# Reconstruction of an *in silico* metabolic model of *Arabidopsis thaliana* through database integration

Karin Radrich[1,2], Yoshimasa Tsuruoka[3], Paul Dobson[4], Albert Gevorgyan[5], Neil Swainston[6], Jean-Marc Schwartz[1,*]

1. Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, UK.

2. Technische Universität München, 80333 München, Germany.

3. National Centre for Text Mining, University of Manchester, Manchester, M1 7DN, UK.

4. School of Chemistry, University of Manchester, Manchester, M1 7DN, UK.

5. Cell Systems Modelling Group, School of Life Sciences, Oxford Brookes University, Oxford OX3 0BP, UK.

6. Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, M1 7DN, UK.

* Corresponding author: jean-marc.schwartz@manchester.ac.uk



## Abstract

The number of genome-scale metabolic models has been rising quickly in recent years, and the scope of their utilization encompasses a broad range of applications from metabolic engineering to biological discovery. However the reconstruction of such models remains an arduous process requiring a high level of human intervention. Their utilization is further hampered by the absence of standardized data and annotation formats and the lack of recognized quality and validation standards.

Plants provide a particularly rich range of perspectives for applications of metabolic modeling. We here report the first effort to the reconstruction of a genome-scale model of the metabolic network of the plant *Arabidopsis thaliana*, including over 2300 reactions and compounds. Our reconstruction was performed using a semi-automatic methodology based on the integration of two public genome-wide databases, significantly accelerating the process. Database entries were compared and integrated with each other, allowing us to resolve discrepancies and enhance the quality of the reconstruction. This process lead to the construction of three models based on different quality and validation standards, providing users with the possibility to choose the standard that is most appropriate for a given application. First, a *core metabolic model* containing only consistent data provides a high quality model that was shown to be stoichiometrically consistent. Second, an *intermediate metabolic model* attempts to fill gaps and provides better continuity. Third, a *complete metabolic model* contains the full set of known metabolic reactions and compounds in *Arabidopsis thaliana.*

We provide an annotated SBML file of our core model to enable the maximum level of compatibility with existing tools and databases. We eventually discuss a series of principles to raise awareness of the need to develop coordinated efforts and common standards for the reconstruction of genome-scale metabolic models, with the aim of enabling their widespread diffusion, frequent update, maximum compatibility and convenience of use by the wider research community and industry.

## Introduction

Metabolism is perhaps the best characterized of all molecular interaction networks in biology. Large amounts of data relating to metabolic reactions are available to date, but despite this wealth of information metabolic phenotypes can still not be accurately predicted (Sweetlove, 2008). Knowledge of plant metabolism in particular remains fractional and efforts to engineer plant metabolism have so far largely failed. A comprehensive understanding of plant metabolism yet has the potential to bring valuable advances in the generation of pharmaceutical products, the production of key secondary metabolites of commercial interest, the improvement of yield and nutritional quality of crops.

Although *Arabidopsis thaliana* has been widely used as a model plant, its metabolic network has not been studied in great details and at a large scale. There has been renewed interest in *A. thaliana* metabolism recently. More than 170 secondary metabolites from seven different classes have been identified in *A. thaliana* (D'Auria & Gershenzon, 2005), whose putative functions cover the defense against pathogens and herbivores, UV protection, resistance to oxidative stress, auxin transport, etc. Glucosinolates are known for their benefits to human nutrition and were found to play a fundamental role in the defense response against microbial and fungal pathogens (Clay *et al*., 2009; Bednarek *et al*., 2009). Biosynthesis pathways of tocochromanols, a group of lipid antioxidants that are essential in human nutrition, have raised promising interest (Dörmann, 2007). *A. thaliana* was also used as a model plant to study polyamine metabolism, which plays an essential role in stress tolerance (Alcázar *et al*., 2006), and flavonoid production, which inhibit or stimulate cell proliferation in different human cancer cell lines (Woo *et al*., 2005).

A major step towards understanding the metabolic phenotypes of an organism is the construction of a comprehensive model of its metabolic network. While metabolic pathways are convenient abstractions to represent the main routes of chemical transformations, their definition is generally arbitrary. The pathway paradigm fails to provide an integrated view of interactions and control mechanisms, which act across and with no regard of pathway boundaries. For that reason, many large-scale metabolic networks have been constructed in recent years, most notably for microorganisms (Heinemann *et al*., 2005; Oh *et al*., 2007; Resendis-Antonio *et al*., 2007; Andersen *et al*., 2008; Herrgård *et al*., 2008; Suthers *et al*., 2009) but also for animals and humans (Sheikh *et al*., 2005; Duarte *et al*., 2007; Ma *et al*., 2007). The applications of such models are plentiful and encompass metabolic engineering studies to design strains overproducing desired products, the prediction of genes responsible for orphan reactions, the determination of active reactions for a given environmental condition, the identification of coupled reaction sets, and evolutionary studies (Feist & Palsson, 2008). Genome-scale metabolic reconstructions were also used to predict potential new antibiotic targets (Lee *et al*., 2009).

