

# Determination of optimal metabolic pathways through a new learning algorithm

C. A. Murthy, Mouli Das, Rajat K. De

*Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India*  
*murthy@isical.ac.in, mouli\_r@isical.ac.in, rajat@isical.ac.in*

Subhasis Mukhopadhyay

*Department of Biophysics, Molecular Biology and Genetics, Calcutta University, Kolkata 700009, India.*  
*smbmbg@caluniv.ac.in*

## Abstract

*In the present article, we introduce a new method for identification of metabolic pathways in constraint based models that consider enzyme and substrate concentrations. It generates data on reaction fluxes based on biomass conservation constraint and then a set of constraints is formulated incorporating weighting coefficients corresponding to concentration of enzymes catalyzing reactions in the pathway. Finally, the rate of yield of the target metabolite, starting with a given substrate, is maximized in order to identify an optimal pathway through these weighting coefficients. In an attempt to solve this problem, we have developed a learning technique that optimizes a given objective function to find the optimal pathways. Finally, we propose a modification of the Newton Raphson method and incorporate it to our proposed methodology, which yields more relevant results from the perspective of biology.*

## 1. Introduction

Cellular metabolism can be thought of as a complex network in which metabolites are linked to each other via reactions. Metabolic networks consist of thousands of molecules that are processed and interconverted by enzymatic reactions. Metabolic pathways are defined as coordinated series of biochemical reactions in which the product of one reaction is the reactant of the subsequent one in the chain [5].

Vast repositories of data concerning enzymology and regulatory features of enzymes, as well as large scale datasets containing information of proteins and metabolites are available on internet. In recent years, an approach known as flux balance analysis has been devel-

oped to describe metabolic physiology in a quantitative manner. It is based on the fundamental law of mass conservation and the application of optimization principles to predict the optimal distribution of metabolic resources within a network. The analysis is performed under steady state conditions and it requires information about the stoichiometry of metabolic pathways and on metabolic demands [4, 1].

Reactions in a metabolic pathway are mostly enzymatic. That is, for a reaction  $A \rightarrow B$  catalyzed by an enzyme E, the rate of production of B depends not only on the concentration of the substrate A but also on the concentration of E that is available for catalyzing the reaction. Assuming that sufficient amount of the substrate A is present, if the concentration of E is low (high) then the rate of production of B will also be low (high). In the extreme pathway analysis (one of the methods under flux balance approach) [6], the authors considered the reaction flux but not the enzyme concentration. This motivates us to develop a new method that considers both the substrate and enzyme concentration, thereby becomes closer to real life situations than what extreme pathway analysis usually offers.

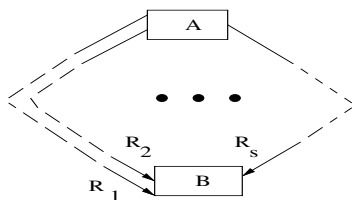
The method, first of all, generates the possible flux vectors. A set of constraints and an objective function incorporating certain weighting coefficients that correspond to enzymes catalyzing biochemical reactions is formulated. The weighting coefficients corresponding to a minimum value of the objective function represent an optimal pathway. These weighting coefficients are determined using a new learning algorithm. The effectiveness of the present method is demonstrated on glycolytic pathways of the two archae *H. salinarum R1* and *N. pharaonis*.

## 2. Proposed methodology

In this section, we introduce a new learning algorithm that determines values of weighting coefficients reflecting enzyme concentrations. These enzyme concentrations are required to get an optimal pathway through which the amount of target metabolite is maximum grown on a given substrate.

### 2.1. Problem definition

A metabolic network is a collection of enzyme-catalyzed reactions and transport processes. A system boundary can be drawn around all these types of reactions, which constitute internal fluxes operating inside the network. Exchange fluxes are allowed to enter or exit the system. Consider a metabolic network (Fig. 1) within a system with the substrate (starting metabolite) A and the final metabolite B.



**Figure 1. A hypothetical biochemical reaction network.**

Let the metabolite B be reached through  $s$  different paths. That is, there are  $s$  biochemical reactions/conversions  $R_1, R_2, \dots, R_s$  in the network involving the metabolite B. Let us also consider that there be  $n$  reactions in the network, *i.e.*,  $n$  internal and exchange fluxes. Now the rate of growth of the metabolite B on the substrate A, which needs to be maximized is obtained by taking algebraic sum of the weighted fluxes of reactions  $R_1, R_2, \dots, R_s$ , and is given by [7]

$$z = \sum_{k=1}^s c_k v_k \quad (1)$$

Here  $v_k$  is the flux of the reaction  $R_k$  involving only the metabolite B. The term  $c_k \in [0,1]$  denotes the weighting factor *i.e.* the level of concentration of the enzyme corresponding to this reaction  $R_k$ .  $c_k = 1$  indicates that the required amount of the enzyme catalyzing the reaction  $R_k$  is present. On the other hand,  $c_k$  closer to 0 indicates that sufficient amount of enzyme is not present to carry out the reaction. Higher the value of  $c_k$ , higher is the concentration of the enzyme and vice-versa. If the concentration of the enzyme increases, the fluxes *i.e.* the

reaction rates also increase and there arises a possibility of obtaining the maximal yield of the target metabolite.

