Introduction

A popular approach to studying cancer metabolomics is to profile the abundances of certain metabolites in a panel of cancer tissue samples. An open problem is exactly *what to do* with this data, i.e. how it may be used to infer the flux through a given pathway. Below, we describe a possible approach, based on elementary flux modes.

Method

What we know

Assume that we have a relatively small metabolic network which we are particularly interested in interrogating. The obvious choice here is central carbon metabolism, including glycolysis, the TCA cycle, and some peripheral pathways like 1C metabolism. We can describe such a network using the conventional stoichiometric matrix **S**, an $m \times n$ matrix of m metabolites and n reactions. Actually, **S** is one of two pieces of data we have *a priori*, the other being the metabolomics data itself, which we can store in a big matrix we will call **D** for data. Let **D** have dimension $m \times s$, where s is the number of samples.

What don't we know

Let's suppose that we can decompose S into a finite and relatively small set of elementary pathways called flux modes, which capture the spectrum of routes that flux may flow. Calculating these flux modes is a standard problem, and is easily accomplished (for small networks). Then, we can write a matrix:

$$\mathbf{F_{f,n}} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{f,1} & a_{m,2} & \cdots & a_{f,n} \end{bmatrix}$$

whose dimension is $f \times n$, with f total elementary flux modes and n total reactions.

Somehow, we need to connect the metabolites that we measure to the flux modes. One way to do this is to appeal to our knowledge of enzyme kinetics. Typically, for a given reaction, the flux v is monotonically dependent on two factors: the amount of catalyzing enzyme, and the amount of substrate. This assumes there is no allosteric regulation, and that the reaction is irreversible. If we can swallow these assumptions, let's further assume that since we do not have data on enzyme abundances/expression, that any changes in flux we observe are determined by changes in metabolite concentration (NB: when people use gene expression data to infer fluxes, they are actually making the converse assumption, that metabolite concentrations contribute nothing).

OK, if we can get to this point, let's make one final assumption: that flux through a reaction is linearly correlated to substrate concentration. This is analogous to saying the enzyme is unsaturated (yes, this is wrong, but again, we have no data and this makes our life easier). If we can do this, let's write one more matrix called **C** for the "catalysis matrix", which has dimension $n \times m$ (# reactions times # of metabolites) which captures the relationship between each reaction and the metabolites which catalyze it:

$$\mathbf{C_{i,j}} = \begin{cases} 1 & S_{i,j} < 0 \\ 0 & \text{otherwise} \end{cases}$$

That is the last piece of information that we need. We would like to get a matrix telling us the activity of each flux mode per sample, let's call this **R** for the "result." Here's how to calculate **R**:

$$\mathbf{R_{f,s}} = \mathbf{F_{f,n}} \times \mathbf{C_{n,m}} \times \mathbf{D_{m,s}}$$

where I have written out the dimensions of each matrix for clarity.