Building and analysing genome-scale metabolic models

David A. Fell†, Mark G. Poolman and Albert Gevorgyan
School of Life Sciences, Oxford Brookes University, Headington, Oxford OX3 0BP, U.K.

Abstract
Reconstructing a model of the metabolic network of an organism from its annotated genome sequence would seem, at first sight, to be one of the most straightforward tasks in functional genomics, even if the various data sources required were never designed with this application in mind. The number of genome-scale metabolic models is, however, lagging far behind the number of sequenced genomes and is likely to continue to do so unless the model-building process can be accelerated. Two aspects that could usefully be improved are the ability to find the sources of error in a nascent model rapidly, and the generation of tenable hypotheses concerning solutions that would improve a model. We will illustrate these issues with approaches we have developed in the course of building metabolic models of Streptococcus agalactiae and Arabidopsis thaliana.

Requirements for a metabolic model
There is increasing recognition that metabolism of cells cannot be simply regarded as the simple addition of the collection of canonical pathways they contain. Many metabolic properties - the response to a mutation that inactivates an enzyme, for example - can only be properly understood by examining the metabolic network as a whole. Inevitably, that means that the network under consideration has to be as comprehensive as possible, since it is difficult to rule out, a priori, that a particular enzyme or reaction will be irrelevant. Under these circumstances, where it is necessary to consider hundreds or thousands of metabolites and reactions, the biochemist’s metabolic map becomes less and less useful: it is tedious to tailor a generic map to represent a specific cell; it is impossible to verify whether it is formally correct, and only very trivial questions can be answered by visual inspection. The solution is to compute the answers, somewhat as a road route between two points is routinely calculated to find the fastest route, the shortest route, the most economical route, or the route avoiding a particular section of road.

How can this be implemented for metabolism? Even for the relatively few organisms extensively studied by biochemists, it is not possible to consult comprehensive lists of the enzymes or reactions they contain. For most other organisms, the only available catalogue of their metabolic repertoire will be derived from their annotated genome sequence, hence the origin of genome-scale metabolic models [1–3]. These depend on the ability to recognize and annotate genes coding for enzymes, and to associate those enzymes with metabolites and reactions. The ideal outcome is thus a list of all the reactions and metabolites that could occur in a cell, linked through enzymes and to their associated genes. (There are as well some spontaneous reactions of metabolites that will occur in the absence of enzymes.) If we could then compute the metabolic capabilities of (or their absence from) the network, we would have made a genotype–phenotype mapping of a complex multi-genic trait. Furthermore, if we learn how to do this with acceptable accuracy on organisms for which we have sufficient biochemical knowledge to check the results, we can hope to predict the metabolism of relatively unknown but sequenced organisms. Hence we would argue that metabolism makes the most straightforward test case for the general feasibility of such functional genomics approaches. However, since the number of sequenced genomes is growing rapidly, and is far ahead of the number of genome-scale metabolic models [4], it is important to accelerate the processes of building and checking the metabolic models.

In the present article, we will describe how the process of building a genome-scale metabolic network can be assisted by computer-aided methods for finding errors in metabolic network models and improving their quality.

Building the model
As an outline, the process of building a metabolic network model starts from a genome annotation in a database; from the annotation we recover enzyme identities in terms of the EC numbers, and, for each EC number present, we obtain a list of the reactions catalysed by the enzyme. We and others [5,6] have reviewed some of the errors that arise because of incomplete annotation and inconsistencies both between and within the different databases used to translate EC numbers to reactions. For the purposes of the present article, we note that, at this point, the list of enzymes is almost certainly incomplete and not necessarily accurate, and not all reactions catalysed by the enzymes will have been added to the list in the form of unambiguously characterized molecular entities. As we shall show later, it is possible and easier to find reactions that have been added to the model that have no function than it is to find the reactions that are missing, hence it is better to err on the
The key to nearly all forms of metabolic modelling is to turn the reaction list into a mathematical object that contains the same information, but which can be a basis for computation. This is the stoichiometry matrix, where each column corresponds to a reaction, and each row to a metabolite, and the entry where a row and column intersect is the number of molecules of the metabolite formed (positive) or consumed (negative) by the reaction. Most of the entries will be zero, and, as the matrix is large, it is best constructed from the reaction list by computer to ensure the non-zero entries are in the correct place.

