Steady States in Metabolism

Eric Lee and Madhusudan Manjunath

This is a term paper written for the course "Non-Linear Algebra" by Prof. Sturmfels in Spring 2014.

1 Introduction

Chemical Reactions networks are widely studied in a variety of fields, including Systems Biology, Medicine, and Chemical Engineering. Analyzing chemical reaction networks is a widely studied problem going back to Guldberg and Waage in the 19th century. Despite being well-studied, many chemical reaction networks still aren't very well understood, partly due to the complexity of naturally occuring reactions that are of interest to researchers. In this respect, creating accurate mathematical models that take physical, chemical, and environmental constraints into account and at the same time amenable for analysis is important. Keeping this in mind, we undertook a detailed study of metabolism, a key chemical reaction network in the human body.

1.1 Chemical Reaction Network Theory

We start with a introduction to the model that we will be working with. A chemical reaction network X has n chemical complexes and k rate constants and m chemical species. Here, a chemical complex is any set of reactants or products, a rate constant k_{ij} is the rate of reaction from chemical complex n_i to chemical complex n_j, and a chemical species is a molecule in the chemical reaction network associated to some chemical complex n_i. The network can be viewed as a weighted, directed graph with n nodes and k edges.
For example, if this chemical equation represents a hypothetical chemical reaction network

$$\mathrm{A} + 2\,\mathrm{B} \longleftrightarrow \mathrm{B} + \mathrm{C} \tag{1}$$

We then separate this reaction into its two components

$$\mathrm{A} + 2\,\mathrm{B} \xrightarrow{\mathrm{K_f}} \mathrm{B} + \mathrm{C}$$
$$\mathrm{A} + 2\,\mathrm{B} \xleftarrow{\mathrm{K_b}} \mathrm{B} + \mathrm{C}$$

Then we get a graph looking like:

This graph has two nodes and two edges, corresponding to two chemical complexes (A+2 B and B+C) and the forward/backward rate constants respectively. Moreover, it contains three chemical species, A, B, and C. Not the distinction between a chemical complex and a chemical species. K_f and K_b are the forward and backwards rate constants respectively (fixed scalars). Note that if a chemical reaction isn't reversible, we set K_b to 0.

Given this transformation of a chemical reaction network into a directed, weighted graph, the model that we will be using, which is introduced in any undergraduate chemistry course, stems from the *Law of Mass Action*. It views the chemical reaction network as a *dynamical system* i.e. the concentrations of the chemical species in the reaction network are functions of time. The Law of Mass Action states that this the derivative of function must be proportional to the rate constants and the concentration of the reactants and products respectively. We refer to the running example (1) for a better understanding of the model, from which we get the following differential equations:

$$\frac{d}{dt}[A] = -K_f[A]^2[B] + K_b[B][C]$$
$$\frac{d}{dt}[B] = -K_f[A]^2[B] + K_b[B][C]$$
$$\frac{d}{dt}[C] = +K_f[A]^2[B] - K_b[B][C]$$

*note that $[x]$ means the chemical concentration of a chemical x

Looking at the derivative of A, we see that is is equal to the difference between the product of the backward rate constant multiplied by the weight on the graph's right node and the product of the forward rate constant multiplied by the weight on the graph's left node. A way of understanding why these equations is the following: standing at the left node and reacting forward, we lost a molecule of A. However, standing at the right node and reacting backward, we gain a molecule of A, hence the differential equation $\frac{d}{dt}[A] = -K_f[A]^2[B] + K_b[B][C]$. So you lose chemical concentration $-K_f[A]^2[B]$ reacting forward, and you gain chemical concentration $K_b[B][C]$ reacting forward. The net change is then our derivative. The same reasoning holds for the other two differential equations we get.

Generalizing this for all chemical reaction networks, we say that for any chemical reaction network R with n chemical complexes, k rate constants and m chemical species, we can create its graph G =graph(R) with n nodes and k edges.

