Computational Biology and Chemistry 59 (2015) 98–112

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data

Brandon E. Barker[a,*], Narayanan Sadagopan[b,1], Yiping Wang[c,1], Kieran Smallbone[d,e], Christopher R. Myers[c,f,g], Hongwei Xi[h], Jason W. Locasale[c,i,**], Zhenglong Gu[c,i,**]

[a] Center for Advanced Computing, Cornell University, 534 Rhodes Hall, Ithaca, NY, USA
[b] Program of Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA, USA
[c] Tri-Institutional Training Program in Computational Biology and Medicine, 1300 York Avenue, Box 194, New York, NY, USA
[d] School of Computer Science, The University of Manchester, Manchester, UK
[e] Manchester Center for Integrative Systems Biology, The University of Manchester, Manchester, UK
[f] Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY, USA
[g] Institute of Biotechnology, Cornell University, Ithaca, NY, USA
[h] Department of Computer Science, Boston University, 111 Cummington Street, Boston, MA, USA
[i] Division of Nutritional Sciences, Cornell University, Savage Hall, Ithaca, NY, USA

ARTICLE INFO

Article history:
Received 12 February 2015
Received in revised form 27 May 2015
Accepted 3 August 2015
Available online 1 September 2015

ABSTRACT

A major theme in constraint-based modeling is unifying experimental data, such as biochemical information about the reactions that can occur in a system or the composition and localization of enzyme complexes, with high-throughput data including expression data, metabolomics, or DNA sequencing. The desired result is to increase predictive capability and improve our understanding of metabolism. The approach typically employed when only gene (or protein) intensities are available is the creation of tissue-specific models, which reduces the available reactions in an organism model, and does not provide an objective function for the estimation of fluxes. We develop a method, flux assignment with LAD (least absolute deviation) convex objectives and normalization (FALCON), that employs metabolic network reconstructions along with expression data to estimate fluxes. In order to use such a method, accurate measures of enzyme complex abundance are needed, so we first present an algorithm that addresses quantification of complex abundance. Our extensions to prior techniques include the capability to work with large models and significantly improved run-time performance even for smaller models, an improved analysis of enzyme complex formation, the ability to handle large enzyme complex rules that may incorporate multiple isoforms, and either maintained or significantly improved correlation with experimentally measured fluxes. FALCON has been implemented in MATLAB and ATS, and can be downloaded from: https://github.com/bbarker/FALCON. ATS is not required to compile the software, as intermediate C source code is available. FALCON requires use of the COBRA Toolbox, also implemented in MATLAB.

1. Introduction

FBA (flux balance analysis) is the oldest, simplest, and perhaps most widely used linear constraint-based metabolic modeling approach (Shestov et al., 2013; Lewis et al., 2012). FBA has become extremely popular, in part, due to its simplicity in calculating reasonably accurate microbial fluxes or growth rates (e.g. Schuetz et al., 2012; Fong and Palsson, 2004); for many microbes, a simple synthetic environment where all chemical species are known suffices to allow proliferation, giving fairly complete constraints on model inputs. Additionally, it has been found that their biological objectives can be largely expressed as linear objectives of fluxes, such as maximization of biomass (Schuetz et al., 2012). Neither of these assumptions necessarily hold for mammalian cells growing *in vitro* or *in vivo*, and in particular the environment is far more complex for mammalian cell cultures, which have to undergo gradual metabolic adaptation via titration to grow on synthetic media (Pirkmajer and Chibalin, 2011). Recently, there have

* Corresponding author.
** Corresponding authors at: Tri-Institutional Training Program in Computational Biology and Medicine, 1300 York Avenue, Box 194, New York, NY, USA.
E-mail addresses: brandon.barker@cornell.edu (B.E. Barker), locasale@cornell.edu (J.W. Locasale), zg27@cornell.edu (Z. Gu).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.compbiolchem.2015.08.002
1476-9271/

been many efforts to incorporate both absolute and differential expression data into metabolic models (Blazier and Papin, 2012). The minimization of metabolic adjustment (MoMA; Segrè et al., 2002) algorithm is the simplest metabolic flux fitting algorithm, and it can be extended in order to allow the use of absolute expression data for the estimation of flux (Lee et al., 2012), which is the approach taken in this study. A different approach for using expression in COBRA, also very simple, is E-flux, which simply uses some function of expression (chosen at the researcher's discretion; typically a constant multiplier of expression) as flux constraints (Colijn et al., 2009). Despite this surprising simplicity, the method has found many successful applications, but the user-chosen parameter and use of expression as hard constraints is, in our opinion, a detraction, and others have had better results taking an approach similar to Lee et al. (2012) (Bogart and Myers, in press).

