6.047/6.878 Lecture 19: Introduction to Steady State Metabolic Modeling

Guest Lecture by James Galagan
Scribed by Meriem Sefta (2011)
Jake Shapiro, Andrew Shum, and Ashutosh Singhal (1910)
Molly Ford Dacus and Anand Oza (1909)
Christopher Garay (1908)

September 10, 2012

Contents

List of Figures

1 Introduction

Metabolic modeling allows us to use mathematical models to represent complex biological systems. This lecture discusses the role of modeling the steady state of biological systems in understanding the metabolic capabilities of organisms. We also briefly discuss how well steady state models are able to replicate in-vitro experiments.

1.1 What is Metabolism?

According to Matthews and van Holde, metabolism is the totality of all chemical reactions that occur in living matter. This includes catabolic reactions, which are reactions that lead to the breakdown of molecules into smaller components, and anabolic reactions, which are responsible for the creation of more complex molecules (e.g. proteins, lipids, carbohydrates, and nucleic acids) from smaller components. These reactions are responsible for the release of energy from chemical bonds and the storage of this energy. Metabolic reactions are also responsible for the transduction and transmission of information (for example, via the generation of cGMP as a secondary messenger or mRNA as a substrate for protein translation).

1.2 Why Model Metabolism?

An important application of metabolic modeling is in the prediction of drug effects. An important subject of modeling is the organism Mycobacterium tuberculosis [15]. The disruption of the mycolic acid synthesis pathways of this organism can help control TB infection. Computational modeling gives us a platform for identifying the best drug targets in this system. Gene knockout studies in *Escherichia coli* have allowed scientists to determine which genes and gene combinations affect the growth of this important model organism [6]. Both agreements and disagreements between models and experimental data can help us assess our knowledge of biological systems and help us improve our predictions about metabolic capabilities. In the next lecture, we will learn the importance of incorporating expression data into metabolic models. In addition, a variety of infectious disease processes involve metabolic changes at the microbial level.

2 Model Building

An overarching goal of metabolic modeling is the ability to take a schematic representation of a pathway and change that it into a mathematical formula modeling the pathway. For example, converting the following pathway into a mathematical model would be incredible useful.

2.1 Chemical Reactions

In metabolic models, we are concerned with modeling chemical reactions that are catalyzed by enzymes. Enzymes work by acting on a transition state of the enzyme-substrate complex that lowers the activation energy of a chemical reaction. The diagram on slide 5 of page 1 of the lecture slides demonstrates this phenomenon. A typical rate equation (which describes the conversion of the substrates S of the enzyme reaction into its products P) can be described by a Michaelis-Menten rate law:

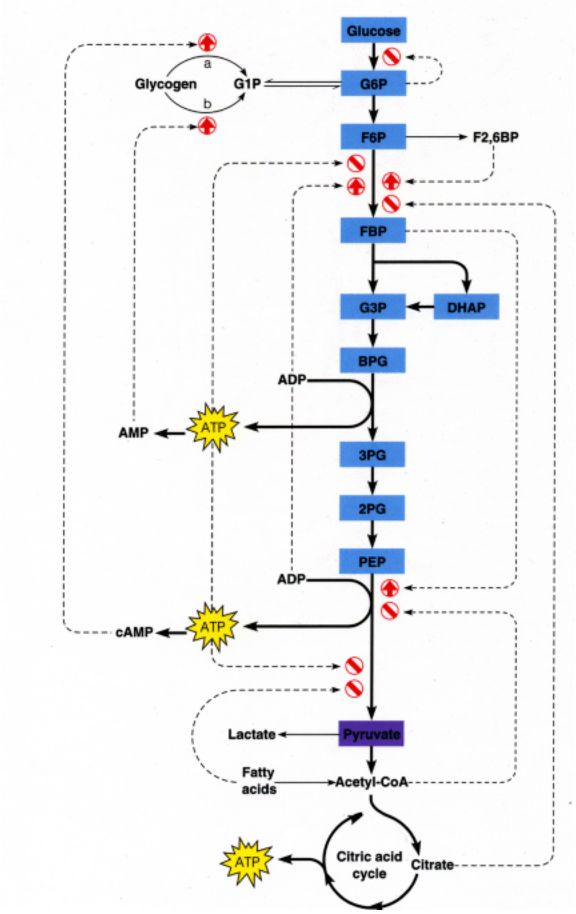


Figure 1: The process leading to and including the citric acid cycle.

$$\frac{V}{V_{max}} = \frac{[S]}{K_m+[S]}$$

In this equation, V is the rate of the equation as a function of substrate concentration [S]. It is clear that the parameters K_m and V_{max} are necessary to characterize the equation.

The inclusion of multiple substrates, products, and regulatory relationships quickly increases the number of parameters necessary to characterize such equations. The figures on slides 1, 2, and 3 of page 2 of the lecture notes demonstrate the complexity of biochemical pathways. Kinetic modeling quickly becomes infeasible: the necessary parameters are difficult to measure, and also vary across organisms [10]. Thus, we are interested in a modeling method that would allow us to use a small number of precisely determined parameters. To this end, we recall the basic machinery of stoichiometry from general chemistry. Consider the chemical equation $A+2B \rightarrow 3C$, which says that one unit of reactant A combines with 2 units of reactant B to form 3 units of reactant C. The rate of formation of the compound X is given by the time derivative of [X]. Note that C forms three times as fast as A. Therefore, due to the stoichiometry of the reaction, we see that the reaction rate (or reaction flux) is given by

$$flux = \frac{d[A]}{dt} = \frac{1}{2}\frac{d[B]}{dt} = \frac{1}{3}\frac{d[C]}{dt}$$

This will be useful in the subsequent sections. We must now state the simplifying assumptions that make our model tractable.

2.2 Steady-State Assumption

The steady state assumption assumes that there is no accumulation of any metabolite in the system. This allows us to represent reactions entirely in terms of their chemistry (i.e. the stoichiometric relationships between the components of the enzymatic reaction). Note that this does not imply the absence of flux through any given reaction. Rather, steady-state actually implies two assumptions that are critical to simplify metabolic modeling. The first is that the internal metabolite concentrations are constant, and the second is that fluxes, ie input and output fluxes, are also constant.

An analogy is a series of waterfalls that contribute water to pools. As the water falls from one pool to another, the water levels do not change even though water continues to flow (see page 2 slide 5). This framework prevents us from being hindered by the overly complicated transient kinetics that can result from perturbations of the system. Since we are usually interested in long-term metabolic capabilities (functions on a scale longer than milliseconds or seconds), the steady state dynamics may give us all the information that we need.