The reconstruction of a large-scale metabolic model remains a long process, requiring a high level of human input. A lot of information about metabolic reactions is available in public databases, compiled through extensive curation of the biochemical literature. However each database has its limitations, and therefore models produced automatically

from one of these databases generally contain numerous inaccuracies and are incompatible with other applications. The most common sources of problems are the non-uniqueness of metabolite identifiers (some compounds being represented by generic classes such as "alcohol"), unbalanced atomic species arising from an incorrect stoichiometry or formula of one or more reactants, incorrect or missing cofactors, and enzymes catalyzing more than one reaction (Poolman *et al*., 2006). For that reason, several sources of data need to be confronted and assembled in order to obtain a more reliable model . The likelihood that the same error would appear in two independent sources should be of an order of magnitude smaller than the frequency of errors in any database. This simple idea has lead us to develop an original methodology for metabolic network reconstruction based on the integration of two databases.

In the case of *Arabidopsis thaliana*, the two most extensive sources of metabolic reaction data are Kegg (Kanehisa *et al*., 2008) and AraCyc (Zhang *et al*., 2005). The concept for the construction of our model is based on the fact that no database currently contains fully accurate information about the enzymes, reactions and metabolites present in an organism. In addition, there are still gaps in the knowledge about many parts of the cellular composition, leading to different lacks in databases. A reconstruction using only one source of data would naturally copy errors and gaps into the reconstructed model. By creating an intersection between two databases, differences can be identified and taken into account for further curation. The resulting process can be defined as semi-automatic, as the extraction and comparison of data between databases can be largely automated, but manual curation remains necessary to analyze and solve discrepancies between them.

Another important and often ignored aspect of metabolic model reconstruction is that different sources of data have different levels of certainty. A reaction consistently described by several independent sources should have a higher degree of reliability than a reaction described by a unique source. For this reason, we here present three different models of *A. thaliana* metabolism corresponding to decreasing levels of confidence (Figure 1). The *core (yellow) model* contains only compounds and reactions which have been reliably identified in both databases and whose description is identical in both of them. The *intermediate (green) model* contains compounds and reactions found in only one database, but with a strong connection to the core model so that the absence of a perfect match is likely to be due to minor inaccuracies. The *complete (blue) model* contains all remaining compounds and reactions.

## Methods

In order to map metabolites between Kegg and AraCyc, several features of the data were taken into account. Two compounds were defined as being identical only if all features were positively matched. These features included compound names,·chemical formulae and structures, and enzymes catalyzing reactions where the considered metabolites participate.

### Compound name similarity

Names of metabolites may differ between databases for several reasons. Many chemical

compounds are commonly known under multiple names, and all of their names are not necessarily indicated in all databases. Furthermore, here is no universal and common identifier between Kegg and AraCyc, as compounds are sometimes referred by their ChEBI (Chemical Entities of Biological Interest; Degtyarenko *et al.*, 2008), CAS (Chemical Abstracts Service; Buntrock, 2001) or PubChem (Austin *et al.*, 2004) identifier. For example, there are five different names listed for the Kegg entry C00022 (Pyruvate, Pyruvic acid, 2-oxopropanoate, 2-oxopropanoic acid, pyroracemic acid), and nine different possibilities for the same compound in AraCyc (pyruvate, BTS, alpha-ketopropionic acid, acetylformic acid, pyroracemic acid, 2-oxopropanoic acid, pyruvic acid, 2-oxopropanoate, 2-oxo-propionic acid). If at least one of these entries is the same in both databases, the identification of matching metabolites is straightforward. However there are many cases where no perfect match can be found. For example the Kegg compound C10434 is named 5-O-Caffeoylshikimic acid, and the same compound in AraCyc has the name caffeoylshikimate.

For this reason a string similarity algorithm, originally developed for the identification of gene/protein name similarity, has been employed to compare metabolite names (Tsuruoka *et al.*, 2007). This algorithm uses logistic regression to compute the similarity between strings by incorporating a variety of features. A training data set has to be supplied in order to teach the program which differences can be treated as similar and which ones are not allowed. It is important to use features that can well characterize a string pair by capturing the similarity between a variety of variations while highlighting the difference between terms which are not synonymous. The considered features are character bigrams, prefixes and suffixes, numbers, acronyms, common and different tokens. An appropriate training set was prepared by processing sets of multiple names of the same metabolite in each of the databases. We found that the characteristic differences that occur between metabolite names are essentially of a very similar nature as those occurring between protein names, allowing the algorithm to perform efficiently. The result of this process consisted in a list of matched names with an associated of their identity.



**Chemical structure similarity**

The formula of a metabolite theoretically confines the search for matching metabolites considerably. However it is not usable as a unique identifier. Two metabolites can have the same global formula and be completely different chemically because of the various possible structures of atoms. For example, 2-carboxy-D-arabinitol 1-phosphate, L-galactose-1-phosphate, D-hexose 6-phosphate and alpha-D-mannose 1-phosphate all have the same formula C6H13O9P. On the other hand, two metabolites that are identical may be represented by slightly different chemical formulae, either because an error is present in one source or because of chemical modifications (such as pH-dependent breakdowns of carboxylic groups).

