Title: Enhancing metabolic models with genome-scale experimental data

Terms for subject index are **bold**

Keywords:

Genome-scale modeling, constraint-based metabolic modeling, flux balance analysis, genome-scale data, transcriptomics, proteomics, metabolomics, shadow prices, machine learning

Enhancing metabolic models with genome-scale experimental data

Kristian Jensen*, Steinn Gudmundsson** & Markus Herrgard*

*The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark

**Center for Systems Biology, University of Iceland

Abstract

Genome-scale metabolic reconstructions have found widespread use in scientific research as structured representations of knowledge about an organism's metabolism and as starting points for metabolic simulations. With few simplifying assumptions, genome-scale models of metabolism can be used to estimate intracellular reaction rates. However, with the rapid increase in the availability of genome-scale data, there is ample opportunity to refine the predictions made by metabolic models, by integrating experimental data. In this chapter, we review different methods for combining genome-scale metabolic models with genome-scale experimental data, such as transcriptomics, proteomics and metabolomics. Integrating experimental data into the models generally results in more precise and accurate simulations of cellular metabolism.

1. Reconstruction and analysis of metabolic networks

To describe and understand the functioning of living cells, it is essential to study metabolism. The chemical conversion of nutrients into energy, biomass and secondary products is one of the main components of the cellular phenotype, and a defining characteristic of life. Since the metabolic capabilities of an organism are ultimately determined by its genotype, advances in genome sequencing technologies during the last two decades have had a substantial impact on our knowledge about metabolism. With a fully annotated whole genome sequence of an organism, it is feasible to compile a database of all the biochemical reactions that can be catalyzed inside the cell. Besides a list of reactions and their stoichiometries, such a database, called a **genome-scale metabolic reconstruction**, often includes information that links each reaction to the genes encoding the enzymes that catalyze it (Price, Reed and Palsson, 2004). The

earliest published genome-scale reconstructions were for organisms with small genomes such as *Haemophilus influenzae* (Schilling and Palsson, 2000) and *Escherichia coli* (Edwards and Palsson, 2000), but reconstructions for more complex organisms including *Saccharomyces cerevisiae* (Förster *et al.*, 2003), *Arabidopsis thaliana* (de Oliveira Dal'Molin *et al.*, 2010) and *Homo sapiens* (Duarte *et al.*, 2007) have followed since. Revised versions of genome-scale metabolic reconstructions are sometimes published when new genes are discovered or annotated functions of known genes are updated.

A genome-scale metabolic reconstruction allows systematic analysis of the metabolic network of an organism, and can even form a starting point for whole-cell simulations (Orth, Thiele and Palsson, 2010; Karr *et al.*, 2012). In order to perform such analyses, the genome-scale reconstruction must be formulated as a mathematical model, e.g. in the form of a system of differential equations,

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{S} \cdot \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{k}) \tag{1}$$

Here $\boldsymbol{S}$ denotes the stoichiometric matrix, derived from the genome-scale reconstruction with element s_{ij} denoting the stoichiometric coefficient of metabolite i in reaction j, and $\boldsymbol{x}$ is a vector of concentrations of all metabolites in the cell. Reaction rates, $\boldsymbol{v}$, are a function of current metabolite concentrations and kinetic parameters, $\boldsymbol{k}$. Given initial metabolite concentrations, the system of differential equations is readily solved numerically. While the formulation is conceptually simple, its use on the genome-scale has been impeded by limited knowledge of the many kinetic parameters (McCloskey, Palsson and Feist, 2013).

To avoid the issue of unknown kinetic parameters, **constraint-based metabolic modeling** methods are often used instead. Constraint-based modeling imposes constraints on the system and finds metabolic reaction rates that are consistent with these constraints. The most central constraint is the assumption of steady-state, where the concentrations of internal metabolites are assumed to be constant. This corresponds to setting the left-hand side of Equation 1 to zero and results in a system of linear equations,

$$\boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0} \tag{2}$$

that can be solved for the reaction rates (also known as fluxes), $\boldsymbol{v}$ (Orth, Thiele and Palsson, 2010). The kinetic parameters are not accounted for explicitly in constraint-based models, which only require the stoichiometric matrix to be known. For most genome-scale reconstructions, the system of equations is underdetermined, meaning

that an infinite number of flux solutions exist. One way to address this issue is to identify a solution that optimizes a specific objective. This is based on an assumption that the cell has evolved to maximize some biological objective, e.g. production of ATP or production of biomass. Production of biomass is modeled through a bulk-reaction that consumes biomass constituents such as nucleotides and amino acids in empirically determined ratios (Orth, Thiele and Palsson, 2010). This method is known as **flux balance analysis** (FBA) and has become the foundation of most work in constraint-based metabolic modeling. Performing flux balance analysis requires the solution of a linear optimization problem. The result is a set of reaction rates that satisfy the constraints of the system and is consistent with the defined biological objective.