The steady-state assumption makes the ability to generalize across species and reuse conserved pathways in models much more feasible. Reaction stochiometries are often conserved across species, since they involve only conservation of mass. The biology of enzyme catalysis, and the parameters that characterize it, are not similarly conserved. These include species-dependent parameters such as the activation energy of a reaction, substrate affinity of an enzyme, and the rate constants for various reactions. However, none of these are required for steady-state modeling.

It is also of interest to note that, since time constants for metabolic reactions are usually in the order of milliseconds, most measurement technologies used today are not able to capture these extremely fast dynamics. This is the case of metabolomics mass spectrometry based measurements for example. In this method, the amounts of all the internal metabolites in a system are measured at a given point in time, but measurements can be taken at best every hour. In the majority of circumstances, all that is ever measured is steady state.

2.3 Reconstructing Metabolic Pathways

There are several databases that can provide the information necessary to reconstruct metabolic pathways in silico. These databases allow reaction stoichiometry to be accessed using Enzyme Commission numbers. Reaction stochiometries are the same in all the organisms that utilize a given enzyme. Among the databases of interest are ExPASy [5], MetaCyc [16], and KEGG [14]. These databases often contain pathways organized by function that can be downloaded in SBML format, making pathway reconstruction very easy for well-characterized pathways.

3 Metabolic Flux Analysis

Metabolic flux analysis (MFA) is a way of computing the distribution of reaction fluxes that is possible in a given metabolic network at steady state. We can place constraints on certain fluxes in order to limit the space described by the distribution of possible fluxes. In this section, we will develop a mathematical formulation for MFA. Once again, this analysis is independent of the particular biology of the system; rather, it will only depend on the (universal) stoichiometries of the reactions in question.

3.1 Mathematical Representation

Consider a system with m metabolites and n reactions. Let x_i be the concentration of substrate i, so that the rate of change of the substrate concentration is given by the time derivative of x_i . Let x be the column vector (with m components) with elements x_i . For simplicity, we consider a system with m = 4 metabolites A, B, C, and D. This system will consist of many reactions between these metabolites, resulting in a complicated balance between these compounds.

Once again, consider the simple reaction $A + 2B \rightarrow 3C$. We can represent this reaction in vector form as (-1 -2 3 0). Note that the first two metabolites (A and B) have negative signs, since they are consumed in the reaction. Moreover, the elements of the vector are determined by the stoichiometry of the reaction, as in Section 2.1. We repeat this procedure for each reaction in the system. These vectors become the columns of the stoichiometric matrix S. If the system has m metabolites and n reactions, S will be a m n matrix. Therefore, if we define v to be the n-component column vector of fluxes in each reaction, the vector S_v describes the rate of change of the concentration of each metabolite. Mathematically, this can be represented as the fundamental equation of metabolic flux analysis:

$$\frac{dx}{dt} = Sv$$

The matrix S is an extraordinarily powerful data structure that can represent a variety of possible scenarios in biological systems. For example, if two columns c and d of S have the property that $c = d$, the columns represent a reversible reaction. Moreover, if a column has the property that only one component is nonzero, it represents in exchange reaction, in which there is a flux into (or from) a supposedly infinite sink (or source), depending on the sign of the nonzero component.

We now impose the steady state assumption, which says that the left size of the above equation is identically zero. Therefore, we need to find vectors v that satisfy the criterion $Sv = 0$. Solutions to this equation will determine feasible fluxes for this system.

3.2 Null Space of S

The feasible flux space of the reactions in the model system is defined by the null space of S, as seen above. Recall from elementary linear algebra that the null space of a matrix is a vector space; that is, given two vectors y and z in the nullspace, the vector $ay + bz$ (for real numbers a, b) is also in the null space. Since the null space is a vector space, there exists a basis b_i, a set of vectors that is linearly independent and spans the null space. The basis has the property that for any flux v in the null space of S, there exist real numbers α_i such that

$$v = \Sigma_i \alpha_i b_i$$

How do we find a basis for the null space of a matrix? A useful tool is the singular value decomposition (SVD) [4]. The singular value decomposition of a matrix S is defined as a representation $S = UEV*$, where U is a unitary matrix of size m, V is a unitary matrix of size n, and E is a mxn diagonal matri$x$, with the (necessarily positive) singular values of S in descending order. (Recall that a unitary matrix is a matrix with orthonormal columns and rows, i.e. $U * U = UU* = I$ the identity matrix). It can be shown that any matrix has an SVD. Note that the SVD can be rearranged into the equation $Sv = \sigma u$, where u and v are columns of the matrices U and V and is a singular value. Therefore, if $\sigma = 0$, v belongs to the null space of S. Indeed, the columns of V that correspond to the zero singular values form an orthonormal basis for the null space of S. In this manner, the SVD allows us to completely characterize the possible fluxes for the system.

3.3 Constraining the Flux Space

The first constraint mentioned above is that all steady-state flux vectors must be in the null space. Also negative fluxes are not thermodynamically possible. Therefore a fundamental constraint is that all fluxes must be positive. (Within this framework we represent reversible reactions as separate reactions in the stoichiometric matrix S having two unidirectional fluxes.)

These two key constraints form a system that can be solved by convex analysis. The solution region can be described by a unique set of Extreme Pathways. In this region, steady state flux vectors v can be described as a positive linear combination of these extreme pathways. The Extreme Pathways, represented in slide 25 as vectors b_i, circumscribe a convex flux cone. Each dimension is a rate for some reaction. In slide 25, the z-dimension represents the rate of reaction for v_3 . We can recognize that at any point in time, the organism is living at a point in the flux cone, i.e. is demonstrating one particular flux distribution. Every point in the flux cone can be described by a possible steady state flux vector, while points outside the cone cannot.

One problem is that the flux cone goes out to infinity, while infinite fluxes are not physically possible. Therefore an additional constraint is capping the flux cone by determining the maximum fluxes of any of our reactions (these values correspond to our V_{max} parameters). Since many metabolic reactions are interior to the cell, there is no need to set a cap for every flux. These caps can be determined experimentally by measuring maximal fluxes, or calculated using mathematical tools such as diffusivity rules.