The MoMA method, framed as a constrained least-squares optimization problem, is typically employed to calculate the flux vector of an *in silico* organism after a mutation by minimizing the distance between the wild-type flux and the mutant flux. The biological intuition is that the organism has not had time to adapt to the restricted metabolic capacity and will maintain a similar flux to the wild-type (WT) except where the perturbations due to the mutation dictate necessary alterations in fluxes (Shlomi et al., 2005). Suppose $\mathbf{a}$ is the WT flux vector obtained by an optimization procedure such as FBA, empirical measurements, or a combination of these. For an undetermined flux vector $\mathbf{v}$ in a model with N reactions the MoMA objective can be expressed as

$$\text{minimize} \sum_{i=1}^{N} (v_i - a_i)^2 \tag{1}$$

subject to the stoichiometric constraints $\mathbf{Sv} = \mathbf{0}$ where $\mathbf{v} = (v_1, \ldots, v_N)^T$ and $\mathbf{S}$ is the stoichiometric matrix (rows correspond to metabolites, columns to reactions, and entries to stoichiometric coefficients). Constant bounds on fluxes are often present, such as substrate uptake limits, or experimental $V_{\max}$ estimates, so we write these as the constraints $\mathbf{v}_{lb} \preceq \mathbf{v} \preceq \mathbf{v}_{ub}$. The objective may be equivalently expressed in the canonical quadratic programming (QP) vector form as min. $(1/2)\mathbf{v}^T\mathbf{v} - \mathbf{a}^T\mathbf{v}$. This assumes that each a_i is measured, but it is also possible and sometimes even more useful to employ this objective when only a subset of the a_i are measured (if a_i is not measured for some i, then we omit $(v_i - a_i)^2$ from the objective). In metabolomics, for instance, it is always the case in experiments with labeled isotope tracers that only a relatively small subset of all fluxes are able to be estimated with metabolic flux analysis (MFA; Shestov et al., 2013). Combining MoMA with MFA provides a technique to potentially estimate other fluxes in the network.

A variant of MoMA exists that minimizes the absolute value of the difference between a_i and v_i for all known a_i. To the best of our knowledge, the following linear program is the simplest version of linear MoMA, which assumes the existence of a constant flux vector $\mathbf{a}$:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} d_i \\ \text{subject to} \quad & \mathrm{Sv} = \mathbf{0} \\ & \mathbf{v}_{lb} \preceq \mathbf{v} \preceq \mathbf{v}_{ub} \\ \forall i: \quad & -d_i \le v_i - a_i \le d_i \\ & d_i \ge 0 \end{aligned} \tag{2}$$

The d_i are just the distances from *a priori* fluxes to their corresponding fitted fluxes. Linear MoMA has the advantage that it is not biased towards penalizing large magnitude fluxes or under-penalizing fluxes that are less than one (Boyd and Vandenberghe, 2004; Shlomi et al., 2005). Additionally, linear programs are often amenable to more alterations that maintain convexity than a quadratic program and tend to have fewer variables take on small values, and it is much easier to interpret the importance of a zero than a small value (Boyd and Vandenberghe, 2004).

We wish to apply MoMA to expression data rather than flux data, but there are two primary problems that must be tackled. First, we must quantify enzyme complex abundance as accurately as possible given the gene expression data. Although there is not a one-to-one correspondence between reactions and enzyme complexes, the correspondence is much closer than that between individual genes and metabolic reactions. In the first part of this work, we employ an algorithm that can account for enzyme complex formation and thus quantify enzyme complex abundance. Second, we must fit real-valued variables (fluxes) to non-negative data (expression), which is challenging to do efficiently. To accomplish this, we build on the original MoMA objective, which must be altered in several ways (also discussed in Lee et al. (2012), which lays the groundwork for the current method). We develop automatic scaling of expression values so that they are comparable to flux units obtained in the optimization routine. This can be an advantage over the prior method as it no longer requires the manual choice of a flux and complex abundance pair with a ratio that is assumed to be representative of every such pair in the system. Related to this, we also implement the sharing of enzyme complex abundance between the reactions that the complex catalyzes, rather than assuming there is no competition between reactions catalyzed by the same complex. Reaction direction assignment enables comparison of fluxes and expression by changing fluxes to non-negative values. We show that batch assignment, rather than serial assignment (Lee et al., 2012) of reaction direction can greatly improve time efficiency while maintaining or slightly improving correlations with experimental fluxes. In addition to several of the methods described so far, we also included in our comparison two methods for *tissue-specific* modeling. In GIMME, the authors remove reactions whose associated gene expression is below some threshold, then add reactions that preclude some user-defined *required metabolic functionalities* in an FBA objective back into the model, and finally use FBA again to obtain fluxes (Becker and Palsson, 2008). The other tissue-specific method we compared with is iMAT, which employs a mixed integer linear programming (MILP) problem to maximize the number of reactions whose activity corresponds to their expression state (again using thresholds, but this time, there are low, medium, and highly expressed genes, and only the lowly and highly expressed genes are included in the objective) all while subject to typical constraints found in FBA (Shlomi et al., 2008).

Finally, we employ several sensitivity analyses and performance benchmarks so that users of the FALCON method and related methods may have a better understanding of what to expect in practice.

2. Methods

Most genome-scale models have attached Boolean (*sans* negation) gene rules to aid in determining whether or not a gene deletion will completely disable a reaction. These are typically called GPR (gene–protein-reaction) rules and are a requirement for FALCON; their validity, like the stoichiometric matrix, is important for generating accurate predictions. Also important are the assumptions and limitations for the process of mapping expression data to complexes so that a scaled enzyme complex abundance (hereafter referred to as complex abundance) can be estimated. We address these in the next section and have attached a flow chart to illustrate the overall process of mapping expression of individual genes