Despite the simple formulation and strong assumptions, FBA has proven useful in a number of metabolic modeling applications, to predict the rates of metabolic reactions, typically called the **flux distribution** (McCloskey, Palsson and Feist, 2013). It can been used for instance to predict essential metabolic genes, i.e. genes that are required for the synthesis of one or more biomass constituents. This is done by simply removing corresponding reactions from the model and performing FBA. If the maximal biomass flux is zero in the knockout model, the gene is expected to be essential. Comparisons with experimental data from single-knockout studies have shown good correspondence with the results of FBA-based essentiality predictions in *E. coli* and other bacteria such as *Pseudomonas aeruginosa* (Edwards and Palsson, 2000; Oberhardt *et al.*, 2008). In other organisms, e.g. *S. cerevisiae*, predictions of essentiality are less accurate, and for multiple knockouts in particular, there is only a very low correlation between experimental data and FBA predictions (Heavner and Price, 2015).

The assumption of maximization of biomass production as a metabolic objective is often reasonable for microorganisms during exponential growth, but it will clearly not hold for most mammalian cells or other multicellular organisms whose evolutionary pressure has selected for far more complex traits than simply growth at the cellular level. In place of FBA, Markov chain Monte Carlo (MCMC) methods can be used to uniformly sample the feasible steady-state flux space described by Equation 2. MCMC methods provide an estimate of the joint probability distribution of fluxes and do not depend on a pre-specified biological objective. The applications of random sampling methods include the analysis of red blood cells under storage conditions (Bordbar *et al.*, 2016), aspirin resistance in platelets (Thomas *et al.*, 2015), transcriptional regulation in human adipocytes (Mardinoglu *et al.*, 2014) and in bacterial communities in the human gut (Shoaie *et al.*, 2013), as well as the metabolic re-wiring that takes place in epithelial

to mesenchymal transition during the development of breast cancer (Halldorsson *et al.*, 2017).

2. Constraining metabolic models with transcriptomics and proteomics data

Although mass balance is an essential principle, metabolism is constrained by other factors and physical principles as well. FBA assumes that the cell can use all metabolic reactions at a given time in the combination that gives the highest biomass production, however, this is contradicted by the fact that only a proportion of an organism's genes will be transcriptionally active at the same time. Thus further constraints can be imposed on the model by leveraging information about the transcriptional state of the cell. This can be used to create context-specific models from generic models, such as the generic human reconstruction Recon1 (Duarte *et al.*, 2007), as well as to improve the accuracy of flux predictions. The simplest realization of this idea utilizes the fact that an enzyme cannot catalyze any reaction flux if its encoding gene is not expressed. Reactions catalyzed by genes with transcript levels below a defined threshold can thus be forced to be inactive by removing them from the model. Flux distributions obtained with such a constrained model were found to be more strongly correlated to experimentally measured fluxes in *S. cerevisiae* compared to an unconstrained model (Åkesson, Förster and Nielsen, 2004). More sophisticated algorithms minimize the difference between the predicted flux distribution and the gene expression data. The Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm (Becker and Palsson, 2008) finds flux values which minimize the utilization of reactions with low expression levels, in order to meet pre-specified metabolic requirements such as growth. The iMAT method developed by Shlomi and coworkers (Shlomi *et al.*, 2008) alleviates the need for a pre-specified cellular objective and is therefore suitable for analyzing mammalian cells and tissues. The method partitions gene expression values into three groups, corresponding to high, moderate and low expression and then maximizes the number of reactions with flux levels in agreement with the expression states. This enabled identification of tissue-specific metabolic activities in different human tissues, and the construction of tissue-specific models of human metabolism. An extension of iMAT was used to construct a model of cancer metabolism from Recon1 and expression data from cancer cell lines in the NCI-60 collection. The cancer model was then used to identify several cytostatic drug targets, and generate a list of potential selective anti-cancer treatments (Folger *et al.*, 2011).

Since Åkesson and coworkers first used gene-expression data to constrain metabolic models, a large number of methods that integrate expression data and flux predictions have been published. An evaluation of many of these methods, by their ability to predict flux distributions in *E.* coli and *S. cerevisiae*, showed that none of them performed significantly better than parsimonious FBA, an extension of FBA that finds the flux distribution with the smallest sum of fluxes that can support the optimal objective value (Machado and Herrgård, 2014). This suggests that gene transcription levels do not correlate strongly with reaction fluxes, at least in microbial cells, which is not surprising considering that translational efficiency, post-translational modifications and allosteric regulation all have an effect on fluxes as well.

