

# Genome-scale models of bacterial metabolism: reconstruction and applications

Maxime Durot, Pierre-Yves Bourguignon &amp; Vincent Schachter

Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, Evry, France

**Correspondence:** Vincent Schachter, Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, 2 rue Gaston Crémieux, CP5706, 91057 Evry, Cedex, France. Tel.: +33 1 60 87 25 92; fax: +33 1 60 87 25 14; e-mail: vs@genoscope.cns.fr

Received 30 July 2008; revised 22 October 2008; accepted 22 October 2008.  
First published online December 2008.

DOI:10.1111/j.1574-6976.2008.00146.x

Editor: Victor de Lorenzo

## Keywords

metabolic network; systems biology; computational methods; genome-scale metabolic models; metabolic engineering; omics data integration.

## Abstract

Genome-scale metabolic models bridge the gap between genome-derived biochemical information and metabolic phenotypes in a principled manner, providing a solid interpretative framework for experimental data related to metabolic states, and enabling simple *in silico* experiments with whole-cell metabolism. Models have been reconstructed for almost 20 bacterial species, so far mainly through expert curation efforts integrating information from the literature with genome annotation. A wide variety of computational methods exploiting metabolic models have been developed and applied to bacteria, yielding valuable insights into bacterial metabolism and evolution, and providing a sound basis for computer-assisted design in metabolic engineering. Recent advances in computational systems biology and high-throughput experimental technologies pave the way for the systematic reconstruction of metabolic models from genomes of new species, and a corresponding expansion of the scope of their applications. In this review, we provide an introduction to the key ideas of metabolic modeling, survey the methods, and resources that enable model reconstruction and refinement, and chart applications to the investigation of global properties of metabolic systems, the interpretation of experimental results, and the re-engineering of their biochemical capabilities.

## Introduction

The flow of genome sequencing, metagenome sequencing and other high-throughput experimental efforts aimed at exploring the space of microbial biochemical capabilities has been steadily growing in recent years. At the time of writing, more than 1800 bacterial genome-sequencing projects have been initiated and nearly 650 have been completed (<http://www.genomesonline.org>, <http://www.ebi.ac.uk/integr8>). Combined with increasingly efficient annotation methods, these set the stage for the systematic identification of most enzymes encoded in the genomes of the corresponding bacterial species. A variety of so-called '-omics' technologies now routinely provide large-scale functional clues on molecular interactions and cellular states, offering snapshots of the dynamic operation of metabolism under specified conditions, and adding to the store of accumulated knowledge on microbial biochemistry and physiology.

Simultaneously, the expected wealth of new biochemical activities, the progress of metabolic engineering techniques

aimed at harnessing these activities, and the perspective of applications to white and green biotechnology have triggered a strong renewed interest in the exploration of bacterial metabolism. In addition to charting the range of naturally evolved chemical transformations, relevant research questions include the following: How does the global metabolism of a bacterium react to changes in its environment? What kind of joint metabolic operation of distinct species can help sustain a bacterial community? How can genomic and biochemical information be best exploited to gain insights into the relationship between an organism's genotype and its phenotype? For instance, can we predict changes in metabolism-related phenotypic traits caused by simple or complex genotype modifications? How did metabolic processes evolve? How can metabolic networks be efficiently reprogrammed for a variety of utilitarian purposes?

Investigations of a bacterium's metabolism are typically fed by knowledge (ultimately from observations) at two different scales of description of the chemistry at work within cells. The larger scale focuses on the physiology of the whole bacterial

cell. For instance, which media is it able to grow on? What are the relative quantities of chemical nutrients it requires for growth? How efficient is the cell at converting chemicals from the environment into its own components? Such metabolic capabilities result from the coordinated action of the enzymes expressed in the respective species, the knowledge of which belongs to the finer, molecular scale. Each of the corresponding biochemical conversions can be identified either directly by performing enzymatic assays, or indirectly, from the genome sequence, through a homology relationship with proteins whose function has been previously elucidated. Together, the reactions that have been demonstrated to potentially occur in the cell form the *metabolic network* of the organism. Metabolic networks can thus be viewed as lists of those molecular mechanisms (reactions) and associated molecular components (enzymes, substrates, and products) that are most directly related to the metabolic capabilities mentioned above.

For a given bacterial species, confronting knowledge from these two scales, molecular vs. cellular, can reveal inconsistencies. For instance, it may happen that no sequence of identified reactions is capable of producing one of the essential cell components from the set of compounds available in a defined growth medium, even though the species is known to grow on that medium. Furthermore, when the two scales are consistent, their relationship can be investigated further in order to enumerate the possible implementations of the physiology that the metabolic network can achieve. Biochemists have traditionally performed such investigations by modularizing the set of reactions into *metabolic pathways*, typically grouping together reactions that allow the conversion of one or more 'input' metabolites into 'output' metabolites. Pathway boundaries are somewhat arbitrary, even though inputs and outputs tend to be metabolites involved in several reactions. Pathway-based analyses are thus focused on the possible fates of a restricted number of compounds, and are amenable to manual expertise thanks to the simplification brought by the modularized view (Huang *et al.*, 1999; Teusink *et al.*, 2005; Risso *et al.*, 2008).

Yet, metabolic pathways typically involve a large number of 'side metabolites' such as cofactors and byproducts of chemical reactions, and metabolism is as much about converting nutrient into cell components as it is about regenerating cofactors and recycling (or secreting) ultimately unused byproducts. The latter transformations typically involve several pathways, and are dependent on the stoichiometry and rates of the reactions. Manual approaches are insufficient to assess their feasibility by a given network for at least two reasons: metabolic networks are too large, and the question requires a quantitative analysis.

Bridging that gap between knowledge of the metabolic network structure and observed metabolic phenotypes is precisely where metabolic models come into play. Generally

speaking, a model of a natural system is one of many possible mathematical representation of that system, explicitly describing some of its features and supporting predictions on some other features, the latter being typically time- or environment dependent. In this particular case, knowledge of the metabolic network alone is not quite sufficient to predict the metabolic capabilities of a cell. Also needed are a structured (mathematical) representation of that network, together with a set of rules and possibly quantitative parameters enabling simulations or predictions on the joint operation of all network reactions in a given environment, and in particular predictions on the values of metabolite fluxes and/or concentrations (Papin *et al.*, 2003). The above, in short, constitutes a metabolic model.

Constraint-based genome-scale models of metabolism (Palsson, 2006) are a category of models precisely aimed at assessing the physiological states achievable by a given metabolic network, and at uncovering their biochemical implementation in terms of metabolic fluxes. They offer an idealized view of the cell as a set of 'pipes,' with metabolites flowing through each pipe, and biochemical conversions taking place at junctions between pipes. Some metabolites can also be exchanged with the environment, flowing in or out of the system through dedicated pipes that can be opened or shut, and may have upper bounds on their throughput. The cell is required to achieve balanced production and consumption of all the intermediate substrates and products involved in its metabolism: what flows in a junction must flow out.

Constraint-based models can help investigate in a systematic manner most of the research questions listed at the start of this introduction, because they provide a way to explore the consequences on the operation of the entire metabolic network of the piecemeal information available on each of its parts. They are especially well suited to 'what if' experiments involving genetic or environmental perturbations, such as: how would the cell behave in an environment with a different chemistry than the ones that have been experimented on? How would one or more deletions affect its metabolic capabilities? Which deletions would maximize the production of both metabolite *x* and biomass?

Before a model for a given species can be used to gain new insights into its metabolic capabilities or evolutionary history, it must first be built from the scattered genomic, biochemical, and physiological information available on that species up to a point where known physiology can be predicted from biochemistry without major mistakes. This process is sometimes known as 'model reconstruction'; its endpoint is a functional genome-scale model, i.e. a structured representation of the current state of knowledge on the metabolism of the respective species (Reed *et al.*, 2006a). The model provides a framework to interpret new experimental data gathered at the cellular or molecular scale. That data may be incompatible with the current model, in

which case either or both should be questioned, leading to possible revisions or improvements. If, on the other hand, data and model are compatible, the new evidence may still narrow down the set of possible metabolic behaviors of the cell, thus enriching the model (Covert *et al.*, 2004).

This review article covers both the reconstruction of genome-scale metabolic models and their applications to basic and applied research in microbiology. Following a primer on constraint-based models, we will review the state of the art in model reconstruction. Next, we will survey the main applications of metabolic models, from phenotype predictions to data interpretation or metabolic engineering. Practical aspects of direct relevance to the working microbiologist will be covered by a sketch of the main dedicated database and software resources. We will conclude the review with a discussion on future directions in the field.

## Foundations of genome-scale metabolic modeling

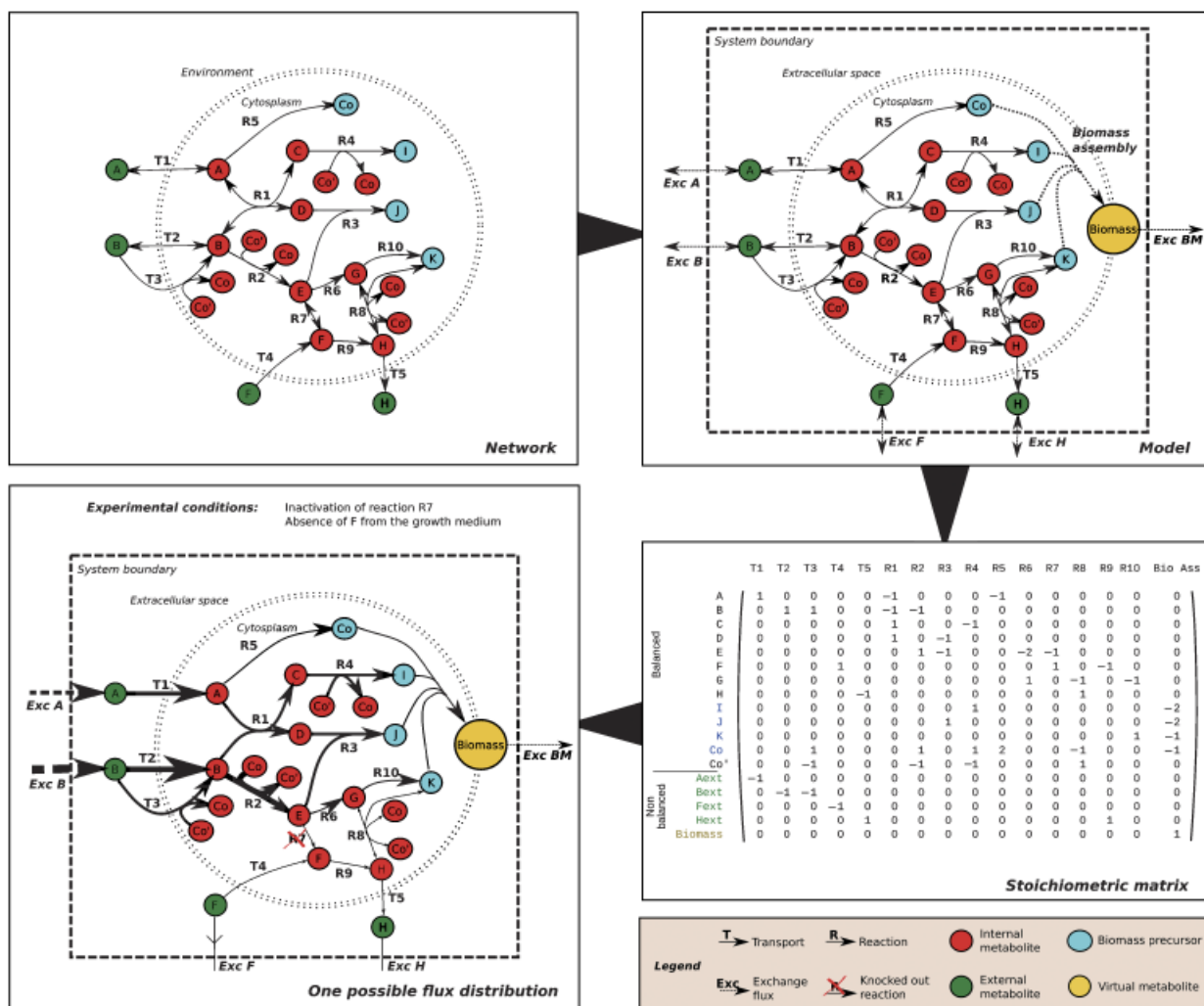
The metabolic state of a cell and its variation over time can be described by metabolite concentrations and reaction rates, which can be viewed as the ‘endpoints’ of metabolic operation. These quantities are related by the law of conservation of matter, which states that the net production rate of a metabolite equals the sum of the rates of the reactions consuming or producing it, weighted by the associated relative stoichiometric coefficients. Conversely, enzyme kinetics express reaction rates as complex functions of metabolite concentrations and enzymatic activities, which vary over time as a result of transcriptional and metabolic regulation (Smallbone *et al.*, 2007). Deriving meaningful predictions from these two types of equations for large metabolic systems is a very challenging proposition, not only because of the mathematics, but also because many of the parameters are not known, difficult to measure, and possibly context dependent. In practice, these pitfalls restrict the use of kinetic modeling to metabolic systems much smaller than ‘whole-cell’ metabolic networks, which typically include hundreds of reactions for a bacterium.

Constraint-based models bypass these difficulties by focusing on the average reaction rates achievable by cells grown in steady or slowly varying environmental conditions. Rates are typically averaged over minutes, fitting with the typical time scale of uptake or secretion rates measurements. Such averages are not affected by transient states because the characteristic relaxation time of metabolic systems – i.e. the time it takes for chemical reactions within the cell to reach a steady state – is much shorter than a minute. Moreover, because environmental changes and variations of enzyme concentrations occur on longer time scales, one need not take into account regulatory changes to assess average reaction rates over minutes. Turnover rates of

most intracellular metabolites are high in bacterial cells (Stephanopoulos *et al.*, 1998). At the time scale considered here, their concentrations have therefore generally reached steady levels, and remain constant as long as environmental conditions do not change. As a consequence, the law of conservation of matter constrains the production and consumption rates of these metabolites to be balanced. These assumptions are usually summarized under the expression *steady-state hypothesis* and the corresponding constraint on reaction rates as a *mass balance* (or stoichiometric) *constraint* (Stephanopoulos *et al.*, 1998). Obviously, this reasoning applies only to metabolites that are neither taken in from an external pool (e.g. nutrients) nor excreted from the cell or accumulated in large quantities (e.g. cell components such as nucleic acids, amino acids, or some lipids). For each metabolite that can be ‘balanced,’ the mass balance constraint can be expressed mathematically by a linear equation relating reaction rates of the form  $\sum s_j v_j = 0$ , where  $s_j$  is the stoichiometric coefficient of the metabolite in reaction  $j$ , and  $v_j$  the rate of reaction  $j$ .

In addition to mass balance constraints, reactions that are known to be thermodynamically irreversible *in vivo* are constrained to have a non-negative reaction rate. Similarly, upper bounds on the reaction rates can be known from measurements or theory and included in the model as additional constraints on the reaction fluxes (Reed & Palsson, 2003).

