Journal of Theoretical Biology 254 (2008) 807–816

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Metabolic networks are NP-hard to reconstruct

Zoran Nikoloski [a,*], Sergio Grimbs [b], Patrick May [b], Joachim Selbig [a,b]

[a] Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, Haus 20, 14476 Potsdam, Germany
[b] Max-Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam, Germany

ARTICLE INFO

Article history:
Received 26 November 2007
Received in revised form
14 July 2008
Accepted 14 July 2008
Available online 22 July 2008

Keywords:
Reconstruction
Metabolic networks
Completeness
Approximation

ABSTRACT

High-throughput data from various omics and sequencing techniques have rendered the automated metabolic network reconstruction a highly relevant problem. Our approach reflects the inherent probabilistic nature of the steps involved in metabolic network reconstruction. Here, the goal is to arrive at networks which combine probabilistic information with the possibility to obtain a small number of disconnected network constituents by reduction of a given preliminary probabilistic metabolic network. We define *automated metabolic network reconstruction* as an optimization problem on four-partite graph (nodes representing genes, enzymes, reactions, and metabolites) which integrates: (1) probabilistic information obtained from the existing process for metabolic reconstruction from a given genome, (2) connectedness of the raw metabolic network, and (3) clustering of components in the reconstructed metabolic network. The practical implications of our theoretical analysis refer to the quality of reconstructed metabolic networks and shed light on the problem of finding more efficient and effective methods for automated reconstruction. Our main contributions include: a completeness result for the defined problem, polynomial-time approximation algorithm, and an optimal polynomial-time algorithm for trees. Moreover, we exemplify our approach by the reconstruction of the sucrose biosynthesis pathway in *Chlamydomonas reinhardtii*.

1. Introduction

The availability of fully sequenced genomes, coupled with the development of effective bioinformatics methods for gene annotation, offers the possibility for reconstructing entire metabolic networks. The problem of metabolic network reconstruction is clearly related to the precise understanding of the genetic basis for metabolic organization and regulation. While preliminary metabolic networks have already been reconstructed solely based on gene annotation (Ma and Zeng, 2003; Romero et al., 2005; Reed et al., 2006), this process may discard some available information: It is often the case that the function of genes is determined by the highest similarity obtained through comparison to other already annotated organisms. However, in such practice, alternative gene functions may result in smaller but still significant similarity (Green and Karp, 2004).

The following steps are crucial for reconstructing a metabolic network based on the genome of a given organism: (I) establishing gene models, (II) sequence similarity search (e.g., BLAST), (III) gene product annotation, with the help of available enzyme databases (e.g., KEGG, Expasy, Brenda), (IV) enzyme–reaction association, with the help of reaction databases (e.g., KEGG LIGAND Goto et al., 2002), and (V) pathway mapping. The outcome from steps (I) to (IV) results in a preliminary metabolic network given by sets of: enzyme–gene relationships, reaction–enzyme relationships, reactions, and metabolites, which make up the metabolic network. Finally, in step (V), the identified reactions are mapped onto a collection of pathways, e.g., from KEGG (Kanehisa et al., 2004) or MetaCyc (Caspi et al., 2006), to obtain a raw metabolic network. Currently, this network is taken to be the reconstructed metabolic network for the organism whose genome is considered as input to the process.

The preliminary metabolic network is furthermore carefully calibrated by the experimental results reported in literature. This iterative manual process can be labor-intensive and time-consuming. Even for fairly simple microorganisms such as *Escherichia coli* (Reed et al., 2004) and *Saccharomyces cerevisiae* (Duarte et al., 2004), the metabolic networks reconstructed with high quality have taken years to assemble. There are ongoing research efforts to use the same reconstruction methodology on the human genome, with variable success directly related to the complexity of this task (Romero et al., 2005; Ma et al., 2007).

Assembling the preliminary metabolic network often employs prediction-based bioinformatics methods and is, therefore, *probabilistic*. For instance, gene annotation is based on prediction,

* Corresponding author. Tel.: +49 33 15678622; fax: +49 33 15678136.
E-mail addresses: nikoloski@mpimp-golm.mpg.de (Z. Nikoloski), grimbs@mpimp-golm.mpg.de (S. Grimbs), may@mpimp-golm.mpg.de (P. May), selbig@mpimp-golm.mpg.de (J. Selbig).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.07.015

yielding enzyme–gene relationships explicitly weighted with the accuracy of prediction (e.g., in the range $(0,1]$). Moreover, there may be an ambiguous relationship between enzymes and reactions in the reaction databases, in the sense that an enzyme in a given organism may not catalyze a reaction which is catalyzed by the same enzyme in another organism (Wu et al., 2006; Wang et al., 2006). Hence, in the absence of precise human-curated knowledge, the enzyme–reaction relationships for a given organism are also weighted with the accuracies of their computational predictions. A threshold can be imposed to include the most relevant relationships. However, it is often the case that only the highest-value predictions are included in the reconstructed metabolic network. Therefore, the possibility that, for instance, a given gene codes for more than one enzyme or that an enzyme catalyzes more than one reaction is often neglected. Finally, in step (V), only a portion of a given pathway may be included, resulting in a disconnected metabolic network. This shortcoming of the reconstruction process points at necessary clustering of connected reactions to show their functional relationships.

