Computational and Structural Biotechnology Journal 11 (2014) 59–65

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Mini Review

Methods for integration of transcriptomic data in genome-scale metabolic models

Min Kyung Kim [a], Desmond S. Lun [a,b,*]

[a] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
[b] Phenomics and Bioinformatics Research Centre and School of Mathematics and Statistics, University of South Australia, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Available online 3 September 2014

Keywords:
Flux balance analysis
Contraint-based model
Omics

ABSTRACT

Several computational methods have been developed that integrate transcriptomic data with genome-scale metabolic reconstructions to infer condition-specific system-wide intracellular metabolic flux distributions. In this mini-review, we describe each of these methods published to date with categorizing them based on four different grouping criteria (requirement for multiple gene expression datasets as input, requirement for a threshold to define a gene's high and low expression, requirement for a priori assumption of an appropriate objective function, and validation of predicted fluxes directly against measured intracellular fluxes). Then, we recommend which group of methods would be more suitable from a practical perspective.

Contents

1. Introduction

Intracellular metabolic reactions provide a cell with basic biochemical building blocks, energy, and a thermodynamically favorable environment to sustain its life. Because of the large connectivity inherent to metabolic networks via metabolites participating in multiple metabolic reactions, determination of system-level changes in intracellular metabolic fluxes of organisms is important for understanding the fundamental mechanisms of their metabolic responses to environmental or genetic perturbations [1,2].

^{13}C metabolic flux analysis (^{13}C-MFA) allows intracellular fluxes to be quantified experimentally. In this approach, cells are grown on ^{13}C-labeled substrates until the cells are at both metabolic steady state (i.e. when concentrations of metabolites remain stable over time) and isotopic steady state (i.e. when the isotope label is distributed throughout the network, and all isotopomer fractions are constant over time). Then the level of ^{13}C enrichment in metabolites of the cells is measured by mass spectrometry (MS) or nuclear magnetic resonance (NMR). Intracellular flux distribution is reconstituted from the ^{13}C enrichment patterns [3–8]. System-wide quantification of intracellular metabolic fluxes using ^{13}C-MFA, however, is challenging not only because of the extensive instrumentation required but also because of the limited number of fluxes and conditions that can be experimentally measured. Typically, ^{13}C-MFA focuses on central carbon metabolism [7–10].

An alternative method that is widely used for system-level studies of metabolism is a computational modeling approach called flux balance analysis (FBA). FBA predicts metabolic flux distributions at steady state by making use of in silico genome-scale metabolic models [11]. These genome-scale metabolic models are assembled and manually-curated from annotated genome, biochemical, genetic, and cell phenotype data [11–13]. To use FBA, a genome-scale metabolic model is converted into a $m \times n$ stoichiometric matrix, S, where the rows in S

* Corresponding author at: Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA. Tel.: +1 856 225 6094; fax: +1 856 225 6624.
E-mail addresses: mk1034@rutgers.edu (M.K. Kim), dslun@rutgers.edu (D.S. Lun).

http://dx.doi.org/10.1016/j.csbj.2014.08.009

correspond to the m metabolites of the metabolic network, and the columns represent the n reactions (Fig. 1a). Each matrix element s_{ij}, indicates a stoichiometric coefficient, that is, the number of molecules of the ith metabolite participating in the jth reaction. $s_{ij} = 0$ means that the ith metabolite is not involved, and a positive or a negative s_{ij} indicates that the ith metabolite is a product or a reactant of the jth reaction, respectively. Under the steady state assumption, the metabolic flux distribution can be represented mathematically by $S \cdot v = 0$, where v is a column vector whose elements are the unknown reaction rates (fluxes) through each of the reactions of S (Fig. 1b). Since genome-scale metabolic models include all possible metabolic reactions implied by the genome annotation regardless of whether the annotated metabolic genes are expressed in a given environment, the resulting system $S \cdot v = 0$, is in general underdetermined [14,15]. Thus, physiologically meaningful flux solutions need to be narrowed down from all the possible flux distributions by imposing additional constraints on the system and by optimizing certain objective functions when performing FBA (Fig. 1c) [16]. The standard FBA involves solving the following linear optimization problem:

$$\max \; f'v \quad \text{subject to} \begin{cases} Sv = 0 \\ lb \le v \le ub \end{cases} \tag{1}$$

where v is a flux vector representing the reaction rates of the n reactions in the network, f is a coefficient vector defining the organism's objective function, S is the stoichiometric matrix, and lb and ub are the minimum and maximum reaction rates through each reaction in v.

If the complete regulatory structure of an organism were known, it would be possible to produce context-specific constraints by computing which cellular components may be expressed in a given condition. However, the regulatory structure is unknown even for the relatively simple and extensively-studied bacterium, *Escherichia coli*, partly due to the lack of comprehensive transcription unit information and because of the lack of information on the relationship between genotype and phenotype [17].