A step closer to the actual reactions than mRNA abundance is protein concentration. A certain correlation between mRNA and protein concentration is to be expected (Gry *et al.*, 2009), and several methods for integrating gene expression data into metabolic models can indeed use protein abundance data with the same algorithms, simply by replacing gene expression thresholds with protein abundance thresholds (Becker and Palsson, 2008; Machado and Herrgård, 2014). However, there have also been attempts to more explicitly incorporate proteomics data into the modeling frameworks. A central component of enzyme kinetics is the concept of the catalytic capacity of an enzyme. Each enzyme molecule can only perform a certain number of conversions per second; an increased flux will therefore require a larger number of enzymes at some point. The maximum possible flux, represented by the V_{max} parameter, can be calculated from the enzyme concentration and catalytic turnover number, k_{cat}

$$V_{max} = k_{cat} \cdot [E] \tag{3}$$

If the catalytic turnover parameters are known, this relationship can be used to constrain fluxes using protein concentration data. In the GECKO modeling framework (Sánchez *et al.*, 2017) a constraint is added for each enzyme, representing the enzyme's degree of utilization, where the upper bound is set to the measured enzyme concentration. The utilization of an enzyme is obtained by summing v/k_{cat} for all reactions catalyzed by that enzyme. Using GECKO with a proteomics dataset for *S. cerevisiae*, Sanchez and coworkers showed that the space of possible fluxes was reduced considerably by excluding all flux distributions that were not consistent with the observed enzyme levels. On the other hand, the fluxes predicted for *S. cerevisiae* grown in glucose limited minimal medium, did not have a significantly smaller error, compared to experimentally measured fluxes, than those predicted with FBA. It is possible however, that the advantage of using proteomics data will be larger in cases where the

assumption of maximal growth is not valid, e.g. under stress conditions or in genetically perturbed strains. GECKO can also be used in the absence of proteomics data by imposing a single overall constraint on the total enzyme mass. This resulted in more accurate predictions of maximal growth rates on a wide range of different carbon sources, for which FBA tends to overestimate growth rate. Another interesting growth effect that was captured by including an overall protein constraint is the shift from respiration to fermentation at high growth rates. This overflow metabolism, also known as the Crabtree effect in yeast (Crabtree, 1929) and the Warburg effect in cancer cells (Warburg, Wind and Negelein, 1927), cannot be captured by FBA, where simply the flux distribution with the highest biomass yield is found, independently of growth rate. The overflow effect is most likely caused by respiratory enzymes having a higher proteome cost than fermentative enzymes (Basan *et al.*, 2015), which means that at high growth rates protein allocation becomes limiting and fermentation becomes more efficient even though it has a lower energy/carbon yield. Overflow metabolism has been modeled e.g. in *E. coli* (Basan *et al.*, 2015), *S. cerevisiae* (Sánchez *et al.*, 2017) and cancer cells (Shlomi *et al.*, 2011), by different models with the common trait of somehow constraining the proteome.

The causes of the Warburg effect in cancer cells were studied using Recon1 by placing a constraint on total enzyme concentration to account for enzyme solvent capacity (Shlomi *et al.*, 2011). To compute the contribution of each enzyme to the total concentration, an estimate of the enzyme turnover number was required. Estimates for 15% of the reactions could be obtained from biochemical databases, the rest was assigned a fixed value of 25/s. Using FBA and random sampling, the Warburg effect was shown to be a consequence of metabolic adaptations to increase biomass productivity. Further analysis revealed the preference of cancer cells to take up glutamine instead of other amino acids.

Resource allocation between cellular processes in *Bacillus subtilis* was recently analyzed using a method that incorporates genome-wide protein quantification data and extracellular nutrient concentrations with a metabolic reconstruction (Goelzer *et al.*, 2015). The method, Resource Balance Analysis (RBA), links flux to enzyme abundance, assuming a relationship similar to Equation 3, while incorporating information on protein activity and protein localization. The use of RBA is fairly involved compared to the methods described earlier and requires specification of a large number of parameters. The parameters were partly obtained from Uniprot and partly inferred from data. RBA accurately predicted the allocation of resources in *B. subtilis* over a wide

range of conditions. In vivo knockouts of enzymes which were expressed but predicted to have zero flux in the model resulted in significantly increased growth (Goelzer *et al.*, 2015). This suggests that the method may be useful for constructing minimal cell factories, e.g. for protein production.

3. Models of metabolism and macromolecular expression

The previously described methods for combining *omics* data and metabolic models are mostly based on heuristically formulated constraints and/or objectives. When the measured quantities – such as mRNA and protein abundances – are not explicitly accounted for in the modeling framework, they cannot be seamlessly integrated into it. To address this problem, an extended modeling framework that explicitly models the expression of macromolecules, such as RNA and protein, has been developed. Construction of such models of metabolism and expression (**ME-models**) began with the reconstruction of the macromolecular expression network of *E. coli*, analogously to the metabolic network (Thiele *et al.*, 2009). Transcription of a given gene to produce mRNA is modeled as a reaction consuming nucleotides in proportions consistent with the specific sequence, and similarly translation is modeled as a reaction consuming charged tRNAs while producing protein and uncharged tRNAs. In order to model how metabolic catalysis is dependent on translation of a specific protein and how translation of a protein is dependent on transcription of its gene to mRNA, these different reactions must be coupled (Thiele *et al.*, 2009; Lerman *et al.*, 2012). A certain quantity of an enzyme can only catalyze a limited reaction flux and Equation 3 can be rearranged to enable calculation of the minimum amount of enzyme required to catalyze a given flux