All chemical structure processing was implemented in Pipeline Pilot workflows (Hassan *et al.*, 2006). Input structures from KEGG and AraCyc were first converted to canonical SMILES, a unique line representation of two-dimensional molecular structure (Weininger *et al.*, 1989). Exact structural matches occurred where SMILES strings were identical.

In previous metabolic network reconstructions it was observed that equivalent metabolites differed across sources in a number of ways. These included stereochemical differences (due to varying levels of detail about chiral centres or configurations about double bonds), tautomeric variants (where proton localization differs), and by charge state. To identify differences of these types SMILES strings were matched after purging stereochemical information from structure, or after calculation of the canonical tautomer, or following recalculation of ionization at pH 7.4.

One further important way in which equivalent metabolites from different sources were noted to differ was by simple structural errors in their representation, such as incorrect bonding or missing functional groups. Providing that the majority of the structure is correctly specified then corrections for errors of this type can be suggested by standard cheminformatics molecular similarity measures (Willett, 2006) (here using the ECFP_4 molecular fingerprint and Tanimoto similarity metric).

Whereas exact matches of the original canonical SMILES strings provides an unambiguous mapping of metabolites across data sources, the other types of matches are only approximate (they may or may not be correct) and so require further manual checking. The approximate matching algorithms do, however, significantly reduce the manual checking workload.



**Complete process**

An iterative approach was adopted to integrate the different features of metabolite identification (Figure 2). The first step consisted in creating a list of mapped metabolites as a starting point. Those metabolites have been allocated by searching for Kegg references in AraCyc. For some of its compounds AraCyc provides the unique Kegg identifier, allowing the undoubted matching of a first set of compounds. The next step consisted in searching for all reactions in both databases that contain those compounds. More specifically, we considered those reactions where all compounds were already identified or only one was missing. Reactions where all compounds were identified could subsequently be compared, and if all compounds and the catalyzing enzymes were the same, the reactions were accepted as being the same. In those reactions where one compound was missing, the known compounds were compared to each other. If two reactions in AraCyc and Kegg had the same number of metabolites and all known metabolites were the same, the respectively unknown compounds in each reaction were accepted as candidates for being identical. If the catalyzing enzyme in both reactions was the same, and the name strings had a high probability of being similar, and the structures or formulae were equal, then two candidate compounds were accepted as being identical.

At the end of the first iteration step, new compounds were added to the list of matched compounds and a new iteration was started. The whole process was repeated several times until no additional matched compounds could be found. During the process, compounds that gave a positive result in some but not all of the before-mentioned features were copied into a separate list that was manually examined. The positive results of these checks were added to the list of compounds in order to improve the outcome of the next iteration.

The final lists of matched compounds and reactions were given new unique identifiers for the new model. These matched compounds and reactions constitute the first or *core (yellow) metabolic model*. Compounds and reactions that were not matched in both databases were assigned to the second or third model depending on an additional process. In the *intermediate (green) model*, we included reactions for which either the full set of substrates or the full set of products belonged to the core model. Such reactions are likely to be valid as they have a strong connection to the core model, but may be insufficiently or inconsistently described resulting in discrepancies between databases or their absence in one of them. Compounds involved in such reactions and not already included in the core model were also added to the intermediate model. All remaining compounds and reactions were included into the *complete (blue) model*.

A general problem in metabolic models is caused by the fact that many compounds have different protonation states depending on the pH. There is no pH consistency in formulae found in databases, and AraCyc does not represent protons in reactions. As a consequence, we neglected differences in proton content between formulae and did not represent protons in reactions either. A strict balance of hydrogen atoms thus cannot be expected in our models.



**Stoichiometric consistency validation**

Incorrect definition of reaction stoichiometries often results in stoichiometric inconsistencies – a common type of modelling error, defined as contradictions between the fundamental physical constraints of mass positivity and mass conservation (Gevorgyan et al., 2008). An example is shown below:

R1 : A ↔ B

R2 : A ↔ B + C

In this network, the metabolite C cannot be assigned any positive molecular mass without violating the mass balance in the whole system. Stoichiometric inconsistencies are often caused by violations of atomic balance, by ambiguous generic metabolite definitions (e.g. “primary alcohol”) and by inclusion of polymers with variable polymerization degrees and units (e.g. starch and protein).

Stoichiometric consistency validation involves inspection of the left null-space (the null-space of the transposed stoichiometry matrix) and may include the following optional steps: Firstly, the stoichiometric consistency of the full network is verified. If the network is inconsistent, the non-conserved metabolites (Nikolaev *et al.*, 2005) are detected (e.g. C in the above example). Further, for each non-conserved metabolite, the minimal inconsistent net stoichiometries involving it are calculated (these are defined as net stoichiometries with empty sets of substrate or products, e.g. by subtracting the reaction R1 from R2, we obtain ∅ ↔ C). Finally, the elementary leakage modes (minimal linear combinations of reactions resulting in inconsistent net stoichiometries) can be detected, e.g. (-R1, R2) in the above example. The localization of such minimal inconsistent subnetworks helps to detect input errors in the reactions.