By The Law of Mass Action, this graph G represents a system of m differential equations, which we call X. This system of differential equations

$$X(C_1...C_m) \tag{2}$$

is a system of polynomial differential equations in $\{C_1....C_m\}$, where C_i is a the chemical concentration of the ith chemical species in the reaction network. This transformation via law of mass action of a dynamical system to a system of polynomial differential equations will be the bridge between chemical reaction network theory and the non-linear algebra and real algebraic geometry we learned in this course.

2 Motivation and Problem Statement

The concentration of each reactant when the chemical reaction system is in equilibrium is called the steady state of a biological network. A study state occurs when $X(C_1...C_m) = 0$, or in other words, when the rate of change of each chemical concentration is 0. A method of finding a steady states has been suggested by Anne Shiu [1]. Shiu asserts that if one is attempting to solve for the steady state, for all intents and purposes, they need not consider the system of polynomial differential equations dependent on time. In other words, Shiu considers (2) as simply a system of polynomials. In Shiu's model, the set of steady states is a real algebraic variety cut out by (2). Because (2) is a dynamical system, it contains some initial condition that corresponds to some steady state in the real algebraic variety. We decided to apply Shiu's method of to an important chemical reaction network.

Madhu and I met with Dave Savage, a chemistry professor who runs the Savage Lab in Berkeley's biochemistry department, and his student, Avi Flamholtz. They suggested that we look at Metabolism, a fundamental reaction in the human body. A key assumption in the approach of Shiu [1] is that the reaction constants k_i of a chemical reaction network are real numbers i.e. are known a priori. However, Dave and Avi revealed that in natural chemical reaction networks, for instance, those modelling metabolism, reaction constants are not known a priori and that there are no reliable experimental methods to determine the reaction constants precisely. What is known experimentally are lower and upper bounds on the rate constants.

Following Dave and Avi's suggestion, we propose the following variant to Shiu's approach: we regard the reactions constants as bounded parameters, in contrast to the model of Shiu where the reaction constants are regarded as real numbers.

