BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2529–2531
doi:10.1093/bioinformatics/btu321

Systems biology

Advance Access publication May 7, 2014

fastGapFill: efficient gap filling in metabolic networks

Ines Thiele*, Nikos Vlassis and Ronan M. T. Fleming

Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg, L-4362

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Genome-scale metabolic reconstructions summarize current knowledge about a target organism in a structured manner and as such highlight missing information. Such gaps can be filled algorithmically. Scalability limitations of available algorithms for gap filling hinder their application to compartmentalized reconstructions.

Results: We present fastGapFill, a computationally efficient tractable extension to the COBRA toolbox that permits the identification of candidate missing knowledge from a universal biochemical reaction database (e.g. Kyoto Encyclopedia of Genes and Genomes) for a given (compartmentalized) metabolic reconstruction. The stoichiometric consistency of the universal reaction database and of the metabolic reconstruction can be tested for permitting the computation of biologically more relevant solutions. We demonstrate the efficiency and scalability of fastGapFill on a range of metabolic reconstructions.

Availability and implementation: fastGapFill is freely available from http://thielelab.eu.

Contact: ines.thiele@uni.lu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 21, 2013; revised on April 29, 2014; accepted on April 30, 2014

1 INTRODUCTION

A biomolecular network reconstruction summarizes biochemical, physiological and genomic knowledge in a mathematically structured electronic format (Palsson, 2006). It can be converted into a computational model, and predictions have been used to accelerate biotechnological and biomedical discoveries (Oberhardt *et al.*, 2010). The predictive capacity and accuracy of a model depend on the comprehensiveness and biochemical fidelity of the reconstruction, with respect to the underlying biochemistry. The comprehensiveness of a genome-scale metabolic reconstruction can be improved by using the model to detect and fill network gaps (Rolfsson *et al.*, 2011). Similarly, reconstruction fidelity can be improved by using the model to detect reconstruction stoichiometry inconsistent with biochemistry (Gevorgyan *et al.*, 2008) or reactions inconsistent with steady state flux (Vlassis *et al.*, 2014).

Existing gap-filling algorithms, reviewed by Orth and Palsson (2010), become intractable in high dimensions. Decompartmentalization of genome-scale compartmentalized metabolic networks reduces their dimension, rendering gap filling tractable (Rolfsson *et al.*, 2011). However, this approach underestimates the amount of missing information because it connects reactions that would normally not co-occur in the same cellular compartment.

We present fastGapFill, the first scalable algorithm capable of efficiently detecting and filling network gaps in compartmentalized genome-scale models. fastGapFill draws on, and extends, fastcore (Vlassis *et al.*, 2014), an algorithm to approximate the cardinality function to identify a compact flux consist ent model, in which all reactions carry a non-zero flux in at least one flux distribution. fastGapFill allows integrating all three notions of model consistency, namely, gap-filling, flux consistency and stoichiometric consistency in a single tool.

2 METHODS

Formulation of the gap-filling problem. In the metabolic gap-filling problem (Reed *et al.*, 2006), one starts with a computational metabolic model, M, that contains at least one *blocked* reaction, which, though desired, does not admit a non-zero steady state flux. From a universal database, e.g. the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), a search is made for at least one reaction that needs to be added to fill at least one *gap* in the model, such that at least one formerly blocked reaction can carry flux. Among other criteria, it may also be desirable to compute a compact flux consistent model, where the number of added universal reactions is minimal. A specific instance of this problem occurs in metabolic modeling, although our algorithm is applicable for any biochemical network model with gaps.

Computing a compact flux consistent model. We repurposed the recently developed fastcore algorithm (Vlassis *et al.*, 2014) to compute a near-minimal set of reactions that need to be added to an input metabolic model M to render it flux consistent. fastcore takes inputs M and a *core* set of reactions $C \subset M$. Then, it greedily expands C by computing a set of modes of M whose overall support contains the whole of C and a minimal set from $M \backslash C$. This is achieved by a series of L_1-norm regularized linear programs that optimize a relaxed version of an (intractable) integer program under cardinality constrains (Vlassis *et al.*, 2014). Our implementation efficiently identifies blocked reactions.

Preprocessing to generate a global model. A cellularly compartmentalized metabolic model (S) without blocked reactions (B), where $S \cup B \equiv M$, is expanded by a universal metabolic database U (e.g. KEGG), such that a copy of U is placed in each cellular compartment of S (including the extracellular space), to generate SU. For each metabolite occurring in a non-cytosolic compartment, a reversible intercompartmental transport reaction is added. For each extracellular metabolite, an exchange reaction is added. The sum of the latter two reaction sets (X) is added to SU to generate a global model, which is extended with *solvable blocked reactions* ($B_s \subset B$), that is, reactions that were previously flux inconsistent but become flux consistent when added to the global model. In the extended global model (SUX), all reactions are flux consistent. Note that not all blocked reactions B may be solvable, and thus, they will not be present in SUX. All reactions of S and B_s represent the core set.

*To whom correspondence should be addressed.

