6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

6.047/6.878 Lecture 22: Metabolic Modeling 2

November 20, 2008

1 Review

In the last lecture, we discussed how to use linear algebra to model metabolic networks with flux-balance analysis (FBA), which depends only on the stoichiometry of the reactions, not the kinetics. The metabolic network is represented as an n × m matrix M whose columns are the m reactions occurring in the network and whose rows are the n products and reactants of these reactions. The entry M(i, j) represents the relative amount of metabolite i consumed or produced by reaction j. A positive value indicates production; a negative value indicates consumption. The nullspace of this matrix consists of all the m × 1 reaction flux vectors that are possible given that the metabolic system is in steady state; i.e. all the sets of fluxes that do not change the metabolite concentrations. The nullspace ensures that the system is in steady state, but it there are also additional constraints in biological systems: fluxes cannot be infinite, and each reaction can only travel in the forward direction (in cases where the backward reaction is also biologically possible, it is included as a separate column in the matrix). After bounding the fluxes and constraining the directions of the reactions, the resulting space of possible flux vectors is called the constrained flux-balance cone. The edges of the cone are known as its extreme pathways. By choosing an objective function—some linear combination of the fluxes—to maximize, we can calculate the optimal values for the fluxes using linear programming and the simplex algorithm. For example, the objective function may be a weighted sum of all the metabolite fluxes that represents the overall growth rate of the cell (growth objective). Alternatively, we can maximize use of one particular product by find the m × 1 flux vector that is in the flux-balance cone and has the largest negative value of that product.

At the end of the last lecture, we also discussed knockout phenotype predictions, a classic application of metabolic modeling. Experimental biologists often gather information about the function of a protein by generating transgenic organisms in which the gene encoding that protein has been disrupted, or knocked out. We can simulate in silico the effect of knocking out a particular enzyme on metabolism by assuming that the reaction catalyzed by that enzyme does not occur at all in the knockout: zeroing out the j^{th} column of M effectively removes the j^{th} reaction from the network. What used to be an optimal solution may now lie outside of the constrained flux-balance cone and thus no longer be feasible in the absence of the j^{th} reaction. We can determine the new constrained flux-balance cone and calculate the new optimum set of fluxes to maximize our objective function, predicting the effect of the knockout on metabolism.

2 Knockout Phenotype Prediction in Eukaryotes

(Forester *et al*, 2003; Famili *et al*, 2003)

The last lecture gave an example of using knockout phenotype prediction to predict metabolic changes in response to knocking out enzymes in *E. coli*, a prokaryote. Eukaryotic cells, including animal cells, have more complex organization and more complex gene regulation and gene expression pathways. In this lecture, a recent attempt to perform similar knockout phenotype prediction in yeast, a eukaryote, was presented (Forster *et al*, 2003 and Famili *et al*, 2003). The authors predicted whether a variety of enzyme knockouts would be able to grow under a few different environmental conditions and compared the predictions to experimental results. They achieved 81.5% agreement between their predictions and experiment, but looking at the data more closely reveals that all of the discrepancies were false positives, in which the yeast were predicted to grow but did not. In fact, the model predicted that almost every knockout would grow. While not bad, these results were not as compelling as those for the prokaryotic case, highlighting the need to incorporate the ability of cells to regulate gene expression in response to environmental and metabolic changes into flux-balance analysis of eukaryotes.

3 Overview of Today's Material: Extensions to FBA

Today we discuss a number of extensions to flux-balance analysis that provide additional predictive power. First, we demonstrate the ability of FBA to give quantitative predictions about growth rate and reaction fluxes given different environmental conditions. We then describe how to use FBA to predict time-dependent changes in growth rates and metabolite concentrations using quasi steady state modeling. Next, we discuss two approaches for taking gene expression changes into account in the FBA model: by building the rules of gene regulation into a Boolean network or by using available expression data from microarray experiments to constrain the flux-balance cone. Finally, we provide an example of using expression data to predict the state of the environment from the metabolic state, rather than the other way around.

4 Quantitative Flux Prediction

(Edwards, Ibarra, & Palsson, 2001)

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 of this was done by Edwards, Ibarra, and Palsson (2001), who 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. 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 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 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. 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 don't know the significance of the size of the residuals.