We can also add input and output fluxes that represent transport into and out of our cells (V_{in} and V_{out}). These are often much easier to measure than internal fluxes and can thus serve to help us to generate a more biologically relevant flux space. An example of an algorithm for solving this problem is the simplex algorithm [1]. Slides 24-27 demonstrate how constraints on the fluxes change the geometry of the flux cone. In reality, we are dealing with problems in higher dimensional spaces.

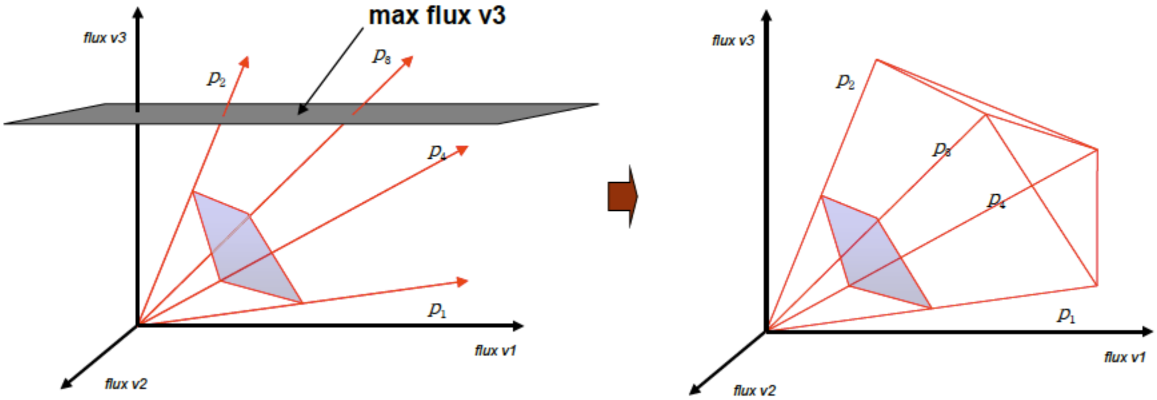


Figure 2: Adding constraints to extreme pathways.

3.4 Linear Programming

Linear programming is a generic solution that is capable of solving optimization problems given linear constraints. These can be represented in a few different forms.

Canonical Form :

- Maximize: $c^T x$
- Subject to: $Ax \leq b$

Standard Form :

- Maximize $\Sigma c_i * x_i$
- Subject to $a_{ij} X_i \leq b_i foralli, j$
- Non-negativity constraint: $X_i \geq 0$

A concise and clear introduction to Linear Programming is available here: http://www.purplemath.com/modules/linprog.htm The constraints described throughout section 3 give us the linear programming problem described in lecture. Linear programming can be considered a first approximation and is a classic problem in optimization. In order to try and narrow down our feasible flux, we assume that there exists a fitness function which is a linear combination of any number of the fluxes in the system. Linear programming (or linear optimization) involves maximizing or minimizing a linear function over a convex polyhedron specified by linear and non-negativity constraints.
We solve this problem by identifying the flux distribution that maximizes an objective function:
The key point in linear programming is that our solutions lie at the boundaries of the permissible flux space and can be on points, edges, or both. By definition however, an optimal solution (if one exists) will lie at a point of the permissible flux space. This concept is demonstrated on slide 30. In that slide, A is the stoichiometric matrix, x is the vector of fluxes, and b is a vector of maximal permissible fluxes.

Linear programs, when solved by hand, are generally done by the Simplex method. The simplex method sets up the problem in a matrix and performs a series of pivots, based on the basic variables of the problem statement. In worst case, however, this can run in exponential time. Luckily, if a computer is available, two other algorithms are available. The ellipsoid algorithm and Interior Point methods are both capable

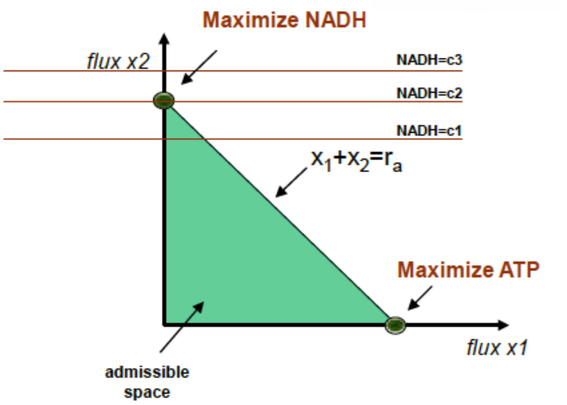


Figure 3: Maximizing two functions with linear programming.

of solving any linear program in polynomial time. It is interesting to note, that many seemingly difficult problems can be modeled as linear programs and solved efficiently (or as efficiently as a generic solution can solve a specific problem).

In microbes such as E. coli, this objective function is often a combination of fluxes that contributes to biomass, as seen in slide 31. However, this function need not be completely biologically meaningful. For example, we might simulate the maximization of mycolates in *M. tuberculosis*, even though this isnt happening biologically. It would give us meaningful predictions about what perturbations could be performed in vitro that would perturb mycolate synthesis even in the absence of the maximization of the production of those metabolites.Flux balance analysis (FBA) was pioneered by Palssons group at UCSD and has since been applied to E. coli, M. tuberculosis, and the human red blood cell [**?**].

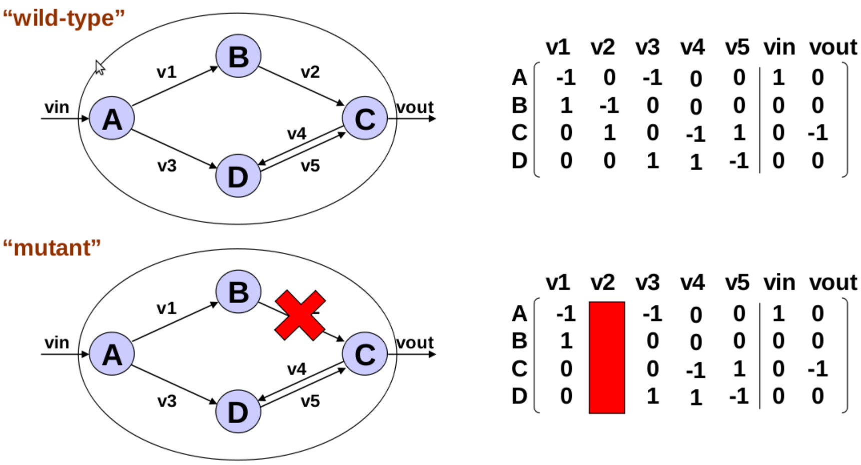
4 Applications

4.1 *In Silico* Detection Analysis

With the availability of such a powerful tool like FBA, more questions naturally arise. For example, are we able to predict gene knockout phenotype based on their simulated effects on metabolism? Also, why would we try to do this, even though other methods, like protein interaction map connective, exist? Such analysis is actually necessary, since other methods do not take into direct consideration the metabolic flux or other specific metabolic conditions.

