Bioinformatics, 33(4), 2017, 608–611
doi: 10.1093/bioinformatics/btw684
Advance Access Publication Date: 24 November 2016
Applications Note

Systems biology

BoostGAPFILL: improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods

Tolutola Oyetunde[1,†,*], Muhan Zhang[2,†], Yixin Chen[2], Yinjie Tang[1] and Cynthia Lo[1,*]

[1]Department of Energy, Environmental and Chemical Engineering and [2]Department of Computer Science and Engineering, Washington University, Saint Louis, MO 63130, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on July 19, 2016; revised on October 23, 2016; editorial decision on October 24 2016; accepted on October 25, 2016

Abstract

Motivation: Metabolic network reconstructions are often incomplete. Constraint-based and pattern-based methodologies have been used for automated gap filling of these networks, each with its own strengths and weaknesses. Moreover, since validation of hypotheses made by gap filling tools require experimentation, it is challenging to benchmark performance and make improvements other than that related to speed and scalability.

Results: We present BoostGAPFILL, an open source tool that leverages both constraint-based and machine learning methodologies for hypotheses generation in gap filling and metabolic model refinement. BoostGAPFILL uses metabolite patterns in the incomplete network captured using a matrix factorization formulation to constrain the set of reactions used to fill gaps in a metabolic network. We formulate a testing framework based on the available metabolic reconstructions and demonstrate the superiority of BoostGAPFILL to state-of-the-art gap filling tools. We randomly delete a number of reactions from a metabolic network and rate the different algorithms on their ability to both predict the deleted reactions from a universal set and to fill gaps. For most metabolic network reconstructions tested, BoostGAPFILL shows above 60% precision and recall, which is more than twice that of other existing tools.

Availability and Implementation: MATLAB open source implementation (https://github.com/Tolutola/BoostGAPFILL)

Contacts: toyetunde@wustl.edu or muhan@wustl.edu.

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

1 Introduction

Genome-scale metabolic reconstructions are the basis of constraint-based analyses, which are finding ever increasing applications in metabolic engineering for industrial, medical and environmental purposes (Bordbar *et al.*, 2014). One of the major reasons for inconsistencies between genome-scale model predictions and experimental measurements is the presence of gaps in the network reconstruction (Palsson, 2015). Knowledge gaps are the result of missing information on genes, proteins, or reactions, while scope gaps occur due to the fact the metabolic network is only one of several integrated cellular networks (e.g. signaling networks). Thus, the consumption and production of a metabolite might not be fully captured by metabolism alone. Moreover, some microbes that depend on communal support of other organisms actually have gaps in their metabolism. Therefore, automated gap filling tools are merely hypotheses generators whose predictions need to be verified experimentally.

Two general approaches to tackle the challenge of network gaps have been reviewed (Orth and Palsson, 2010). The first involves the use of algorithms based on network topology and genomic data. These are mostly concerned with finding gene candidates for orphan reactions. The second seeks to find missing reactions by minimizing the difference between computation and experiments. Gap-filling algorithms serve a dual benefit of model refinement and discovery of new biological capabilities (Orth and Palsson, 2010). Thus, efficient and robust gap-filling algorithms would prove invaluable in the development of high fidelity metabolic network reconstructions (Latendresse *et al.*, 2012). Newer approaches have sought to uncover inherent patterns in metabolic networks and have shown promise in predicting diverse network functions (Ganter *et al.*, 2014). However, some of the predictions based on these methods might not be biologically realizable. Constraint-based methods, on the other hand, may not capture the information embedded in the network topology.

It is difficult to test the accuracy of gap filling algorithms because verification usually involves experimentation to examine the biological relevance of suggested reactions. Thus, it is important to develop benchmark tests for gap filling algorithms to increase confidence in their use.