5 Quasi Steady State Modeling

(Varma & Palsson, 1994)

The previous example used FBA to make quantitative growth predictions under specific environmental conditions (point predictions), but is it also possible to predict time-dependent changes in response to varying environmental or cellular conditions? Varma and Palsson (1994) demonstrate the use of FBA to predict time-dependent changes in *E. coli* metabolism. Their quasi steady state model is based on the idea that time-dependent changes in fluxes can be modeled as transitions through individual time points that are themselves in steady state. In other words, quasi steady state modeling divides time into differential slices of size Δt and assumes that the state of the system is constant within each interval. This assumption requires that metabolism can adjust to changes more rapidly than the changes occur. To model a time profile of the system under the quasi steady state assumption, we use FBA to calculate the fluxes within each time interval. Because fluxes represent the derivatives of the metabolite concentrations, we can assume that the derivatives are constant over each Δt and integrate to find the starting metabolite concentrations at the next time interval. So quasi steady state modeling is an iterative process in which we calculate the optimal fluxes and growth rate for the system at one time point, and then use those optimal fluxes to derive the initial environmental conditions for the next time point.

Varma and Palsson (1994) first predicted growth rate, oxygen uptake, and acetate secretion for specified glucose uptake rates, using quantitative flux prediction as in the previous example. Their FBA model incorporated experimentally-determined values of maximum oxygen utilization rate, maximum aerobic glucose utilization rate, maximum anaerobic glucose utilization rate, and growth- and non-growth-associated maintenance requirements. Their predictions were again found to be very similar to experimental results.

The researchers then 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 or a slow continuous glucose supply. They predicted available glucose concentration and acetate secretion over time, and their predictions again matched very well with experimental measurements: their model anticipated both the smooth behavior (e.g. decreasing glucose concentration, increasing acetate secretion over time) and also sudden transitions (e.g. the sudden decrease in acetate secretion when available glucose concentration reached zero). 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 that occurs experimentally when *E. coli* activates alternate carbon utilization pathways only in the absence of glucose.

6 Incorporating Regulation into Metabolic Models

Metabolic pathways are regulated at many different levels: the metabolite itself can be regulated, while transcriptional, translational, and post-translational regulation control the availability of active enzyme. These regulatory processes have been studied extensively by biologists, and many of the discrepancies between FBA predictions and experimental results can be explained by using existing knowledge about gene regulation. As stated in Covert *et al* (2001):

"...FBA leads to incorrect predictions in situations where regulatory effects are a dominant influence on the behavior of the organism. ... Thus, there is a need to include regulatory events within FBA to broaden its scope and predictive capabilities."

Incorporating known regulatory information into metabolic models in order to improve prediction is an important area of current research.

6.1 Regulation as Boolean Logic
(Covert, Schilling, & Palsson, 2001)

The first attempt to include regulation in an FBA model was published by Covert, Schilling, and Palsson in 2001. 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 does or does not occur 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 can be used to determine whether enzymes will be expressed or not, depending on the currently expressed genes and the current environmental conditions. For example, transcription of the enzyme (E) might only occur if the appropriate gene (G) is available for transcription and if a repressor (B) is not 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.

An application of this method from the study by Covert *et al* 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. This information can be represented by the Booleans:

RPc1 = IF (Carbon1)
tTc2 = IF NOT (RPc1)

Covert *et al* found that this approach gave predictions about metabolism that matched results from experimentally-induced diauxic shift.

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. Given recent experimental advances, it has become very easy to measure the expression of large numbers of genes using microarrays, and we no longer need to depend on rules to calculate gene expression levels over time.

6.2 Modeling Metabolism with Expression Data

As discussed previously in lecture, it is now possible to measure mRNA levels for thousands of genes at a time by microarray experiments. Historically, data from microarray experiments have been analyzed by clustering, and unknown genes are hypothesized to function similarly to known genes within the same cluster. This analysis can be faulty, however, as genes with similar functions may not always cluster together. Incorporating microarray expression data into FBA provides another way of interpreting the data.