More precisely, whereas [1] considers rate constants

$$K_i$$

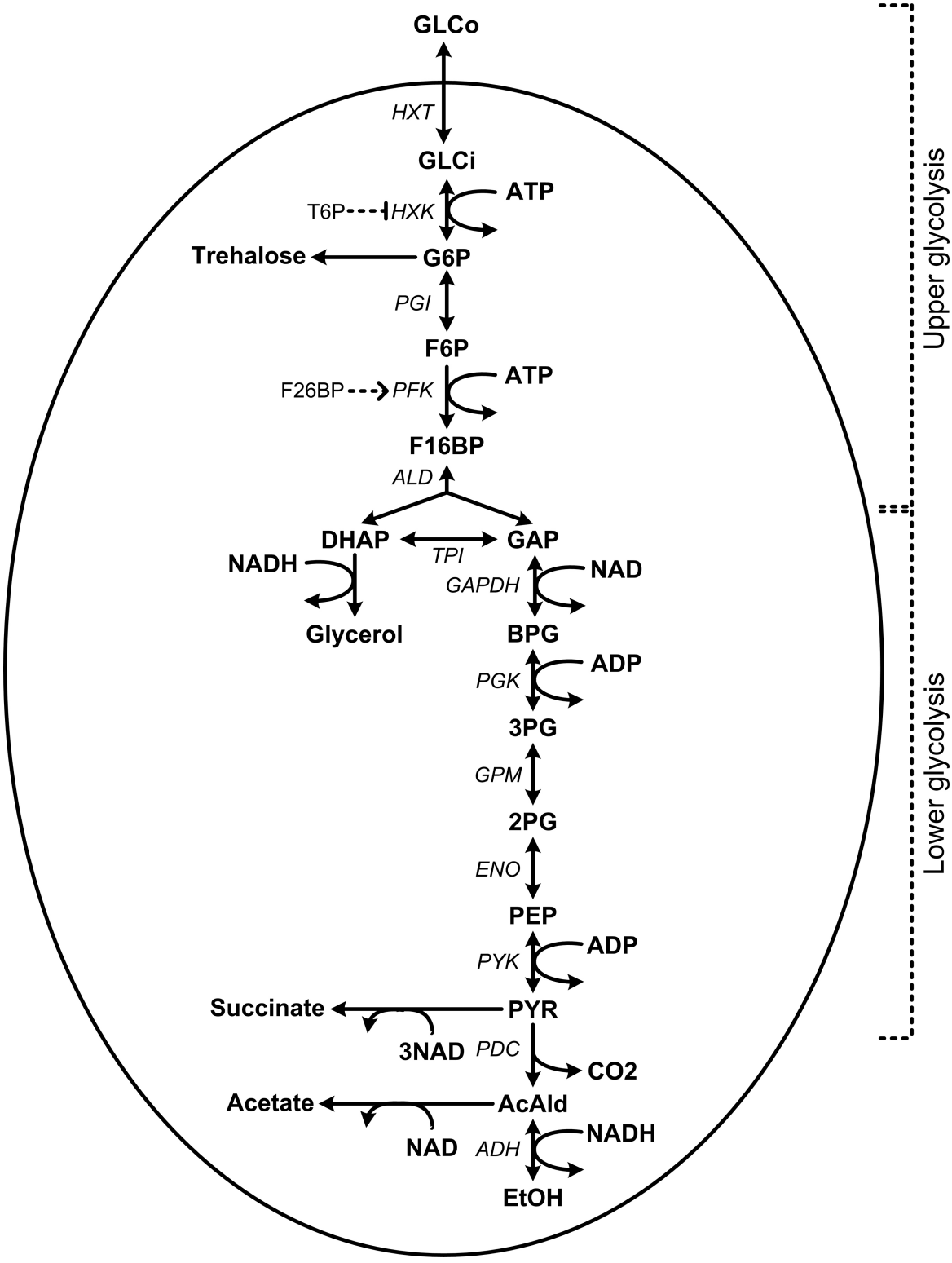
as fixed scalars, we say that

$$L_i \leq K_i \leq U_i$$

where $L_i, U_i \in \mathbb{R}^+$ represent lower and upper bounds to the rate constants respectively. We want to apply this modification of [1] to Glycosis, the most important sub-network of metabolism which has a well-known steady states measured by experimental chemists, thereby allowing us to compare the steady state we calculate to the experimentally determined steady state.

3 Metabolism as a Chemical Reaction Network

We now undertake a detailed study of Glycosis using our modification of [1]. Glycosis is a chemical reaction system that converts Glucose into pyruvate to ATP and NADH, which are the molecules that "power" the human body. The graph of Glycosis is given below:


GLCo
HXT
GLCi
ATP
T6P
HXK
Trehalose
G6P
PGI
F6P
ATP
F26BP
PFK
F16BP
ALD
DHAP
TPI
GAP
NADH
NAD
GAPDH
Glycerol
BPG
ADP
PGK
3PG
GPM
2PG
ENO
PEP
ADP
PYK
Succinate
PYR
3NAD
PDC
CO2
Acetate
AcAld
NAD
NADH
ADH
EtOH
Upper glycolysis
Lower glycolysis

Note that for Glycosis, if a reaction is reversible, its backwards reaction rate is the same as its forward reaction rate. We can then take get system of differential equations looking like:

$$\frac{d}{dt}[GICi] = +k_{HXK}[G6P] - (k_{transport} + k_{HXK})[GLCi]$$
$$\frac{d}{dt}[G6P] = k_{HXK}[GICi] + k_{PGI}[F6P] - (k_{HXK} + k_{PGI})[G6P]$$
$$\frac{d}{dt}[F6P] = k_{PGI}[G6P] - (k_{PGI} + k_{PFK})[F6P]$$
$$\frac{d}{dt}[F16BP] = k_{PFK}[F6P] - k_{ALD}[GAP]$$
$$\frac{d}{df}[GAP] = k_{ALD}[F16BP] + k_{GAPDH}[BPG] - (k_{ALD} + k_{GAPDH})[GAP]$$
$$\frac{d}{dt}[BPG] = k_{GAPDH}[GAP] + k_{PGK}[3PG] - (k_{PGK} + k_{GAPDH})[BPG]$$
$$\frac{d}{dt}[3PG] = k_{PGK}[BPG] + k_{GPM}[2PG] - (k_{PGK} + k_{GPM})[3PG]$$
$$\frac{d}{dt}[2PG] = k_{GPM}[3PG] + k_{ENO}[PEP] - (k_{GPM} + k_{ENO})[2PG]$$
$$\frac{d}{dt}[PEP] = k_{ENO}[2PG] + k_{PYK}[PYR] - (k_{PYK} + k_{ENO})[PEP]$$
$$\frac{d}{dt}[PYR] = k_{PDC}[AcAld] + k_{PYK}[PEP] - (k_{PDC} + k_{PYK})[PYR]$$
$$\frac{d}{dt}[AcAld] = k_{PDC}[PYR] + k_{ADH}[EtOH] - (k_{EtOH} + k_{Acetate})[AcALD]$$
$$\frac{d}{dt}[NADH] = k_{ADH}[AcAld] - k_{glycerol}[Glycerol]$$
$$\frac{d}{dt}[ATP] = -k_{HXK}[G6P] + k_{HXK}[GLc_{in}] + k_{PFK}[F6P]$$