Knocking out a gene in an experiment is simply modeled by removing one of the columns (reactions) from the stochiometric matrix. (A question during class clarified that a single gene can knock out multiple columns/reactions.) Thereby, these knockout mutations will further constrain the feasible solution space by removing fluxes and their related extreme pathways. If the original optimal flux was outside is outside the new space, then new optimal flux is created. Thus the FBA analysis will produce different solutions. The solution is a maximal growth rate, which may be confirmed or disproven experimentally. The growth rate at the new solution provides a measure of the knockout phenotype. If these gene knockouts are in fact lethal, then the optimal solution will be a growth rate of zero.

Studies by Edwards, Palsson (1900) explore knockout phenotype prediction use to predict metabolic changes in response to knocking out enzymes in E. coli, a prokaryote [**?**]. In other words, an in silico metabolic model of E.coli was constructed to simulate mutations affecting the glycolysis, pentose phosphate, TCA, and electron transport pathways (436 metabolites and 719 reactions included). For each specific condition, the optimal growth of mutants was compared to non-mutants. The in vivo and in silico results were then compared, with 86% agreement. The errors in the model indicate an underdeveloped model (lack of knowledge). The authors discuss 7 errors not modeled by FBA, including mutants inhibiting stable RNA synthesis and producing toxic intermediates.



	v1	v2	v3	v4	v5	vin	vout
A	-1	0	-1	0	0	1	0
B	1	-1	0	0	0	0	0
C	0	1	0	-1	1	0	-1
D	0	0	1	1	-1	0	0

	v1	v2	v3	v4	v5	vin	vout
A	-1	🟥	-1	0	0	1	0
B	1	🟥	0	0	0	0	0
C	0	🟥	0	-1	1	0	-1
D	0	🟥	1	1	-1	0	0

Figure 4: Removing a reaction is the same as removing a gene from the stoichiometric matrix.

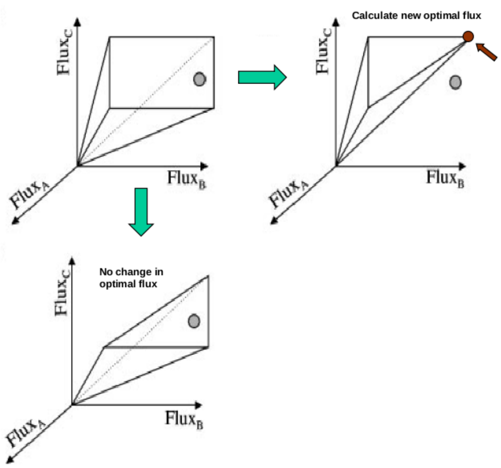


Figure 5: Constraining the feasible solution space may create a new optimal flux.

4.2 Quantitative Flux *In Silico* Model Predictions

Can models quantitatively predict fluxes, growth rate? We demonstrate the ability of FBA to give quantitative predictions about growth rate and reaction fluxes given different envi- ronmental conditions. More specifically, prediction refers to externally measurable fluxes as a function of controlled uptake rates and environmental conditions. Since FBA maximizes an objective function, resulting in a specific value for this function, we should in theory be able to extract quantitative information from the model.

An early example by Edwards, Ibarra, and Palsson (1901), predicted the growth rate of E. coli in culture given a range of fixed uptake rates of oxygen and two carbon sources (acetate and succinate), which they could control in a batch reactor [6]. They assumed that E. coli cells adjust their metabolism to maximize growth (using a growth objective function) under given environmental conditions and used FBA to model the metabolic pathways in the bacterium. The input to this particular model is acetate and oxygen, which is labeled as V_{IN}.

The controlled uptake rates fixed the values of the oxygen and acetate/succinate input fluxes into the network, but the other fluxes were calculated to maximize the value of the growth objective.

The growth rate is still treated as the solution to the FBA analysis. In sum, optimal growth rate is predicted as a function of uptake constraints on oxygen versus acetate and oxygen versus succinate. The basic model is a predictive line and may be confirmed in a bioreactor experimentally by measuring the uptake and growth from batch reactors (note: experimental uptake was not constrained, only measured).

This model by Palsson was the first good proof of principle in silico model. The authors quantitative growth rate predictions under the different conditions matched very closely to the experimentally observed growth rates, implying that E. coli do have a metabolic network that is designed to maximize growth. It had good true positive and true negative rates. The agreement between the predictions and experimental results is very impressive for a model that does not include any kinetic information, only stoichiometry. Prof. Galagan cautioned, however, that it is often difficult to know what good agreement is, because we dont know the significance of the size of the residuals. The organism was grown on a number of different nutrients. Therefore, the investigators were able to predict condition specific growth. Keep in mind this worked, since only certain genes are necessary for some nutrients, like fbp for gluconeogenesis. Therefore, knocking out fbp will only be lethal when there is no glucose in the environment, a specific condition that resulted in a growth solution when analyzed by FBA.

4.3 Quasi Steady State Modeling (QSSM)

We're now able describe how to use FBA to predict time-dependent changes in growth rates and metabolite concentrations using quasi steady state modeling. The previous example used FBA to make quantitative growth predictions under specific environmental conditions (point predictions). Now, after growth and uptake fluxes, we move on to another assumption and type of model.

Can we use a steady state model of metabolism to predict the time-dependent changes in the cell or environments? We do have to make a number of quasi steady state assumptions (QSSA):

1. The metabolism adjusts to the environmental/cellular changes more rapidly than the changes themselves

2. The cellular and environmental concentrations are dynamic, but metabolism operates on the condition that the concentration is static at each time point (steady state model).

Is it possible to use QSSM to predict metabolic dynamics over time? For example, if there is less acetate being taken in on a per cell basis as the culture grows, then the growth rate must slow. But now, QSSA assumptions are applied. That is, in effect, at any given point in time, the organism is in steady state.