There are then two types of method that use the information in the matrix. Graph analysis methods use it as an 'adjacency matrix', indicating which metabolites are one, two etc. reactions away from a given metabolite, and use this to construct potential routes through the network. This is similar to the way route-finding programs produce an itinerary between two points on a road network. However, metabolism is not like a road route in an important way; if you travel along an itinerary in your car, you do not expect that parts of the car and its contents would be removed or exchanged on the way as a systematic consequence of travelling along particular segments of road. Were that to be the case, you would want to know whether or not you and your original car would arrive at the destination. This is, of course, what happens to metabolites as they are transformed by reactions, and though modified graph methods can take some account of this, we have argued elsewhere that metabolic routes found in this way will contain sequences that cannot result in any sustained mass transfer from the starting metabolite to the end point [18]. The other group of methods interprets the stoichiometry matrix as a set of equations that enforce a steady-state condition, where the production and consumption of each intermediate along any predicted route is exactly balanced, so there can be no leak of mass into, or gain from, the rest of the network. This means that the concentrations of all metabolites, apart from the 'externals' (which are mostly nutrients, waste products and polymeric biomass) will remain constant, and the routes correspond to the net transport of mass from inputs to outputs, as expressed in a stoichiometrically balanced equation for the overall transformation.

In the remainder of the present article, we refer only to modelling approaches of this second type, which include metabolic flux analysis [19], linear programming [10], flux balance analysis [20], elementary modes analysis [21], null-space analysis [22] and extreme pathway analysis [23]. In all these, the stoichiometry matrix is used to generate a set of equations that represent relationships, or constraints, between the relative rates of the reactions that are compatible with steady state; without additional information, it is not possible to predict the actual state of metabolism, but it is possible to place limits on what is attainable and to exclude states that cannot exist. Within these feasible solutions, ones that are minimal or extreme, in the sense of reducing the numbers of participating reactions between specified inputs and outputs, can be regarded as the basic set of routes through the network (elementary modes analysis [12]; extreme pathway analysis [24]). Outlines of the mathematics involved have been described previously [22,25–28].

The stoichiometry matrix can also be formulated as a set of equations that express relationships between the metabolites. If the network represents only a part of metabolism, this can identify groups of metabolites that cannot be totally synthesized or degraded by the network reactions (perhaps because they share a component, termed a conserved moiety, that is untouched by metabolism). For example, a model of central carbohydrate metabolism alone would not account for synthesis and degradation of an adenine moiety, so ATP, ADP and AMP would form a conserved group in the model whose total concentration could not be changed by the network, even though they participate in the reactions [25,29–31].
Powerful though these analytical methods can be, the results can only be reliable if the stoichiometric matrix accurately and consistently reflects the properties of the metabolic network. If the results are incorrect, finding the source of the error simply by human inspection is extremely difficult in a system with numbers of metabolites and reactants both of the order of a thousand and is likely to be a limiting step in the development of genome-scale models. Hence we will consider how the results, when wrong, can be exploited to point to the source of the error.

**Mass conservation**

If every reaction in a metabolic model was perfectly stoichiometrically balanced, with every metabolite in the model having a single empirical formula, then the model as a whole should conserve mass with no reactions causing unexplained gain or loss of matter. Unfortunately, as we have reported previously, the databases that are the sources for the reaction equations contain errors. At first sight, it would seem that this could be simply and automatically checked by counting the number of atoms on each side of the reaction equation, and this does indeed identify a proportion of the errors. However, there are many reactions where one or more reactants do(es) not have a defined atomic composition, and then this check is not feasible (such as ferredoxin, or folate with a polyglutamate tail of undefined length). The solution to this problem is to realize that we do not actually need to know the atomic composition, or even the molecular mass, to know whether a list of reactions is mass-balanced. It suffices that it is possible to assign some single, positive mass value to each metabolite so that every reaction equation is mass-balanced with respect to its declared reactants and products; whether such a solution exists can be checked by the technique of linear programming subject to the constraint that the net mass change in every reaction is zero. If the model fails this test, and only the most carefully curated models will pass it, the problem is to find the source of the problem.