Making this system a bit easier on the eyes by renaming concentrations and rate constants,

$$\frac{d}{dt}[x_1] = k_2[x_2] - (k_1 + k_2)[x_1]$$
$$\frac{d}{dt}[x_2] = k_2[x_1] + k_3[x_3] - (k_2 + k_3)[x_2]$$
$$\frac{d}{dt}[x_3] = k_3[x_2] - (k_3 + k_4)[x_3]$$
$$\frac{d}{dt}[x_4] = k_4[x_3] - k_5[x_5]$$
$$\frac{d}{df}[x_5] = k_5[x_4] + k_6[x_6] - (k_5 + k_6)[x_5]$$
$$\frac{d}{dt}[x_6] = k_6[x_5] + k_7[x_7] - (k_6 + k_7)[x_6]$$
$$\frac{d}{dt}[x_7] = k_7[x_6] + k_8[x_8] - (k_7 + k_8)[x_7]$$
$$\frac{d}{dt}[x_8] = k_8[x_7] + k_9[x_9] - (k_8 + k_9)[x_8]$$
$$\frac{d}{dt}[x_9] = k_9[x_8] + k_{10}[x_{10}] - (k_{10} + k_9)[x_9]$$
$$\frac{d}{dt}[x_{10}] = k_{11}[x_{11}] + k_{10}[x_9] - (k_{11} + k_{10})[x_{10}]$$
$$\frac{d}{dt}[x_{11}] = k_{11}[x_{10}] + k_{12}[x_{14}] - (k_{12} + k_{14})[x_{11}]$$
$$\frac{d}{dt}[x_{12}] = k_{12}[x_{11}] - k_{13}[x_{15}]$$
$$\frac{d}{dt}[x_{13}] = -k_2[x_2] + k_2[x_1] + k_4[x_3]$$

For simplicity's sake, we will call this system corresponding to Glycosis G' Note that instead of a system of polynomial equations, G' is a system of *linear* equations, linear with respect to the x_i's. Then we may write down a matrix corresponding to this system of linear equations which we denote as M_G:

$$\begin{pmatrix}
(-k1-k2) & k2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
k2 & (-k2-k3) & k3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & k3 & (-k3-k4) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & k4 & 0 & -k5 & 0 & 0 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & k5 & (-k5-k6) & k6 & 0 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & 0 & k6 & (-k6-k7) & k7 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & k7 & (-k7-k8) & k8 & 0 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & k8 & (-k8-k9) & -k9 & 0 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & k9 & (-k10-k9) & k10 & 0 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & k10 & (-k10-k11) & k11 & 0\,0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & k11 & (-k12-k14) & 0\,0 & k12 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & k12 & 0\,0 & 0 & -k13 \\
k2 & -k2 & k4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\,0 & 0 & 0
\end{pmatrix}$$

Note that M_G has some nice properties; it is semi-banded, with the exception of a few entries, and sparse.