What values does one get as a solution to the FBA problem? There are fluxes the growth rate. We are predicting rate and fluxes (solution) where VIN/OUT included. Up to now we assumed that the input and output are infinite sinks and sources. To model substrate/growth dynamics, the analysis is performed a bit differently from prior quantitative flux analysis. We first divide time into slices δt. At each time point t, we use FBA to predict cellular substrate uptake (Su) and growth (g) during interval δt. The QSSA means these predictions are constant over δt. Then we integrate to get the biomass (B) and substrate concentration (Sc) at the next time point $t + \delta t$. Therefore, the new VIN is calculated each time based on points δt in-between time. Thus we can predict the growth rate and glucose and acetate uptake (nutrients available in the environment). The four step analysis is:

1. The concentration at time t is given by the substrate concentration from the last step plus any additional substrate provided to the cell culture by an inflow, such as in a fed batch.

2. The substrate concentration is scaled for time and biomass (X) to determine the substrate availability to the cells. This can exceed the maximum uptake rate of the cells or be less than that number.

3. Use the flux balance model to evaluate the actual substrate uptake rate, which may be more or less than the substrate available as determined by step 2.

4. The concentration for the next time step is then calculated by integrating the standard differential equations:

$$\frac{dB}{dt} = gB \rightarrow B = B_o e^{gt}$$

$$\frac{dSc}{dt} = -SuB \rightarrow Sc = Sc_o \frac{X}{g}(e^{gt} - 1)$$

The additional work by Varma et al. (1994) specifies the glucose uptake rate a priori [17]. The model simulations work to predict time-dependent changes in growth, oxygen uptake, and acetate secretion. This converse model plots uptake rates versus growth, while still achieving comparable results in vivo and in silico. The researchers used quasi steady state modeling to predict the time-dependent profiles of cell growth and metabolite concentrations in batch cultures of *E. coli* that had either a limited initial supply of glucose (left) or a slow continuous glucose supply (right diagram). A great fit is evident.

The diagrams above show the results of the model predictions (solid lines) and compare it to the experimental results (individual points). Thus, in *E. coli*, quasi steady state predictions are impressively accurate even with a model that does not account for any changes in enzyme expression levels over time. However, this model would not be adequate to describe behavior that is known to involve gene regulation. For example, if the cells had been grown on half-glucose/half-lactose medium, the model would not have been able to predict the switch in consumption from one carbon source to another. (This does occur experimentally when *E. coli* activates alternate carbon utilization pathways only in the absence of glucose.)

4.4 Regulation via Boolean Logic

There is a number of levels of regulation through which metabolic flux is controlled at the metabolite, transcriptional, translational, post-translational levels. FBA associated errors may be explained by incorporation of gene regulatory information into the models. One way to do this is Boolean logic. The following table describes if genes for associated enzymes are on or off in presence of certain nutrients (an example of incorporating *E. coli* preferences mentioned above):

ON	ON
no glucose(0)	acetate present(1)
ON	OFF
glucose present(1)	acetate present(1)

Therefore, one may think that the next step to take is to incorporate this fact into the models. For example, if we have glucose in the environment, the acetate processing related genes are off and therefore absent from the S matrix which now becomes dynamic as a result of incorporation of regulation into our model. In the end, our model is not quantitative. The basic regulation then describes that if one nutrient-processing enzyme is on, the other is off. Basically it is a bunch of Boolean logic, based on presence of enzymes, metabolites, genes, etc. These Boolean style assumptions are then used at every small change in time dt to evaluate the growth rate, the fluxes, and such variables. Then, given the predicted fluxes, the V_{IN} ,the V_{OUT} , and the system states, one can use logic to turn genes off and on, effectively a δS per δt. We can start putting together all of the above analyses and come up with a general approach in metabolic modeling. We can tell that if glycolysis is on, then gluconeogenesis must be off.

The first attempt to include regulation in an FBA model was published by Covert, Schilling, and Palsson in 1901 [7]. The researchers incorporated a set of known transcriptional regulatory events into their analysis of a metabolic regulatory network by approximating gene regulation as a Boolean process. A reaction was said to occur or not depending on the presence of both the enzyme and the substrate(s): if either the enzyme that catalyzes the reaction (E) is not expressed or a substrate (A) is not available, the reaction flux will be zero:

$$\text{rxn} = \text{IF (A) AND (E)}$$

Similar Boolean logic determined whether enzymes were expressed or not, depending on the currently expressed genes and the current environmental conditions. For example, transcription of the enzyme (E) occurs only if the appropriate gene (G) is available for transcription and no repressor (B) is present:

$$\text{trans} = \text{IF (G) AND NOT (B)}$$

The authors used these principles to design a Boolean network that inputs the current state of all relevant genes (on or off) and the current state of all metabolites (present or not present), and outputs a binary vector containing the new state of each of these genes and metabolites. The rules of the Boolean network were constructed based on experimentally determined cellular regulatory events. Treating reactions and enzyme/metabolite concentrations as binary variables does not allow for quantitative analysis, but this method can predict qualitative shifts in metabolic fluxes when merged with FBA. Whenever an enzyme is absent, the corresponding column is removed from the FBA reaction matrix, as was described above for knockout phenotype prediction. This leads to an iterative process:

1. Given the initial states of all genes and metabolites, calculate the new states using the Boolean network;

2. perform FBA with appropriate columns deleted from the matrix, based on the states of the enzymes, to determine the new metabolite concentrations;

3. repeat the Boolean network calculation with the new metabolite concentrations; etc. The above model is not quantitative, but rather a pure simulation of turning genes on and off at any particular time instant.

On a few metabolic reactions, there are rules about allowing organism to shift carbon sources (C1, C2).

An application of this method from the study by Covert et al.[7] was to simulate diauxic shift, a shift from metabolizing a preferred carbon source to another carbon source when the preferred source is not available. The modeled process includes two gene products, a regulatory protein RPc1, which senses (is activated by) Carbon 1, and a transport protein Tc2, which transports Carbon 2. If RPc1 is activated by Carbon 1, Tc2 will not be transcribed, since the cell preferentially uses Carbon 1 as a carbon source. If Carbon 1 is not available, the cell will switch to metabolic pathways based on Carbon 2 and will turn on expression of Tc2. The Booleans can represent this information:

$$\text{RPc1} = \text{IF(Carbon1) Tc2} = \text{IF NOT(RPc1)}$$

Covert et al. found that this approach gave predictions about metabolism that matched results from experimentally induced diauxic shift. This diauxic shift is well modeled by the in silico analysis see above figure. In segment A, C1 is used up as a nutrient and there is growth. In segment B, there is no growth as C1 has run out and C2 processing enzymes are not yet made, since genes have not been turned on (or are in the process), thus the delay of constant amount of biomass. In segment C, enzymes for C2 turned on and the biomass increases as growth continues with a new nutrient source. Therefore, if there is no C1, C2 is used up. As C1 runs out, the organism shifts metabolic activity via genetic regulation and begins to take up C2. Regulation predicts diauxie, the use of C1 before C2. Without regulation, the system would grow on both C1 and C2 together to max biomass.