**Construction of SBML model**

An SBML version of the core model was constructed using libSBML (Bornstein *et al.*, 2008). The model is MIRIAM compliant (Le Novère *et al.*, 2005), with all compartments, species and reactions annotated with ontological terms, allowing their unambiguous identification and interpretation by third party software tools. Metabolites have been annotated with ChEBI and Kegg terms, along with InChI strings (Stein *et al.*, 2003), while all enzymes are provided with an annotation linking them to the appropriate gene in the TAIR database (Huala *et al.*, 2001).

**Network visualization and analysis**

A network was constructed from the complete (blue) model for topological analysis and visualization. In this network representation, compounds are nodes and reactions are edges. All substrates of a given reaction were connected to all products of that reaction. Isolated compounds were not included in the network representation. The Cytoscape software was used for network visualization (Shannon *et al.*, 2003), and the NetworkAnalyzer plugin for Cytoscape was used for topological analysis (Assenov *et al.*, 2008). Network properties of the three models are summarized in Table 1. The meaning of the parameters shown in Table 1 is briefly explained hereafter (Dong & Horvath, 2007).

The number of *connected components* indicates how many disjoint subnetworks the network is broken into. A *self-loop* is a node connected to itself. The number of *shared neighbors* between two nodes is the number of nodes that are neighbors of both of them. The *shortest path length*, also called the *distance* between two nodes is the smallest number of edges that have to be crossed to go from one edge to the other. The *characteristic path length* is the average distance and the *network diameter* is the largest distance between two nodes in the network.

The *connectivity* of a node is the number of edges connected to it. The *network density* is a measure of how densely the network is populated with edges. A network that contains only isolated nodes has a density of 0, whereas a clique has a density of 1. The *network centralization* is a measure of how strongly a network is focused around central nodes. Networks resembling a star have a centralization close to 1, whereas decentralized networks have a centralization close to 0. The *network heterogeneity* measures the variance of the connectivity and reflects the tendency of a network to contain hub nodes. The *clustering coefficient* of a given node $n$ is a ratio between the number of edges between the neighbors of $n$ and the maximum number of edges that could possibly exist between them.

The *betweenness centrality* $C_b(n)$ of a node $n$ is computed as follows:

$$C_b(n)=\frac{2}{(N-1)(N-2)}\sum_{s\neq n\neq t}\frac{\sigma_{st}(n)}{\sigma_{st}},$$

where $s$ and $t$ are nodes in the network different from $n$, $\sigma_{st}$ is the number of shortest paths from $s$ to $t$, $\sigma_{st}(n)$ is the number of shortest paths from $s$ to $t$ that $n$ lies on, and $N$ is the total number of nodes in the connected component that $n$ belongs to. The betweenness centrality of each node is a number between 0 and 1. It reflects the amount of control that a node exerts over the interactions between other nodes in the network. A node acting as a

bridge between different communities has a high betweenness centrality, while a node that lies inside a community has a low one.

The *closeness centrality* $C_c(n)$ measures how close a node $n$ is to others in the same connected component. It is defined as follows:

$$C_c(n) = \frac{N-1}{\sum_{m \neq n} L(m,n)},$$

where $L(n, m)$ is the length of the shortest path between $n$ and $m$, and $N$ is the total number of nodes in the connected component that $n$ belongs to. The closeness centrality is a measure of how fast information can spread from a given node to other reachable nodes in the network.

## Results

### Metabolic models of *A. thaliana*

We present three metabolic models of *A. thaliana* corresponding to different levels of confidence. The core (yellow) metabolic model only contains compounds and reactions that have been unambiguously identified and matched in the two databases (Supplementary File 1). These metabolites and reactions are expected to have been well characterized and experimentally observed, and are more likely to play an important role in the plant. This model is expected to cover most of the core metabolism of *Arabidopsis thaliana* (Table 1).

The intermediate (green) model additionally contains reactions from both databases where either the full set of substrates or the full set of products belongs to the core model. Compounds involved in these reactions and not already included in the core model were also added to the intermediate model. Such reactions and compounds have a strong connection to the core model, but their confidence status is lower since they have not been unambiguously matched between in both databases. The fact that a compound or reaction was found in only one database may be due to several factors: (i) our reconstruction algorithm may have failed to find the corresponding compound or reaction in the second database; (ii) a metabolite may be represented by a generic class in one database but by a specific compound in the other; (iii) one database may contain an inaccuracy, so that a metabolite or a cofactor is missing or incorrect in a reaction; (iv) a reaction or compound may indeed be absent from one database. Some of these causes might lead to the occurrence of double entries in the intermediate model.

All remaining reactions and compounds were added to the complete (blue) metabolic model, so that a comprehensive data set of all metabolic reactions and compounds described in *A. thaliana* and reactions contained is presented. This model should be considered as a development set whose validity needs to be confirmed from additional sources.

For a better understanding of the different cases that result in the attribution of compounds and reactions to different models, several examples are shown in Figure 3.