Therefore, we can conclude that preliminary metabolic networks, taken as the reconstructed counterparts, are often incomplete, since a large portion of available information is ignored by overlooking its probabilistic nature. As a result, much manual validation and correction is needed. To allow for inclusion of information with varying accuracy of prediction, here, we address the problem of automated reconstruction of metabolic networks. We believe that there is a need for formal definition of metabolic network reconstruction, whose analysis may result in new insights of how to approach and resolve the problem at hand.

The existing approaches for reconstructing metabolic networks include (constraint-based) elementary modes (Stelling et al., 2002) and flux balance analysis (FBA) (Edwards and Palsson, 2000; Price et al., 2003). Elementary modes correspond to the smallest subnetworks that can operate in steady state. FBA uses linear programming to obtain a single (not necessarily unique) solution to an optimization problem (e.g., with growth per substrate uptake as a function to be maximized) and can be used in the analysis of specific cell behaviors. On the other hand, elementary modes allow for investigation of the space of all meaningful physiological states, and can be used to define control-effective fluxes via their respective efficiencies (relating a mode's output to the cost for establishing the mode). In addition, elementary modes can address cellular regulation and can characterize some aspects of cellular behavior from metabolic network topology. We point out that both approaches are structural in the sense that they require the topology of a putative metabolic network together with its stoichiometry in order to elucidate mutant phenotypes, analyze network robustness, and to quantitatively predict functional features of genetic regulation. The approach described here aims at metabolic network reconstruction which satisfies the biochemical balance constraints and relies solely on graph-theoretic concepts.

Contributions and organization: We define the automated reconstruction of metabolic networks as an optimization problem in Section 2. Our approach is exemplified in Section 3 by the reconstruction of the sucrose biosynthesis pathway for *Chlamydomonas reinhardtii*. The results regarding the hardness of the problem are presented in Section 4. The practical implications of our theoretical analysis refer to the quality of reconstructed metabolic networks and shed light on the problem of finding more efficient and effective methods for automated reconstruction. Such methods can result in biologically relevant networks that may speed up the computational analysis, but still require considerable effort for experimental validation. An optimal polynomial-time algorithm for the problem restricted to trees is described and analyzed in Section 5, while approximation results are shown in Section 6.

2. Problem definition

For the purpose of defining the formalism for metabolic network reconstruction, we require the assembly of a preliminary metabolic network, which we call *raw metabolic network*. One technique for obtaining the raw metabolic network includes the steps described in Section 1: after genes have been determined in step (I) and their similarity to genes from other organisms has been established in step (II), the function of genes can be assigned in step (III). Step (III), in fact, results in a set of enzymes that can catalyze a set of reactions. By using existing pathway databases, one can then identify to which pathway(s) the found reactions belong. Often, the existing gene annotation may cover a portion of the pathways, i.e., only few of the pathways' reactions are initially included in the raw metabolic network. Other reactions may be included based on different approaches: usage of experts' knowledge or taxonomic distance between enzymes on pathways from the used database (Peregrin-Alvarez et al., 2003). In order to allow for stoichiometrically balanced reconstructed network, the raw metabolic network should not include stoichiometrically unbalanced reactions proceeding from public databases. Moreover, based on metabolomic studies, the raw metabolic network can be extended to include previously not present metabolites. In the latter case, the raw metabolic network can include reactions that use these metabolites together with the corresponding enzymes and known genes.

Here, the raw metabolic network is represented by a graph G, irrespective of the methods used in its assembly. The node set of G is a union of pairwise disjoint node sets (partitions) representing: genes, enzymes, reactions, and metabolites. The edge set of G is a union of pairwise disjoint edge sets describing gene–enzyme, enzyme–reaction, and reaction–metabolite relationships. Each edge has a weight, representing the accuracy of prediction for a particular relationship (or, its certainty). We assume that the accuracy is given by a real number from the interval $(0,1]$. Some possible methods to obtain the edge-weights include transformation of the E-value or the BLAST score on the interval $(0,1]$ (Green and Karp, 2004) or usage of recent databases for biochemical substructures and prediction of reaction–metabolite relationships (Kotera et al., 2008). However, the formulation of our problem and the proposed approximations are independent of the employed methods for the edge-weights. Edges of weight 0 are not included in the graph G.

In addition, if a reaction is spontaneous or is included without gene evidence, the raw metabolic network is extended to include dummy gene and enzyme nodes corresponding to the reaction. We point out that reactions included from public databases may not be chemically balanced (Poolman et al., 2006). In this case, the raw metabolic network may still include some of the chemically unbalanced reactions upon an expert's opinion and in accordance with biochemical knowledge. However, the formalism presented here does not aim at resolving this known issue of the publically available human-curated databases.

We have chosen the four-partite graph representation as it provides the minimum number of different entities sufficient for metabolic network reconstruction. The included entities are sufficient for our task since the measurement of their respective quantities (e.g., gene expression, fluxes, or metabolite concentrations) yields the minimum effort for validation of the reconstructed network. The graph-theoretic representation employed here can be easily extended to include other biologically relevant entities, such as mRNA, by providing an additional node-partition