Mass balance, irreversibility and upper-bound constraints result from the application of simple laws of physics to individual reactions or metabolites from the network. These constraints propagate from reaction to reaction throughout the metabolic network; the constraint-based modeling framework is designed to automatically compute the resulting balance. To that end, it makes use of a succinct mathematical representation of all reaction stoichiometries: the *stoichiometric matrix* (see Fig. 1). In this matrix, columns represent reactions and rows metabolites. The stoichiometric coefficient of a metabolite within a reaction is included at the intersection of the corresponding row and column (see Fig. 1). Reaction rates are represented in constraint-based models by single numbers, the *reaction fluxes*, which are normalized by the weight of the cells harboring the reactions to account for the size of the colony (a reaction flux is typically expressed with the Unit  $\text{mmol h}^{-1} \text{g}^{-1} \text{dry wt}$ ). Because the goal is to describe the joint operation of many metabolic reactions, it is convenient to define a *flux distribution* as a collection of reaction fluxes covering the entire system. Under the steady-state approximation, the concentrations of balanced metabolites being constant, a flux distribution carries sufficient information to completely describe a state of the system. Using the stoichiometric matrix, a simple matrix equation – summarizing all mass balance equations shown above – can then be used to



**Fig. 1.** Genome-scale modeling of metabolism. A metabolic network (top left) is transformed into a model by defining the boundaries of the system, a biomass assembly reaction, and exchange fluxes with the environment (top right). Using the corresponding stoichiometric matrix (bottom right), the achievable flux distributions compatible with enforced constraints can be found (a particular one is depicted in the bottom left figure).

enforce the mass balance constraints on all reactions fluxes:  $S \cdot v = 0$ , where  $S$  is the stoichiometric matrix and  $v$  the flux distribution represented as a vector.

A precise definition of the boundary of the system to be modeled is also needed to formulate an explicit mathematical representation. The system typically includes the whole cell and its vicinity, in order to encompass all the exchanges of matter between the cell and its environment. Transport reactions that allow for exchange of specific metabolites with the extracellular space through the membrane are also included in the model. Environmental conditions are then modeled by acting on the balance of the external metabolites: metabolites that are available from the environment can be taken up by transporters while the others can only be excreted.

A flux distribution that is compatible with all the constraints in a given environment is considered achievable (or

feasible) by the cell, whereas a distribution that violates at least one of these constraints is not. The simplicity of the system of linear equations that represent constraints is one of the main strengths of the framework, because it permits fast assessments of the feasibility of a flux distribution using a computer and standard algorithms.

The simplicity of constraint-based models comes at the expense of a number of limitations in their predictive capabilities. Such models focus solely on reaction fluxes, and completely ignore the influence of metabolites and enzymes. In reality, however, enzyme kinetics, and transcriptional or metabolic regulation may significantly influence reaction fluxes. Regulation can for instance limit the use of a pathway by downregulating some of its enzymes when particular environmental conditions are met. These mechanisms, if they could somehow be taken into account,

would eliminate flux distributions otherwise allowed by constraint-based models. In other words, models may allow 'false-positive' metabolic states, which respect the enforced metabolic constraints but are inconsistent with other biological processes. Several attempts have been made to extend the constraint-based modeling framework, in order to account for regulatory interactions (Covert *et al.*, 2001), signaling processes (Lee *et al.*, 2008b), the first and second laws of thermodynamics (Beard *et al.*, 2002, 2004), or metabolite concentrations (Kümmel *et al.*, 2006b; Henry *et al.*, 2007). Nevertheless, these extensions require the inclusion of additional experimental data and may result in more complex mathematical formulation hindering their practical use.

Some predictions of constraint-based models may be wrong in cases where modeling assumptions do not hold. For instance, some metabolites do accumulate in the cell, and the mass balance assumption clearly does not hold for these. In general, the concentration of specific metabolites may be high enough relatively to the fluxes they are involved in for the mass balance approximation to become clearly false.

In practice, many of the analytical methods that have been developed for constraint-based models focus on defining and characterizing sets of feasible flux distributions. Others focus on a single distribution. The diversity of flux distributions compatible with constraints in a given environment can be viewed as reflecting the diversity of the metabolic states the cell may find itself in. Nevertheless, the space of feasible flux distributions features biologically informative properties whose determination requires adequate techniques; these will be introduced in the next sections of this review.

## Building the models

The level of detail necessary to build a constraint-based model of a bacterium's metabolism is relatively low; the only information required is the precise reaction stoichiometries and directions, in order to account for mass balance and irreversibility constraints. To reflect the global biochemical capabilities of the organism, the model also needs to encompass the complete set of metabolic activities that can occur within it – or a reasonable approximation thereof. This comprehensiveness requirement and the high number of metabolic reactions make the actual construction of such models a challenging task in itself. In this section, we will review the main methods and resources helping in this task. We will first show how information from genome annotation can be used to infer biochemical reactions at large scale, a task commonly called *metabolic network reconstruction*. We will then review the techniques commonly used to assess the consistency of reconstructed models, and show how missing biochemical activities can be identified to complete the model.

## Initial reconstruction of metabolic models

The most reliable evidence from which the presence of a metabolic reaction in a species can be inferred is experimental proof of the respective biochemical activity. Such biochemical results have been accumulated for several decades, mostly from dedicated experiments targeting well-defined activities. As a consequence, the corresponding reactions have often been precisely and reliably characterized. Exploiting these results to reconstruct the whole metabolism of an organism is a labor-intensive task, however, as it requires processing a high volume of literature. Most existing metabolic models have been reconstructed in this manner and for extensively studied organisms. For instance, the most complete bacterial model available to date – namely iAF1260, the latest model of *Escherichia coli* metabolism – includes references to more than 320 articles (Feist *et al.*, 2007). Two types of databases centralize biochemical knowledge: enzyme-centric ones, which collect functional information acquired on enzymes, for example BRENDA (Barthelmes *et al.*, 2007) or SwissProt (Boutet *et al.*, 2007); and pathway databases, aimed at describing the biochemistry of metabolic processes, for example EcoCyc for *E. coli* metabolism (Karp *et al.*, 2007) or UM-BDD for microbial biodegradation pathways (Ellis *et al.*, 2006) (see Table 1).

These biochemical clues are typically incomplete relatively to the set of all possible activities, especially for less studied organisms. In addition, while technologies aiming at high-throughput characterization of biochemical activities are improving, they are not yet mature enough to provide reasonably good coverage. Genes corresponding to enzymes that have been experimentally characterized have nevertheless been identified. Their homologues in the genome of such species can be identified using comparative genomics methods, thereby indicating the presence of the associated biochemical activities.

The traditional path to inferring metabolic reactions from the genome of an organism is gene-centric, at least in its first steps. Nearly all available genome sequences are now systematically processed through automated annotation pipelines, which identify coding sequences and infer functional annotations. Covering all relevant methods would be beyond the scope of this article, but thorough reviews can be found elsewhere (Médigue & Moszer, 2007). Basically, coding sequences are first identified using highly efficient gene-finding algorithms [such as GENEMARK (Besemer *et al.*, 2001), GLIMMER (Delcher *et al.*, 1999), or AMIGENE (Bocs *et al.*, 2003)], which discard the ORFs that are not likely to be coding for a protein. Functional annotations are then sought for each gene using complementary approaches: sequence homology with proteins of known function [stored for instance in UniprotKB (UniProt, 2008)], conservation of genomic structure with annotated species (e.g. synteny), and prediction of functional domains (Apweiler *et al.*, 2000; Claudel-Renard *et al.*, 2003).

**Table 1.** Data sources for metabolic model reconstruction and refinement

<b>DNA sequence and genome annotation databases</b>		
DDBJ	<a href="http://www.ddbj.nig.ac.jp/">http://www.ddbj.nig.ac.jp/</a>	General nucleotide sequence database
EMBL	<a href="http://www.ebi.ac.uk/embl/">http://www.ebi.ac.uk/embl/</a>	General nucleotide sequence database
GenBank	<a href="http://www.ncbi.nlm.nih.gov/Genbank/">http://www.ncbi.nlm.nih.gov/Genbank/</a>	General nucleotide sequence database
Integr8	<a href="http://www.ebi.ac.uk/integr8/">http://www.ebi.ac.uk/integr8/</a>	Integrated information on complete genomes
CMR	<a href="http://cmr.jcvi.org/">http://cmr.jcvi.org/</a>	Integrated information on complete prokaryotic genomes
IMG	<a href="http://img.jgi.doe.gov/">http://img.jgi.doe.gov/</a>	Integrated system for analysis and annotation of microbial genomes
SEED	<a href="http://seed-viewer.theseed.org/">http://seed-viewer.theseed.org/</a>	Integrated system for analysis and annotation of genomes using functional subsystems
<b>Protein and enzyme databases</b>		
BRENDA	<a href="http://www.brenda-enzymes.info/">http://www.brenda-enzymes.info/</a>	Comprehensive enzyme information system gathering data collected from the literature by curators
ENZYME	<a href="http://www.expasy.ch/enzyme/">http://www.expasy.ch/enzyme/</a>	Enzyme nomenclature database providing extensive information on all enzymes with an associated EC number
UniProt	<a href="http://www.ebi.ac.uk/uniprot/">http://www.ebi.ac.uk/uniprot/</a>	Universal Protein Resource gathering protein sequences and annotations from SwissProt (manually reviewed), trEMBL (computer annotated), and PIR
TransportDB	<a href="http://www.membranetransport.org/">http://www.membranetransport.org/</a>	Predictions of membrane transport proteins for fully sequenced genomes
PSORTdb	<a href="http://db.psort.org/">http://db.psort.org/</a>	Repository of experimentally determined and predicted protein localizations
Prolinks	<a href="http://prolinks.mbi.ucla.edu/">http://prolinks.mbi.ucla.edu/</a>	Database of predicted functional links between proteins
STRING	<a href="http://string.embl.de/">http://string.embl.de/</a>	Database of known and predicted protein–protein interactions
<b>Metabolic databases</b>		
CheBI	<a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a>	Database on small molecules of biological interest
Pubchem	<a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>	Database on small molecules
LipidMaps	<a href="http://www.lipidmaps.org/">http://www.lipidmaps.org/</a>	Database on lipid metabolites
Reactome	<a href="http://www.reactome.org/">http://www.reactome.org/</a>	Curated database of biological pathways
KEGG	<a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a>	Suite of databases comprising information on compounds, reactions, pathways, genes/proteins
BioCyc	<a href="http://www.biocyc.org/">http://www.biocyc.org/</a>	Collection of organism-specific pathway/genome databases, including a curated multiorganism pathway database: MetaCyc
UniPathway	<a href="http://www.grenoble.prabi.fr/obiwarehouse/unipathway/">http://www.grenoble.prabi.fr/obiwarehouse/unipathway/</a>	Curated resource of metabolic pathways linked to UniProt enzyme database
UM-BBD	<a href="http://umbbd.msi.umn.edu/">http://umbbd.msi.umn.edu/</a>	Database on microbial biocatalytic reactions and biodegradation pathways
<b>Experimental data repositories</b>		
IntAct	<a href="http://www.ebi.ac.uk/intact/">http://www.ebi.ac.uk/intact/</a>	Repository of reported protein interactions
DIP	<a href="http://dip.doe-mbi.ucla.edu/">http://dip.doe-mbi.ucla.edu/</a>	Database of experimentally determined interactions between proteins
Array Express	<a href="http://www.ebi.ac.uk/aerep/">http://www.ebi.ac.uk/aerep/</a>	Public repository of microarray data
GEO	<a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>	Public repository of microarray data
ASAP	<a href="http://asap.ahabs.wisc.edu/">http://asap.ahabs.wisc.edu/</a>	Repository of results of functional genomics experiments for selected bacterial species
<i>E. coli</i> multi-omics DB	<a href="http://ecoli.iab.keio.ac.jp/">http://ecoli.iab.keio.ac.jp/</a>	Comprehensive dataset of transcriptomic, proteomic, metabolomic, and fluxomic experiments for <i>E. coli</i> K12
Systomonas	<a href="http://www.systomonas.de/">http://www.systomonas.de/</a>	Repository of 'omics' datasets and molecular networks for pseudomonads species
PubMed	<a href="http://www.pubmed.org/">http://www.pubmed.org/</a>	Database on biomedical literature
<b>Metabolic model repositories</b>		
BiGG	<a href="http://bigg.ucsd.edu/">http://bigg.ucsd.edu/</a>	Repository of reconstructed genome-scale metabolic models
BioModels	<a href="http://www.ebi.ac.uk/biomodels/">http://www.ebi.ac.uk/biomodels/</a>	Database of mathematical models of biological systems

Combining the above methods and information sources increases the reliability of the annotation transfers from proteins of known function to new genes. Current annotation pipelines succeed at assigning a function to 50–80% of the genes (Serres *et al.*, 2004). A number of databases provide such automatically generated annotations for most sequenced bacterial genomes (see Table 1).

In order to build a metabolic model, it is necessary to identify the specific chemical conversions catalyzed by each

enzyme, together with the corresponding stoichiometries. Functional annotations of enzymes therefore need to be translated into appropriate chemical equations. The Enzyme Commission (EC) numbers classification offers an unambiguous way to identify enzyme-catalyzed reactions. When provided by the enzyme annotations, these numbers directly specify which reactions they catalyze. Several enzyme and metabolic databases provide the correspondence between EC numbers and reactions (see Tables 1 and 2). These

**Table 2.** Type of information provided by each data source

Type of information	DBU	EMBL	GenBank	Integr8	CMR	IMG	SEED	BRENDA	ENZYME	UniProt	TransportDB	PSORTdb	Prolinks	STRING	ChEBI	Pubchem	Lipidmaps	Reactome	KEGG	BioCyc	UniPathway	UM-BBD	IntAct	DIP	Array Express	GEO	ASAP	E. coli multi-omics DB	PubMed	
Biochemical activities																														
Enzyme specificity																														
Enzyme localization																														
Reaction equation																														
Reaction direction																														
Metabolite formula																														
GPR association <sup>1</sup>																														
Biomass composition																														
Experimental observations																														

metabolic databases are often comprehensive catalogues of known biochemical reactions with the associated chemical information, including stoichiometry: they include most of the reference knowledge needed to build metabolic models.

Several issues hinder this translation process. First, enzymatic activities that have been identified only recently are usually not included in the EC classification. Furthermore, full EC numbers are not always systematically assigned in the annotation process. As a result, many annotations retrieved from protein databases are only textual (as in UniProtKB) or ontology based [as in Gene Ontology (Ashburner *et al.*, 2000)] and do not provide the required metabolic information directly. To address this shortcoming, PATHOLOGIC, the metabolic network reconstruction software tied to the BioCyc metabolic databases, includes an algorithm performing the identification of gene-reaction links from textual annotations (Karp *et al.*, 2002) (see Table 3). This procedure relies on a dictionary of synonyms, however, and may fail at recognizing an explicit reaction when uncommon terms are used. An expert curation step is thus necessary, for which metabolic pathway databases provide useful guidance. Recent initiatives specifically aim at solving this issue: for instance, textual annotations in UniProtKB/SwissProt are being progressively replaced by direct references to reactions from UniPathway, a metabolic database in which all reaction steps are specified up to the chemical level (see Table 1).