3.1 Bounding the Rate Constants

Bar-Even et al. [4] suggests that there exist numerous physical constraints on reaction rates that lead to bounds on rate constants.

$$\{0 \leq k_i \leq C_i, \forall i\}$$

Some bounds on the rate constants in the paper are explicitly given. For the others, We attempt to retrieve these bounds via the measured flux. More specifically, flux is defined as the rate constant times the chemical concentration divided by reaction area, which is the "size" of the chemical reaction network. Then we have a reasonable way of obtaining an upper bound for reaction rates by finding the flux and dividing by an estimated area.

$$flux_{(i)} \approx \frac{k_i * [x_i]}{A_i}$$

Here, k_i is the ith rate constant, $[x_i]$ is the ith chemical concentration, and A_i is the ith reaction area.

The fluxes and concentrations we will use come from [4] and [2], where they were experimentally determined. Furthermore, we may estimate the reaction area as the surface area of a cell, since Glycosis occurs throughout a cell. A strict lower bound on the rate constants is 0. Then compiling all the information,

$$0 < k_{transport} < 2.2612 \ (1.19)$$
$$0 < k_{HXK} < 0.768 \ (0.08)$$
$$0 < k_{PGI} < 2.5479 \ (1.4)$$
$$0 < k_{ALD} < 1.1304 (0.3)$$
$$0 < k_{GraPH} < 1.1304 \ (0.21)$$
$$0 < k_{PGK} < 0.9235$$
$$0 < k_{PGM} < 1.5029 \ (1.2)$$
$$0 < k_{ENO} < 0.4523 \ (0.04)$$
$$0 < k_{PYK} < 2.26 * 10^{-4} \ (0.14)$$

$$0 < k_{PDC} < 2.1925 \quad (4.33)$$
$$0 < k_{ADH} < 2.3372$$
$$0 < k_{Succinate} < 1.939$$
$$0 < k_{Glycerol} < 1.939$$

*All values are given in moles per minute.

**Note that even though strict inequalities are given, we also consider equalities for convenience, which will be important later.

(note that the textbook values of the rate constants, if available, are listed in parenthesis)

3.2 Solving for the Steady-State

M_G, which is a 13 by 15 matrix, was calculated to have rank 12 via Mathematica. Verification in Macaulay2 by calculating the 13 by 13 minors was done. Therefore, for a generic vector of rate constants $\{k_1, k_2....k_{14}\}$, there corresponds a linear system that has a three-dimensional linear solution space as a variety. Because each k_i is some real interval, the k's considered together cut out a rectangular parallelepiped in $\mathbb{R}^+_{\geq 0} \times \mathbb{R}^+_{\geq 0} \times .....\mathbb{R}^+_{\geq 0}$ (the first "quadrant" in $\mathbb{R}^{14}$), which we call P_K. Then each point in P_K corresponds to some variety. Considering all points in P_K, we then get a bundle of linear spaces corresponding to possible steady states for Glycosis. Now from this bundle of linear spaces of possible steady states, we must find one best suited to the data we are given.