The key to this approach is to determine how the mRNA expression level correlates with the flux through a reaction catalyzed by the encoded enzyme. For the reaction:

$$S + E \longleftrightarrow SE \longrightarrow E + P$$

with substrate S, product P, and enzyme E, the concentration of the enzyme directly determines the flux only if there is an excess of substrate, i.e. the reaction is enzyme-limited. By Michaelis-Menten kinetics, the flux is given by:

$$v = \frac{v_{max}[S]}{[S] + K_m} = \frac{k_c[E_{tot}][S]}{[S] + K_m}$$

where K_m is the concentration of substrate S that gives a flux equal to half the maximum flux, v_{max}, and $[E_{tot}]$ is the concentration of available enzyme. The specific flux depends both on the concentration of enzyme and the concentration of substrate. As a simple example of the importance of substrate concentration in reaction systems, consider a system of two enzyme-catalyzed reactions:

$$S_1 \xrightarrow{[E_1]=0} S_2 \xrightarrow{[E_2]=10} P$$

Even though enzyme E_2 is present, the flux through the second reaction is zero because no S_2 can be produced in the absence of E_1. So the enzyme concentration cannot simply be taken as proportional to the reaction flux in our model of the system. However, the enzyme concentration can be treated as a *constraint* on the maximum possible flux, given that [S] 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 (Prof. Galagan referred students to the lecture notes from last year, when he discussed a study by Zaslaver *et al* (2004) that found that genes required in an amino acid biosynthesis pathway are transcribed sequentially as needed). 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.

Unpublished data from Caroline Colijn, Aaron Brandes, and Jeremy Zucker provide an example of using mRNA expression levels as constraints on maximum flux values. Microarray data from *E. coli* growing on two different carbon sources, glucose and acetate, show significant differences in gene expression between the two conditions. For example, the glyoxylate shunt is a pathway that normally has very low flux but that is required when using carbon sources with only 2 carbons instead of 6. During growth on acetate, expression of the glyoxylate shunt enzymes was upregulated 20-fold and the flux through this pathway significantly increased in experimental measurements. Using the expression levels as flux constraints in an FBA model resulted in calculated flux values that predicted a similar increase in the glyoxylate shunt pathway under acetate growth conditions.

Prof. Galagan's group has been using this combined FBA and microarray data approach to predict the state of metabolic pathways in the tuberculosis (TB) bacterium under various drug treatments. For example, several TB drugs target the biosynthesis of mycolic acid, a cell wall constituent that is not present in animal cells. In 2005, Raman *et al* published an FBA model of mycolic acid biosynthesis, consisting of 197 metabolites and 219 reactions. Microarray expression data is also available for thousands of TB genes in the presence of 75 different drugs, drug combinations, and growth conditions (published in Boshoff *et al*, 2004). Prof. Galagan's group combined these expression data with the published FBA model to predict the effect of each tested condition on mycolic acid synthesis. Their approach was to use mycolic acid production as the objective function in the FBA algorithm, in order to calculate the maximum possible amount of mycolic acid that the TB bacterium could produce under each condition. The experimentally-determined mRNA expression levels were used to constrain the maximum reaction fluxes, and the FBA results from each experimental condition were compared to the results from a control condition in the absence of drug to determine the change in mycolic acid synthesis attributable to the presence of each drug. The group then did additional calculations in order to characterize the significance and specificity of their results. To determine the significance of the size of the calculated changes in mycolic acid synthesis, they repeated the same analysis on each of the different control conditions (different growth conditions but all in the absence of drugs) to see how mycolic acid synthesis varies as a result of normal variations in environmental conditions. The dotted lines in the "Significance and Specificity" figure from lecture represent the 95% confidence interval for normal mycolic acid variation that is not due to the effect of drugs. In addition, to determine whether observed effects were specific to the 30 genes in the mycolic acid biosynthesis pathway versus whether the drugs affected cellular metabolism as a whole, the group repeated the calculations using the expression levels of other sets of 30 randomly-selected genes from the microarray experiment. If the same effect was observed for randomly-selected genes as for the mycolic acid pathway genes, then the drug must not be specific for mycolic acid biosynthesis. The blue 95% error bars in

the "Significance and Specificity" figure from lecture represent the specificity of the effect for mycolic acid biosynthesis.