$$[E] \geq \frac{v}{k_{cat}} \tag{4}$$

Equation 4 represents a constraint that can be used to couple metabolic reactions to the enzymes that catalyze them. Identical constraints can be formulated for ribosomes and mRNA in translation reactions and for RNA-polymerase in transcription reactions. A constraint-based modeling framework, however, does not model concentrations of metabolites (or enzymes) and is thus not directly compatible with such constraints. To circumvent this it is necessary to account for growth-related dilution. In a growing cell, metabolite pools are continuously diluted, because of the expanding intracellular volume, by a rate equal to the product of the growth rate and metabolite concentration. This means that in steady-state, catalysis of a reaction requires that the catalyzing enzyme be produced at a rate proportional to the growth rate. Enzymatic conversion of compound *A* into compound *B* by enzyme *E* thus becomes (Lloyd *et al.*, 2017):

$$A + \frac{\mu}{k_{cat}} E \rightarrow B \quad (5)$$

In FBA the requirement of enzyme production is modeled through the composition of the biomass reaction, but since this reaction is determined *a priori¸* FBA cannot model how biomass composition changes under different growth rates and conditions. With ME-models the empirical biomass reaction is replaced by explicitly modeling the relationship between metabolism and macromolecular expression. ME-models can thus directly predict the expression levels of different proteins, which can be compared with –*omics* datasets. A ME-model of the thermophilic bacterium *Thermotoga maritima* (Lerman *et al.*, 2012), found moderate correlations between predicted and experimentally measured mRNA profiles (r= 0.54), protein expression profiles (r = 0.57), as well as proteome amino acid composition (r = 0.79). A ME-model of *E. coli* showed improved prediction of growth rates in different nutrient conditions compared to FBA (Thiele *et al.*, 2012), and could accurately predict several internal fluxes (O'Brien *et al.*, 2013). Additionally, since ME-models explicitly include the cost of producing the enzymes required for various pathways, they implicitly limit the total proteome size and thus also capture metabolic overflow effects, such as the acetate overflow metabolism in *E. coli* (O'Brien *et al.*, 2013).

Whereas traditional constraint-based metabolic models include, and can thus directly predict, growth rate, uptake and secretion rates and internal fluxes, ME-models can additionally predict expression profiles and proteome composition, and thus they can also be directly constrained by expression and proteomics data. Because of this, ME-models represent an intuitive and theoretically justified method of integrating transcriptomics and proteomics data into metabolic models. They have not yet found broad usage in the metabolic modeling community, presumably because of the time it takes to run simulations (several orders of magnitude higher than with FBA), and the lack of related model and software infrastructure, but these issues are continuously being addressed (Yang *et al.*, 2016; Lloyd *et al.*, 2017).

4. Augmenting models with metabolomics data

In a discussion of data integration in metabolic models, it is impossible not to mention metabolomics. Different analytical methods, e.g. enzymatic assays, chromatography and mass spectrometry, can be used to take snapshots of the cellular metabolism with varying resolution, coverage, precision and throughput. However, they all provide useful information about the concentrations of metabolite pools in the cell. One of the earliest uses of metabolomics data to improve metabolic modeling was **metabolic flux analysis**

(MFA), which utilizes time-course metabolite concentration data from cultures fed with isotopically labeled substrates to infer flux values in the metabolic network (Stephanopoulos, 1999; Sauer, 2006). This is done by monitoring how the isotopes, e.g. ^{13}C or ^{15}N, spread to downstream metabolite pools over time. The advantage of this method is that the resulting fluxes can be used directly to constrain metabolic models or to compare the validity of different simulation methods. However, MFA is labor- and cost intensive and works best on a smaller subset of the entire metabolic network, typically just the central carbon metabolism (Antoniewicz, 2015; Gopalakrishnan and Maranas, 2015).

Changes in extracellular metabolite concentrations over time can be used to estimate uptake and secretion rates and constrain the flux space. However, since constraint-based modeling frameworks model fluxes under an assumption of steady-state, internal metabolite concentration data at a single time point without isotopic labeling cannot be directly utilized. Despite this, metabolomics data can still be used to either constrain the models or to provide new insights in combination with the simulation results. In order to model cells that are not in steady-state, such as human blood cells undergoing physiological changes during storage, Bordbar and coworkers devised a method called unsteady-state FBA (Bordbar *et al.*, 2017). Using time-course metabolomics they determined the rate of accumulation or depletion for internal metabolites, which was then modeled by adding source and sink reactions to the metabolic model. These reactions were then constrained to have fluxes corresponding to the experimentally determined rates of concentration changes. Subsequent MFA revealed that the fluxes predicted with this method were more accurate than those obtained by regular FBA.