The broad specificity of some enzymes may also significantly increase the number of distinct reactions they can catalyze. For instance, enzymes annotated with alcohol dehydrogenase activity (EC 1.1.1.1) may catalyze the degradation of several distinct alcohols. Similarly, enzymes acting on lipids are often not specific to the length of their carbon chain. In such cases, functional annotations often report the activity using generic metabolites (e.g. 'an alcohol' or 'a fatty acid') representing the entire set of possible substrates. Instantiating reactions with specific metabolites is required when building a metabolic model, however, as accounting for the mass balance constraint requires that all metabolites should be well defined. It is thus necessary to identify for each generic compound the corresponding set of specific compounds, as much for primary substrates as for cofactors. This task is complicated by the combinatorial effect, because the number of substrate combinations may significantly increase the number of specific reactions. To address this issue, databases of chemical species can be used to identify all metabolites of a given chemical category (see Tables 1 and 2). In order to determine which metabolites are preferentially recognized by enzymes, processing the literature or browsing information collected in enzyme databases such as BRENDA (Barthelme *et al.*, 2007) is often necessary. Metabolites involved in metabolic pathways that have already been inferred may also help in selecting the most relevant substrates.

**Table 3.** Methods for model reconstruction

<b>Metabolic model reconstruction (beyond the use of dedicated metabolic databases)</b>	
Identification of metabolic reactions from textual gene annotations	Karp <i>et al.</i> (2002)
Direct inference of metabolic reactions from genome sequence	Sun & Zeng (2004), Arakawa <i>et al.</i> (2006), Notebaart <i>et al.</i> (2006)
Use of metabolic context to complete pathways	Karp <i>et al.</i> (2002), Arakawa <i>et al.</i> (2006), DeJongh <i>et al.</i> (2007)
<b>Metabolic model consistency checks</b>	
Flux variability analysis: identification of reactions that are predicted to never carry any flux	Mahadevan & Schilling (2003)
Identification of dead-end metabolites, which can never be produced or consumed.	Segrè <i>et al.</i> (2003), Ebenhöf <i>et al.</i> (2004), Imielinski <i>et al.</i> (2005), Kumar <i>et al.</i> (2007)
Assessment of thermodynamic consistency and assignment of reaction directions.	Yang <i>et al.</i> (2005), Kümmel <i>et al.</i> (2006a, b)
<b>Gap filling and model expansion</b>	
Graph-based metabolic network expansion using shortest metabolic paths	Arita (2003), Boyer & Viari (2003)
GapFill: optimization-based network expansion and reaction reversibility changes to solve dead-end metabolite inconsistencies	Kumar <i>et al.</i> (2007)
Optimization-based metabolic network expansion to resolve inconsistent growth phenotypes	Reed <i>et al.</i> (2006a, b)
Network-based identification of candidate genes for orphan metabolic activities	Osterman & Overbeek (2003), Green & Karp (2004), Chen & Vitkup (2006), Kharchenko <i>et al.</i> (2006), Fuhrer <i>et al.</i> (2007)

Alternative approaches to metabolic network reconstruction bypass the classical annotation step altogether, taking instead advantage of the curated links between enzyme-encoding gene sequences and reactions [or EC numbers, as in the Genome-Based Modeling (GEM) system (Arakawa *et al.*, 2006)] provided by some metabolic databases. Orthology relationships are sought between reference sequences from these databases and the coding sequences from the new genome. While these methods [e.g. AUTOGRAPH (Notebaart *et al.*, 2006), or IDENTICS (Sun & Zeng, 2004), see Table 3] simplify the reconstruction process, they usually do not benefit from advanced annotation techniques, such as those derived from structural genomics or domains recognition, and are more difficult to combine with expert annotation. They are also conditioned on the availability of curated gene-reaction associations for a set of reference organisms.

The reconstruction of the metabolism of a new organism can also benefit from the knowledge of complete pathways in related organisms. Metabolic databases often group reactions into pathways or modules that indicate known co-occurrence relationships between reactions that hold across several organisms. Three main resources provide this type of information: MetaCyc (Caspi *et al.*, 2006), KEGG Modules (Kanehisa *et al.*, 2007), and SEED (Overbeek *et al.*, 2005) (see Tables 1 and 2). Metabolic model reconstruction procedures tied to such databases can exploit the known co-occurrences of reactions across reference organisms whose metabolism has been extensively studied (Arakawa *et al.*, 2006). An instance of a reconstruction procedure taking advantage of this notion of metabolic context is again PATHOLOGIC, which infers the presence of pathways rather than that of single reactions when possible. A reconstruction procedure based on the SEED database was also proposed

recently (DeJongh *et al.*, 2007); it includes a check that the inferred pathways can be properly connected to form a 'working' model. By leveraging a specific form of 'guilt-by-association,' approaches of this type may be able to retrieve reactions catalyzed by enzymes that cannot be correctly identified using current methods. In addition, the presence of spontaneous reactions in the organism may be identified by the occurrence of neighboring reactions in reference metabolic pathways.

In addition to their equations, the reversibility and localization of reactions need to be determined for metabolic models. Few metabolic or enzyme databases report on the reversibility of reactions in *in vivo* conditions (see Table 2). When not found in the literature, reversibility is therefore often determined using simple thermodynamic considerations based on the reaction Gibbs energy, if it is known, or on basic rules depending on the energy equivalents (e.g. NADH or ATP) involved in the reactions (Ma & Zeng, 2003; Kümmel *et al.*, 2006a). Even though very few compartments divide bacterial cells (with periplasm and cytoplasm as the only main compartments in gram-negative bacteria), the presence of such physical separation between metabolites need to be included in their metabolic models. Enzymes present in one compartment cannot interact with metabolites present in another one. To properly model the effect of compartments, the localization of enzymes and the transport of metabolites need to be determined. Information on the localization of enzymes and reactions is seldom included in metabolic databases. Curated versions of BioCyc databases, especially MetaCyc, are a welcome exception, however (Caspi *et al.*, 2006). When not found in the literature, localization can be inferred using *ab initio* predictions from enzyme sequences (Schneider & Fechner, 2004),



or determined experimentally, for example using fluorescence microscopy (Meyer & Dworkin, 2007). Transport of metabolites can be inferred using comparative genomics tools that identify transport enzymes [e.g. TransportDB (Ren *et al.*, 2004)]. Yet, such methods hardly determine the specificity of transporters; knowledge of transported metabolites is therefore often completed using direct information on the microorganism's physiology and the metabolites it was shown to utilize in growth experiments.

Overall, reconstructing a constraint-based model for an organism's metabolism involves collecting various types of information. A summary of the respective contributions of each data source to the model construction is shown in Table 2.

### Checking the consistency of reconstructed models

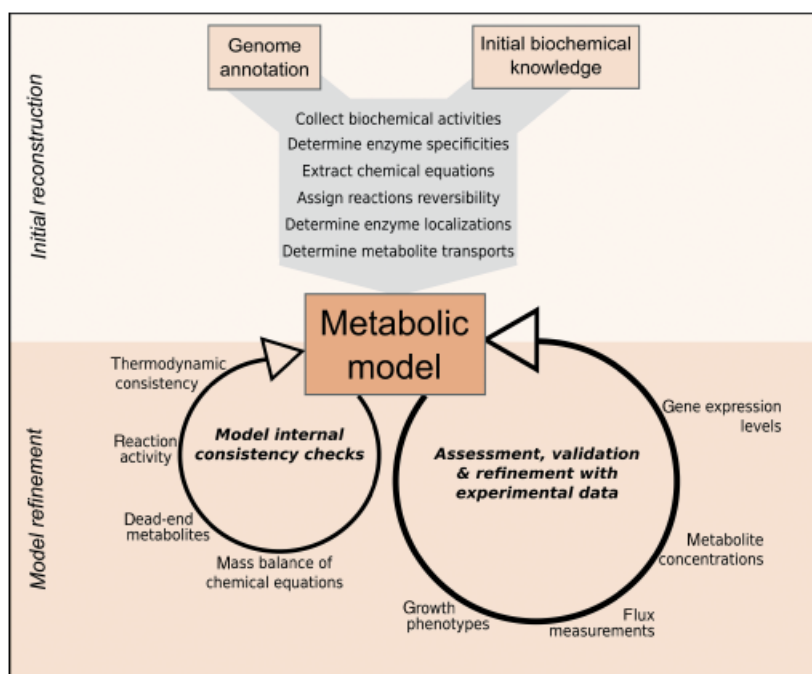
Once a draft metabolic model is obtained, its consistency can be checked using a set of simple tests (see Fig. 2): is the model chemically and physically coherent? Are there remaining 'dead-ends' in metabolic pathways or reactions bound to be inactive? Is the model able to produce essential metabolites from a known growth medium?

Constraint-based metabolic models fundamentally rely on reaction stoichiometries to properly account for the mass balance in metabolism at steady state. It is therefore crucial that all chemical equations are correctly balanced to avoid unrealistic creation or destruction of matter. To control the correctness of the reaction stoichiometries, the atom balance of each reaction can be checked using the chemical formulae

of the metabolites, which are typically found in databases of chemical compounds (see Table 1). For cases where the formula is not available for all metabolites, a method was recently introduced to detect such balance errors in metabolic models by solely comparing chemical equations – for instance, reactions  $A \rightarrow B$  and  $A \rightarrow B+C$  would be identified by this method as 'stoichiometrically inconsistent,' because balancing both equations would require that at least one of the metabolites has a null or negative mass (Gevorgyan *et al.*, 2008).

The assumptions on which constraint-based models are founded do not enforce thermodynamic consistency on the fluxes. Flux distributions obeying conservation of mass can still include internal cycles that violate thermodynamic laws, allowing for instance the artificial generation of high-energy cofactors. To prevent models from predicting such unrealistic metabolic modes, extensions of the modeling framework were proposed that directly enforce these laws (Beard *et al.*, 2002). Their nonlinear nature entails costly computations, however, which hinder the use of such modeling extensions in practice. In order to provide thermodynamically consistent models without including such extensions, methods have been developed to detect inconsistent cyclic modes in draft metabolic models, and propose changes in reaction reversibility that would avoid those modes from being predicted (Yang *et al.*, 2005; Kümmel *et al.*, 2006a).

Before one can reap the benefits of having a model, the model should be functional, i.e. it should be checked that non-null fluxes can actually be predicted. This relates to the completeness of the model, because for instance a missing



**Fig. 2.** Pipeline for model reconstruction and refinement. An initial model is reconstructed from genome annotations and from preexisting knowledge on the species' biochemistry and physiology. Besides collecting the biochemical activities, this task includes several additional key steps. The resulting model is then iteratively corrected and refined, according to internal consistency criteria and by comparing its predictions to experimental data.

reaction in a linear pathway would prevent any non-null flux from being predicted in it at steady state, thereby inactivating all other reactions in the pathway. Metabolites that are never consumed or never produced, so-called 'dead-ends,' typically reveal that reactions are missing. In order to help investigate and correct these so-called 'metabolic gaps,' methods have been developed that assess whether reactions can be active in the model (Reed & Palsson, 2004), identify dead-end metabolites (Kumar *et al.*, 2007) or directly track the producibility of metabolites from source metabolites (Segrè *et al.*, 2003; Ebenhöf *et al.*, 2004). In case the model is later used to predict growth phenotypes (see Applications of metabolic models), the producibility of biomass precursors and the completeness of their biosynthetic pathways should be especially checked beforehand. Dedicated procedures have been designed to systematically perform these checks for newly reconstructed models (Segrè *et al.*, 2003; Imielinski *et al.*, 2005; Senger & Papoutsakis, 2008). Solving such inconsistencies often involves filling metabolic gaps or completing the network with additional metabolic pathways.

The methods presented in this section check the consistency of the reconstructed model with respect to a set of basic rules (see Table 3). We will review in the section on model applications how model predictions can also be confronted with experimental data, providing consistency checks of the model with respect to diverse additional experimental evidence. Interpreting and solving identified inconsistencies of either type are key to improving the quality of the metabolic model.

### Targeted searches for missing metabolic activities

Consistency checks (either internal to the model or relative to experimental datasets) may show that the reconstructed model is incomplete and lacks some metabolic reactions. Resolving these metabolic gaps entails expanding the model by identifying and including missing biochemical activities. This process basically consists of two steps: (1) identifying plausible candidate reactions that could complete the model and (2) finding genes that could catalyze the hypothesized activities.

Reactions contained in metabolic databases are the primary source of information for completing the metabolic model (see Table 1). The search for candidate reactions within these databases can be facilitated using knowledge of existing pathways (as in MetaCyc, SEED, or UM-BBD, see Table 1) or computational methods (Arita, 2003; Boyer & Viari, 2003; Kumar *et al.*, 2007) (see Table 3). In the latter category, the GapFill method was specifically developed to identify dead-ends in models, and correct them by adding reaction from a global repository of reactions, changing the reversibility status of reactions, or adding transporters (Kumar *et al.*, 2007). The addition of reactions to the model is guided by an optimization step minimizing the number of

reactions. Similarly, Reed *et al.* (2006b) proposed a method which drives the expansion of the metabolic model to account for the utilization of additional external compounds. For metabolites experimentally shown to be used by the organism but not predicted as such by the model (see Applications of metabolic models on growth phenotype predictions for methods to perform these predictions), their method automatically proposes minimal sets of reactions from a repository of reactions that, if added, would allow the model to exploit the external metabolites.

The set of reactions referenced in metabolic databases is far from being comprehensive: the right candidates for completing the model may not yet be known. Computational and experimental approaches have been proposed to extend this 'universe of possible reactions.' On the computational side, several methods originating from the field of chemo-informatics have been designed to infer chemical transformations (Gasteiger, 2005). Some of them have been more specifically adapted to biochemical transformations, using rules on enzymatic conversions to infer new conversions for biologically relevant metabolites (Klopman *et al.*, 1994; Arita, 2000; Hatzimanikatis *et al.*, 2005; Ellis *et al.*, 2008).

Numerous experimental methods are also being developed to explore the range of possible biochemical reactions. MS and nuclear magnetic resonance (NMR) techniques are able to identify and quantify large sets of metabolites at high throughput (Dunn *et al.*, 2005; Dettmer *et al.*, 2007). Computational methods have been proposed to infer reactions from MS data, by analyzing mass differences between related metabolites (Breitling *et al.*, 2006) or correlations between metabolite concentrations across distinct conditions (Steuer, 2006). They do not provide direct evidence for biochemical transformations, however: their predictions should be treated as clues to be confirmed by additional information. Although mostly used to determine metabolic fluxes, atom-labeling experiments could also become powerful tools to elucidate novel metabolic pathways (Sauer, 2006). They can advantageously complement computational *ab initio* pathway inference methods by selecting candidate pathways that are compatible with observed isotopic patterns. Finally, untargeted enzyme activity screenings have recently been performed to identify the substrates of enzymes of unknown function and discover novel activities (Saghatelian *et al.*, 2004; Saito *et al.*, 2006). The availability of large-scale libraries of ORF clones (Kitagawa *et al.*, 2005) should increase the likelihood of such methods expanding the store of known reactions.