The sucrose phosphate phosphohydrolase reaction (a) is entirely yellow because it is identically described in both databases and all compounds were successfully matched to each other. So is the ribulose bisphosphate carboxylase reaction (b), even though it is described with two protons on the right hand side in Kegg but without in Aracyc, because we decide to ignore protonation states. The pyruvate kinase reaction with GTP/GDP as cofactors (c) is green even though all its substrates and products are yellow, because only Kegg describes the possibility of GTP/GDP involvement. Another pyruvate kinase reaction involving ATP/ADP is described in both databases and is included in the core (yellow) model. The acetyl-CoA synthetase reaction and acetyl adenylate (d) were not found in AraCyc, they are included into the intermediate (green) model because both acetyl-CoA and AMP are unambiguously identified. The carbon-sulfur lyase reaction (e) is blue because neither substrate nor product are known by AraCyc. The sphingolipid biosynthesis reaction (f) is blue because it uses generic metabolite classes in AraCyc that cannot be matched to specific metabolites in Kegg.

**Detailed coverage of metabolic pathways**

We investigated the distributions of enzymes belonging to the three metabolic models among Kegg pathways (Supplementary File 2). In most of the carbohydrate metabolism pathways, the core (yellow) metabolic model covers between 70% and 80% of all enzymes attributed by Kegg to these pathways. This proportion generally rises above 85% in the intermediate (green) model. In nucleotide metabolism pathways, 87% of the enzymes are covered by the core model and 91% by the intermediate model. For amino acid metabolism and secondary metabolites biosynthesis, these values are most of the time between 50% and 70% in the core model and 75% in the intermediate model. Lipid metabolism has a lower coverage with 40% to 60% of enzymes being in the core model and around 70% in the intermediate model. It is not surprising that core metabolic pathways are generally better covered, as these pathways should have been the most intensively analyzed experimentally and most accurately described.

As an illustration of the different levels of quality of our reconstructed models, we describe the case of the citrate and glyoxylate cycles in detail (Figure 4). Most parts of the citrate cycle and the glyoxic shift have been reliably identified through the semi-automatic process and were included into the core (yellow) model. The remaining gaps are filled in the intermediate (green) model. The factors leading to some reactions and compounds not being included into the core model are detailed below:

(i) The succinate dehydrogenase reaction between succinate and fumarate was not included into the core model due to an ambiguity between various forms of ubiquinones and ubiquinols acting as cofactors. These compounds are represented by generic classes in Kegg, but by specific ubiquinone-8 and ubiquinol-8 compounds in AraCyc.

(ii) The transition between 2-oxoglutarate and succinyl-CoA is represented by a direct alpha-ketoglutarate dehydrogenase reaction in AraCyc. This reaction is not present in Kegg, which instead represents the transition by three different steps involving 3-carboxy-1-hydroxy-propyl-thiamine diphosphate and succinyldihydro-lipoamide-E.

(iii) Similarly, the transition between isocitrate and 2-oxoglutarate via oxalosuccinate is

represented in Kegg but not in AraCyc. The direct isocitrate dehydrogenase reaction does appear in both databases.

**Stoichiometric consistency validation**

The intermediate (green) and complete (blue) metabolic models unsurprisingly contain many stoichiometric inconsistencies because these models contain generic metabolite classes (e.g. "alcohols"). In the core (yellow) metabolic model however, the only non-conserved metabolite detected was molecular hydrogen. This inconsistency inevitably follows from skipping protons from reaction definitions and currently cannot be resolved, given the inaccuracies in the input data. Since hydrogen interconversions are generally not relevant for applications of metabolic models, we conclude that the stoichiometric consistency validation of the core model was successful.

**Network properties**

We analyzed the topological properties of the reconstructed metabolic models in order to verify whether they were compatible with those of previously reported models of other species (Table 1 and Figure 5). Different network representations can be used to represent systems of metabolic reactions, and the values of network parameters depend on the chosen representation. In this work metabolites were represented as nodes and reactions as edges. As the directionality of metabolic reactions is generally subject to ambiguity, edges were set to be undirected. Two metabolites were connected by an edge if they participate as substrate and product respectively in the same reaction. Common small molecules such as ATP, NADH, water, etc, were not removed from the network representation.

Distributions of the most important network properties are plotted in Figure 5 for the complete (blue) metabolic model. The intermediate (green) and core (yellow) metabolic models exhibit similar distributions. Average network parameters are shown for the three networks in Table 1. The connectivity distribution of our metabolic model has the same allure, resembling a power-law, as universally observed in metabolic networks (Jeong *et al.*, 2000; Wagner & Fell, 2001; Almaas, 2007). The average clustering coefficient is our models is close to 0.2, to be compared with reported values of 0.20 for *E. coli*, 0.23 for *S. cerevisiae*, and 0.28 for *H. pylori* based on the same network representation (Almaas, 2007). The average path length in our models is close to 3, which is the same as observed in many other metabolic networks based on the same network representation (Jeong *et al.*, 2000). It is worth noticing that this value becomes significantly higher when common small molecules are removed from the network (Ma & Zeng, 2003) or when an atomic representation of metabolism is adopted (Arita, 2004).