Aside from enforcing steady state, a commonly used constraint in constraint-based models is to force certain fluxes to only go in one direction. This is straightforward for some reactions whose thermodynamics make it practically irreversible under biological conditions. Other reactions are closer to equilibrium and can go in both directions depending on specific conditions. The spontaneous direction of a reaction can be calculated by the formula

$$\Delta_r G = \Delta_r G° + RT\ log(Q) \tag{6}$$

If the left-hand side (the **reaction Gibbs free energy**) is negative, the reaction will proceed spontaneously in the forward direction, while it will proceed spontaneously in the reverse direction if the reaction Gibbs free energy is positive. $\Delta_r G°$ is the reaction Gibbs free energy under standard conditions, RT is the gas constant times the absolute temperature and Q is the reaction quotient, containing the concentrations of the

reaction products and substrates. The standard Gibbs free energy must in principle be determined experimentally, but in most cases it can be calculated from the structure of the participating metabolites and already known reaction Gibbs free energies for other reactions (Noor *et al.*, 2013). This means that a dataset of metabolite concentrations can be used to constrain reactions to a specific direction depending on the specific metabolic conditions, reducing the space of feasible fluxes significantly (Soh and Hatzimanikatis, 2014). In many simulated growth conditions, it can be sufficient simply to constrain reaction directionalities according to the most common mode of operation without regard to actual metabolite concentrations. Some reactions however, occur in the unconventional direction under extreme conditions, such as very high CO_2 concentrations. In such cases using thermodynamics and metabolite data to inform reaction directionalities will be particularly beneficial and can lead to more accurate simulations (Soh, Miskovic and Hatzimanikatis, 2012).

Constraint-based simulations can also be combined with metabolomics data in another way. In addition to calculating a flux distribution, simulating a constraint-based model also provides so-called **shadow prices**. Each shadow price is linked to a metabolite and reflects how much the objective function, e.g. growth, could be improved if the model were allowed to get some of that metabolite "for free". In other words a shadow price is a measure of how limiting a given metabolite's mass balance is for the objective function. Depending on the algorithm used to solve the FBA problem, shadow prices are either a byproduct of the solution process or can be obtained with modest computational effort.

Zampieri and coworkers investigated the evolution of antibiotic resistance in *E. coli* using adaptive laboratory evolution (Zampieri *et al.*, 2017). By maximizing and minimizing flux through each reaction in the model and calculating the shadow prices, the authors could identify reactions, which, when maximized or minimized, resulted in shadow prices that were consistent with the observed patterns of metabolite concentration changes. Those reactions were hypothesized as being targets of evolution, whose flux should be increased in order to increase antibiotic resistance.

Besides constraint-based modeling, the most common way to simulate cellular metabolism is with kinetic models. This involves the solution of the system of differential equations shown in Eq. 1 from given initial metabolite concentrations. As previously described, one of the challenges with this approach is the requirement of knowing the values of all the kinetic parameters of the system. For small biochemical systems, the kinetic parameters can sometimes be determined individually through *in*

vitro experiments, but for genome-scale models this is not feasible. Additionally there is no guarantee that the *in vitro* kinetic parameters are representative of how an enzyme functions *in vivo* (Teusink *et al.*, 2000). Instead of the bottom-up approach of experimentally determining each parameter, a top-down approach may be used, where the model parameters might initially be estimated from prior information, such as *in vitro* data, but are predominantly selected by fitting simulation results to genome-scale experimental data. This has long been done for small-scale networks, using metabolomics and MFA data (Jamshidi and Palsson, 2008; Srinivasan, Cluett and Mahadevan, 2015), however with continual increases in dataset sizes and computing power, it has also become feasible to do this for genome-scale networks. Recently a genome-scale kinetic model of *E. coli* was published along with estimated values for all kinetic parameters (Khodayari and Maranas, 2016). The model parameters were fitted using experimental flux data and model predictions were validated against metabolomics data. In addition the model could quantitatively predict product yields of 24 different compound in 320 mutant strains, which was considerably better than the constraint-based simulation methods it was tested against. In another study kinetic models of human red blood cells were used to investigate individual variations in susceptibility to side effects of the hepatitis B drug Ribavirin (Bordbar *et al.*, 2015). By measuring intracellular metabolite levels in red blood cells of 24 patients, they could determine individual kinetic parameter values for each of the patients, and show that those parameters were predictive of whether the patient was sensitive to side effects. Furthermore, the identified relationships between kinetic parameters and sensitivity to drug side-effect were consistent with known mechanisms of Ribavirin side effects. These results show that kinetic modeling frameworks have the potential to significantly outperform constraint-based simulations, and that with modern *omics* technologies and computer power, it is feasible to parametrize them sufficiently to predict metabolic behavior (Saa and Nielsen, 2017).