The search for candidate genes for orphan metabolic activities is in some ways the reverse of the classical genome annotation problem (i.e. searching the function of identified genes). Yet, many of the tools developed to determine gene functions can be adapted for this purpose. Sequence homology to already characterized genes is central to most methods for candidate gene detection, but combining it

with additional types of evidence can significantly improve performance. For instance, several approaches exploit functional links, such as gene neighborhood, gene co-expression, protein interaction, or phylogenetic co-occurrence, to relate candidate genes with genes involved in the same metabolic pathways or close in the metabolic network (Osterman & Overbeek, 2003; Green & Karp, 2004; Chen & Vitkup, 2006; Kharchenko *et al.*, 2006; Fuhrer *et al.*, 2007). Databases such as STRING (von Mering *et al.*, 2007) or Prolinks (Bowers *et al.*, 2004) compile large sets of functional links across a wide range of organisms. On the experimental side, enzyme activity screenings are used to validate the generated candidates. Furthermore, when the orphan activity is associated to a specific phenotype, screens of systematic knockout mutant phenotypes can help in identifying candidates (Aghaie *et al.*, 2008).

The two types of methods – finding candidate reactions or candidate genes – benefit from being used in combination, as identifying genes for putative reactions can help in selecting the proper reactions to include.

## Applications of metabolic models

A wealth of computational methods has been developed to help analyze biological properties revealed by reconstructed metabolic models. Not only would a comprehensive and technical description exceed the scope of this review, but these methods have been extensively covered elsewhere, either on the technical side (Price *et al.*, 2004) or for applications on a specific organism, i.e. *E. coli* (Feist & Palsson, 2008). We will provide here the reader with a review on the main applications for which constraint-based models have been most successful and are mostly promising for bacterial species. We will distinguish four main types of applications: (1) analysis of network properties of metabolism, (2) prediction and analysis of bacterial growth phenotypes, (3) model-based interpretation of experimental data, and (4) metabolic engineering.

### Analysis of network properties

The principle of constraint-based modeling consists in studying the set of reaction fluxes – namely flux distributions – that are achievable at steady state given the constraints imposed on the system. Reaction fluxes can vary inside a continuous set of possible values. This set can encompass significant variability at the level of individual pathway or reaction fluxes. A wide range of methods have been designed to explore that variability and analyze specific properties of metabolites and reactions which emerge from the flux constraints.

One approach consists in sampling the set of achievable flux distributions (Almaas *et al.*, 2004; Reed & Palsson, 2004; Wiback *et al.*, 2004). Methods that provide a uniform sampling of the possible states have been proposed (Almaas

*et al.*, 2004; Wiback *et al.*, 2004). By sampling a significant number of metabolic states, these approaches offer an overview of the range of flux distributions that can occur in the metabolic network at steady state. The ‘uniform’ nature of the sampling is based only on the mathematical description of the set of possible flux distributions, avoiding any prior assumption on which metabolic states are most likely to be selected *in vivo*. For instance, these sampling methods have been used to evaluate the relative occurrence of reactions within the set of possible flux distributions and across several environmental conditions (Almaas *et al.*, 2004). This analysis showed that a few reactions are active in many sampled flux distributions and carry high fluxes – forming a so-called high-flux metabolic backbone – while many others are active in few sampled flux distributions and carry low fluxes. Similar methods were also used to evaluate the correlation of flux values between pairs of reactions across sampled metabolic states (Reed & Palsson, 2004; Becker *et al.*, 2007) and thereby determine metabolic dependencies between reactions. From a more theoretical angle, sampling was also used to evaluate the size of the set of possible flux distributions (Wiback *et al.*, 2004; Braunstein *et al.*, 2008). When computed for distinct (genetic perturbation  $\times$  environmental condition) pairs, the relative sizes of the corresponding flux distribution sets were interpreted as indicators of the respective diversity of metabolic states in the tested conditions (Wiback *et al.*, 2004).

The diversity of achievable metabolic fluxes can also be evaluated locally for each reaction. Flux variability analysis was designed for this purpose: an optimization procedure computes the minimal and maximal allowed flux of each reaction independently (Mahadevan & Schilling, 2003). This procedure identifies reactions that do not carry any flux, or conversely those that carry non-null flux in all possible metabolic states. Flux variability analysis has been broadly used to predict the activity of reactions for specific sets of metabolic constraints (Mahadevan & Schilling, 2003; Reed & Palsson, 2004; Teusink *et al.*, 2006; Feist *et al.*, 2007; Henry *et al.*, 2007; Shlomi *et al.*, 2007a).

Flux sampling or flux variability approaches only provide partial description of the set of possible flux distributions. To get a comprehensive picture of the possibilities, methods which compute elementary modes (Schuster *et al.*, 2000) and extreme pathways (Schilling *et al.*, 2000) have been developed. These notions differ only slightly in their mathematical formulation (Klamt & Stelling, 2003; Papin *et al.*, 2004): the main idea is to determine the set of elementary and independent metabolic routes that can occur in the metabolic model. These elementary routes are flux distributions that (1) respect all assumed constraints, including steady state and irreversibility, and (2) are elementary in the sense that they are composed of a minimal set of active reactions. This second condition ensures that the flux

distribution is not decomposable into a combination of smaller elementary routes. It can be shown that any achievable flux distribution can be expressed as a combination of such elementary routes. This property, together with the fact that the set of elementary routes is unique, independently of the method used to compute it (Klamt & Stelling, 2003), has inspired numerous applications. This subfield is also known as metabolic pathway analysis. For instance, elementary modes and extreme pathways have been used to exhaustively describe the independent metabolic routes occurring in newly reconstructed models, often sorted by metabolic function (Schilling & Palsson, 2000; Van Dien & Lidstrom, 2002; Papin *et al.*, 2002). The redundancy of routes can be assessed and the respective yields of routes of conversion can be compared (Papin *et al.*, 2002). Conversely, the relative importance of reactions in metabolism was scored using elementary routes, reactions involved in many routes being likely to be key players in metabolism (Stelling *et al.*, 2002). Finally, metabolic dependencies between reactions which are stronger than those determined only by analyzing the correlation of fluxes in sampled distributions can be deduced from knowing elementary routes. Reactions that always appear jointly in elementary routes are bound to operate together (Pfeiffer *et al.*, 1999). The main obstacle in metabolic pathway analysis is the size and complexity of the metabolic models, as the number of elementary routes dramatically increases with the size of the model (Yeung *et al.*, 2007). The computation of all routes is currently only tractable for medium-size models, although significant progresses have been made recently (Terzer & Stelling, 2008).

Alternative approaches have been developed in order to explore metabolic dependencies in models of larger size. One of them, flux coupling analysis, has become a popular analytical tool (Burgard *et al.*, 2004). Flux coupling analysis identifies all pairs of reactions whose fluxes are always coupled at steady state. It has been used in a wide range of studies, and the resulting sets of coupled reactions were for instance compared with correlations observed in the transcriptional states of enzymes (Reed & Palsson, 2004; Notebaart *et al.*, 2008), interpreted with respect to the structure of the metabolic regulation (Notebaart *et al.*, 2008), and used to study the horizontal transfer of genes during bacterial evolution (Pál *et al.*, 2005a, b). Similar methods were developed to study metabolic relationships between metabolites, either by simply examining the co-occurrence of metabolites in reactions (Becker *et al.*, 2006) or by determining conservation relations between metabolites (Nikolaev *et al.*, 2005; Imielinski *et al.*, 2006). This last type of method was applied to determine coupling relationships between metabolite concentrations, identify metabolite pools sharing conserved chemical moieties (Nikolaev *et al.*, 2005), and exhaustively predict distinct minimal growth media for *E. coli* (Imielinski *et al.*, 2006).

## Prediction of growth phenotypes

One of the primary uses of genome-scale metabolic models is the prediction of growth phenotypes (Price *et al.*, 2004; Palsson, 2006). Because these models aim at comprehensiveness, they are able to account for all main metabolic processes contributing to growth, i.e. the production of energy and biomass precursors from external metabolites. Growth phenotypes can therefore be predicted by examining to which extent metabolic requirements for growth, in terms of energy generation and biomass precursors synthesis, can be fulfilled by flux distributions from the model. Growth phenotypes can be predicted either in a qualitative manner (prediction of the mere ability to grow) by checking piecemeal for the producibility of each biomass precursor metabolite (Imielinski *et al.*, 2005), or in a quantitative manner (prediction of growth performance) by including a biomass reaction consuming them in proportion to their ratio in biomass composition and studying the flux values it can attain (Price *et al.*, 2004). Determining biomass composition is therefore a necessary prerequisite to growth phenotype predictions. This is often achieved by examining the relevant literature or adapting known biomass compositions of related organisms. The Flux Balance Analysis (FBA) method was specifically designed to predict quantitative growth phenotypes (Varma & Palsson, 1994b; Price *et al.*, 2004). It computes the maximal growth yield achievable in the metabolic model by maximizing the biomass reaction flux (representing the growth rate) given a set of bounded intake rates for external substrates. FBA relies on the strong assumption that bacteria have optimized their growth performance in a subset of possible environments during their evolution, thereby making the maximization of biomass production a driving principle for metabolic operation (Varma & Palsson, 1994b). This assumption has been confirmed by experiments in several cases (Edwards *et al.*, 2001). Using FBA, global quantitative relationships can be predicted between the input rates of nutrients, the output rates of byproducts, and the growth rate (Stephanopoulos *et al.*, 1998; Edwards *et al.*, 2002; Price *et al.*, 2004).

The global energy consumption of the cell can significantly influence the outcome of quantitative growth phenotype predictions. Two ATP hydrolysis fluxes are added to the models in order to properly account for it. One is constant and models the non-growth-associated maintenance, which represents the fraction of the energy demand necessary for the cell survival that is independent from its growth rate, for example to maintain the right ionic strength (Stouthamer & Bettenhausen, 1973). The second flux is proportional to the growth rate and corresponds to the energy demand associated with growth beyond the mere requirements of metabolic pathways – which are already directly accounted for in the model – for example energy for cell division or assembly

**Table 4.** Existing genome-scale metabolic models for bacterial organisms

Organism	Reference	Genes	Reactions*	Metabolites <sup>†</sup>	Experimental assessment		
					Wild-type growth phenotypes	Knockout mutant growth phenotypes	Quantitative growth measures
<i>Acinetobacter baylyi</i>	Durot et al. (2008)	774	875	701	173/190 (91%)	1138/1208 (94%)	–
<i>Bacillus subtilis</i>	Oh et al. (2007)	844	1020	988	200/271 (74%)	720/766 (94%)	–
<i>Clostridium acetobutylicum</i>	Lee et al. (2008a)	432	502	479	10/11 (91%)	–	X
<i>Clostridium acetobutylicum</i>	Senger & Papoutsakis (2008)	474	552	422	–	–	–
<i>Escherichia coli</i> <sup>‡</sup>	Feist et al. (2007)	1260	2077	1039	129/170 (74%)	1152/1260 (92%)	X
<i>Geobacter sulfurreducens</i>	Mahadevan et al. (2006)	588	523	541	–	–	X
<i>Haemophilus influenza</i>	Schilling & Palsson (2000)	412	461	367	–	–	–
<i>Helicobacter pylori</i> <sup>§</sup>	Thiele et al. (2005)	341	476	485	–	54/72 (75%)	–
<i>Lactobacillus plantarum</i>	Teusink et al. (2006)	721	643	531	–	–	X
<i>Lactococcus lactis</i>	Oliveira et al. (2005)	358	621	422	–	–	X
<i>Mannheimia succiniciproducens</i>	Hong et al. (2004)	335	373	332	–	–	–
<i>Mycobacterium tuberculosis</i>	Beste et al. (2007)	726	849	739	–	547/705 (78%)	X
<i>Mycobacterium tuberculosis</i>	Jamshidi & Palsson (2007)	661	939	828	–	132/237 (56%)	X
<i>Neisseria meningitidis</i>	Baart et al. (2007)	555	496	471	–	–	X
<i>Pseudomonas aeruginosa</i>	Oberhardt et al. (2008)	1056	883	760	78/95 (82%)	893/1056 (85%)	–
<i>Pseudomonas putida</i>	Nogales et al. (2008)	746	950	710	84/90 (93%)	665/746 (89%) <sup>¶</sup>	X
<i>Rhizobium etli</i>	Resendis-Antonio et al. (2007)	363	387	371	–	–	–
<i>Staphylococcus aureus</i>	Becker & Palsson (2005)	619	641	571	–	–	–
<i>Staphylococcus aureus</i>	Heinemann et al. (2005)	551	774	712	–	8/14 (57%)	–
<i>Streptomyces coelicolor</i>	Borodina et al. (2005)	700	700	500	54/58 (93%)	11/12 (92%)	X

First two columns of experimental assessment show the number of correct predictions among all experimentally determined qualitative growth phenotypes. Last column specifies whether the model has been assessed against quantitative growth rate measurements.

\*Number of distinct reactions including transport processes.

<sup>†</sup>Number of biochemically distinct metabolites.

<sup>‡</sup>This model is an update of two earlier models for *E. coli* (Edwards & Palsson, 2000; Reed et al., 2003).

<sup>§</sup>This model is an update of an earlier model for *H. pylori* (Schilling et al., 2002).

<sup>¶</sup>Using gene essentiality data for *Pseudomonas aeruginosa*.

of higher order cell structures. These two parameters are usually determined by fitting growth yield predictions derived using FBA to measured growth yields provided by growth monitoring experiments (Reed et al., 2006a). Measurements of growth yields for distinct growth rates are sufficient to fit both growth-associated and non-growth-associated maintenance parameters (Varma & Palsson, 1994a). The values of these parameters were determined using experimental growth measurements for a significant proportion of reconstructed models (see Table 4).

Once fitted, and assuming these parameters remain constant across environments, the model can be used to predict growth rates on different media (Edwards et al., 2001). Predicted growth yields revealed to be consistent with observed ones on a significant number of media for *E. coli* (Edwards et al., 2001). Inconsistencies between predicted and observed growth yields can have multiple interpretations. First, the assumption of optimal substrate utilization can be questionable for growth predictions on environments that are not commonly encountered by the organism (Ibarra et al., 2002; Schuster et al., 2008). Using an adaptive

evolution experiment on *E. coli* cells grown in glycerol minimal medium, Ibarra and colleagues actually observed that, while the initial growth yield was suboptimal, it progressively evolved to reach the optimal value predicted by the model. Other biological constraints, such as regulation or capacity constraints, may also prevent the organism from using optimal flux distributions (Oliveira et al., 2005; Feist et al., 2007). Comparing predictions of growth phenotypes with experimental measures may also help in refining the model. A model component that is often refined using quantitative growth predictions is the stoichiometry of proton translocation that occurs in reactions of electron transport systems, such as the respiratory chain. These stoichiometries are often hard to determine *a priori*, yet they impact directly the P/O ratio and the efficiency of energy generation (Reed et al., 2006a). With the help of a metabolic model and growth yield measurements on several distinct media, Feist et al. (2006) studied the unknown proton translocation stoichiometry of such a reaction in *Methanosarcina barkeri* by determining for each media the model maintenance parameters that provided the best growth yield

predictions for different hypothesized values of the stoichiometry. Assuming that maintenance should not significantly change across media, they selected the stoichiometry that triggered the smallest variation among the determined maintenance parameters across the environments. Other studies investigated the stoichiometry of proton translocation in the respiratory chain by directly exploiting measured ratios of electron acceptor (e.g. oxygen, or Fe(III) in *Geobacter sulfurreducens*) consumption rate vs. carbon source consumption rate and growth rate (Heinemann *et al.*, 2005; Mahadevan *et al.*, 2006).