The top ten hubs for the three metabolic models are listed in Table 2. The connectivity of these hubs logically increases from the core to the complete models, but their ranking is only marginally affected by the difference in confidence levels between models. Water remains the most highly connected molecule in all cases, and nine out of ten molecules consistently appear the top ten ranking for all three models. These hubs include most of the ubiquitous small molecules found in other metabolic models, when they are not

removed.

A graphical network representation of the complete (blue) metabolic model is provided in Figure 6.

## Discussion

The number of published genome-scale metabolic models has grown rapidly in recent years. After the first models reconstructions were published for *E. coli* and *S. cerevisiae* (Edwards & Palsson, 2000; Förster *et al.*, 2003) the number of such reconstruction has been growing quickly in recent years, covering many microorganisms, animals and human. A comprehensive description of the motivations and applications of such reconstructions has been presented by Feist and Palsson (2008). These applications include network property analysis, metabolic engineering, biological discovery, phenotypic assessment, and evolutionary analysis.

No large-scale reconstruction of the metabolic network of a plant has been undertaken previously, and yet many of the applications mentioned before take even higher relevance in plants. Metabolic engineering is of particular significance in plants and offers promising perspectives to improving production yields, enhancing the nutritional value of crops, and generating valuable molecules for pharmacology and energy production. High-quality and comprehensive models of plant metabolism will be crucial to allow these applications to be developed. The metabolic networks of plants are of a higher complexity than those of most other living species; it is therefore both relevant and timely to start investing efforts in the construction of such models.

However many issues presently hamper the use of genome-scale metabolic models by the wider research community and industry. These issues include:

(i) A limited usage of standardized formats, nomenclatures and annotations. Most metabolic models are published in spreadsheet format, using customs identifiers and nomenclature. This is a considerable obstacle for the transfer of these models to other applications and for the comparison of different models.

(ii) Limited coordination between different reconstruction efforts. One such coordination has recently lead to the publication of a consensus metabolic model of *S. cerevisiae* (Herrgård *et al.*, 2008), and a similar effort is currently under way for the human metabolic network (Mo & Palsson, 2009).

(iii) The absence of update mechanisms enabling the integration of the latest scientific discoveries by the wider research community into existing models.

(iii) The absence of universal quality and validation standards.

Although this work does not claim to address all these issues, we introduced in this work a few principles seeking to propose avenues for solutions and to raise awareness about current limitations. First, we provide an annotated SBML version of the core (yellow) metabolic of *A. thaliana* (Supplementary File 3). SBML has become the *de facto* standard for systems biology models and allows them to be used by the widest range of tools. Standardized annotations following SBO specifications ensure that metabolites and

enzymes can be easily identified and linked with existing databases. However, for such formats to be universally adopted by the biochemical research community, efficient and user-friendly tools will need to be developed allowing the easy input and conversion of models to a well annotated and standardized format.

While large international meetings have proven successful to confront and integrate different existing metabolic models, mechanisms allowing a convenient integration of models as they are developed would be more efficient. We showed that by confronting and integrating two independent sources, we were able to semi-automatically reconstruct a core metabolic model of *A. thaliana*, whose quality is comparable to existing manually reconstructed models of other species. Such mechanisms could be generalized by the use of common repositories, following the models used for gene sequences or protein structures, allowing users to deposit new models and enhance existing ones through a seamless integration process.

Last, all applications using genome-scale metabolic models do not necessarily require the same level of data quality. For network analysis, a relatively straightforward or automatic reconstruction may be sufficient, while for metabolic engineering or experimental design a highly accurate and well-annotated model is generally necessary. It is therefore important to keep track of the sources and validation level of data used in reconstructions, so that users are able to select the data with the appropriate level of confidence for their application. As a first step towards such a process, we here provide three models with different levels of confidence. The core (yellow) model has the highest confidence level and was proven to be stoichiometrically consistent, but has some gaps. For applications such as Flux Balance Analysis, a more continuous model can be preferable even though some reactions might be of a lower confidence level. The intermediate (green) model attempts to suit such needs by filling gaps through the inclusion of partial information. The complete (blue) model eventually contains the largest amount of available information, but with the restriction that some reactions may be unconfirmed and the risk of duplications.



## Conclusion

In this work, we simultaneously introduced the first large-scale model of the metabolism of the plant *Arabidopsis thaliana* and a methodology allowing an efficient semi-automatic construction of metabolic models via the integration of different data. The integration of different data sources significantly enhances the quality of a reconstructed model and leads to quality standards that are comparable to manual reconstructions. A long-term and coordinated international effort will be desirable to provide comprehensive and accurate genome-scale metabolic models of plants, and to provide the standards and infrastructure allowing a widespread diffusion, frequent update, maximum compatibility and convenience of use of such models by the widest research community and industry.

## Acknowledgements

PD is funded by BBSRC grant BB/F006012/1. NS thanks the EPSRC and BBSRC for funding of the Manchester Centre for Integrative Systems Biology. The authors thank Ralf Steuer,

Hans Westerhoff, Pedro Mendes, Mark Poolman and David Fell for fruitful comments and discussions.