5. Combining metabolic models and machine learning methods

The term **machine learning** covers a broad range of methods where large datasets are used to infer relationships between variables or to predict various outcomes from given input data. Often this is done without much consideration of specific mechanisms of the studied phenomena. Such data-driven methods can of course be applied to metabolic data, but with limited connection to biological mechanisms, the results are often difficult to interpret. Instead, machine learning methods can be combined with domain-specific biological knowledge, such as the information encoded within a genome-scale

reconstruction, to create hybrid methods that also take advantage of the metabolic network structure.

Plaimas and coworkers predicted gene essentiality in *E. coli* using a hybrid method (Plaimas *et al.*, 2008). Instead of using FBA to predict essentiality as described previously, they defined a set of features for each reaction, including metrics of network topology, gene expression data and predicted FBA fluxes. These features were fed into a support vector machine classifier together with labels from experimental essentiality data (Baba *et al.*, 2006). The predictive accuracy of gene essentiality was 92%, compared to 85% for FBA. Furthermore, the genes where essentiality was not correctly predicted were retested experimentally, and in several cases the authors identified errors in the original experimental dataset. By removing single features from the input data one at a time, the authors could also identify which features were most important for accurately predicting essentiality. Prediction with FBA suffers mainly from two problems, namely that the metabolic network might be incomplete, and that the assumption of growth optimality does not always hold (O'Brien, Monk and Palsson, 2015). A hybrid method can instead learn from data, utilizing the biological context, e.g. in the form of a metabolic network, only when it improves prediction performance. A similar method was recently used to predict drug side effects (Shaked *et al.*, 2016). A list of drugs known to inactivate one or more enzymatic reactions was used as training data, with features corresponding to the minimum and maximum possible FBA flux for each reaction after deactivating the drug's target reaction(s) in the Recon1 model. Support vector machine classifiers were then trained to predict which (if any) side effects the drug would have. Using a feature selection method it was also possible to find the features that were most strongly associated with a given side effect. Many of the results were found to be consistent with the published literature of these drug side effects.

A third example of a combination of machine learning with metabolic network data was used to predict novel drug-reaction interactions for cancer therapy (Li *et al.*, 2010). The method requires the construction of a reaction flux similarity matrix. This matrix was obtained using the GIMME algorithm to predict reaction fluxes from gene expression data in 59 cancer cell lines. Reactions with the same flux profile across the cell lines were said to have a high similarity, while reactions with different flux profiles had a low similarity. The reaction flux similarity matrix was combined with knowledge of existing drug-reaction interactions, using a K-nearest neighbors algorithm, to predict new interactions.

Where purely model-based algorithms may suffer from lack of biological knowledge, the use of machine learning methods in biomedical research is often hampered by difficulties in interpreting the results. The examples above show that the two methodologies can be combined to achieve results that are informed by experimental data, while maintaining biologically relevant relationships between variables. Such hybrid methods can be used to build accurate predictive models, while also providing new biological insights and will without doubt find widespread use in the future.

References

Åkesson, M., Förster, J. and Nielsen, J. (2004) 'Integration of gene expression data into genome-scale metabolic models', *Metabolic Engineering*, 6(4), pp. 285–293. doi: 10.1016/j.ymben.2003.12.002.

Antoniewicz, M. R. (2015) 'Methods and advances in metabolic flux analysis: a mini-review', *Journal of Industrial Microbiology and Biotechnology*, 42(3), pp. 317–325. doi: 10.1007/s10295-015-1585-x.

Baba, T. *et al.* (2006) 'Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection', *Molecular Systems Biology*, 2. doi: 10.1038/msb4100050.

Basan, M. *et al.* (2015) 'Overflow metabolism in Escherichia coli results from efficient proteome allocation', *Nature*, 528(7580), pp. 99–104. doi: 10.1038/nature15765.

Becker, S. a. and Palsson, B. O. (2008) 'Context-specific metabolic networks are consistent with experiments', *PLoS Computational Biology*, 4(5). doi: 10.1371/journal.pcbi.1000082.

Bordbar, A. *et al.* (2015) 'Personalized Whole-Cell Kinetic Models of Metabolism for Discovery in Genomics and Pharmacodynamics', *Cell Systems*. Elsevier Inc., 1(4), pp. 283–292. doi: 10.1016/j.cels.2015.10.003.

Bordbar, A. *et al.* (2016) 'Identified metabolic signature for assessing red blood cell unit quality is associated with endothelial damage markers and clinical outcomes', *Transfusion*, 56(4), pp. 852–862. doi: 10.1111/trf.13460.

Bordbar, A. *et al.* (2017) 'Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics', *Scientific Reports*. Nature Publishing Group, 7(December 2016), p. 46249. doi: 10.1038/srep46249.

Crabtree, H. G. (1929) 'Observations on the carbohydrate metabolism of tumours', *The*

Biochemical journal, 23(3), pp. 536–45. doi: 10.1042/bj0230536.

Duarte, N. C. *et al.* (2007) 'Global reconstruction of the human metabolic network based on genomic and bibliomic data', *Proceedings of the National Academy of Sciences*, 104(6), pp. 1777–1782. doi: 10.1073/pnas.0610772104.