Models can readily predict the effect of gene deletion on growth phenotypes. To that end, a layer of Gene Protein Reaction associations – usually called GPR (Reed *et al.*, 2003) – is added to the model to predict the effect of gene deletion on reaction activity. Each reaction is associated to its enzyme-encoding genes by a Boolean rule: genes encoding for subunits of an enzymatic complex are linked with an AND rule, while genes encoding for alternative enzymes are linked with an OR rule. Using GPR rules, gene deletions are translated into ‘blocked’ reactions, which are then inactivated in the model by constraining their fluxes to zero. FBA can be applied to predict growth phenotypes of gene knockout mutants. Nevertheless, the hypothesis of optimal growth is largely debatable for such genetically engineered mutants, as their metabolism was not exposed to evolutionary pressure. Basing on the assumption that metabolism in a knockout mutant operates as closely as possible to metabolism in the wild-type strain, two specific methods were introduced. They predict knockout mutant growth phenotypes by minimizing either the overall flux change [MoMA (Segrè *et al.*, 2002)] or the number of regulatory changes [ROOM (Shlomi *et al.*, 2005)] between the wild-type strain and the mutant strain (see Table 5). Both methods were shown to provide slightly better predictions than FBA.

The throughput of experiments evaluating qualitative growth phenotypes – i.e. described simply as *viable* or *lethal* – has increased dramatically in the last few years. Phenotype Microarrays from Biolog Inc. typically report growth phenotypes for several hundreds of media in a single experiment (Bochner *et al.*, 2001). In parallel to this, collections of knockout mutants are being built for a growing number of bacteria (Akerley *et al.*, 2002; Jacobs *et al.*, 2003; Kobayashi *et al.*, 2003; Baba *et al.*, 2006; Liberati *et al.*, 2006; Suzuki *et al.*, 2006; de Berardinis *et al.*, 2008). The systematic assessment of growth phenotypes of knockout mutants provides a significant resource for exploring the metabolic capabilities of organisms and investigating their gene functions (Carpenter & Sabatini, 2004), but their direct interpretation is made difficult by the complexity and size of metabolic networks (Gerdes *et al.*, 2006). These results can be readily compared with model predictions, however, providing a way to interpret them and assess the model correctness. Given the qualitative nature of these growth

phenotypes, two types of inconsistencies may arise: *false viable* predictions – growth was predicted yet not observed experimentally – and *false lethal* predictions – growth was not predicted yet observed experimentally. On the one hand, these inconsistencies may be caused by limitations of the model or cases where the modeling assumptions do not hold. Regulation may for instance trigger a lethal phenotype by blocking an alternate pathway, which would not be predicted as blocked in the merely metabolic model. On the other hand, examining the inconsistencies may identify errors in the model and lead to its refinement. All model components may comprise errors, including the GPR associations, the metabolic network itself, and the stated biomass requirements. False lethal predictions are often clues that some biomass component is actually not essential, or that the model lacks an alternative gene or pathway that would allow it to survive in the given experimental conditions. Conversely, false viable predictions can help detect missing essential biomass components, genes falsely annotated as encoding isozymes or reactions that were wrongly assigned or are inactive in the experimental conditions (Duarte *et al.*, 2004; Joyce *et al.*, 2006). Growth phenotype predictions have been evaluated for a significant proportion of reconstructed models, whenever experimental data were available (see Table 4). Interpretation of inconsistent cases by expert examination led to several annotation and model refinements, some of which were supported by the results of targeted experiments (Covert *et al.*, 2004; Duarte *et al.*, 2004; Joyce *et al.*, 2006; Reed *et al.*, 2006b). Automated methods were recently introduced to systematically look for interpretations of inconsistencies and possible modifications in the model. Corrections of the GPR associations can be systematically proposed that match the gene essentiality observation with predicted reaction essentiality (M. Durot *et al.*, unpublished data). With regard to the metabolic network itself, metabolic gap filling approaches have been adapted to propose network corrections that resolve wrongly predicted growth phenotypes (Reed *et al.*, 2006b). Finally, valuable insights into the determination of essential biomass precursors can be provided by methods that analyze correlations between lethality and metabolite production (Imielinski *et al.*, 2005; Kim *et al.*, 2007). All these methods act independently on distinct components of the model. A unifying method integrating all types of corrections, which is yet to come, could lead to an integrated platform for the systematic interpretation of upcoming growth phenotyping results.

Models can actually predict growth phenotypes for any environmental condition and any combination of gene deletions, which is beyond reach of experiments. Given the combinatorial complexity of mixing several gene deletions, dedicated methods have been designed to analyze the effects of multiple deletions and applied to identify epistatic interactions between genes (Klamt & Gilles, 2004; Deutscher *et al.*,

**Table 5.** Main analytical methods for genome-scale models sorted by type of application

<b>Analysis of network properties</b>	
Flux sampling: random sampling of flux distribution among the set of possible metabolic states	Almaas <i>et al.</i> (2004), Reed & Palsson (2004), Wiback <i>et al.</i> (2004)
Flux variability analysis: examination of flux variability for each reaction	Mahadevan & Schilling (2003)
Metabolic pathway analysis, elementary modes/extreme pathways: comprehensive description of all independent metabolic modes achievable in the metabolic network	Schilling <i>et al.</i> (2000), Schuster <i>et al.</i> (2000), Klamt & Stelling (2003)
Flux coupling: identification of reaction pairs whose fluxes are coupled	Burgard <i>et al.</i> (2004)
Metabolite coupling/evaluation of conserved metabolite pools	Nikolaev <i>et al.</i> (2005), Becker <i>et al.</i> (2006), Imielinski <i>et al.</i> (2006)
<b>Prediction and interpretation of bacterial growth phenotypes</b>	
Producibility analysis of biomass precursors	Imielinski <i>et al.</i> (2005)
FBA: quantitative prediction of growth yield by maximization of growth rate given bounded nutrient input rates	Varma & Palsson (1994a, b)
MOMA: prediction of gene deletion mutant flux distribution by minimizing overall flux changes with wild type	Segrè <i>et al.</i> (2002)
ROOM: prediction of gene deletion mutant growth by minimizing regulatory changes with wild type	Shlomi <i>et al.</i> (2005)
Identification of multiple gene deletion essentialities	Klamt & Gilles (2004), Deutscher <i>et al.</i> (2006), Imielinski & Belta (2008)
<b>Model-based interpretation of experimental data</b>	
<i>Metabolic flux measurements</i>	
Metabolic Flux Analysis using labeled metabolites: prediction of attainable reaction fluxes given observed metabolite isotopic patterns	Wiechert (2001), Sauer (2006)
Global prediction of reaction activities using metabolic flux measurements on subsets of reactions	Herrgård <i>et al.</i> (2006a, b)
Identification of metabolic objectives best describing observed fluxes	Burgard & Maranas (2003), Schuetz <i>et al.</i> (2007)
<i>Metabolite concentrations</i>	
Comparison of model coverage with experimentally detected metabolites	Oh <i>et al.</i> (2007)
NET analysis and TMFA: application of thermodynamic constraints to reaction directions using metabolite concentrations	Kümmel <i>et al.</i> (2006a, b), Henry <i>et al.</i> (2007)
<i>Gene expression</i>	
Identification of metabolic pathways correlated with gene expression levels	Schwartz <i>et al.</i> (2007)
Refinement of flux distribution predictions by blocking reactions corresponding to unexpressed genes	Akesson <i>et al.</i> (2004)
Evaluation of consistency of gene expression levels with metabolic objectives	Becker & Palsson (2008)
rFBA and SR-FBA: prediction of gene expression states using Boolean regulatory rules	Covert <i>et al.</i> (2001), Barrett <i>et al.</i> (2005), Barrett & Palsson (2006), Shlomi <i>et al.</i> (2007a, b)
<b>Metabolic engineering</b>	
Systematic identification of gene deletions enhancing metabolite production yield	Burgard <i>et al.</i> (2003), Patil <i>et al.</i> (2004), Alper <i>et al.</i> (2005a, b)
OptStrain: systematic identification of reaction additions enabling the production of novel metabolites	Pharkya <i>et al.</i> (2004)
Prediction of adjustments of enzyme expression levels enhancing metabolite production yield	Pharkya & Maranas (2006), Lee <i>et al.</i> (2007)

2006, 2008; Imielinski & Belta, 2008). Prediction of growth phenotypes have also been used to automatically assign condition-dependent roles to genes (Shlomi *et al.*, 2007b), investigate the causes of gene dispensability (Papp *et al.*, 2004; Kuepfer *et al.*, 2005), or study bacterial evolution (Pál *et al.*, 2005a, 2006). These two latter studies on bacterial evolution used an *E. coli* model to analyze the effect of changing growth environments on the acquisition of new metabolic capabilities by horizontal gene transfer (Pál *et al.*, 2005a) and to simulate the reductive evolution of metabolism in specific environmental conditions (Pál *et al.*, 2006).

### Model-based interpretation of experimental data

The recent development of experimental techniques has enabled measurements at genome-scale of several types of quantities, generating so-called 'omics' datasets. These datasets provide partial yet comprehensive snapshots of cellular mechanisms (Ishii *et al.*, 2007a), but their interpretation is made difficult by the volume of data. Computational methods are thus needed if meaningful biological results are to be extracted (Joyce & Palsson, 2006). A variety of

methods have been developed to exploit experimental data related to metabolic states, for example measurements of metabolic fluxes, metabolite concentrations, enzyme levels, or gene expression, in the light of genome-scale models. Two cases generally arise: either experimental observations are directly comparable to model predictions, or these observations lead to the imposition of additional constraints that refine the set of predicted metabolic states. Observations falling in the second category allow for instance the selection of those metabolic routes that are compatible with the experimental observations, or help predict quantitative values for the fluxes. When directly comparable to model predictions, experimental data may be used to assess model correctness and assumptions, identify inconsistencies, and target improvements, as illustrated above with growth phenotypes (Reed *et al.*, 2006b). We will review such integration methods in the following sections for three types of experimental data: measurement of (1) reaction fluxes, (2) metabolite concentrations, and (3) gene expression levels.

Refining the model with experimental data increases its correctness with respect to the observations but may decrease its predictive power. Predictions performed with a refined model should actually be interpreted with care to avoid circular reasoning: data that have been directly used to improve the model can no more be considered as predictions, they are part of the evidences on which the model is based to perform predictions. For instance, a model whose maintenance parameters have been determined using growth rate measurement can no more *predict* the growth rate for the environmental condition. This problem can become serious when models are extensively fitted with experimental data, as they then become more descriptive than predictive. Nevertheless, some refinement processes applied to genome-scale models involve finding additional biological evidence that supports the refinement, thereby breaking the circular reasoning. For instance, corrections of inconsistent growth phenotype predictions by additions of alternate enzymes often involve finding additional proofs that the introduced enzymes possess the right activity.

### Metabolic flux measurements

One of the most direct experimental accesses to metabolic fluxes is provided by atom-labeling experiments (Wiechert, 2001; Sauer, 2006). By analyzing the fate of labeled metabolites, valuable information can be deduced about the reactions that are actually taking place. The most common technique for this consists in analyzing the stable isotope patterns (mostly using  $^{13}\text{C}$ ) found in products of metabolism given known isotope patterns in nutrient metabolites (Wiechert, 2001; Sauer, 2006). These data can be properly interpreted only using a metabolic model that includes information about atom mappings for each reaction (Zupke & Stephanopoulos,

1994; Wiechert *et al.*, 1999; Antoniewicz *et al.*, 2007a). Such models have been built for a few organisms, often using existing constraint-based models as a basis (Antoniewicz *et al.*, 2007b; Suthers *et al.*, 2007). While atom mappings for reactions are currently mostly inferred using cheminformatics methods (Raymond *et al.*, 2002; Arita, 2003; Hattori *et al.*, 2003), this information will likely be made accessible in dedicated databases in the coming years.

By qualitatively examining isotope patterns in nutrients and products, information can already be extracted about the possible routes of conversion (van Winden *et al.*, 2001; Sauer, 2006; Kuchel & Philp, 2008). Patterns in products actually depend on their biosynthetic pathways. Observed patterns that are inconsistent with the predicted possible patterns are clues that other pathways may occur *in vivo*. This approach was for instance recently used to evaluate the model of *G. sulfurreducens*: an inconsistent isotope pattern for isoleucine led to the discovery of an isoleucine biosynthesis pathway previously uncharacterized in this bacteria (Risso *et al.*, 2008).

Quantitative interpretation of isotope patterns together with measurement of extracellular metabolite fluxes can help determine the value of intracellular reaction fluxes using Metabolic Flux Analysis (Zupke & Stephanopoulos, 1994; Stephanopoulos *et al.*, 1998; Wiechert *et al.*, 1999; Sauer, 2006; Antoniewicz *et al.*, 2007a). Known flux values can then be directly exploited in models to characterize which metabolic pathways are operating and quantify their fluxes. As an application, Herrgård *et al.* (2006a) introduced the optimal metabolic network identification method, which combines flux measurements for a fraction of the reactions with the assumption of optimal growth from FBA to globally infer which reactions are active. This method has been for instance used to identify bottleneck reactions that limit the growth in engineered strains, and discard putative reactions from newly reconstructed models (Herrgård *et al.*, 2006a).

Observed fluxes were also used to determine relevant objective functions to choose when predicting metabolic states with FBA (Burgard & Maranas, 2003). By evaluating the match of predicted fluxes with observed ones, these studies could identify those metabolic objectives that provided the best fit. Distinct objectives, including maximization of ATP or biomass yields, were identified for instance in *E. coli* depending on the environmental conditions (Schuetz *et al.*, 2007). Observed metabolic fluxes, however, often show that metabolism does not necessarily operate according to optimality principles (Fischer *et al.*, 2004), especially when regulatory constraints are overlooked.

### Metabolite concentrations

High-throughput measurement of intracellular metabolite concentrations is becoming common practice thanks to



recent developments in MS and NMR technologies (Dunn *et al.*, 2005; Dettmer *et al.*, 2007). Metabolite profiling experiments commonly detect thousands of peaks, among which hundreds can usually be exploited to identify metabolites and determine their concentrations, using for instance known spectra of reference metabolites (Dunn *et al.*, 2005). These datasets, while not fully comprehensive, provide significant information on metabolites present in the cell.