So far we have discussed using this combined FBA-Boolean network approach to model regulation at the transcriptional/translational level, and it will also work for other types of regulation. The main limitation is for slow forms of regulation, since this method assumes that regulatory steps are completed within a single time interval (because the Boolean calculation is done at each FBA time step and does not take into account previous states of the system). This is fine for any forms of regulation that act at least as fast as transcription/translation. For example, phosphorylation of enzymes (an enzyme activation process) is very fast and can be modeled by including the presence of a phosphorylase enzyme in the Boolean network. However, regulation that occurs over longer time scales, such as sequestration of mRNA, is not taken into account by this model. This approach also has a fundamental problem in that it does not allow actual experimental measurements of gene expression levels to be inputted at relevant time points.

We do not need our simulations to artificially predict whether certain genes are on or off. Microarray expression data allows us to determine which genes are being expressed, and this information can be incorporated into our models.

4.5 Coupling Gene Expression with Metabolism

In practice, we do not need to artificially model gene levels, we can measure them. As discussed previousky, it is possible to measure the expressions levels of all the mRNAs in a given sample. Since mRNA expression data correlates with protein expression data, it would be extremely useful to incorporate it into the FBA. Usually, data from microarray experiments is clustered, and unknown genes are hypothesised to have function similar to the function of those known genes with which they cluster. This analysis can be faulty, however, as genes with similar actions may not always cluster together. Incorporating microarray expression data into FBA could allow an alternate method of interpretation of the data. Here arises a question, what is the relationship between gene level and flux through a reaction?

Say the reaction $A \rightarrow B$ is catalyzed by an enzyme. If a lot of A present, increased expression of the gene for the enzyme causes increased reaction rate. Otherwise, increasing gene expression level will not increase reaction rate. However, the enzyme concentration can be treated as a constraint on the maximum possible flux, given that the substrate also has a reasonable physiological limit.

The next step, then, is to relate the mRNA expression level to the enzyme concentration. This is more difficult, since cells have a number of regulatory mechanisms to control protein concentrations independently of mRNA concentrations. For example, translated proteins may require an additional activation step (e.g. phosphorylation), each mRNA molecule may be translated into a variable number of proteins before it is degraded (e.g. by antisense RNAs), the rate of translation from mRNA into protein may be slower than the time intervals considered in each step of FBA, and the protein degradation rate may also be slow. Despite these complications, the mRNA expression levels from microarray experiments are usually taken as upper bounds on the possible enzyme concentrations at each measured time point. Given the above relationship between enzyme concentration and flux, this means that the mRNA expression levels are also upper bounds on the maximum possible fluxes through the reactions catalyzed by their encoded proteins. The validity of this assumption is still being debated, but it has already performed well in FBA analyses and is consistent with recent evidence that cells do control metabolic enzyme levels primarily by adjusting mRNA levels. (In 1907, Professor Galagan discussed a study by Zaslaver et al. (1904) that found that genes required in an amino acid biosynthesis pathway are transcribed sequentially as needed [2]). This is a particularly useful assumption for including microarray expression data in FBA, since FBA makes use of maximum flux values to constrain the flux balance cone.

Colijn et al. address the question of algorithmic integration of expression data and metabolic networks [3]. They apply FBA to model the maximum flux through each reaction in a metabolic network. For example, if microarray data is available from an organism growing on glucose and from an organism growing on acetate, significant regulatory differences will likely be observed between the two datasets. V_{max} tells us what the maximum we can reach. Microarray detects the level of transcripts, and it gives an upper boundary of V_{max}.

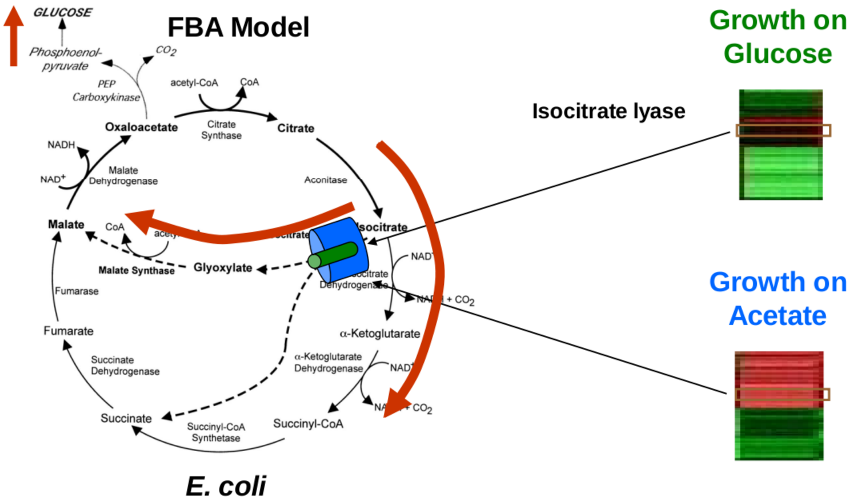


Figure 6: Model of Coljin et. al [3]

In addition to predicting metabolic pathways under different environmental conditions, FBA and microarray experiments can be combined to predict the state of a metabolic system under varying drug treatments. For example, several TB drugs target mycolic acid biosynthesis. Mycolic acid is a major cell wall constituent. In a 1904 paper by Boshoff et al., researchers tested 75 drugs, drug combinations, and growth conditions to see what effect different treatments had on mycolic acid synthesis [9]. In 1905, Raman et al. published an FBA model of mycolic acid biosynthesis, consisting of 197 metabolites and 219 reactions [13].

The basic flow of the prediction was to take a control expression value and a treatment expression value for a particular set of genes, then feed this information into the FBA and measure the final effect on the treatment on the production of mycolic acid. To examine predicted inhibitors and enhancers, they examined significance, which examines whether the effect is due to noise, and specificity, which examines whether the effect is due to mycolic acid or overall supression/enhancement of metabolism. The results were fairly encouraging. Several known mycolic acid inhibitors were identified by the FBA. Interesting results were also found among drugs not specifically known to inhibit mycolic acid synthesis. 4 novel inhibitors and 2 novel enhancers of mycolic acid synthesis were predicted. One particular drug, Triclosan, appears to be an enhancer according to the FBA model, whereas it is currently known as an inhibitor. Further study of this particular drug would be interesting. Experimental testing and validation are currently in progress.