Edwards, J. S. and Palsson, B. O. (2000) 'The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities', *Proceedings of the National Academy of Sciences*, 97(10), pp. 5528–5533. doi: 10.1073/pnas.97.10.5528.

Folger, O. *et al.* (2011) 'Predicting selective drug targets in cancer through metabolic networks', *Molecular Systems Biology*. Nature Publishing Group, 7(1), pp. 527–527. doi: 10.1038/msb.2011.63.

Förster, J. *et al.* (2003) 'Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae.', *Omics : a journal of integrative biology*, 7(2), pp. 193–202. doi: 10.1089/153623103322246584.

Goelzer, A. *et al.* (2015) 'Quantitative prediction of genome-wide resource allocation in bacteria', *Metabolic Engineering*. Elsevier, 32, pp. 232–243. doi: 10.1016/j.ymben.2015.10.003.

Gopalakrishnan, S. and Maranas, C. D. (2015) '13C metabolic flux analysis at a genome-scale', *Metabolic Engineering*. Elsevier, 32, pp. 12–22. doi: 10.1016/j.ymben.2015.08.006.

Gry, M. *et al.* (2009) 'Correlations between RNA and protein expression profiles in 23 human cell lines', *BMC Genomics*, 10(1), p. 365. doi: 10.1186/1471-2164-10-365.

Halldorsson, S. *et al.* (2017) 'Metabolic re-wiring of isogenic breast epithelial cell lines following epithelial to mesenchymal transition', *Cancer Letters*, 396, pp. 117–129. doi: 10.1016/j.canlet.2017.03.019.

Heavner, B. D. and Price, N. D. (2015) 'Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction', *PLoS Computational Biology*, 11(11), pp. 1–26. doi: 10.1371/journal.pcbi.1004530.

Jamshidi, N. and Palsson, B. Ø. (2008) 'Formulating genome-scale kinetic models in the post-genome era.', *Molecular systems biology*, 4(171), p. 171. doi: 10.1038/msb.2008.8.

Karr, J. R. *et al.* (2012) 'A whole-cell computational model predicts phenotype from genotype', *Cell*, 150(2), pp. 389–401. doi: 10.1016/j.cell.2012.05.044.

Khodayari, A. and Maranas, C. D. (2016) 'A genome-scale Escherichia coli kinetic

metabolic model k-ecoli457 satisfying flux data for multiple mutant strains', *Nature Communications*, p. 13806. doi: 10.1038/ncomms13806.

Lerman, J. A. *et al.* (2012) 'In silico method for modelling metabolism and gene product expression at genome scale', *Nature Communications*. Nature Publishing Group, 3(May), p. 929. doi: 10.1038/ncomms1928.

Li, L. *et al.* (2010) 'Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines.', *BMC bioinformatics*, 11(1), p. 501. doi: 10.1186/1471-2105-11-501.

Lloyd, C. J. *et al.* (2017) 'COBRAme: A Computational Framework for Building and Manipulating Models of Metabolism and Gene Expression', *bioRxiv*, p. 106559. doi: http://dx.doi.org/10.1101/106559.

Machado, D. and Herrgård, M. (2014) 'Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism', *PLoS Computational Biology*, 10(4). doi: 10.1371/journal.pcbi.1003580.

Mardinoglu, A. *et al.* (2014) 'Integration of clinical data with a genome-scale metabolic model of the human adipocyte', *Molecular Systems Biology*. Nature Publishing Group, 9(1), pp. 649–649. doi: 10.1038/msb.2013.5.

McCloskey, D., Palsson, B. Ø. and Feist, A. M. (2013) 'Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli.', *Molecular systems biology*, 9(1), p. 661. doi: 10.1038/msb.2013.18.

Noor, E. *et al.* (2013) 'Consistent Estimation of Gibbs Energy Using Component Contributions', *PLoS Computational Biology*, 9(7). doi: 10.1371/journal.pcbi.1003098.

O'Brien, E. J. *et al.* (2013) 'Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction', *Molecular Systems Biology*, 9(1), pp. 693–693. doi: 10.1038/msb.2013.52.

O'Brien, E. J., Monk, J. M. and Palsson, B. O. (2015) 'Using genome-scale models to predict biological capabilities', *Cell*. Elsevier Inc., 161(5), pp. 971–987. doi: 10.1016/j.cell.2015.05.019.

Oberhardt, M. A. *et al.* (2008) 'Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1', *Journal of Bacteriology*, 190(8), pp. 2790–2803. doi: 10.1128/JB.01583-07.

de Oliveira Dal'Molin, C. G. *et al.* (2010) 'AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis', *Plant Physiology*, 152(2), pp. 579–589. doi:

10.1104/pp.109.148817.

Orth, J. D., Thiele, I. and Palsson, B. Ø. (2010) 'What is flux balance analysis?', *Nature biotechnology*. Nature Publishing Group, 28(3), pp. 245–248. doi: 10.1038/nbt.1614.