Merely comparing the set of detected metabolites to the set of metabolites present in the model already help in assessing the comprehensiveness of the model. For example, in the reconstruction process of *Bacillus subtilis* metabolic model, Oh *et al.* (2007) evaluated the overlap between model metabolites and intracellular metabolites identified in a metabolomics dataset; among 350 intracellular metabolites identified, only 160 were present in the model. No previously known biochemical activities could be associated with the remaining metabolites, illustrating the fact that a large part of *B. subtilis* metabolism remains unknown. These unaccounted metabolites can guide further investigations on missing activities, leading to expansion of the model's metabolite scope consequently.

By extending the constraint-based modeling framework to encompass thermodynamic constraints on Gibbs energies of reactions, knowledge of absolute metabolite concentrations can be translated into constraints on flux directions (Kümmel *et al.*, 2006b; Henry *et al.*, 2007). A first application is to check the consistency of metabolomic datasets with respect to metabolic fluxes predicted by the model. Methods and software have been developed to pinpoint inconsistent concentration measures (Zamboni *et al.*, 2008). Conversely, metabolomic-derived constraints refine the characterization of metabolic fluxes within the model; their integration has allowed the prediction of ranges of concentrations for unmeasured metabolites, reaction directions, and ranges of Gibbs energies of reactions, identifying thereby potentially regulated reactions (Kümmel *et al.*, 2006b).

Thermodynamic constraints merely enforce link between the concentrations of metabolites and the directions of reactions. Taking reaction kinetics into consideration could reinforce that link and make it more quantitative. Extending models to handle kinetics is still an open issue (Famili *et al.*, 2005; Yugi *et al.*, 2005; Ishii *et al.*, 2007b; Smallbone *et al.*, 2007; Covert *et al.*, 2008; Jamshidi & Palsson, 2008), all the more challenging because of the potential influence of regulation, the scarcity of kinetic parameter values and the lack of scalable analytical methods.

## Gene expression data

Thanks to technological advances, gene expression levels are among the most widely accessible type of 'large-scale' experimental data. While such datasets provide a global overview of

the level of expression of enzymes, deriving information on reaction fluxes from gene expression levels is hindered by the numerous biological processes intervening between them. Changes in rates of translation or mRNA and enzyme degradation may significantly modify the quantity of enzymes available from a given amount of transcript. In addition, changes in substrate/product concentrations or metabolic regulations can influence the reaction fluxes irrespective to the enzyme quantities. As a consequence, no simple correlations are necessarily observed between gene expression levels and reaction fluxes (Gygi *et al.*, 1999; ter Kuile & Westerhoff, 2001; Yang *et al.*, 2002; Akesson *et al.*, 2004).

Some approaches have nonetheless been developed to exploit information from gene expression data using models. In the vein of pathway- or module-based methods interpreting changes of gene expressions at the level of pathways or biological processes (Hanisch *et al.*, 2002; Draghici *et al.*, 2003; Yang *et al.*, 2004), methods relying on a graph representation of metabolism (Patil & Nielsen, 2005) or on a decomposition of metabolic models into elementary modes (Schwartz *et al.*, 2007) were introduced to correlate expression levels with possible metabolic states. These approaches are merely descriptive: the model provides a suitable metabolic context to interpret the experimental data. Gene expression data have also been used to refine the characterization of metabolic fluxes in models. For instance, by blocking reactions corresponding to unexpressed genes, metabolic fluxes could be characterized more precisely in a yeast model (Akesson *et al.*, 2004). In the same spirit, a method was recently introduced to evaluate the consistency of gene expression datasets with metabolic objectives, and identify subsets of active reactions that best correlate with expressed genes and metabolic objectives (Becker & Palsson, 2008). Even though these methods only rely on a limited dependency between gene expression level and reaction flux – reactions catalyzed by unexpressed genes should have low fluxes – they succeed in somewhat improving the characterization of metabolic states, or in assessing the consistency of the model with the experimental data.

As an attempt to account for transcriptional regulation, regulatory interactions were introduced in models by translating them into Boolean rules (Covert *et al.*, 2001). In such joint regulatory-metabolic model, Boolean variables qualitatively describe the transcription state of genes, including genes coding for enzymes and transcription factors, while Boolean rules determine their regulatory dependencies. Metabolic reactions are then allowed to have a nonzero flux only if the transcriptional state of their enzymes is *true*. Several methods have been developed to study these joint models. Regulatory FBA (rFBA) simulates time courses of gene expression states: at each time step, the new transcriptional state is computed from the metabolic state predicted at the previous time step, and is used to constrain FBA

prediction of the current metabolic state (Covert *et al.*, 2001). A specific representation scheme was later developed to encode the sequence of expression states predicted by rFBA in a unified manner, in order to compare regulatory responses across various environments (Barrett *et al.*, 2005). Another type of method has been recently developed to determine joint steady states of gene expression and metabolic fluxes. Examining these steady states contributed to the identification of redundantly expressed enzymes and the quantification of the effect of transcriptional regulation in determining flux activity in *E. coli* (Shlomi *et al.*, 2007a). Finally, two studies compared experimental expression levels with predicted expression states to assess the correctness of joint regulatory-metabolic models of *E. coli* and yeast (Covert *et al.*, 2004; Herrgård *et al.*, 2006b). A significant proportion of inconsistent expression states could be corrected in these models by searching for missing interactions (Covert *et al.*, 2004; Herrgård *et al.*, 2006b). In the same vein, a method was recently designed to automate the identification of experiments that are likely to bring most information on potentially missing regulatory interactions (Barrett & Palsson, 2006).

### Using genome-scale models for metabolic engineering

The use of microbial organisms for industrial purposes has grown considerably in the past few years, with potential applications ranging from the production of valuable metabolites to the degradation of pollutants and the generation of renewable energy (Janssen *et al.*, 2005; Ro *et al.*, 2006; Peng *et al.*, 2008; Rittmann, 2008). The field of metabolic engineering aims at designing and improving industrial microorganisms through the rational design of genetic manipulations leading to enhanced performance (Bailey, 1991; Stephanopoulos *et al.*, 1998). With the advent of genome-scale experimental technologies, the set of metabolic engineering methods is progressively expanding to include systems-wide analyses, enabling for instance to study the operation of regulatory and metabolic networks at large scale (Park *et al.*, 2008). In this respect, genome-scale metabolic models provide to engineers an effective toolbox to investigate the metabolic behavior of their strain of interest and target improvements (Kim *et al.*, 2008).

As a first class of applications, all analytical methods presented in the previous sections can be directly applied to engineering purposes. Such methods may help for instance to evaluate the maximum theoretical efficiencies of pathways or determine appropriate host strains by predicting their metabolic capabilities from their reconstructed models. More importantly, metabolic models can help in characterizing the actual metabolism operation of engineered strains, especially when experimental data have been

acquired on them. Metabolic Flux Analysis provides for instance quantitative values for intracellular fluxes, which may be used to determine the actual pathway utilization and pinpoint bottleneck reactions (Stephanopoulos *et al.*, 1998). Such information is of high significance for the metabolic engineers, as it may help them in designing further metabolic modifications.

Metabolic models also provide the ability to formulate hypotheses and evaluate *in silico* the potential of genetic modifications. A common cause of low production yields lies in the presence of pathways that divert fluxes to the production of undesirable byproducts or compete for the utilization of precursors and cofactors. While such pathways may be identified manually, their direct removal through gene deletion may cause side effects, for example alter the regeneration of cofactors, the redox balance, or the energy balance (Kim *et al.*, 2008). Genome-scale models can predict the effect of gene deletions on metabolic phenotypes. Several methods were designed with the aim of selecting those gene deletions that would provide the greatest benefit for a given metabolite production goal. Alper *et al.* (2005a) developed a procedure that sequentially screen the effect of single and multiple gene deletions in order to select those enabling the best product yields while maintaining sufficient growth rates. They successfully applied their method to enhance the yield of a lycopene producing *E. coli* strain (Alper *et al.*, 2005b). Screening *in silico* the high number of combinations of multiple gene deletions may turn out to be costly and practically impossible. Optimization methods based on genetic (Patil *et al.*, 2005) or linear programming (Burgard *et al.*, 2003) algorithms were introduced to circumvent this issue. The second optimization method, called OptKnock, specifically searches gene deletions coupling the production of a targeted metabolite with growth rate; the rationale being that improving the growth rate by adaptive evolution would jointly improve the metabolite production rate and that this coupling would make the engineered strain more evolutionary stable (Burgard *et al.*, 2003). Gene deletions proposed by this method were tested experimentally to enhance lactic acid production in an *E. coli* strain (Fong *et al.*, 2005). Adaptive evolution experiments performed on the engineered strains actually showed that lactic acid production was coupled to growth and achieved increased secretion rates of the product. In addition to gene deletions, metabolic models can explore the effect of adding new pathways, and help select the most appropriate ones. In this aim, the OptStrain method was designed to systematically suggest additions of reactions to produce novel metabolites (Pharkya *et al.*, 2004). OptStrain relies on a comprehensive database of biochemical reactions and may propose alternative solutions. A last set of methods consists in designing suitable up- or downregulations of metabolic enzymes. Intervening on gene expression levels is indeed a powerful

tool to tune metabolism operation, but the specific effects of such interventions are often hardly predictable (Kim *et al.*, 2008). In a study involving a L-threonine producing strain of *E. coli*, Lee *et al.* (2007) made use of its metabolic model to predict gene expression changes enhancing the strain yield. Specifically, they predicted flux values of key reactions leading to optimal L-threonine production and compared them with measured fluxes. They then used the relative difference between them to guide the tuning of the expression of the corresponding genes. A more systematic approach was introduced with the OptReg method, which identifies at genome-scale the relative changes of flux values with respect to the wild-type flux distribution that provide the best production yield (Pharkya & Maranas, 2006). Results of OptReg can be used to identify candidate enzymes for up- or downregulation.

Yet, two main issues limit the predictive capabilities of metabolic models. First, while regulation may play a central role in controlling the efficiency of product synthesis, it is completely overlooked in metabolic models. Studying regulatory interactions – using for instance models of regulatory networks – may actually provide useful insights, for example to remove feedback inhibitions or fine-tune transcriptional regulatory circuits commanding the product biosynthesis (Kim *et al.*, 2008). Not accounting for enzyme quantities but only reaction fluxes imposes a second limitation to genome-scale models. Implementing changes in flux values – suggested for instance by metabolic model optimization methods – by altering the quantity of enzymes is a difficult task, as enzyme kinetics and metabolite concentrations may significantly influence the flux change. In order to determine the effect of enzyme quantity changes on metabolic fluxes, more detailed approaches are required, for example metabolic control analysis (Fell, 1992).

## Resources, databases, and tools

At the time of this review, genome-scale models have been reconstructed for at least 17 bacteria (see Table 4). For all of them, extensive manual curation was required in order to integrate information from the literature on their biochemistry and physiology with functional information from genome annotation. These models are therefore of high quality on average, and mostly complete with respect to the current knowledge of their metabolism. An increasing subset is being assessed and corrected against large-scale experimental data (see Table 4), and an impressive array of analytical studies has been applied to the most popular ones, for example *E. coli* (Feist & Palsson, 2008).

Models used to be made available independently by their authors, under a variety of naming conventions and formats. This is a significant obstacle to their reusability, as significant effort is required to adapt them to modeling

software other than the ones they were constructed with. Differences in reaction and metabolite names also hamper direct comparisons between different models. Fortunately, some attempts to address these issues are under way. The general-purpose SBML format (Systems Biology Markup Language) (Hucka *et al.*, 2003) is often used to exchange constraint-based models, thus playing the role of a ‘default’ standard for models. While SBML can be imported by many modeling tools, it is not fully adapted to the specifics of models; this may result in information or functionality loss during exchange. In addition to providing a standard format, SBML supports the association of model components with external references, such as reaction and metabolite identifiers in universal metabolic databases, using MIRIAM annotations (Le Novère *et al.*, 2005). If widely used, this feature should facilitate model reuse and comparison.

In order to facilitate model reuse and comparison, dedicated model repositories have been developed. Perhaps the most widely adopted initiative of this type is the Biomodels.net repository (Le Novère *et al.*, 2006) which stores biochemical models of any type in SBML format. Because of its focus on more detailed dynamic models and the related generic format choice, the repository is not fully compatible with constraint-based models and qualitative predictions, as illustrated by the current low number of such models included. Agreements with several journals make it mandatory for authors to deposit models mentioned in their manuscripts in Biomodels.net, where they are checked for syntactic correctness. On some models, a more elaborate test on the compatibility between model predictions and results presented in the associated paper is also performed.

Currently, the only freely accessible (to academic users) repository dedicated to constraint-based models is the BiGG database (<http://bigg.ucsd.edu>). Its unified dictionary of metabolite and reaction names enables direct comparisons between its metabolic models.

Relatively few software tools have been specifically developed to handle genome-scale constraint-based models, compared with the number of tools developed for kinetic modeling. As the modeling framework relies primarily on linear algebra and linear programming, general purpose mathematical software platforms, for example MATLAB (<http://www.mathworks.com/>) and MATHEMATICA (<http://www.wolfram.com/>), or optimization modeling packages, for example GAMS (<http://www.gams.com/>), are well suited. Specialized optimization packages can be added for greater efficiency. In addition, modules dedicated to constraint-based modeling have been developed for MATLAB: FLUXANALYZER (Klamt *et al.*, 2007), the COBRA TOOLBOX (Becker *et al.*, 2007), or METATOOL (von Kamp & Schuster, 2006) for elementary mode analysis are good representatives. Libraries for importing SBML models within these programs are also provided by the SBML developer community (Bornstein

*et al.*, 2008). Among the software tools that are stand-alone, one should mention the SYSTEMS BIOLOGY RESEARCH TOOLBOX (Wright & Wagner, 2008), SCRUMPY (Poolman, 2006), METAFUXNET (Lee *et al.*, 2003), or FLUXEXPLORER (Luo *et al.*, 2006), each with their own specific strengths. Interestingly, very few programs focus or even support the model reconstruction process by providing the analytical capabilities for consistency checks: the commercial SYMPHENY platform (<http://www.genomatica.com/>) associates a metabolic database with several analytical methods, while YANASQUARE (Schwarz *et al.*, 2007) facilitates the reconstruction of models from KEGG and performs selected structural analyses (e.g. elementary modes). Very recently, web-based tools have been released to enable on-line analyses on specific metabolic models (Beste *et al.*, 2007; Durot *et al.*, 2008). Given the need for faster and better reconstruction, we expect more progress in that direction.

### Concluding remarks and future directions

Constraint-based genome-scale metabolic models can be viewed as 'systems-level' analytical layers which enable computation and reasoning on the consequences of the accumulated knowledge on the biochemistry encoded in a given genome, and confrontation of that knowledge with the known physiology of the corresponding species or with additional experimental evidence. These models thus bridge the gap between genotype and phenotype and enable a wide spectrum of analyses and *in silico* experiments, providing a solid foundation for systems analyses and metabolic engineering.

The systematic and automated reconstruction of genome-scale models from genomes and additional high-throughput data may seem like a natural extension of genome annotation (Reed *et al.*, 2006a), but remains beyond the reach of current methods. While genome-scale models can be reconstructed using only sequence and qualitative functional information, gaining the additional predictive and analytical power of models still requires significant effort and expertise. Genome annotations must first be translated into a network, which must then be turned into a model with the help of additional information, and systematically checked with respect to biochemical consistency rules and experimental observations. Only after a model is complete enough to enable meaningful predictions at the phenotypic level can it be used to predict phenotypes or other properties beyond those that can be immediately verified.