We have proposed the following method to track the error to a small subset of the total reactions of the model [32]. It works because each reaction can be regarded as a statement about the composition of its reactant molecules. For example, the reaction $A \rightarrow B$ states that the mass and composition of $A$ is identical with that of $B$, whereas the occurrence of the reaction $A \rightarrow B + C$ states that the composition of $A$ equals that of $B$ plus $C$; these reaction statements are inconsistent unless $C$ has zero mass, which is obviously not possible if $C$ is a real compound. The technique of mixed integer linear programming can be used to maximize the number of metabolites that can be assigned a consistent positive mass, and a greedy algorithm can be later used to find a maximal subset of consistent reactions.

The reactions and metabolites left out of this subset of consistent reactions are the ones harbouring the error, but there may still be too many of them for the source of the problem to be obvious. We examine the subset of inconsistent reactions with a mixed integer linear programming technique too involved to explain in the present article, but described previously [32], which extracts the the non-reducible sets of reactions that generate a leakage of mass into or out of the system. At this point, the numbers of reactions and metabolites involved with an inconsistency should be a handful, small enough for the correction to be obvious to the modeller.

### Refining the reaction list

It is common to characterize metabolic models by the numbers of reactions and metabolites included, but these can be misleading measures without qualification. One reason is that there is a case during model building for adding more reactions than can be connected to the network. This will arise if a gene is annotated to an enzyme with a broad substrate range (e.g. alcohol dehydrogenase, EC 1.1.1.1), since it will not be clear while the model is incomplete as to which alcohols and aldehydes will be available by being produced or consumed by other enzymes in the network. A simple solution is to add more than is likely to be necessary and to remove at a later stage the ones that are not functional.

What criteria do we have for reactions that are non-functional? One test is where a metabolite is only involved in a single reaction in the network. In this case, the metabolite cannot reach a steady state because it is impossible to have its balanced production and consumption. Furthermore, the associated reaction can never be active with a non-zero flux, since otherwise it would be a perpetual source or sink of matter. Hence there is an argument for removing such reactions from the reaction list, and the argument is strengthened because in enzyme [33] and metabolism databases [34,35] up to to 40% of metabolites appear to have a single occurrence [6], so it is probable that such metabolites and their reactions will arise in the model with a significant frequency. Identifying the relevant metabolites is easy, since they will have a single non-zero entry in their row of the stoichiometry matrix. The argument against removing these metabolites and reactions is that the error may not be in their addition, but in the absence of a balancing reaction, either because of a missing or erroneous annotation, or because the metabolite can exchange with the environment through a transporter or by diffusion across the membrane (i.e. an exchange reaction). Experimental evidence on which metabolites are taken up from or excreted into the medium is useful in deciding whether it is legitimate to add an exchanger; inspecting the exchangers and counting their number in a genome-scale metabolic model can indicate whether such additions have been made too readily.

Having surveyed the model for metabolites with single connectivity and removed them and their associated reactions, it is necessary to repeat the exercise because the first round of reaction deletions may have removed the balancing reactions for other metabolites. The process terminates after a few rounds, leaving a core of connected metabolites. Lack of connectivity of metabolites, however, is not the only reason a reaction may be inactive at steady state; it is necessary that the network can deliver a sustained supply of substrates and remove the products. (This is the same reason that graph analysis techniques can predict non-functional pathways that...
are not found by stoichiometric analysis, as discussed previously [18].) Another test that does examine stoichiometric feasibility involves calculation from the stoichiometry matrix of its ‘null space’; the null space contains, in compressed form, information about the relative rates of the reactions in all feasible steady states of the network. If a reaction has no non-zero values associated with it in the null space, it cannot carry a flux under any steady-state conditions. This test can be lax in some respects, and too harsh in others. It is lax in that it does not consider irreversibility of reactions, and might therefore suggest a reaction can be active when thermodynamic constraints would prevent it. It also allows reactions that are only involved in ‘internal cycles’, i.e. cyclic pathways that are not linked to any net mass transfer from inputs to outputs and therefore have no thermodynamic driving force. It can be severe in that the answer applies for the specific set of exchanges allowed with the environment in the network model, and may therefore suggest a reaction is inactive when there are environmental conditions when it might be used.