Plaimas, K. *et al.* (2008) 'Machine learning based analyses on metabolic networks supports high-throughput knockout screens', *BMC Systems Biology*, 2(1), p. 67. doi: 10.1186/1752-0509-2-67.

Price, N. D., Reed, J. L. and Palsson, B. Ø. (2004) 'Genome-scale models of microbial cells: evaluating the consequences of constraints', *Nature Reviews Microbiology*, 2(11), pp. 886–897. doi: 10.1038/nrmicro1023.

Saa, P. A. and Nielsen, L. K. (2017) 'Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks', *Biotechnology Advances*. Elsevier, (March), pp. 0–1. doi: 10.1016/j.biotechadv.2017.09.005.

Sánchez, B. J. *et al.* (2017) 'Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints', *Molecular Systems Biology*, 13(8), p. 935. doi: 10.15252/msb.20167411.

Sauer, U. (2006) 'Metabolic networks in motion: 13C-based flux analysis', *Molecular Systems Biology*, 2, pp. 1–10. doi: 10.1038/msb4100109.

Schilling, C. H. and Palsson, B. Ø. (2000) 'Assessment of the Metabolic Capabilities of Haemophilus influenzae Rd through a Genome-scale Pathway Analysis', *Journal of Theoretical Biology*, 203(3), pp. 249–283. doi: 10.1006/jtbi.2000.1088.

Shaked, I. *et al.* (2016) 'Metabolic Network Prediction of Drug Side Effects', *Cell Systems*. Elsevier Inc., 2(3), pp. 209–213. doi: 10.1016/j.cels.2016.03.001.

Shlomi, T. *et al.* (2008) 'Network-based prediction of human tissue-specific metabolism', *Nature Biotechnology*, 26(9), pp. 1003–1010. doi: 10.1038/nbt.1487.

Shlomi, T. *et al.* (2011) 'Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect', *PLoS Computational Biology*, 7(3), pp. 1–8. doi: 10.1371/journal.pcbi.1002018.

Shoaie, S. *et al.* (2013) 'Understanding the interactions between bacteria in the human gut through metabolic modeling', *Scientific Reports*, 3(1), p. 2532. doi: 10.1038/srep02532.

Soh, K. C. and Hatzimanikatis, V. (2014) 'Constraining the Flux Space Using Thermodynamics and Integration of Metabolomics Data', in Krömer, J. O., Nielsen, L. K.,

and Blank, L. M. (eds) *Metabolic Flux Analysis: Methods and Protocols*. New York, NY: Springer New York, pp. 49–63. doi: 10.1007/978-1-4939-1170-7_3.

Soh, K. C., Miskovic, L. and Hatzimanikatis, V. (2012) 'From network models to network responses: Integration of thermodynamic and kinetic properties of yeast genome-scale metabolic networks', *FEMS Yeast Research*, 12(2), pp. 129–143. doi: 10.1111/j.1567-1364.2011.00771.x.

Srinivasan, S., Cluett, W. R. and Mahadevan, R. (2015) 'Constructing kinetic models of metabolism at genome-scales: A review', *Biotechnology Journal*, 10(9), pp. 1345–1359. doi: 10.1002/biot.201400522.

Stephanopoulos, G. (1999) 'Metabolic Fluxes and Metabolic Engineering', *Metabolic Engineering*, 1(1), pp. 1–11. doi: 10.1006/mben.1998.0101.

Teusink, B. *et al.* (2000) 'Can yeast glycolysis be understood terms of vitro kinetics of the constituent enzymes? Testing biochemistry', *European Journal of Biochemistry*, 267(17), pp. 5313–5329. doi: 10.1046/j.1432-1327.2000.01527.x.

Thiele, I. *et al.* (2009) 'Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization', *PLoS Computational Biology*, 5(3). doi: 10.1371/journal.pcbi.1000312.

Thiele, I. *et al.* (2012) 'Multiscale Modeling of Metabolism and Macromolecular Synthesis in E. coli and Its Application to the Evolution of Codon Usage', *PLoS ONE*, 7(9). doi: 10.1371/journal.pone.0045635.

Thomas, A. *et al.* (2015) 'Network reconstruction of platelet metabolism identifies metabolic signature for aspirin resistance', *Scientific Reports*, 4(1), p. 3925. doi: 10.1038/srep03925.

Warburg, O., Wind, F. and Negelein, E. (1927) 'The metabolism of tumors in the body', *The Journal of General Physiology*, 8(6), pp. 519–530. doi: 10.1085/jgp.8.6.519.

Yang, L. *et al.* (2016) 'solveME: fast and reliable solution of nonlinear ME models', *BMC Bioinformatics*. BMC Bioinformatics, 17(1), p. 391. doi: 10.1186/s12859-016-1240-1.

Zampieri, M. *et al.* (2017) 'Metabolic constraints on the evolution of antibiotic resistance', *Molecular Systems Biology*, 13(3), p. 917. doi: 10.15252/msb.20167028.