## Authors' contributions

KR developed the methodology and implemented algorithms for model development. YT, PD, AG, NS implemented algorithms for model development and validation. JMS designed and coordinated the study and analyzed results. KR, PD, AG, NS, JMS wrote the manuscript. All authors read and approved the final manuscript.

## References


- Alcázar R, Marco F, Cuevas JC, Patron M, Ferrando A, Carrasco P, Tiburcio AF, Altabella T (2006) Involvement of polyamines in plant response to abiotic stress. *Biotechnol Lett* **28**: 1867-76.
- Almaas E (2007) Biological impacts and context of network theory. *J Exp Biol* **210**: 1548-58.
- Andersen MR, Nielsen ML, Nielsen J (2008) Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Mol Syst Biol* **4:** 178.
- Arita M (2004) The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* **101:** 1543-47.
- Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht A (2008) Computing topological parameters of biological networks. *Bioinformatics* **24:** 282-4.
- Austin CP, Brady LS, Insel TR, Collins FS (2004) NIH Molecular Libraries Initiative. *Science* **306:** 1138-39.
- D'Auria JC, Gershenzon J (2005) The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr Opin Plant Biol* **8:** 308-16.
- Bednarek P, Pislewska-Bednarek M, Svatos A, Schneider B, Doubsky J, Mansurova M, Humphry M, Consonni C, Panstruga R, Sanchez-Vallet A, Molina A, Schulze-Lefert P (2009). A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. *Science* **323:** 101-6.
- Buntrock RE (2001) Chemical registries - in the fourth decade of service. *J Chem Inf Comput Sci* **41:** 259-63.
- Clay NK, Adio AM, Denoux C, Jander G, Ausubel FM (2009) Glucosinolate metabolites required for an *Arabidopsis* innate immune response. *Science* **323:** 95-101.
- Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36:** D344–50.
- Dong J, Horvath S (2007) Understanding network concepts in modules. *BMC Syst Biol* **1:** 24.

- Dörmann P (2007) Functional diversity of tocochromanols in plants. *Planta* **225:** 269-76.
- Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA* **6:** 1777-82.
- Edwards JS, Palsson BØ (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* **97:** 5528-33.
- Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli. Nat Biotechnol* **26:** 659-67.
- Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* **13**: 244-53.
- Gevorgyan A, Poolman MG, Fell DA (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics* **24**: 2245-51.
- Hassan M, Brown RD, Varma-O'Brien S, Rogers D (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* **10:** 283-99.
- Heinemann M, Kümmel A, Ruinatscha R, Panke S (2005) In silico genome-scale reconstruction and validation of the *Staphylococcus aureus* metabolic network. *Biotechnol Bioeng* **92:** 850-64.
- Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY *et al.* (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* **26:** 1155-60.
- Huala E, Dickerman A, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang J, Huang W, Mueller L, Bhattacharyya D, Bhaya D, Sobral B, Beavis B, Somerville C, Rhee SY (2001) The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* **29**: 102-5.
- Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* **407:** 651-4.
- Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36:** D480-4.
- Lee DS, Burd H, Liu J, Almaas E, Wiest O, Barabási AL, Oltvai ZN, Kapatral V (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel anti-microbial drug targets. *J Bacteriol*, online preprint.
- Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**: 1509-15.

- Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I (2007) The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* **3:** 135.
- Ma H, Zeng AP (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**: 270-7.
- Mo ML, Palsson BØ (2009) Understanding human metabolic physiology: a genome-to-systems approach. *Trends Biotechnol* **27**: 37-44.
- Nikolaev EV, Burgard AP, Maranas CD (2005) Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophys J* **88**: 37-49.
- Oh YK, Palsson BØ, Park SM, Schilling CH, Mahadevan R (2007) Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* **282**: 28791-9.
- Poolman MG, Bonde BK, Gevorgyan A, Patel HH, Fell DA (2006) Challenges to be faced in the reconstruction of metabolic networks from public databases. *Syst Biol* **153:** 379-84.
- Resendis-Antonio O, Reed JL, Encarnación S, Collado-Vides J, Palsson BØ (2007) Metabolic reconstruction and modeling of nitrogen fixation in *Rhizobium etli*. *PLoS Comput Biol* **3**:1887-95.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498-504.
- Sheikh K, Förster J, Nielsen LK (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnol Prog* **21:** 112-21.
- Stein SE, Heller SR, Tchekhovskoi D (2003) An open standard for chemical structure representation: The IUPAC chemical identifier. In *Proceedings of the 2003 International Chemical Information Conference* (Nîmes), *Infonortics,* 131-143.
- Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CD (2009) A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. *PLoS Comput Biol* **5**:e1000285.
- Sweetlove LJ, Fell DA, Fernie AR (2008) Getting to grips with the plant metabolic network. *Biochem J* **409:** 27-41.
- Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S (2007) Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics* **23:** 2768-74.
- Wagner A, Fell DA (2001) The small world inside large metabolic networks. *Proc R Soc Lond B Biol Sci* **268**: 1803-10.
- Weininger D, Weininger A, Weininger JL (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* **29**, 97-101.
- Willett P (2006) Similarity-based virtual screening using 2D fingerprints. *Drug*

*Discov Today* **11:** 1046-53.