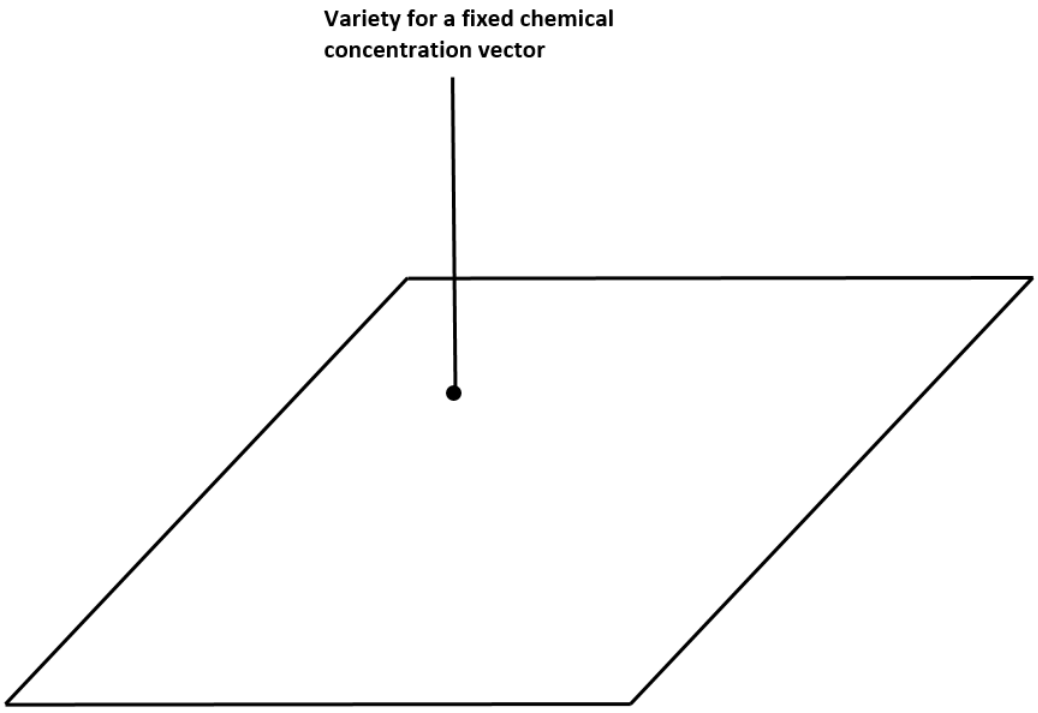


Figure 1: A visualization of P_K. For each chemical concentration vector i.e. fixed point in P_K, we get a linear system that cuts out some three dimensional linear space. When we consider all points in P_K, we get a bundle of linear spaces

Kuepfer et al. [3] suggests that Glycosis seeks to optimize some chemical concentrations. This logic is quite reasonable because Glycosis is a chemical reaction network dedicated to producing two molecules, ATP and NADH, which are fundamental to the human body; human cells break apart ATP and NADH for energy. We then applied this principle of Glycosis maximizing the amount of ATP and NADH it produces to our steady state problem. More precisely, we formulate an optimization problem that finds the "most ideal" steady state in our bundle of linear spaces. We introduce an objective function maximizing the concentrations of ATP and NADH corresponding to the principle introduced in [3].

$$maximize : [NADH] + [ATP] \tag{3}$$

subject to:

$$M_G * C = 0$$
$$C(1) = 1$$
$$\forall i, k_i = (a_i, b_i) \in \mathbb{R}^+$$
$$C \in \mathbb{R}^{\geq^+ \times \mathbb{R}^+_{\geq} \times \ldots\ldots \times \mathbb{R}^+_{\geq}}$$

Our objective function (3) assumes that the "weight" of ATP and NADH is the same. C is our vector of chemical concentrations. Our first and third constraints require that we be inside the bundle of linear spaces mentioned before. The second constraint sets the concentration of the first chemical species to one, which normalizes our optimization problem. For comparison, we will rescale our vector afterwards by the experimentally measured steady state of the first measured chemical to account for the initial condition given in [2]. Our last condition requires all chemical concentrations to be positive, because negative chemical concentrations are not physically defined. We solved this two ways (or rather, tried to solve this two ways).

3.2.1 The Numerical Method

The Numerical Method is the quick and dirty way we used to find a *local* maxima via linear algebra. Each point on our P_K is a linear program. We create an initial grid G and then solve the linear programs at each grid point. Then we pick a subgrid S_G, refine it, and solve the linear programs at each subgrid point. We repeat until a breaking condition is met.