The results of this approach were encouraging. Of the 7 known mycolic acid inhibitors, 6 were correctly identified as inhibitors by the model. Their specificity was also predicted correctly (for example, PA-824 was correctly predicted as a non-specific inhibitor). Interestingly, the 7th known inhibitor, triclosan, was also identified by the model but was predicted to be an enhancer rather than an inhibitor, suggesting that additional studies of this drug may reveal multiple effects under different conditions. 4 novel inhibitors and 3 novel enhancers of mycolic acid synthesis were also predicted by the model.

One could argue that predictions about the effects of these drugs on mycolic acid biosynthesis could have been made on the basis of the expression data alone, without using the FBA model. However, there are three reasons why the combined FBA approach is preferable. First, the known inhibitors of mycolic acid biosynthesis have different mechanisms of action and don't cluster together when the microarray data are clustered by traditional methods, suggesting that expression data alone would not give accurate predictions about their functions. FBA also has the advantage of not requiring a labeled training set; in this case, a training set for enhancers of mycolic acid biosynthesis was not available, since there are no known enhancers. Finally, the FBA model contains additional information and can answer more questions than clustering analysis; in this case, for example, it provided information on the specificity and strength of each drug's effect on mycolic acid biosynthesis.

7 Predicting Nutrient Source

The examples above have used modeling to predict the metabolic state of an organism given known environmental conditions. But now that we can obtain information about the metabolic state of an organism from microarray expression data, we can imagine the converse: using modeling to predict the state of the environment given a known metabolic 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). 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.

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 the lecture slides, 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 correct one was always accurately predicted by the model, 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.

8 Summary and Additional Approaches

This lecture has taken us from simple flux-balance analysis, which gave steady-state reaction flux predictions based on the stochiometry of metabolic reactions, through improvements that allow quantitative and time-dependent predictions, to incorporation of gene expression changes that now allow modeling of complex cellular behaviors such as carbon source switching. This is an active area of research and many other improvements are possible. For example, Palsson's group has recently incorporated microarray data into the Boolean model of gene regulation by turning genes off in the Boolean network whenever their measured expression levels fall below some threshold value. Another group recently used an FBA approach that maximizes the number of reactions in the metabolic network whose fluxes match the measured expression levels of the corresponding enzymes; in other words, the number of reactions whose enzymes are not expressed and have zero flux plus the number of reactions whose enzymes are expressed and have significant flux. The class suggested that a new approach to identifying drug targets in a pathogen could be to calculate in silico which enzyme and reaction, if knocked out, would have the largest negative effect on flux through a pathway of interest in a pathogen. The converse of this is the synthetic biology approach: trying to add reactions into a network (first in silico, by searching the space of possible models) to see which additional reaction would most improve flux through a pathway of interest.

9 References

Boshoff, H.I.M., Myers, T.G., Copp, B.R., McNeil, M.R., Wilson, M.A., and Barry, C.E.III. "The Transcriptional Responses of *Mycobacterium tuberculosis* to Inhibitors of Metabolism." *J Biol Chem* 279(38):40174-40184 (2004).

Covert, M., Schilling, C. H., and Palsson, B.Ø. "Regulation of Gene Expression in Flux Balance Models of Metabolism." *Journal of Theoretical Biology* 213(1):73-88 (2001).

Edwards, J.S., Ibarra, R.U., and Palsson, B.Ø. "In silico predictions of Escherichi coli metabolic capabilities are consistent with experimental data." *Nature Biotechnology* 19:125-130 (2001).

Famili,I., Förster, J., Nielsen, J., and Palsson, B.Ø. "Saccharomyces cerevisiae Phenotypes can be Predicted using Constraint-based Analysis of a Genome-scale Reconstructed Metabolic Network." *PNAS*, 100:13134-13139 (2003).

Förster, J., Famili, I., Fu, P., Palsson, B.Ø., and Nielsen, J. "Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network." *Genome Research*, 13(2):244-253 (2003).

Raman, K., Rajagopalan, P., and Chandra, N. "Flux Balance Analysis of Mycolic Acid Pathway: Targets for Anti-Tubercular Drugs." *PLoS Comput Biol* 1(5):e46 (2005).

Varma, A. and Palsson, B.Ø. "Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild Type Escherichia coli W3110." *Appl Environ Microbiol* 60(10):3724-3731 (1994).

Zaslaver, A., Mayo, A.E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M.G., and Alon, U. "Just-in-time transcription program in metabolic pathways." *Nat Genet* 36(5):486-91 (2004).