Clustering may also be ineffective in identifying function of various treatments. Predicted inhibitors, and predicted enhancers of mycolic acid synthesis are not clustered together. In addition, no labeled training set is required for FBA-based algorithmic classification, whereas it is necessary for supervised clustering algorithms.

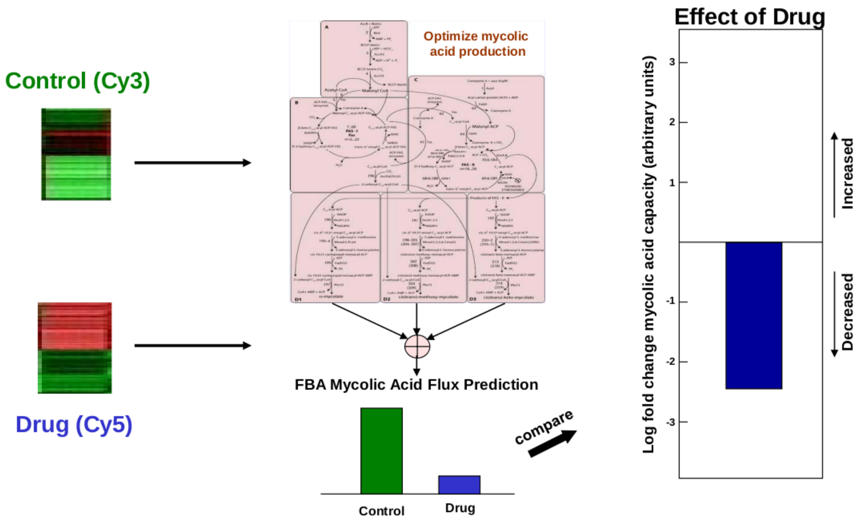

Control (Cy3)
Drug (Cy5)
Optimize mycolic acid production
FBA Mycolic Acid Flux Prediction
Control
Drug
compare
Effect of Drug
Log fold change mycolic acid capacity (arbitrary units)
3
2
1
-1
-2
-3
Increased
Decreased

Figure 7: Basic flow in predicting state of a metabolic system under varing drug treatments

4.6 Predicting Nutrient Source

Now, we get the idea of predicting the nutrient source that an organism may be using in an environment, by looking at expression data and looking for associated nutrient processing gene expression. This is easier, since we cant go into the environment and measure all chemical levels, but we can get expression data rather easily. That is, we try to predict a nutrient source through predictions of metabolic state from expression data, based on the assumption that organisms are likely to adjust metabolic state to available nutrients. The nutrients may then be ranked by how well they match the metabolic states.

The other way around could work too. Can I predict a nutrient given a state? Such predictions could be useful for determining the nutrient requirements of an organism with an unknown natural environment, or for determining how an organism changes its environment. (TB, for example, is able to live within the environment of a macrophage phagolysosome, presumably by altering the environmental conditions in the phagolysosome and preventing its maturation.)

We can use FBA to define a space of possible metabolic states and choose one. The basic steps are to:

- Start with max flux cone (representing best growth with all nutrients available in environment). Find optimal flux for each nutrient.
- Apply expression data set (still not knowing nutrient). This will allow you to constrain the cone shape and figure out the nutrient, which is represented as one with the closest distance to optimal solution.

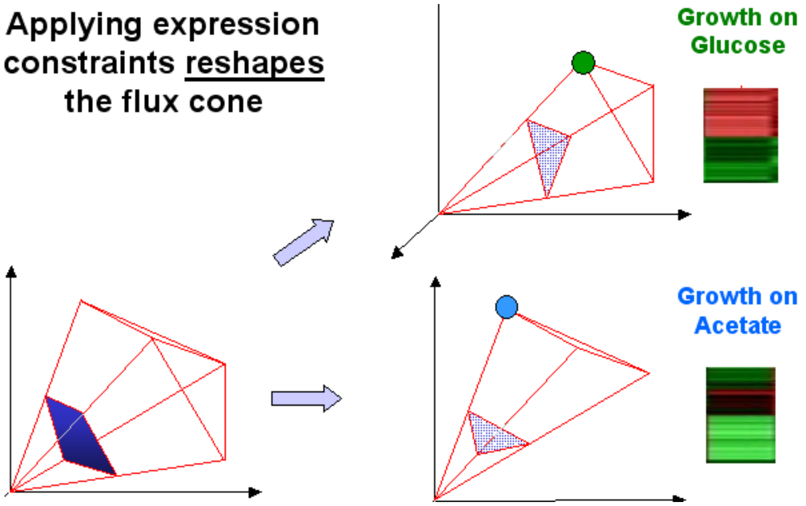


Figure 8: Applying expression data set allows constraining of cone shape.

In Figure 8, you may see that the first cone has a number of optimals, so the real nutrient is unknown. However, after expression data is applied, the cone is reshaped. It has only one optimal, which is still in feasible space and thereby must be that nutrient you are looking for.

As before, the measured expression levels provide constraints on the reaction fluxes, altering the shape of the flux-balance cone (now the expression-constrained flux balance cone). FBA can be used to determine the optimal set of fluxes that maximize growth within these expression constraints, and this set of fluxes can be compared to experimentally-determined optimal growth patterns under each environmental condition of interest. The difference between the calculated state of the organism and the optimal state under each condition is a measure of how sub-optimal the current metabolic state of the organism would be if it were in fact growing under that condition.

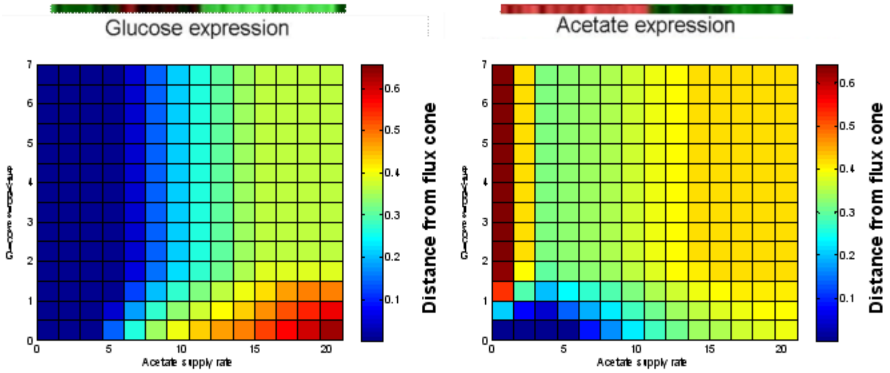


Figure 9: Results of nutrient source prediction experiment.