- Woo HH, Jeong BR, Hawes MC (2005) Flavonoids: from cell cycle regulation to biotechnology. *Biotechnol Lett* **27:** 365-74.
- Zhang P, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol* **138:** 27-37.

## Tables

**Table 1: Network properties of the three metabolic networks.**

The meaning of the parameters given in this table is explained in the Methods section.

| | Core (yellow) model | Intermediate (green) model | Complete (blue) model |
|---|---|---|---|
| Number of nodes | 770 | 1207 | 2288 |
| Number of edges | 2255 | 3792 | 6547 |
| Network density | 0.008 | 0.005 | 0.002 |
| Network heterogeneity | 2.223 | 2.623 | 3.362 |
| Number of self-loops | 0 | 8 | 31 |
| Clustering coefficient | 0.215 | 0.233 | 0.189 |
| Connected components | 6 | 5 | 28 |
| Network diameter | 8 | 8 | 10 |
| Network centralization | 0.288 | 0.301 | 0.271 |
| Average path length | 3.114 | 3.158 | 3.286 |
| Average connectivity | 5.857 | 6.270 | 5.696 |

**Table 2: Hubs of the three metabolic networks.**

| Core (yellow) model | | Intermediate (green) model | | Complete (blue) model | |
|---|---|---|---|---|---|
| Metabolite | Degree | Metabolite | Degree | Metabolite | Degree |
| Water | 227 | Water | 369 | Water | 628 |
| ATP | 117 | Oxygen | 169 | Oxygen | 270 |
| ADP | 107 | NADP | 160 | ATP | 229 |
| NADPH | 89 | NADPH | 158 | NADP | 219 |
| Orthophosphate | 88 | ATP | 155 | NADPH | 218 |
| NADP | 84 | ADP | 128 | Carbon dioxide | 192 |
| Carbon dioxide | 81 | Carbon dioxide | 118 | Diphosphate | 182 |
| Oxygen | 79 | Orthophosphate | 102 | ADP | 159 |
| Diphosphate | 77 | Diphosphate | 101 | NAD | 143 |
| NAD | 60 | NAD | 89 | NADH | 140 |

## Figure legends

**Figure 1: Presentation and size of the three metabolic models.**

**Figure 2: Steps of the model building process.**

**Figure 3: Examples of reactions and their attribution to different confidence levels.**

(a) Sucrose phosphate phosphohydrolase is identically described in both databases. (b) Ribulose bisphosphate carboxylase has a discrepancy in hydrogen content but protons involvement is ignored. (c) All substrates and product are validated in both database, but the pyruvate kinase reaction with GTP/GDP as cofactors is only described in one database. (d) Acetyl adenylate and the acetyl-CoA synthetase reaction are only found in one database. (e) Both substrate and product are only found in one database. (f) Both sides of the reaction involve generic compound classes which are only used by one database.

**Figure 4: Attribution of reactions and compounds in the tricarboxylic acid cycle.**

Most reactions and compounds in the tricarboxylic acid cycle belong to the core (yellow) model. Succinate dehydrogenase, ketoglutarate dehydrogenase, and the transition between isocitrate and 2-oxoglutarate via oxalosuccinate belong to the intermediate (green) model. The transition between 2-oxoglutarate and succinyl-CoA via succinyldihydro-lipoamide-E belongs to the complete (blue) model.

**Figure 5: Topological properties of the complete (blue) metabolic model.**

(a) Node degree distribution. (b) Average clustering coefficient distribution. (c) Betweenness centrality. (d) Closeness centrality. (e) Shared neighbors distribution. (f) Shortest path length distribution. See methods section for an explanation of network parameters.

**Figure 6: Graphical network representation of the complete model.**

Nodes belonging to the core model are colored in yellow. Nodes added in the intermediate model are colored in green. Nodes added in the complete model are colored in blue.

## Supplementary Files

### Supplementary File 1: Compound and reaction data of the core (yellow) metabolic model.

The first sheet contains the list of compounds and the second sheet the list of reactions. Each compound is identified by a local identifier consisting of "Ath_C" followed by a four-digit number, its Kegg identifier and AraCyc name. Each reaction is identified by a local identifier consisting of "Ath_R" followed by a four-digit number, its Kegg identifier and AraCyc name. The stoichiometry column describes the reaction using local compound identifier. Substrates and products are separated by the equal ("=") sign. The stoichiometry is always explicitly written even when it is one. The enzyme column lists the enzymes catalyzing each reaction by their EC number.

### Supplementary File 2: Distribution of enzymes in the three metabolic models for each Kegg pathway.

The first two columns give the Kegg identifier and name of each pathway. The columns in yellow give the number of enzymes from this pathway attributed to the core metabolic model and its percentage in relation to the total number of enzymes contained in the pathway. The columns in green give the number of enzymes attributed to the intermediate metabolic model and its percentage in relation to the total number of enzymes. The blue column gives the number of enzymes contained in the complete model, which is equal to total number of enzymes contained in the pathway.

### Supplementary File 3: Annotated SBML file of the core (yellow) metabolic model.