Table 1. Gap filling of metabolic reconstructions on a standard desktop computer (Dell, Intel Core i5, 16 GB RAM, 64 bit)

Model name	*Thermotoga maritima* (Zhang *et al*., 2009)	*Escherichia coli* (Feist *et al*., 2007)	Synechocystis sp. (Nogales *et al*., 2012)	sIEC (Sahoo and Thiele, 2013)	Recon 2 (Thiele *et al*., 2013)
S[a]	418 × 535	1501 × 2232	632 × 731	834 × 1260	3187 × 5837
SUX[a]	14 020 × 31 566	21 614 × 49 355	28 174 × 62 866	48 970 × 109 522	58 672 × 132 622
Comp[b]	2	3	4	7	8
B	116	196	132	22	1603
B_s	84	159	100	17	490
Number of gap-filling reactions	87	138	172	14	400
$t_{preprocessing}$ (s)[c]	52	237	344	1003	5552
$t_{fastGapFill}$ (s)	21	238	435	194	1826

[a]The dimensions are given as metabolites × reactions.
[b]Comp, compartments.
[c]Preprocessing includes computing the flux consistent metabolic model, merging of UX for all compartments of S and adding solvable blocked reactions B_s.
Note: Equal weighting of all reactions was used. See Supplementary Table S1 for candidate gap-filling solutions.

Computing a compact flux consistent subnetwork of a global model. fastGapFill computes a subnetwork of SUX, consisting of all core reactions, plus a minimal number of reactions from UX, such that all reactions in the resulting compact subnetwork are flux consistent. This is achieved by using a slightly modified version of fastcore, in which a vector of linear weightings prioritizes the addition of reactions within UX. For instance, one may prioritize the addition of metabolic reactions from U over transport reactions from X, or, by varying the weightings on non-core reactions, alternate compact sets of gap-filling reactions may be identified.

Optional analysis of gap-filling reactions. We provide the option to compute a flux vector that maximizes the flux through each blocked reaction in turn, while minimizing the Euclidean norm of flux through the subnetwork of SUX computed by one call to fastGapFill. Note that flux through more than one solvable blocked reaction may be necessary to fill a gap, and that the computed flux vector may not be of minimum cardinality.

Computing stoichiometric consistency. Many reaction databases contain *stoichiometric inconsistencies* (Gevorgyan *et al*., 2008), where the stoichiometry for at least two reactions is inconsistent with conservation of mass. For instance, the reactions $A \rightleftharpoons B$ and $A \rightleftharpoons B + C$ are stoichiometrically inconsistent, as no positive molecular mass can be assigned to A, B and C, such that the mass on both sides of both reactions is equal. fastGapFill allows to identify stoichiometrically inconsistent reactions from filling gaps, by using the scalable approach for approximate cardinality maximization used within fastcore, to compute a maximal set of metabolites in U that are involved in reactions that conserve mass.

3 IMPLEMENTATION

An open source, MATLAB (Mathworks, Inc.), implementation of fastGapFill is available as a cross-platform desktop computer extension to the openCOBRA toolbox (Schellenberger *et al*., 2011).

4 DISCUSSION

We applied fastGapFill to five metabolic models (Table 1), demonstrating its broad applicability and scalability for various sizes of the gap-filling problem. Alternate gap-filling solutions can be computed by changing weightings on non-core reactions in the preprocessed problem. Note that all candidate metabolic and transport reactions are hypotheses requiring experimental validation (Rolfsson *et al*., 2011). Our implementation provides an openCOBRA (Schellenberger *et al*., 2011) compatible version of the KEGG reaction database; however, any other universal reaction database could be used with fastGapFill, so long as the same input format is maintained and care is taken to correctly identify identical metabolites. fastGapFill is the first scalable approach to identify candidate missing knowledge in compartmentalized metabolic reconstructions, and the approach is applicable to any form of biochemical network gap-filling problem.

Funding: I.T. was supported by an ATTRACT program grant (FNR/A12/01) from the Luxembourg National Research Fund (FNR). R.F. was supported by the Interagency Modeling and Analysis Group, Multi-scale Modeling Consortium U01 awards from the National Institute of General Medical Sciences, award GM102098-01, and U.S. Department of Energy, Office of Science, Biological and Environmental Research Program, award ER65524.

Conflict of Interest: none declared.

REFERENCES

Feist,A.M. *et al.* (2007) A genome–scale metabolic reconstruction for *Escherichia coli* K–12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Gevorgyan,A. *et al.* (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, **24**, 2245–2251.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Nogales,J. *et al.* (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proc. Natl Acad. Sci. USA*, **109**, 2678–2683.

Oberhardt,M.A. *et al.* (2010) Metabolic network analysis of *Pseudomonas aeruginosa* during chronic cystic fibrosis lung infection. *J. Bacteriol.*, **192**, 5534–5548.

Orth,J. and Palsson,B.Ø. (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.*, **107**, 403–412.

Palsson,B.Ø. (2006) *Systems Biology: Properties of Reconstructed Networks*. Cambridge Univ Press, New York.

Reed,J.L. *et al.* (2006) Systems approach to refining genome annotation. *Proc. Natl Acad. Sci. USA*, **103**, 17480–17484.

Rolfsson,O. *et al.* (2011) The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *BMC Syst. Biol.*, **5**, 155.

Sahoo,S. and Thiele,I. (2013) Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. *Hum. Mol. Genet.*, **22**, 2705–2722.

Schellenberger,J. *et al.* (2011) Quantitative prediction of cellular metabolism with constraint–based models: the COBRA Toolbox v2.0. *Nat. Protoc.*, **6**, 1290–1307.

Thiele,I. *et al.* (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **31**, 419–425.

Vlassis,N. *et al.* (2014) Fast reconstruction of compact context-specific metabolic network models. *PLoS Comp. Biol.*, **10**, e1003424.

Zhang,Y. *et al.* (2009) Three–dimensional structural view of the central metabolic network of *Thermotoga maritima*. *Science*, **325**, 1544–1549.