In this work, we present a novel gap-filling framework, BoostGAPFILL, which integrates constraint-based and pattern-based methods Zhang *et al.* (2016) for metabolic network refinement. Our framework is inspired by machine learning methods developed for the Netflix prize (Koren *et al.*, 2009). We test the robustness of the gap-filling algorithms using artificial gaps (i.e. metabolites that cannot be produced or consumed at steady state) to simulate poorly characterized biochemistry. The gaps are introduced by randomly deleting reactions from the network. We then rank the algorithms on their ability to predict the actual deleted reactions from a universal reactions database and unblock blocked metabolites (i.e. gaps).

2 Methods

Our novel algorithm combines machine learning and constraint-based methods to identify possible candidates for missing reactions. We use machine learning to characterize the topology of the incomplete metabolic network and predict a set of possible reactions. The preliminary predictions are integrated with standard constraint-based gap filling in two ways: (i) using the preliminary predictions as weighting factors in constraint-based algorithms and (ii) solving the pattern-based problem simultaneously with the standard gap filling formulation (Maranas and Zomorrodi, 2016). Details of this are described in the supplementary file. The basic concepts of the pattern module of our algorithm are shown in Figure 1. The mathematical details are presented in Box S1 of the supplementary file.

2.1 Step A: conversion of incomplete stoichiometric matrix to metabolite adjacency matrix

The binary incidence matrix, $\widehat{S}$, can be derived from the stoichiometric matrix, S, by simply placing a one if the corresponding entry in the stoichiometric matrix is not zero, and a zero if otherwise. Post multiplying $\widehat{S}$ with its transpose gives an m by m metabolite adjacency matrix, **A**, where m is the number of metabolites. **A** provides information about the relationship between the different metabolites. Each entry gives the number of reactions in which the two metabolites jointly participate.

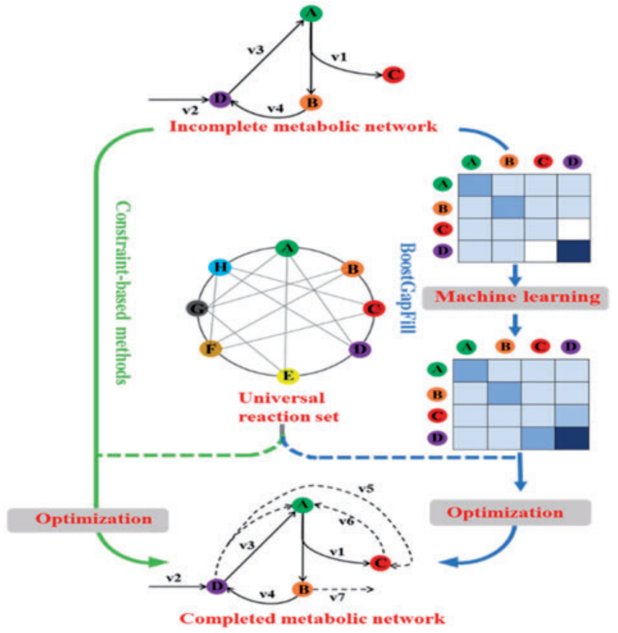

Fig. 1. Basic concepts of the pattern-based module of BoostGAPFILL (right) contrasted with constraint-based procedures (left). In BoostGAPFILL, the partial adjacency matrix is derived from the incomplete stoichiometric matrix. The partial adjacency matrix is completed using matrix factorization models. Then reactions are selected from a universal database. The selection is formulated as an integer least squares problem in which the difference between the completed adjacency matrix is transformed to the stoichiometric matrix. In constraint-based procedures, the reactions are selected directly from the universal reactions database using an optimization criterion, such as minimum number of reactions required to fill the gaps in the network

2.2 Step B: completion of metabolite adjacency matrix using matrix factorization

The entries of **A** conceptually represent the ranking of the relationship between metabolites. **A** is incomplete and we employ the standard matrix factorization model (Koren *et al.*, 2009) as implemented in the free tool libFM (Rendle, 2012) for its completion. Slight modifications are discussed in Box 1 of the supplementary file.