### 2.2. Generation of flux vectors

Here we describe a method [6] for generating flux vectors that satisfy approximately the quasi-steady state condition. That is, we generate those  $\mathbf{v}$  which satisfy

$$\mathbf{S} \cdot \mathbf{v} \approx \mathbf{0} \quad (2)$$

and the inequalities described later in this section.  $\mathbf{S}$  is the  $m \times n$  stoichiometric matrix with  $m$  as the number of metabolites and  $n$  as the number of reactions. Note that  $\mathbf{S}$  can be computed from a reaction database. The flux vectors  $\mathbf{v}$  form the null space of  $\mathbf{S}$ . Since in practical situations,  $n > m$ , equation (2) is under determined. We generate  $l$  number of basis vectors  $\mathbf{v}_b$  that form the null space of the stoichiometric matrix  $\mathbf{S}$ . Subsequently we generate  $l$  number of non-negative random numbers  $a_p$ ,  $p = 1, 2, \dots, l$  to generate a vector  $\mathbf{v} = \sum_{p=1}^l a_p \mathbf{v}_{bp}$  satisfying the following inequality constraints. All the internal fluxes are non-negative yielding:  $v_i \geq 0, \forall i$ . The constraints on the exchange fluxes depending on their direction can be expressed as  $\alpha_j \leq b_j \leq \beta_j$  where  $\alpha_j \in \{-\infty, 0\}$  and  $\beta_j \in \{0, \infty\}$ , based on the direction of the exchange flux. Thus we generate a large number of flux vectors that form the data set.

### 2.3. Constraints

Equation (2) describes the quasi-steady state condition, which assumes that the concentration of the enzymes catalyzing various reactions in the network are present in the system at the required level. So the genes that produce these enzymes need to be expressed at the required level. But in real systems, the genes that produce these enzymes may not be expressed at the required level. This imposes restrictions on the system, and for this purpose, we define a new set of constraints as

$$\mathbf{S}(\mathbf{C}\mathbf{v}) = \mathbf{0} \quad (3)$$

where  $\mathbf{C}$  is an  $n \times n$  diagonal matrix whose  $i$ -th diagonal element is  $c_i$  for each  $i$ . That is, if  $\mathbf{C} = [\gamma_{ij}]_{n \times n}$ , then  $\gamma_{ij} = \delta_{ij} c_i$ , where  $\delta_{ij}$  is the Kronecker delta. Note that  $c_i$  is the weighting factor corresponding to the  $i$ th reaction in the network, irrespective of whether the reaction involves the metabolite B or not.

Thus the problem of determining a metabolic pathway yielding maximum rate of production of a metabolite B starting from a substrate A, boils down to a maximization problem, where  $z$  is maximized with respect to  $\mathbf{c}$ , subject to satisfying the constraint given in equation(3) along with the inequality constraints.

## 2.4. New Learning algorithm for estimating $c_i$

Combining equations (1) and (3), we can reformulate the objective function as

$$y = 1/z + \mathbf{\Lambda}^T \cdot (\mathbf{S} \cdot (\mathbf{C} \cdot \mathbf{v})) \quad (4)$$

that needs to be minimized with respect to the weighting factors  $c_i$  for all  $i$ . Here  $z$  is sum of fluxes leading towards and from the objective metabolite, the second term is related to the enzyme concentration. The term  $\mathbf{\Lambda} = [\lambda_1, \lambda_2, \dots, \lambda_m]^T$  is the regularizing parameter. For the sake of simplicity, we have considered here  $\lambda_1 = \dots = \lambda_m = \lambda$  (say). Initially, a set of random values in  $[0, 1]$  corresponding to  $c_i$ 's are generated. The values of  $c_i$ s are modified iteratively by a new learning algorithm, where

$$\Delta c_i = -\frac{\partial y}{\partial c_i} / \left| \frac{\partial^2 y}{\partial c_i^2} \right| \quad (5)$$

