-zero flux). These reactions, less the essential ones, are components of pathways that are redundant for growth. Preliminary assessment of the redundant reactions revealed that they are mostly found in parallel or cyclic pathways. Firstly, a large number of these reactions involve transhydrogenation, whereby two or more reactions can be assembled into a

---

* Cytoplasmic reactions in the human genome-scale model
pathway that produces a net transfer of redox equivalent from one cofactor to another (e.g., NAD⁺, NADP⁺, FAD, ferredoxin). In KEGG PATHWAY, the ambiguity in cofactor usage of a particular gene product often lead to a duplication of the same reaction but with different cofactors involved. Secondly, the GSM reflect a generic cell and has the potential of cells in various tissues and in varying states. For example, the network contains pathways for both biosynthesis and catabolism of a large number of biomass constituents, while in reality these pathways would be temporally and/or spatially separated. The importance of this redundancy will only be realized in an organismal-level model. Undoubtedly, the level of redundancy would be greatly reduced, if reactions were filtered based on genes actually transcribed in a given cell, e.g., using transcriptomics.

Figure 2. Visualization of the interaction network of metabolites. Metabolites are presented as nodes, while reactions are presented as edges. Metabolites with greater degree of connectivity are shown as larger node with larger label. Left figure contain all metabolites in the GSM, while the right figure has highly connected nodes (H₂O, H⁺, O₂, ATP, NADPH, NADH, ADP, pyrophosphate) removed. The figures are produced using Cytoscape 2.6.0 (http://www.cytoscape.org). They are drawn using the spring-embedded layout, which tend to distribute singletons toward the peripheral space of the interaction map (see supplementary for high-resolution colour figure).

Almost half the reactions in the current model are directly or indirectly tied up to singleton (dead-end) metabolites, which account for 950 out of 2104 metabolites. Some of the singleton metabolites results from the non-specification of minor components in biomass (e.g., spermidine), which means their net synthesis must be zero. These are readily resolved by including a biomass synthesis reaction. Many singleton metabolites, however, are not connected to core metabolism in terms of carbon, only in terms of interaction with H₂O, H⁺, ATP and/or redox cofactors (Fig. 2, right). This would be true of many xenobiotics metabolized in the liver, which are taken up, undergo a few reaction steps before being secreted again. Finally, some singletons are undoubtedly the result of wrongly or poorly annotated genes leading to inclusion of reactions not found in the mouse.
Using our approach, singleton metabolites will be gradually resolved as a more
detailed biomass composition is considered, as better transporter annotation tools
becomes available, and as secondary pathway annotation is improved. Importantly, the
model remains fully functional in terms of predicting major metabolic activity despite the
presence of unresolved singletons.

5. Discussion

A few critical observations were made from our experience with the large-scale
reconstruction and subsequent manual curation of the *M. musculus* GSM. We
demonstrated that the GSM was able to simulate basic growth and metabolic function
without engaging extensive curation efforts. This validates the value of online genome
and proteome databases, reinforcing the fact that the metabolic coverage by KEGG
PATHWAY is, to some extent, complete, and that sub-cellular localization annotations
derived from UniProtKB were accurate and readily usable. The ability to perform
automated large-scale metabolic reconstruction facilitates the on-going reconciliation of
our GSM with new genome annotations.

As expected, core metabolic pathways are portrayed in greater detail, and do not
require extensive curation. On the other hand, curation efforts tend to be directed at
singletons [5], to establish some form of network connectivity of the peripheral
pathways, which are mostly discontinuous. While undoubtedly valuable, the large effort
should be balanced against the returns. Where not directly linked to our needs, we are
happy to let the model automatically evolve as the research community collectively
improves the underlying databases. It has made possible to automate the curation tasks
given a suitable reference model [12], using similar procedure outlined in our approach.

Instead we are focusing our effort at extending the existing scope of the sub-cellular
compartmentalization to include nucleus, endoplasmic reticulum, peroxisome and so
forth. For example, the oxidation of fatty acid should be differentiated into the
peroxisomal and mitochondrial pathways. We have also commenced work on the
ultimate challenge of capturing metabolic interactions between tissues and organs [5]. A
feature missing from current work is metabolic regulation. The imposing regulatory
network is complex, but is necessary to reflect metabolic changes for different growth
conditions [4].

In conclusion, we have developed a reproducible approach for the reconstruction of
*M. musculus* GSM. The approach is readily adopted because it employs generic tools for
data extraction, storage and flux simulation. The development of the *M. musculus* GSM
is on-going. One of the many developmental milestones is to capture and validate the
gene–protein–reaction associations, and present these associations in a suitable
hierarchical format [11]. This is necessary to support integration of heterogeneous
datasets in future.
References