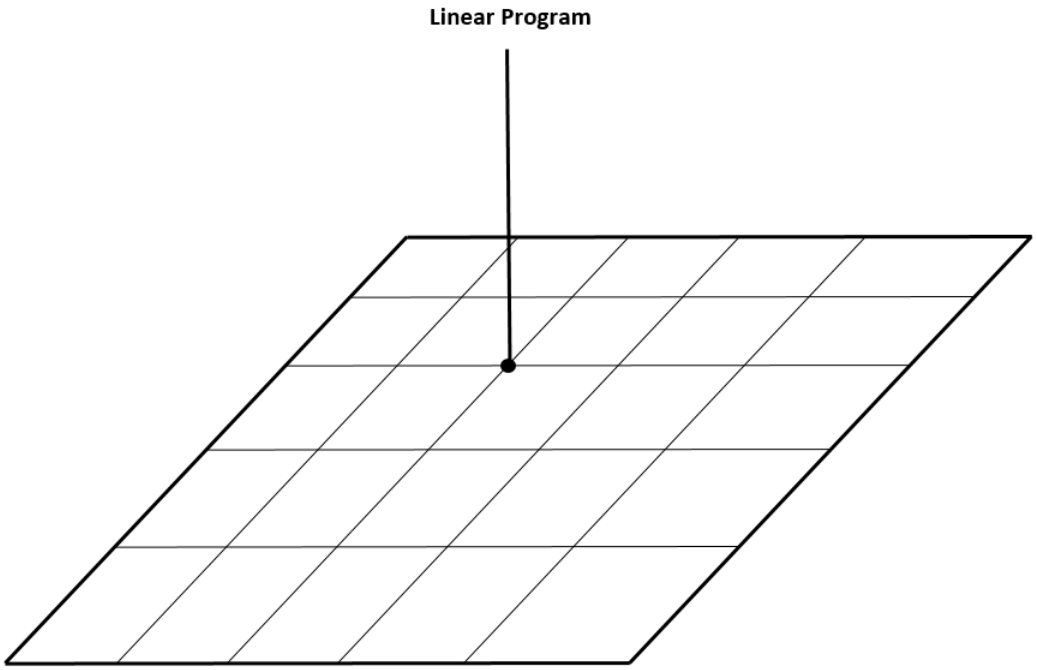


Figure 2: This is an illustration of the Numerical Method. First, the grid G^* is generated on the rectangular parallelepided P_K and then a linear program is solved on each grid point

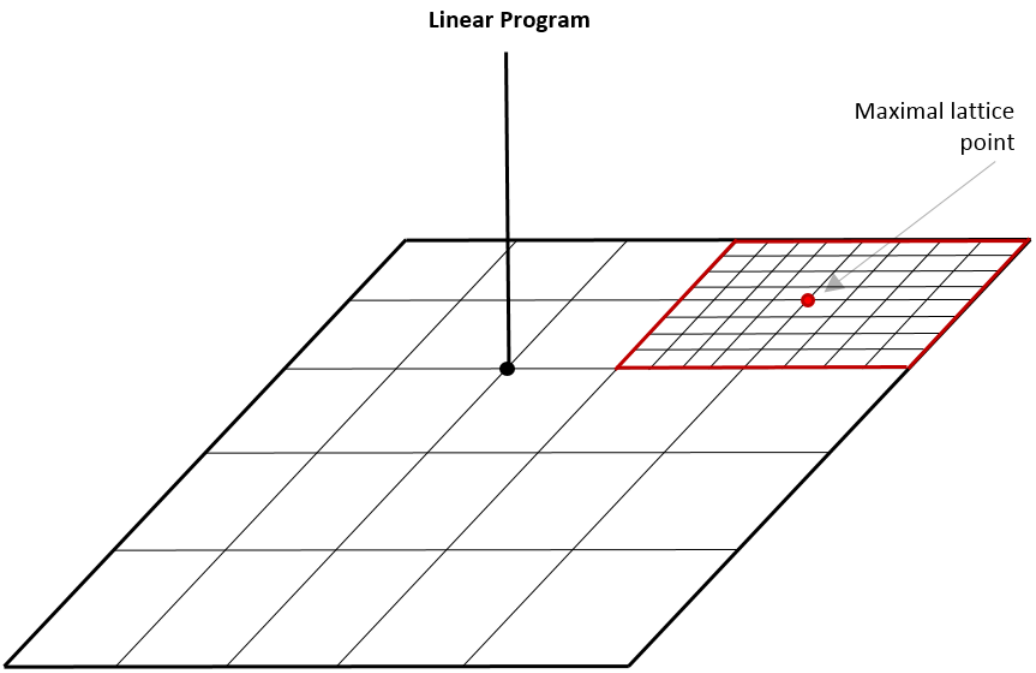


Figure 3: The grid point with the highest optimal value is picked, then the subgrid S_G around that grid point is picked. S_G is refined by increasing the number of grid points within itself. We then set G equal to the refined S_G. This process is repeated with an arbitrary number of iterations. We stop iterating when our optimal point at the ith iteration and our optimal point at the jth iteration are equal (given some error bound).