Thus the modified value of  $c_i$  is given by

$$c_i(t+1) = c_i(t) + \Delta c_i, \quad \forall i, t = 0, 1, 2, \dots$$

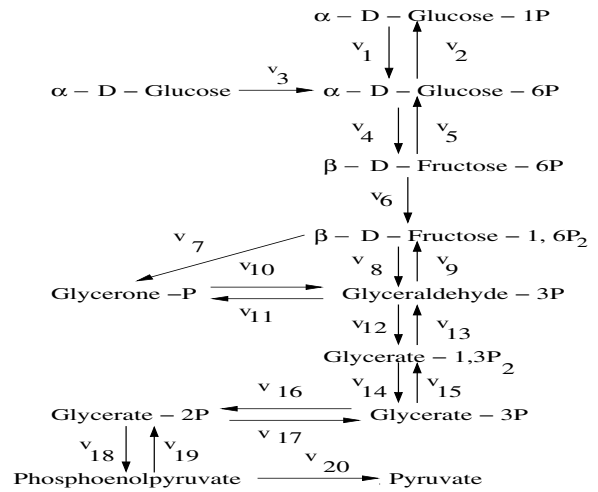
$c_i(t+1)$  is the value of  $c_i$  at iteration  $(t+1)$ , which is computed based on the  $c_i$ -value at the iteration  $t$ . Equation (5) is a modification of Newton Raphson method. Unlike the usual gradient descent techniques, equation (5) does not need any learning parameter. Regularization parameter  $\lambda$  is chosen empirically. Here we are varying the value of  $\lambda$  from 0.1 to 1.0 in steps of 0.1. For each value of  $\lambda$  as we are increasing the number of iterations, the value of  $y$  gradually decreases and the  $c_i$ -values corresponding to the flux vector  $\mathbf{v}$  are observed. The value of  $\lambda$  for which  $y$ -value becomes minimum is finally considered. Then the corresponding  $c_i$ -values indicate an optimal pathway through which the rate of yield of metabolite B becomes maximum being grown on the substrate A.

## 3. Results

We have applied the proposed method on a number of various metabolic pathway in different organisms. In most of the cases, the proposed method yields more biologically relevant results as compared to that of extreme pathway analysis. In order to restrict the size of the article, we have included here the results on two pathways, e.g. *H. salinarum R1* and *N. pharaonis*. All these real life pathways are obtained from KEGG database (<http://www.genome.jp/kegg/pathway.html>).

## 3.1. Analysis of the results

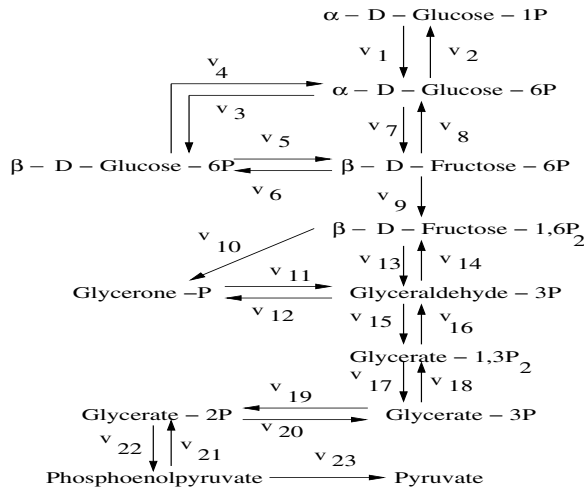
Considering the glycolytic pathway in *H. salinarum R1* (Fig. 2), there are 20 internal fluxes and 12 metabolites. Here we are maximizing the yield of pyruvate obtained from the starting metabolite  $\alpha$ -D-Glucose-1P. We associate the weighting factors  $c_1, c_2, \dots, c_{20}$  corresponding to the enzymes catalyzing these reactions respectively and generate 20 dimensional flux vectors. The objective function  $y$  (in Equation 4) is obtained by replacing  $z$  using  $z = c_{20}v_{20}$ . Following the present method, we have obtained  $\alpha - D - Glucose - 1P \rightarrow \alpha - D - Glucose - 6P \rightarrow \beta - D - Fructose - 6P \rightarrow \beta - D - Fructose - 1, 6P_2 \rightarrow Glyceraldehyde - 3P \rightarrow Glycerate - 1, 3P_2 \rightarrow Glycerate - 3P \rightarrow Glycerate - 2P \rightarrow phosphoenolpyruvate \rightarrow pyruvate$  as an optimal pathway. The extreme pathway analysis [6] results in  $\alpha - D - Glucose - 1P \rightarrow \alpha - D - Glucose - 6P \rightarrow \beta - D - Fructose - 6P \rightarrow \beta - D - Fructose - 1, 6P_2 \rightarrow Glycerone - P \rightarrow Glyceraldehyde - 3P \rightarrow Glycerate - 1, 3P_2 \rightarrow Glycerate - 3P \rightarrow Glycerate - 2P \rightarrow Phosphoenolpyruvate \rightarrow Pyruvate$  as an optimal pathway. Note that the difference can be observed on reaching the intermediate metabolite  $\beta$ -D-Fructose-1,6 $P_2$ .



