Computers and Chemical Engineering 73 (2015) 64–69

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

A new algorithm to find all alternate optimal flux distributions of a metabolic network

Ehsan Motamedian*

Biotechnology Group, Chemical Engineering Department, Tarbiat Modares University, P.O. Box 14155-4838, Tehran, Iran

ARTICLE INFO

Article history:
Received 17 June 2014
Received in revised form 17 October 2014
Accepted 27 November 2014
Available online 6 December 2014

Keywords:
Flux balance analysis
Metabolic network
Multiple optimal solutions
Flux variability
Mixed-integer linear programming

ABSTRACT

Finding all optimal solutions for a metabolic model is the challenge of metabolic modeling, but there is no practical algorithm for large scale models. A two-phase algorithm is proposed here to systematically identify all optimal solutions. In phase 1, the model is reduced using the FVA approach; in phase 2, all optimal solutions are searched by the addition of a binary variable to convert the model to an MILP problem. The proposed approach proved itself to be a more tractable method for large scale metabolic models when compared with the previously proposed algorithm. The algorithm was implemented on a metabolic model of *Escherichia coli* (iJR904) to find all optimal flux distributions. The model was reduced from 1076 to 80 fluxes and from 998 to 54 equations and the MILP problem was solved, resulting in 30,744 various flux distributions. For the first time, this number of optimal solutions has been reported.

1. Introduction

Linear programming (LP) has been successfully used to predict flux distribution in different microorganisms by applying flux balance analysis (FBA) (Kauffman et al., 2003; Llaneras and Picó, 2008; Orth et al., 2010; Raman and Chandra, 2009). A metabolic model is first constructed by assuming a pseudo steady-state condition using the biochemical reaction network of microorganisms to calculate mass balance on each metabolite (compound). This usually results in a set of underdetermined algebraic equations with the fluxes of metabolic reactions as the unknown variables and mass balance equations as the constraints.

A suitable objective function, such as cell growth, ATP production, or product formation, is defined and optimized using LP (Feist and Palsson, 2010; Kauffman et al., 2003; Varma and Palsson, 1994) to obtain the optimal flux distribution within the cell. The optimal solution of an LP problem lies on a vertex of the feasible solution region. Cases occur that optimal flux distributions in the cell network are not unique (Lee et al., 2000; Mahadevan and Schilling, 2003; Reed and Palsson, 2004). In other words, multiple optimal flux distributions (solutions) may exist that result in the same optimal value for an objective function. This creates an optimal solution region (optimal hyperplane) enclosed by multiple optimal vertices (Bazaraa et al., 1990; Motamedian and Naeimpoor, 2011; Taha, 2006). Although this optimal hyperplane consists of infinite optimal solutions, finding all multiple optimal solutions and optimal flux distributions conventionally means finding the optimal vertices. Each convex combination of optimal vertices results in an optimal solution on the optimal hyperplane. Optimal vertices present the simplest different uses of metabolism to achieve the optimal value of objective function.

Lee et al. (2000) proposed a recursive mixed integer linear programming (MILP) approach to find all optimal vertices of an LP problem with the addition of two types of binary variable for each non-zero basic variable to the LP problem and its conversion to an MILP problem. In small metabolic networks, there are a low number of optimal vertices, but a large number of optimal vertices may exist in large scale networks (especially in genome-scale) and hence, the MILP approach can be computationally time-consuming and intractable (Mahadevan and Schilling, 2003; Reed and Palsson, 2004). Thus, Mahadevan and Schilling (2003) introduced flux variability analysis (FVA) to study multiple optimal solutions. This approach begins with determination of the optimal value of the objective function by solving the LP problem. Using this solution, the range of variability of each flux in the network can be calculated using a series of LP problems wherein the value of the objective function is fixed at its optimal value and each variable flux is maximized and subsequently minimized. FVA provides only a subset and not necessarily all possible optimal vertices, as mentioned by Mahadevan and Schilling (2003). In addition, for a metabolic model with n reactions, FVA must solve $2n$ LP problems from scratch (Gudmundsson and Thiele, 2010), requiring a lengthy computation

* Tel.: +98 2182883947; fax: +98 2182884931.
E-mail address: motamedian@modares.ac.ir

http://dx.doi.org/10.1016/j.compchemeng.2014.11.006
0098-1354/

time, especially for large scale problems. To speed up FVA, Gudmundsson and Thiele (2010) proposed fastFVA which solves each new problem using the data obtained in the last optimal solution instead of solving the $2n$ new LP problems from scratch. They showed that fastFVA decreases running time considerably over direct FVA implementation.

The existence of optimal flux distributions can be an indication of the flexibility of a metabolic network in the choice of an internal state and can represent biologically meaningful different use of biochemical reactions. Among these optimal vertices, one or more better solutions could exist that can optimize other objective functions such as decreased genetic manipulation, decreased by-product secretion and proposing a more appropriate metabolic engineering plan. For example, all optimal vertices found by Lee et al. (2000) have the same objective function value of 0, but present different ways for carbon to pass through an *Escherichia coli* network lacking pyruvate kinase. Among all optimal vertices, there is a solution that gives rise to lower specific glucose uptake rate (4.01 mmol/gDCW/h) and higher biomass yield.