Obstacles to automating this process include technical difficulties in translating annotations into proper biochemical activities, and also the fact that methods for model refinement have been designed and applied separately for each type of experimental data. There is increasing pressure for this situation to evolve, however, as the boost in the throughput of experimental techniques and the advent of

'multi-omics' datasets (Ishii *et al.*, 2007a) promises a wealth of information that will be exploitable only by computer-assisted interpretation, with the help of models. At the same time, the field of metabolic modeling is now approaching the level of maturity necessary for several data integration methods to be used together as components in integrated model reconstruction and refinement strategies.

Significant benefits could result from the availability of a wider spectrum of bacterial metabolic models. They would provide an integrated view of metabolic pathways across the tree of life, thereby enabling so-called transverse approaches to annotation, and a variety of comparative metabolic analysis. To that end, the notion of pathway – defined unambiguously as the conversion between specified sets of input compounds (reactants) and output compounds (products) – can bring a useful decomposition of metabolism into basic biochemical functional units, in the spirit pioneered by SEED (Overbeek *et al.*, 2005), KEGG Modules (Kanehisa *et al.*, 2007), or MetaCyc (Caspi *et al.*, 2006). The field of bacterial evolution is poised to benefit as well: for instance, the availability of models for several bacteria along the phylogenetic tree would allow more comprehensive studies on the constraints implied by bacteria's metabolic capabilities and their evolution. While this type of study has been pioneered with a few selected models (Pál *et al.*, 2005a, 2006), working with a larger set of models will undoubtedly bring different insights (see (Kreimer *et al.*, 2008) for an example with networks). Modeling can also help in studying bacterial communities, as chemical interactions occurring between bacteria often need to be understood within the context of their metabolisms. Indeed, models have already been reconstructed and analyzed for small communities (Stolyar *et al.*, 2007); progress on that front may prove very useful in studying metabolic interactions in more complex communities, assisting in the functional interpretation of metagenome sequences. Last but not least, metabolic engineering applications would clearly benefit from the availability of a large set of bacterial models, as these would constitute a repository of characterized metabolic pathways, facilitating the combinatorial design of new catalytic systems, providing solid bases to test hypothetical genetic constructions, and helping with the selection of relevant strains for specific engineering objectives.

### Acknowledgements

We would like to thank the two anonymous reviewers for their numerous suggestions, which helped improve the manuscript. We are grateful for the support of the European Networks of Excellence BIOSAPIENS (contract no. LSHG-CT-2003-503265) and ENFIN (contract no. LSHG-CT-2005-518254).

## Statement

Re-use of this article is permitted in accordance with the Creative Commons Deed, Attribution 2.5, which does not permit commercial exploitation.

## References

- Aghaie A, Lechaplais C, Sirven P *et al.* (2008) New insights into the alternative D-glucarate degradation pathway. *J Biol Chem* **283**: 15638–15646.
- Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N & Mekalanos JJ (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *P Natl Acad Sci USA* **99**: 966–971.
- Akesson M, Förster J & Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. *Metab Eng* **6**: 285–293.
- Almaas E, Kovacs B, Vicsek T, Oltvai ZN & Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**: 839–843.
- Alper H, Jin Y-S, Moxley JF & Stephanopoulos G (2005a) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng* **7**: 155–164.
- Alper H, Miyaoku K & Stephanopoulos G (2005b) Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* **23**: 612–616.
- Antoniewicz MR, Kelleher JK & Stephanopoulos G (2007a) Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab Eng* **9**: 68–86.
- Antoniewicz MR, Kraynie DF, Laffend LA, González-Lergier J, Kelleher JK & Stephanopoulos G (2007b) Metabolic flux analysis in a nonstationary system: fed-batch fermentation of a high yielding strain of *E. coli* producing 1,3-propanediol. *Metab Eng* **9**: 277–292.
- Apweiler R, Attwood TK, Bairoch A *et al.* (2000) InterPro – an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**: 1145–1150.
- Arakawa K, Yamada Y, Shinoda K, Nakayama Y & Tomita M (2006) GEM system: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* **7**: 168.
- Arita M (2000) Metabolic reconstruction using shortest paths. *Simulat Pract Theory* **8**: 109–125.
- Arita M (2003) *In silico* atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. *Genome Res* **13**: 2455–2466.
- Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* **25**: 25–29.
- Baart G, Zomer B, de Haan A, van der Pol L, Beuvery EC, Tramper J & Martens D (2007) Modeling *Neisseria meningitidis* metabolism: from genome to metabolic fluxes. *Genome Biol* **8**: R136.
- Baba T, Ara T, Hasegawa M *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**: 2006.0008.
- Bailey JE (1991) Toward a science of metabolic engineering. *Science* **252**: 1668–1675.
- Barrett CL & Palsson BO (2006) Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. *PLoS Comput Biol* **2**: e52.
- Barrett CL, Herring CD, Reed JL & Palsson BO (2005) The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *P Natl Acad Sci USA* **102**: 19103–19108.
- Barthelme J, Ebeling C, Chang A, Schomburg I & Schomburg D (2007) BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. *Nucleic Acids Res* **35**: D511–D514.
- Beard DA, Liang S-D & Qian H (2002) Energy balance for analysis of complex metabolic networks. *Biophys J* **83**: 79–86.
- Beard DA, Babson E, Curtis E & Qian H (2004) Thermodynamic constraints for biochemical networks. *J Theor Biol* **228**: 327–333.
- Becker SA & Palsson BO (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol* **5**: 8.
- Becker SA & Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol* **4**: e1000082.
- Becker SA, Price ND & Palsson BØ (2006) Metabolite coupling in genome-scale metabolic networks. *BMC Bioinformatics* **7**: 111.
- Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ & Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat Protoc* **2**: 727–738.
- Besemer J, Lomsadze A & Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607–2618.
- Beste D, Hooper T, Stewart G *et al.* (2007) GSMN-TB: a web-based genome scale network model of *Mycobacterium tuberculosis* metabolism. *Genome Biol* **8**: R89.
- Bochner BR, Gadzinski P & Panomitros E (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res* **11**: 1246–1255.
- Bocs S, Cruveiller S, Vallenet D, Nuel G & Médigue C (2003) AMIGene: Annotation of microbial genes. *Nucleic Acids Res* **31**: 3723–3726.
- Bornstein BJ, Keating SM, Jouraku A & Hucka M (2008) LibSBML: an API library for SBML. *Bioinformatics* **24**: 880–881.
- Borodina I, Krabben P & Nielsen J (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res* **15**: 820–829.
- Boutet E, Lieberherr D, Tognolli M, Schneider M & Bairoch A (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol Biol* **406**: 89–112.

- Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO & Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**: R35.
- Boyer F & Viari A (2003) *Ab initio* reconstruction of metabolic pathways. *Bioinformatics* **19**(suppl 2): ii26–ii34.
- Braunstein A, Mulet R & Pagnani A (2008) Estimating the size of the solution space of metabolic networks. *BMC Bioinformatics* **9**: 240.
- Breitling R, Ritchie S, Goodenow D, Stewart ML & Barrett MP (2006) *Ab initio* prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics* **2**: 155–164.
- Burgard AP & Maranas CD (2003) Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng* **82**: 670–677.
- Burgard AP, Pharkya P & Maranas CD (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* **84**: 647–657.
- Burgard AP, Nikolaev EV, Schilling CH & Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* **14**: 301–312.
- Carpenter AE & Sabatini DM (2004) Systematic genome-wide screens of gene function. *Nat Rev Genet* **5**: 11–22.
- Caspi R, Foerster H, Fulcher CA *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **34**: D511–D516.
- Chen L & Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol* **7**: R17.
- Claudel-Renard C, Chevalet C, Faraut T & Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**: 6633–6639.
- Covert MW, Schilling CH & Palsson B (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **213**: 73–88.
- Covert MW, Knight EM, Reed JL, Herrgard MJ & Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**: 92–96.
- Covert MW, Xiao N, Chen TJ & Karr JR (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* **24**: 2044–2050.
- de Berardinis V, Vallenet D, Castelli V *et al.* (2008) A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Mol Syst Biol* **4**: 174.
- DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M & Best A (2007) Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* **8**: 139.
- Delcher AL, Harmon D, Kasif S, White O & Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**: 4636–4641.
- Dettmer K, Aronov PA & Hammock BD (2007) Mass spectrometry-based metabolomics. *Mass Spectrom Rev* **26**: 51–78.
- Deutscher D, Meilijson I, Kupiec M & Ruppin E (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet* **38**: 993–998.
- Deutscher D, Meilijson I, Schuster S & Ruppin E (2008) Can single knockouts accurately single out gene functions? *BMC Syst Biol* **2**: 50.
- Draghici S, Khatri P, Martins RP, Ostermeier GC & Krawetz SA (2003) Global functional profiling of gene expression. *Genomics* **81**: 98–104.
- Duarte NC, Herrgard MJ & Palsson BO (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14**: 1298–1309.
- Dunn WB, Bailey NJC & Johnson HE (2005) Measuring the metabolome: current analytical technologies. *Analyst* **130**: 606–625.
- Durot M, Le Fevre F, de Berardinis V *et al.* (2008) Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst Biol* **2**: 85.
- Ebenhöh O, Handorf T & Heinrich R (2004) Structural analysis of expanding metabolic networks. *Genome Infor* **15**: 35–45.
- Edwards JS & Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *P Natl Acad Sci USA* **97**: 5528–5533.
- Edwards JS, Ibarra RU & Palsson BO (2001) *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* **19**: 125–130.
- Edwards JS, Ramakrishna R & Palsson BO (2002) Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng* **77**: 27–36.
- Ellis LBM, Roe D & Wackett LP (2006) The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res* **34**: D517–D521.
- Ellis LBM, Gao J, Fenner K & Wackett LP (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res* **36**: W427–W432.
- Famili I, Mahadevan R & Palsson BO (2005) k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophys J* **88**: 1616–1625.
- Feist AM & Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* **26**: 659–667.
- Feist AM, Scholten JCM, Palsson BØ, Brockman FJ & Ideker T (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol Syst Biol* **2**: 2006.0004.
- Feist AM, Henry CS, Reed JL *et al.* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3**: 121.
- Fell DA (1992) Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J* **286**: 313–330.

- Fischer E, Zamboni N & Sauer U (2004) High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived  $^{13}\text{C}$  constraints. *Anal Biochem* **325**: 308–316.
- Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD & Palsson BO (2005) *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* **91**: 643–648.
- Fuhrer T, Chen L, Sauer U & Vitkup D (2007) Computational prediction and experimental verification of the gene encoding the NAD<sup>+</sup>/NADP<sup>+</sup>-dependent succinate semialdehyde dehydrogenase in *Escherichia coli*. *J Bacteriol* **189**: 8073–8078.
- Gasteiger J (2005) Chemoinformatics: a new field with a long tradition. *Anal Bioanal Chem* **384**: 57–64.
- Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R & Osterman A (2006) Essential genes on metabolic maps. *Curr Opin Biotech* **17**: 448–456.
- Gevorgyan A, Poolman MG & Fell DA (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics* **24**: 2245–2251.
- Green ML & Karp PD (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* **5**: 76.
- Gygi SP, Rochon Y, Franza BR & Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**: 1720–1730.
- Hansch D, Zien A, Zimmer R & Lengauer T (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics* **18**(suppl 1): S145–S154.
- Hattori M, Okuno Y, Goto S & Kanehisa M (2003) Heuristics for chemical compound matching. *Genome Infor* **14**: 144–153.
- Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD & Broadbelt LJ (2005) Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**: 1603–1609.
- 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**: 850–864.
- Henry CS, Broadbelt LJ & Hatzimanikatis V (2007) Thermodynamics-based metabolic flux analysis. *Biophys J* **92**: 1792–1805.
- Herrgård MJ, Fong SS & Palsson BØ (2006a) Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* **2**: e72.
- Herrgård MJ, Lee B-S, Portnoy V & Palsson BØ (2006b) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res* **16**: 627–635.
- Hong SH, Kim JS, Lee SY *et al.* (2004) The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat Biotechnol* **22**: 1275–1281.
- Huang M, Oppermann-Sanio FB & Steinbüchel A (1999) Biochemical and molecular characterization of the *Bacillus subtilis* acetoin catabolic pathway. *J Bacteriol* **181**: 3837–3841.
- Hucka M, Finney A, Sauro HM *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**: 524–531.
- Ibarra RU, Edwards JS & Palsson BO (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**: 186–189.
- Imielinski M & Belta C (2008) Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Syst Biol* **2**: 40.
- Imielinski M, Belta C, Halasz A & Rubin H (2005) Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities. *Bioinformatics* **21**: 2008–2016.
- Imielinski M, Belta C, Rubin H & Halász A (2006) Systematic analysis of conservation relations in *Escherichia coli* genome-scale metabolic network reveals novel growth media. *Biophys J* **90**: 2659–2672.
- Ishii N, Nakahigashi K, Baba T *et al.* (2007a) Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* **316**: 593–597.
- Ishii N, Nakayama Y & Tomita M (2007b) Distinguishing enzymes using metabolome data for the hybrid dynamic/static method. *Theor Biol Med Model* **4**: 19.
- Jacobs MA, Alwood A, Thaipisuttikul I *et al.* (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *P Natl Acad Sci USA* **100**: 14339–14344.
- Jamshidi N & Palsson B (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. *BMC Syst Biol* **1**: 26.
- Jamshidi N & Palsson BØ (2008) Formulating genome-scale kinetic models in the post-genome era. *Mol Syst Biol* **4**: 171.
- Janssen DB, Dinkla IJT, Poelarends GJ & Terpstra P (2005) Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ Microbiol* **7**: 1868–1882.
- Joyce AR & Palsson BØ (2006) The model organism as a system: integrating ‘omics’ data sets. *Nat Rev Mol Cell Bio* **7**: 198–210.
- Joyce AR, Reed JL, White A *et al.* (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J Bacteriol* **188**: 8259–8271.
- Kanehisa M, Araki M, Goto S *et al.* (2007) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**: D480–D484.
- Karp PD, Paley S & Romero P (2002) The pathway tools software. *Bioinformatics* **18**(suppl 1): S225–S232.
- Karp PD, Keseler IM, Shearer A *et al.* (2007) Multidimensional annotation of the *Escherichia coli* K-12 genome. *Nucleic Acids Res* **35**: 7577–7590.
- Kharchenko P, Chen L, Freund Y, Vitkup D & Church GM (2006) Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* **7**: 177.