Subgrids were refined by halving the spacing between each rate constant. For example, if our subgrid were the unit cube in three space, refinement would divide our unit cube into $8 = 2^3$ cubes in three space with width 0.5. Since any subgrid of our rectangular parallelepiped is in 14 space, we refinement leads to 2^14 divisions. Because we had limited computational power for solving linear programs at tens of thousands of grid points (we were limited to desktop computers), we set the error tolerance to

$$|max(iteration(i)) - max(iteration(i-1))| < 10^{-2}$$

10^-2 is a small error compared to the measured steady states (which are are on the order of integers).

This method only works if have some kind of local continuity in the objective function above P_K. A (very)rough proof of the continuity requirement follows: We consider the map

$$\text{f: } P_K \rightarrow \text{R} \tag{4}$$

One may view the (4) as a map from a fixed vector of rate constants on P_k, which we call lower-case p_k to the solution of the linear program corresponding to that vector. Then small perturbations in M_G, which is a part of the linear program, lead to small perturbations in the solution of the linear program. That is to say, there exists some local neighborhood of p_k where we get continuity in our map (4).
Computation was done in MATLAB, which has a built-in linear program solver.

3.2.2 The Non-Linear Method

This method attempts to find the critical points of the objective function (3) inside P_K (credit goes to Jose Rodriguez, who suggested this method). It uses the fact that the critical points of (3) occur when the gradient of (3) is in the row space of the Jacobian of M_G i.e. $\nabla \mathrm{f} \in rowspace(Jac(M_G))$. In other words, if $Jac(M_G)$ has rank k, then its augmented matrix A also has rank k. The augmentated matrix A is created by concantenating $Jac(M_G)$ on top by ∇f. This occurs if and only if the $k+1$ by $k+1$ minors of our concantenated matrix vanish. These minors generate an ideals I whose variety is all possible critical points of (3). The other constraints are then factored in by saturating ideals J_i out of I until the variety cut out by I is 0-dimensional i.e. finite. We then evaluate critical points with the objective function (3) and pick out the one with maximal value.
Computation was attempted in Macaulay2. The problem with this method is that saturating and eliminating ideals is very costly process. Our Augmented Matrix A is 14 by 30 and our desired rank was 12, so the ideal generated by the 13 by 13 minors of a 14 by 30 matrix is quite large. This leads to slow saturation and elimination, to the point where Macualay2 didn't yield an output for two days. We therefore attempted to fix a large number of rate constants to speed up computation at the expense of accuracy.

4 Results

With our numerical method, we found the steady state as

$$C = \begin{bmatrix} 3.3629 \\ 1.0171 \\ 3.2571 \\ 0.9741 \\ 3.7171 \\ 1.3939 \\ 0.7864 \\ 1.0043 \\ 2.4642 \\ 1.8932 \\ 11.4066 \\ 3.3233 \\ 2.3411 \\ 2.1989 \\ 3.6688 \end{bmatrix}$$

Comparison with the steady state in [2] yielded an error of over 100 percent. In other words, if C^* denotes the experimentally measured steady state in [2], then $\|C - C^*\| > \|C^*\|$.
As mentioned earlier, attempting to use non-linear algebra proved too computationally difficult. We therefore tried fixing as many as nine of the rate constants found from different sources. With this relaxation, computation time was greatly reduced and we ended up with an ideal cutting out a zero dimensional variety.

5 References

[1] Anne Shiu, Algebraic methods for biochemical reaction network theory, Spring 2010, Doctoral Thesis, UC Berkeley.
[2] Bas Teusink et al., Can yeast glycosis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing Biochemistry, European Journal of Biochemistry, February 2000.
[3] Lars Kuepfer et al., Efficient classification of complete parameter regions based on semidefinite programming, BMC Bioinformatics, 15 Jaunary 2007.
[4] Arren Bar-Even et al., Rethinking glycosis: on the biochemical logic of metabolic pathways, Nature Chemical Biology, 17 May 2010