Expression data from growth and metabolism may then be applied to predict the carbon source being used. For example, consider E. coli nutrient product. We can simulate this system for glucose versus acetate. The color indicates the distance from the constrained flux cone to the optimal flux solution for that nutrient combo (same procedure described above). Then, multiple nutrients may be ranked, prioritized according to expression data. Unpublished data from Desmond Lun and Aaron Brandes provide an example of this approach.

They used FBA to predict which nutrient source E. coli cultures were growing on, based on gene expression data. They compared the known optimal fluxes (the optimal point in flux space) for each nutrient condition to the allowed optimal flux values within the expression-constrained flux-balance cone. Those nutrient conditions with optimal fluxes that remained within (or closest to) the expression-constrained cone were the most likely possibilities for the actual environment of the culture.

Results of the experiment are shown in Figure 9, where each square in the results matrices is colored based on the distance between the optimal fluxes for that nutrient condition and the calculated optimal fluxes based on the expression data. Red values indicate large distances from the expression-constrained flux cone and blue values indicate short distances from the cone. In the glucose-acetate experiments, for example, the results of the experiment on the left indicate that low acetate conditions are the most likely (and glucose was the nutrient in the culture) and the results of the experiment on the right indicate that low glucose/medium acetate conditions are the most likely (and acetate was the nutrient in the culture). When 6 possible nutrients were considered, the model always predicted the correct one, and when 18 possible nutrients were considered, the correct one was always one of the top 4 ranking predictions. These results suggest that it is possible to use expression data and FBA modeling to predict environmental conditions from information about the metabolic state of an organism.

This is important because TB uses fatty acids in macrophages in immune systems. We do not know which ones exactly are utilized. We can figure out what the TB sees in its environment as a food source and proliferation factor by analyzing what related nutrient processing genes are turned on at growth phases and such. Thereby we can figure out the nutrients it needs to grow, allowing for a potential way to kill it off by not supplying such nutrients or knocking out those particular genes.

It is easier to get expression data to see flux activity than see whats being used up in the environment by analyzing the chemistry on such a tiny level. Also, we might not be able to grow some bacteria in lab, but we can solve the problem by getting the expression data from the bacteria growing in a natural environment and then seeing what it is using to grow. Then, we can add it to the laboratory medium to grow the bacteria successfully.

5 Current Research Directions

6 Further Reading

- Becker, S. A. and B. O. Palsson (1908). Context-Specific Metabolic Networks Are Consistent with Experiments. PLoS Computational Biology 4(5): e1000082.
  - If gene expression lower than some threshold, turn the gene off in the model.
- Shlomi, T., M. N. Cabili, et al. (1908). Network-based prediction of human tissue-specific metabolism. Nat Biotech 26(9): 1003-1010.
  - Nested optimization problem.
  - First, standard FBA
  - Second, maximize the number of enzymes whose predicted flux activity is *consistent with their measured expression level*

7 Tools and Techniques

- Kegg
- BioCyc
- Pathway Explorer (pathwayexplorer.genome.tugraz.at)
- Palssons group at UCSD (http://gcrg.ucsd.edu/)
- www.systems-biology.org
- Biomodels database (www.ebi.ac.uk/biomodels/)
- JWS Model Database (jjj.biochem.sun.ac.za/database/index.html)

8 What Have We Learned?

References

[1]

[2] Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, and Alon U. Just-in-time transcription program in metabolic pathways. *Nat. Gen*, 36:486–491, 2004.

[3] Caroline Coljin. Interpreting expression data with metabolic flux models: Predicting mycobacterium tuberculosis mycolic acid production. *PLoS Computational Biology*, 5(8), Aug 2009.

[4] Price N. D., Reed J. L., Papin J.A, Famili I., and Palsson B.O. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys J.*, 84(2):794–804, Feb 2003.

[5] Gasteiger E., Gattiker A., Hoogland C. andIvanyi I., Appel R.D., , and Bairoch A. Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*, 31(13):3784–3788.

[6] J.S. Edwards, R. U. Ibarra, and B.O. Palsson. In silico predictions of e coli metabolic capabilities are consis ent with experimental data. *Nat Biotechnology*, 19:125–130, 2001.

[7] Covert M et al. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213:73–88, Nov 2001.

[8] J. Forster, I. Famili, B.O. Palsson, and J. Nielsen. Large-scale evaluation of in silico gene deletions in saccharomyces cerevisiae. *OMICS*, 7(2):193–202, 2003. PMID: 14506848.

[9] Boshoff H.I., Myers T.G., Copp B.R., McNeil M.R., Wilson M.A., and Bary C.E. The transcriptional response of mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. *J Biol Chem*, 279:40174–40184, Sep 2004.

[10] Holmberg. On the practical identifiability of microbial-growth models incorporating michaelis-menten type nonlinearities. *Mathematical Biosciences*, 62(1):23–43, 1982.

[11] Edwards J.S. and Palsson B.O. volume 97, pages 5528–5533. Proceedings of the National Academy of Sciences of the United States of America, May 2000. PMC25862.

[12] Edwards J.S., Covert M., , and Palsson B. Metabolic modeling of microbes: the flux balance approach. *Environmental Microbiology*, 4(3):133–140, 2002.

[13] Raman Karthik, Preethi Rajagopalan, and Nagasuma Chandra. Flux balance analysis of mycolic acid pathway: Targets for anti-tubercular drugs. *PLoS Computational Biology*, 1, Oct 2005.

[14] Kanehisa M., Goto S., Kawashima S., and Nakaya. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34, 2006.

[15] Jamshidi N. and Palsson B. Investigating the metabolic capabilities of mycobacterium tuberculosis h37rv using the in silico strain inj661 and proposing alternative drug targets. *BMC Systems Biology*, 26, 2007.

[16] Caspi R., Foerster H., Fulcher C.A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S.Y., Shearer A.G., Tissier C., Walk T.C. ZhangP., and Karp P. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 36(Suppl), 2008.

[17] A. Varma and B. O. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and Environmental Microbiology*, 60:3724–3731, Oct 1994.