Recent advances in omics technologies have enabled quantitative monitoring of the abundance of biological molecules at various levels in a high-throughput manner [18]. In the absence of complete information on regulatory rules, omics data can be integrated with genome-scale metabolic models to improve their predictive power [19,20]. For this purpose, transcriptomic data, i.e. genome-wide mRNA expression profiling data, is useful in some points compared to other omics platforms. Fluxomics (i.e. ^{13}C-MFA) is the most direct measurement of metabolic phenotype, but has the disadvantages in that it is difficult to make measurements and only a limited number of fluxes can be determined as mentioned above. Metabolomics can also be useful, but typically fluxes are more informative than metabolite concentrations themselves, and it is challenging to determine fluxes from metabolite concentrations partly because each metabolite participates in multiple metabolic reactions. Similar to fluxes, specific classes of metabolites such as lipids or labile chemicals easily metabolized are still demanding to measure [21,22]. Unlike the first two omics data that cover a small share of all reactions in a genome-scale model, transcriptomics and proteomics are the platforms where a quantitative snapshot of molecular species at system-level is currently possible [23]. However, proteomics is a relatively immature technology compared to transcriptomics. The accuracy with which protein concentrations can be determined is much lower than that with which mRNA concentrations can be determined. On the other hand, RNA amount changes can be precisely measured in a highly automated process at low cost in comparison with the amount of data gathered [24,25]. By integrating transcriptomics data

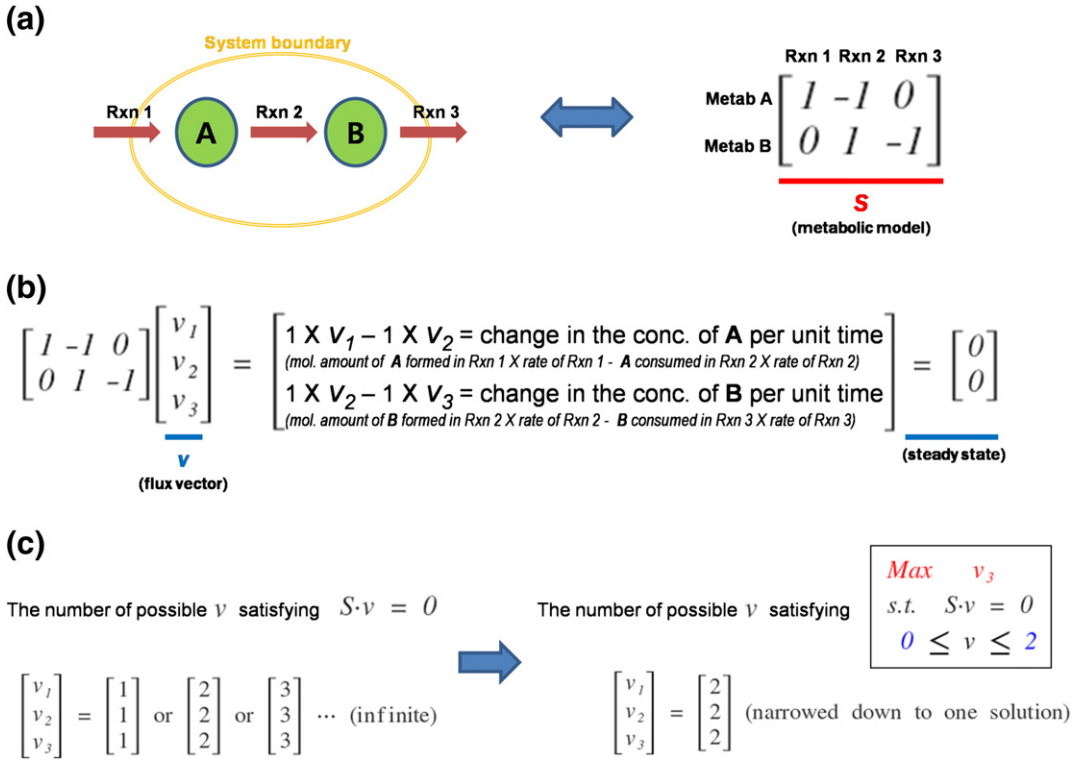

Fig. 1. Flux balance analysis (FBA). This figure illustrates how FBA works with an example of the simple network below consisting of two metabolites, A and B, and three metabolic reactions. (a) To use FBA, the network is converted into a stoichiometric matrix, S, where the rows in S correspond to the metabolites of the metabolic network, and the columns represent the reactions. Each matrix element s_{ij}, indicates a stoichiometric coefficient, that is, the number of molecules of the ith metabolite participating in the jth reaction. $s_{ij} = 0$ means that the ith metabolite is not involved, and a positive or a negative s_{ij} indicates that the ith metabolite is a product or a reactant of the jth reaction, respectively. (b) Under the steady state assumption, the metabolic flux distribution can be represented mathematically by $S \cdot v = 0$, where v is a column vector whose elements are the unknown reaction rates (fluxes) through each of the reactions of S.(c) Since the resulting system, $S \cdot v = 0$, is usually underdetermined, physiologically meaningful flux solutions need to be narrowed down from all the possible flux distributions by imposing additional constraints on the system (e.g. $0 \le v \le 2$ in the figure) and by optimizing certain objective functions (e.g. *Max* v_3 in the figure).