- Kim HU, Kim TY & Lee SY (2008) Metabolic flux analysis and metabolic engineering of microorganisms. *Mol Biosyst* **4**: 113–120.
- Kim P-J, Lee D-Y, Kim TY, Lee KH, Jeong H, Lee SY & Park S (2007) Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *P Natl Acad Sci USA* **104**: 13638–13642.
- Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H & Mori H (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res* **12**: 291–299.
- Klamt S & Gilles ED (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics* **20**: 226–234.
- Klamt S & Stelling J (2003) Two approaches for metabolic pathway analysis? *Trends Biotechnol* **21**: 64–69.
- Klamt S, Saez-Rodriguez J & Gilles ED (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol* **1**: 2.
- Klopman G, Dimayuga M & Talafous J (1994) META. 1. A program for the evaluation of metabolic transformation of chemicals. *J Chem Inf Comput Sci* **34**: 1320–1325.
- Kobayashi K, Ehrlich SD, Albertini A *et al.* (2003) Essential *Bacillus subtilis* genes. *P Natl Acad Sci USA* **100**: 4678–4683.
- Kreimer A, Borenstein E, Gophna U & Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. *P Natl Acad Sci USA* **105**: 6976–6981.
- Kuchel PW & Philp DJ (2008) Isotopomer subspaces as indicators of metabolic-pathway structure. *J Theor Biol* **252**: 391–401.
- Kuepfer L, Sauer U & Blank LM (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res* **15**: 1421–1430.
- Kumar VS, Dasika MS & Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8**: 212.
- Kümmel A, Panke S & Heinemann M (2006a) Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* **7**: 512.
- Kümmel A, Panke S & Heinemann M (2006b) Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol* **2**: 2006.0034.
- Lee J, Yun H, Feist A, Palsson B & Lee S (2008a) Genome-scale reconstruction and *in silico* analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. *Appl Microbiol Biot* **80**: 849–862.
- Lee JM, Gianchandani EP, Eddy JA & Papin JA (2008b) Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol* **4**: e1000086.
- Lee KH, Park JH, Kim TY, Kim HU & Lee SY (2007) Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol* **3**: 149.
- Lee SY, Lee D-Y, Hong SH, Kim TY, Yun H, Oh Y-G & Park S (2003) MetaFluxNet, a program package for metabolic pathway construction and analysis, and its use in large-scale metabolic flux analysis of *Escherichia coli*. *Genome Inform* **14**: 23–33.
- Le Novère N, Finney A, Hucka M *et al.* (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23**: 1509–1515.
- Le Novère N, Bornstein B, Broicher A *et al.* (2006) BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res* **34**: D689–D691.
- Liberati NT, Urbach JM, Miyata S *et al.* (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *P Natl Acad Sci USA* **103**: 2833–2838.
- Luo R, Liao S, Zeng S, Li Y & Luo Q (2006) FluxExplorer: a general platform for modeling and analyses of metabolic networks based on stoichiometry. *Chin Sci Bull* **51**: 689–696.
- Ma H & Zeng A-P (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**: 270–277.
- Mahadevan R & Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**: 264–276.
- Mahadevan R, Bond DR, Butler JE *et al.* (2006) Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl Environ Microb* **72**: 1558–1568.
- Médigue C & Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol* **158**: 724–736.
- Meyer P & Dworkin J (2007) Applications of fluorescence microscopy to single bacterial cells. *Res Microbiol* **158**: 187–194.
- Nikolaev EV, Burgard AP & Maranas CD (2005) Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophys J* **88**: 37–49.
- Nogales J, Palsson BO & Thiele I (2008) A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. *BMC Syst Biol* **2**: 79.
- Notebaart RA, van Enckevort FHJ, Francke C, Siezen RJ & Teusink B (2006) Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* **7**: 296.
- Notebaart RA, Teusink B, Siezen RJ & Papp B (2008) Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol* **4**: e26.
- Oberhardt MA, Puchałka J, Fryer KE, dos Santos VAPM & Papin JA (2008) Genome-scale metabolic network analysis of the opportunistic pathogen *Pseudomonas aeruginosa* PAO1. *J Bacteriol* **190**: 2790–2803.
- Oh Y-K, Palsson BO, 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**: 28791–28799.
- Oliveira AP, Nielsen J & Förster J (2005) Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiology* **5**: 39.



- Osterman A & Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. *Curr Opin Chem Biol* **7**: 238–251.
- Overbeek R, Begley T, Butler RM *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–5702.
- Pál C, Papp B & Lercher MJ (2005a) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37**: 1372–1375.
- Pál C, Papp B & Lercher MJ (2005b) Horizontal gene transfer depends on gene content of the host. *Bioinformatics* **21** (suppl 2): 222–ii223.
- Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG & Hurst LD (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**: 667–670.
- Palsson BO (2006) *Systems Biology. Properties of Reconstructed Networks*. Cambridge University Press, New York, NY, USA.
- Papin JA, Price ND, Edwards JS & Palsson BO (2002) The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J Theor Biol* **215**: 67–82.
- Papin JA, Price ND, Wiback SJ, Fell DA & Palsson BO (2003) Metabolic pathways in the post-genome era. *Trends Biochem Sci* **28**: 250–258.
- Papin JA, Stelling J, Price ND, Klamt S, Schuster S & Palsson BO (2004) Comparison of network-based pathway analysis methods. *Trends Biotechnol* **22**: 400–405.
- Papp B, Pál C & Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**: 661–664.
- Park JH, Lee SY, Kim TY & Kim HU (2008) Application of systems biology for bioprocess development. *Trends Biotechnol* **26**: 404–412.
- Patil KR & Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *P Natl Acad Sci USA* **102**: 2685–2689.
- Patil KR, Akesson M & Nielsen J (2004) Use of genome-scale microbial models for metabolic engineering. *Curr Opin Biotech* **15**: 64–69.
- Patil KR, Rocha I, Förster J & Nielsen J (2005) Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics* **6**: 308.
- Peng R-H, Xiong A-S, Xue Y *et al.* (2008) Microbial biodegradation of polyaromatic hydrocarbons. *FEMS Microbiol Rev* **32**: 927–955.
- Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, Montero F & Schuster S (1999) METATOOL: for studying metabolic networks. *Bioinformatics* **15**: 251–257.
- Pharkya P & Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng* **8**: 1–13.
- Pharkya P, Burgard AP & Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* **14**: 2367–2376.
- Poolman MG (2006) ScrumPy: metabolic modelling with Python. *Syst Biol (Stevenage)* **153**: 375–378.
- Price ND, Reed JL & Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2**: 886–897.
- Raymond JW, Gardiner EJ & Willett P (2002) Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J Chem Inf Comput Sci* **42**: 305–316.
- Reed JL & Palsson BO (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J Bacteriol* **185**: 2692–2699.
- Reed JL & Palsson BO (2004) Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res* **14**: 1797–1805.
- Reed JL, Vo TD, Schilling CH & Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* **4**: R54.
- Reed JL, Famili I, Thiele I & Palsson BO (2006a) Towards multidimensional genome annotation. *Nat Rev Genet* **7**: 130–141.
- Reed JL, Patel TR, Chen KH *et al.* (2006b) Systems approach to refining genome annotation. *P Natl Acad Sci USA* **103**: 17480–17484.
- Ren Q, Kang KH & Paulsen IT (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* **32**: D284–D288.
- Resendis-Antonio O, Reed JL, Encarnación S, Collado-Vides J & Palsson BO (2007) Metabolic reconstruction and modeling of nitrogen fixation in *Rhizobium etli*. *PLoS Comput Biol* **3**: e192.
- Risso C, Van Dien SJ, Orloff A, Lovley DR & Coppi MV (2008) Elucidation of an alternate isoleucine biosynthesis pathway in *Geobacter sulfurreducens*. *J Bacteriol* **190**: 2266–2274.
- Rittmann BE (2008) Opportunities for renewable bioenergy using microorganisms. *Biotechnol Bioeng* **100**: 203–212.
- Ro D-K, Paradise EM, Ouellet M *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**: 940–943.
- Saghatelian A, Trauger SA, Want EJ, Hawkins EG, Siuzdak G & Cravatt BF (2004) Assignment of endogenous substrates to enzymes by global metabolite profiling. *Biochemistry* **43**: 14332–14339.
- Saito N, Robert M, Kitamura S *et al.* (2006) Metabolomics approach for enzyme discovery. *J Proteome Res* **5**: 1979–1987.
- Sauer U (2006) Metabolic networks in motion: <sup>13</sup>C-based flux analysis. *Mol Syst Biol* **2**: 62.
- Schilling CH & Palsson BO (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J Theor Biol* **203**: 249–283.
- Schilling CH, Edwards JS, Letscher D & Palsson BO (2000) Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng* **71**: 286–306.

- Schilling CH, Covert MW, Famili I, Church GM, Edwards JS & Palsson BO (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J Bacteriol* **184**: 4582–4593.
- Schneider G & Fechner U (2004) Advances in the prediction of protein targeting signals. *Proteomics* **4**: 1571–1580.
- Schuetz R, Kuepfer L & Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* **3**: 119.
- Schuster S, Fell DA & Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* **18**: 326–332.
- Schuster S, Pfeiffer T & Fell DA (2008) Is maximization of molar yield in metabolic networks favoured by evolution? *J Theor Biol* **252**: 497–504.
- Schwartz J-M, Gauguain C, Nacher JC, de Daruvar A & Kanehisa M (2007) Observing metabolic functions at the genome scale. *Genome Biol* **8**: R123.
- Schwarz R, Liang C, Kaleta C *et al.* (2007) Integrated network reconstruction, visualization and analysis using YANASquare. *BMC Bioinformatics* **8**: 313.
- Segrè D, Vitkup D & Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *P Natl Acad Sci USA* **99**: 15112–15117.
- Segrè D, Zucker J, Katz J *et al.* (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *OMICS* **7**: 301–316.
- Senger RS & Papoutsakis ET (2008) Genome-scale model for *Clostridium acetobutylicum*: part I. Metabolic network resolution and analysis. *Biotechnol Bioeng* **101**: 1036–1052.
- Serres MH, Goswami S & Riley M (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res* **32**: D300–D302.
- Shlomi T, Berkman O & Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *P Natl Acad Sci USA* **102**: 7695–7700.
- Shlomi T, Eisenberg Y, Sharan R & Ruppin E (2007a) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* **3**: 101.
- Shlomi T, Herrgard M, Portnoy V, Naim E, Palsson BØ, Sharan R & Ruppin E (2007b) Systematic condition-dependent annotation of metabolic genes. *Genome Res* **17**: 1626–1633.
- Smallbone K, Simeonidis E, Broomhead DS & Kell DB (2007) Something from nothing – bridging the gap between constraint-based and kinetic modelling. *FEBS J* **274**: 5576–5585.
- Stelling J, Klamt S, Bettenbrock K, Schuster S & Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**: 190–193.
- Stephanopoulos GN, Aristidou AA & Nielsen J (1998) *Metabolic Engineering. Principles and Methodologies*. Academic Press, Elsevier Science, San Diego, CA, USA.
- Steuer R (2006) Review: on the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform* **7**: 151–158.
- Stolyar S, Van Dien SJ, Hillesland KL, Pinel N, Lie TJ, Leigh JA & Stahl DA (2007) Metabolic modeling of a mutualistic microbial community. *Mol Syst Biol* **3**: 92.
- Stouthamer AH & Bettenhausen C (1973) Utilization of energy for growth and maintenance in continuous and batch cultures of microorganisms. A reevaluation of the method for the determination of ATP production by measuring molar growth yields. *Biochim Biophys Acta* **301**: 53–70.
- Sun J & Zeng A-P (2004) IdentiCS – identification of coding sequence and *in silico* reconstruction of the metabolic network directly from unannotated low-coverage bacterial genome sequence. *BMC Bioinformatics* **5**: 112.
- Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien SJ, Keasling JD & Maranas CD (2007) Metabolic flux elucidation for large-scale models using <sup>13</sup>C labeled isotopes. *Metab Eng* **9**: 387–405.
- Suzuki N, Okai N, Nonaka H, Tsuge Y, Inui M & Yukawa H (2006) High-throughput transposon mutagenesis of *Corynebacterium glutamicum* and construction of a single-gene disruptant mutant library. *Appl Environ Microb* **72**: 3750–3755.
- ter Kuile BH & Westerhoff HV (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett* **500**: 169–171.
- Terzer M & Stelling J (2008) Large scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* **24**: 2229–2235.
- Teusink B, van Enckevort FHJ, Francke C, Wiersma A, Wegkamp A, Smid EJ & Siezen RJ (2005) *In silico* reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. *Appl Environ Microb* **71**: 7253–7262.
- Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ & Smid EJ (2006) Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *J Biol Chem* **281**: 40041–40048.
- Thiele I, Vo TD, Price ND & Palsson BØ (2005) Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an *in silico* genome-scale characterization of single- and double-deletion mutants. *J Bacteriol* **187**: 5818–5830.
- UniProt C (2008) The universal protein resource (UniProt). *Nucleic Acids Res* **36**: D190–D195.
- Van Dien SJ & Lidstrom ME (2002) Stoichiometric model for evaluating the metabolic capabilities of the facultative methylophilic *Methylobacterium extorquens* AM1, with application to reconstruction of C(3) and C(4) metabolism. *Biotechnol Bioeng* **78**: 296–312.
- van Winden WA, Heijnen JJ, Verheijen PJ & Grievink J (2001) A priori analysis of metabolic flux identifiability from (<sup>13</sup>C)-labeling data. *Biotechnol Bioeng* **74**: 505–516.
- Varma A & Palsson BO (1994a) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microb* **60**: 3724–3731.

- Varma A & Palsson BO (1994b) Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* **12**: 994–998.
- von Kamp A & Schuster S (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics* **22**: 1930–1931.
- von Mering C, Jensen LJ, Kuhn M *et al.* (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35**: D358–D362.
- Wiback SJ, Famili I, Greenberg HJ & Palsson BO (2004) Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *J Theor Biol* **228**: 437–447.
- Wiechert W (2001) <sup>13</sup>C metabolic flux analysis. *Metab Eng* **3**: 195–206.
- Wiechert W, Möllney M, Isermann N, Wurzel M & de Graaf AA (1999) Bidirectional reaction steps in metabolic networks: III. Explicit solution and analysis of isotopomer labeling systems. *Biotechnol Bioeng* **66**: 69–85.
- Wright J & Wagner A (2008) The systems biology research tool: evolvable open-source software. *BMC Syst Biol* **2**: 55.
- Yang C, Hua Q & Shimizu K (2002) Integration of the information from gene expression and metabolic fluxes for the analysis of the regulatory mechanisms in *Synechocystis*. *Appl Microbiol Biot* **58**: 813–822.
- Yang F, Qian H & Beard DA (2005) Ab initio prediction of thermodynamically feasible reaction directions from biochemical network stoichiometry. *Metab Eng* **7**: 251–259.
- Yang HH, Hu Y, Buetow KH & Lee MP (2004) A computational approach to measuring coherence of gene expression in pathways. *Genomics* **84**: 211–217.
- Yeung M, Thiele I & Palsson B (2007) Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* **8**: 363.
- Yugi K, Nakayama Y, Kinoshita A & Tomita M (2005) Hybrid dynamic/static method for large-scale simulation of metabolism. *Theor Biol Med Model* **2**: 42.
- Zamboni N, Kümmel A & Heinemann M (2008) anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. *BMC Bioinformatics* **9**: 199.
- Zupke C & Stephanopoulos G (1994) Modeling of isotope distribution and intracellular fluxes in metabolic networks using atom mapping matrices. *Biotechnol Prog* **10**: 489–498.