More information can be derived from the null space by calculation of reaction correlation coefficients [36]. For each pair of reactions in the network, these are the correlation coefficients between their fluxes taken over all steady-state solutions allowed by the null space. A value of 1 (or −1) indicates that the reaction fluxes are always in the same ratio, corresponding to reaction subsets [37]. A value of 0 corresponds to independence between the steady-state reaction fluxes, which is caused, in the majority of cases, by their being in independent subsystems. Hence any network reactions that have reaction correlation coefficients of zero with respect to all inputs and outputs from the network cannot have any role in the metabolism associated with these mass flows; if they are not inactive by the simple null space test, then they must be solely involved in non-driven internal cycles. In our model of Arabidopsis metabolism [38], we identified 77 inoperative reactions out of the 855 remaining after reactions associated with unbalanced metabolites had been removed.

Inspection suggests that published genome-scale models include a significant fraction of reactions that are inactive by the tests above. If the models are regarded as databases of potential reactions in the cell that might occur under some as yet undefined conditions, this is acceptable. However, for other types of analysis, they should first be removed. Remembering that reactions are not enzymes, or genes, removing inactive reactions from the model starts to have consequences when all the reactions associated with an enzyme (gene) have been removed; at that point, the enzyme (gene) is ‘unemployed’. This implies either that the original annotation of the gene was wrong, as it leads to an enzyme activity that cannot be connected to the network, or that other enzymes (or transporters) are missing from the annotation and are needed to complete the network. We would suggest that the number of unemployed genes can be used as a measure of the quality of an annotation, and that minimizing the number is a worthwhile objective. It can be, for example, that annotation methods may not produce a unique suggestion for the identity of a gene. Using the PRIAM database of enzyme signatures leads in a number of cases to multiple matches for the gene, and it is not easy to decide which is the most probable match, as the signatures are different lengths, so relative ‘E’ values of the matches are not a certain guide.

We encountered the problem of unemployed enzymes when building a model of the metabolism of Streptococcus agalactiae. At the same time, the initial model could not account for proline synthesis and lactose utilization. It is not feasible to test all possible combinations of all annotations for those genes with multiple assignments, but by using the number of unemployed genes as a target to be minimized, and making random exchanges of annotations in a simulated annealing approach, we were able to find combinations of gene assignments that reduced the number of unemployed genes and resulted in metabolic networks able to utilize lactose and synthesize proline (and lost no other metabolic functions) [39]. In effect, this automatic procedure generates hypotheses about possible different assignments for genes originally ascribed different functions, and ultimately this needs to be tested experimentally. Other methods to expand metabolic models based on network function have been described [40,41]. Unlike pathway-oriented gap-filling methods [42,43], which search for genes that may encode enzymes classed as missing by comparison with reference pathways, these network-orientated approaches recognize that different organisms may implement metabolic functions in different ways, and also consider the potential negative impact of re-assigning a gene annotation from one enzyme to another.

**Conclusions**

We argued at the beginning of the present article that reconstructing an organism’s metabolic network from its annotated genome sequence represents a potentially achievable, but non-trivial, case study for the prediction of phenotype from sequence data. This is not the only motivation: among the others are prediction of the systemic effects of mutations or drugs that inactivate particular reactions and the design of metabolic engineering strategies. However, the issues discussed above illustrate that there are problems still to be resolved in addition to the usual one of incomplete data. The overall process requires that it is possible to reason rigorously over the large amounts of assembled information, and this in turn depends on two important preconditions: first that the problem can be represented in a suitable mathematical form for appropriate tools to be able to determine solutions, and secondly that the known biological data is recorded in a manner that is compatible with such approaches. Even for the case of metabolism, there is still much to be achieved on both counts.

**Funding**

This work was supported by the Biotechnology and Biological Sciences Research Council [grant numbers 332/CFB17702 and BB/E00203X/1] and Oxford Brookes University.
References


Received 27 April 2010
doi: 10.1042/BST03811197