**Figure 2. Glycolytic pathway in *H. salinarum R1*.**

For the glycolytic pathway in *N. pharaonis* (Fig. 3), there are 23 internal fluxes and 12 metabolites. The starting metabolite is  $\alpha$ -D-Glucose-1P and the target is pyruvate. Applying the proposed method we have obtained  $\alpha - D - Glucose - 1P \rightarrow \alpha - D - Glucose - 6P \rightarrow \beta - D - Fructose - 6P \rightarrow \beta - D - Fructose -$

$1,6P_2 \rightarrow \text{Glycerone} - P \rightarrow \text{Glyceraldehyde} - 3P \rightarrow \text{Glycerate} - 1,3P_2 \rightarrow \text{Glycerate} - 3P \rightarrow \text{Glycerate} - 2P \rightarrow \text{phosphoenolpyruvate} \rightarrow \text{pyruvate}$  as the optimal path. The standard extreme pathway analysis method yields a different path as  $\alpha - D - \text{Glucose} - 1P \rightarrow \alpha - D - \text{Glucose} - 6P \rightarrow \beta - D - \text{Glucose} - 6P \rightarrow \beta - D - \text{Fructose} - 6P \rightarrow \beta - D - \text{Fructose} - 1,6P_2 \rightarrow \text{Glycerone} - P \rightarrow \text{Glyceraldehyde} - 3P \rightarrow \text{Glycerate} - 1,3P_2 \rightarrow \text{Glycerate} - 3P \rightarrow \text{Glycerate} - 2P \rightarrow \text{phosphoenolpyruvate} \rightarrow \text{pyruvate}$ .



**Figure 3. Glycolytic pathway in *N. pharaonis*.**

### 3.2. Biological relevance and validation

Here we demonstrate how the results obtained by the present method are biologically more relevant than those obtained by the extreme pathway analysis. Genome analyses have confirmed that archaea share many features with eukaryotes, and therefore can serve as streamlined models for understanding eukaryotic biology. Glycolysis is responsible for sugar metabolism, in particular glucose, and its conservation in the evolution of the organisms suggests that glucose is a major source of energy for many organisms [8]. Using *H. salinarum R1* and *N. pharaonis* as haloarchaea models, the main metabolic pathway i.e. the glycolytic pathway has been characterised in the last few years. The glycolytic pathway of *H. salinarum R1* and *N. pharaonis* has been reviewed in detail in [2]. The glycolytic pathway in *H. salinarum R1* as obtained from our proposed methodology has been observed in [3]. In archaea, the best study about glycolytic pathway has been carried out in

*H. salinarum R1*. These studies have revealed a modified glycolytic pathway which presents novel enzymes involved in the catabolism of glucose.

### 4. Conclusions

Here we have developed a simple method for identifying an optimal metabolic pathway by introducing a new learning algorithm. The method involves formulation of the rate of yield of a metabolite incorporating weighting coefficients indicating the concentration levels of enzymes catalyzing biochemical reactions in the pathway. The method can suitably be used using reaction databases without going into complex mathematical calculations, and without using various kinetic parameters that are hard to estimate. It has been found that the method is able to produce biologically more relevant results than extreme pathway analysis. This method could be of a great interest for the scientific community, as current pathway identification methods, e.g. determination of elementary flux modes and extreme pathways cannot be applied to many real life models due to their numerical complexity. Moreover, the method can also be used in certain problems of metabolic engineering.

### References

- [1] J. S. Edwards and B. O. Palsson. Metabolic flux balance analysis and the in silico analysis of escherichia coli k-12 gene deletions. *BMC Bioinformatics*, 1(1), July 2000.
- [2] M. Falb, K. Muller, L. Konigsmair, T. Oberwinkler, P. Horn, S. Gronau, O. Gonzalez, F. Pfeiffer, E. Bornberg-Bauer, and D. Oesterhelt. Metabolism of halophilic archaea. *Extremophiles*, 12:177–196, 2008.
- [3] O. Gonzalez, S. Gronau, M. Falb, F. Pfeiffer, E. Mendoza, R. Zimmer, and D. Oesterhelt. Reconstruction, modeling and analysis of halobacterium salinarum r-1 metabolism. *Molecular Biosystems*, 4:148–159, 2008.
- [4] K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14:491–496, 2003.
- [5] J. A. Papin, N. D. Price, S. J. Wiback, and B. O. Palsson. Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences*, 28(5):250–258, May 2003.
- [6] C. H. Schilling, D. Letscher, and B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203(3):229–248, April 2000.
- [7] T. Shlomi, O. Berkman, and E. Ruppin. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *PNAS*, 102(21):7695–7700, May 2005.
- [8] L. Stryer, J. M. Berg, J. L. Tymoczko, and N. D. Clarke. *Biochemistry*. W H Freeman, New york, USA, 1998.