Despite the important of optimal vertices, all optimal flux distributions (vertices) of a genome-scale metabolic model have not been calculated because there is no practical algorithm to do so. Reed and Palsson (2004) used the MILP approach to calculate a subset of optimal growth solutions in an expanded genome-scale network of *E. coli* (Reed et al., 2003) for a large number of mediums. The number of alternate optimal solutions was large and it was computationally difficult to enumerate all optimal vertices; thus, they calculated only the first 500 optimal solutions (incomplete MILP solutions) for each environmental condition. From these, they identified the partial flux variability which was then compared with the results obtained using FVA. They claimed that the partial solutions obtained by MILP distinguished all variable fluxes obtained using FVA. A comparison of the results obtained using FVA and MILP was not performed since only an incomplete set of MILP optimal vertices were calculated.

The present study proposes a new algorithm that can practically find all optimal vertices. The algorithm consists of two phases: problem reduction and finding optimal vertices. For problem reduction, only variable fluxes are considered and the number of constraints is reduced as much as possible. The remaining constraints and variables define the optimal hyperplane. All vertices of the optimal hyperplane are specified using a new MILP algorithm that is a combination of the FVA and modified MILP (Lee et al., 2000) approaches. The algorithm has been implemented on a genome-scale metabolic model (Reed et al., 2003) modified for *E. coli* BW25113 Δpta to find all flux distributions that result in maximum lactate production under suboptimal anaerobic growth conditions. Castano-Cerezo et al. (2009) used experimental data to show that the Δpta strain grows at a lower rate and with a lower biomass yield than the wild type under anaerobic growth on glucose. Acetate (the main by-product of wild type) and ethanol were produced at lower rates in the new strain than in the wild type strain. The Δpta strain produced excessive amounts of lactate instead of acetate and ethanol. In the present study, the variable fluxes and all possible optimal flux distributions were determined when the mutants produced maximum rates of lactate to enumerate the various pathways that cause the maximum production of lactate instead of acetate.

2. Materials and methods

2.1. Genome scale model

The genome scale stoichiometric model (iJR904) used in this study for *E. coli* MG1655 includes 988 metabolites, 1020 reactions, 904 genes and 2 cytosolic and extracellular compartments (Reed et al., 2003). Upper bound of all intracellular reaction fluxes was limited to 1000 mmol/gDCW/h and lower bound of intracellular reversible and irreversible reaction fluxes was limited to −1000 mmol/gDCW/h and zero, respectively.

The model was modified to accommodate genetic differences between MG1655 and BW25113. Since the araBAD, rhaBAD, and lacZ genes are absent from the BW25113 strain, the associated metabolic reactions were removed (Joyce et al., 2006). The upper limits of the glucose and oxygen uptake rates were set to 10 and 0 mmol/gDCW/h, respectively, to simulate anaerobic growth on minimal glucose medium. The secretion of hydrogen, a product of *E. coli* BW25113, was added to the model. Hydrogen was allowed to freely leave the network with a maximum production rate of 1000 mmol/gDCW/h.

In the first *in silico* experiment, biomass formation was used as the objective function to be maximized. Maximal lactate secretion was calculated for suboptimal conditions for a fixed growth rate of 95% optimal growth. FVA was used to determine the variable fluxes and range of variability for maximal lactate secretion and optimal solutions were sought using the new algorithm. Calculations were made in MATLAB software using the COBRA toolbox. MALTAB was linked to GAMS/CPLEX to solve LP and MILP problems.

2.2. Method

A metabolic network is defined by a set of metabolites (compounds) and a set of biochemical reactions interconnecting these metabolites. After reconstruction of a metabolic network (Thiele and Palsson, 2010), the stoichiometric coefficients of the reactions are determined. The stoichiometric coefficients of a metabolic network with m metabolites and n reactions can be presented by a stoichiometric matrix (S) in which the rows and columns correspond to the metabolites and reactions, respectively. Thus, $S_{i,j}$ array of the stoichiometric matrix is the stoichiometric coefficient for metabolite i with respect to reaction j. If a metabolite is formed by the reaction j, the coefficient has a positive sign; if it is consumed by the reaction j, the stoichiometric coefficient appears with a negative sign. All other rows (corresponding to metabolites that do not participate in the reaction j) are zero.

Once the stoichiometric matrix has been determined, mass balances for the intracellular metabolites (assuming a pseudo-steady-state condition) can be represented by a set of linear equations (Edwards et al., 2002; Kauffman et al., 2003; Llaneras and Picó, 2008; Raman and Chandra, 2009) as in Eq. (1):

$$S \cdot r = b \tag{1}$$

where r is the flux vector and b is the right-hand side vector determined by known reaction fluxes. This set of algebraic equations is usually underdetermined and, therefore, infinite solutions (feasible solution region) exist. The feasible solution region is a convex polyhedral space due to the linearity of equations (Fig. 1). In order to obtain a unique solution within this feasible region, FBA can be used by defining an objective function, such as growth or product formation, and optimizing it to achieve an optimal solution (Kauffman et al., 2003). The LP formulation of the metabolic model shown in Eq. (2):

$$\begin{aligned} &\text{Max } Z = cr \\ &\text{Such that } S \cdot r = b \\ &r_{\min} \le r \le r_{\max} \end{aligned} \tag{2}$$

where c is the objective function vector and $r_{\max}$ and $r_{\min}$ are the vectors containing the upper and lower limits of the variable fluxes, respectively. FBA results in an optimal solution that always lies on a vertex of the feasible solution region. This optimal vertex presents