2.3 Step C: prediction of new reactions from a universal reaction set

Next, we attempt to recover the completed **S** by an integer least squares optimization in which we select reactions from a universal set that best match the completed **A**. The integer least squares optimization is relaxed to avoid long computational times associated with integer optimization problems. The result is a ranking of all reactions. Selections are made based on the top percent threshold or the top number of reactions. This step (of selecting reactions from a set based on some constraints) is common to standard gap filling tools, and is the step where we integrate standard constraints.

2.4 Modes of running BoostGAPFILL

BoostGAPFILL can be run in three modes (shown in Supplementary Fig. S1). Mode 1: the tool is run as described above. Thus, the

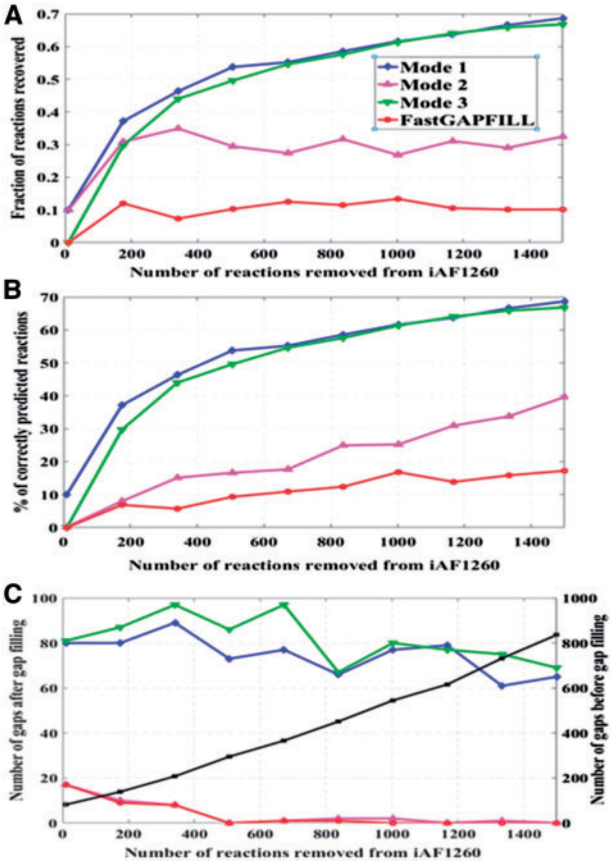

Fig. 2 Comparison of the performance of gap-filling algorithms on *E.coli* model iAF1260. (**A**) Reactions are selected at random and deleted from iAF1260. The number of reactions deleted is shown on the *x* axis. Gap-filling algorithms are then used to predict possible candidates to complete the network. The number of reactions correctly predicted as a fraction of the number of reactions deleted is shown on the *y* axis. For BoostGAPFILL run in mode 1 and 3, the number of predicted reactions is the same as the number of deleted reactions (this can be manually set in the algorithm). For other algorithms the number of reactions predicted vary and cannot be directly set. (**B**) For the same simulation described above, the number of reactions removed from iAF1260 is shown on the *x*-axis, and the number of correctly predicted reactions is shown as a percentage of the total number of reactions predicted by the algorithm. (**C**) The number of gaps in the network before (shown as a black line) and after gap filling is shown on the right and left *y* axes respectively. Note that the model before gap filling has a certain number of reactions deleted (as seen on the *x* axis). Both mode 2 of BoostGAPFILL and FASTGAPFILL completely fill all the gaps

predictions are based solely on the inherent metabolite patterns in the incomplete network. This mode is very accurate at capturing the topological information in the network as seen in Figure 2 but does not fill all the gaps. Mode 2: The pattern based module is used to weight reactions in the universal database for use in FASTGAPFILL. Thus, BoostGAPFILL is used as a preprocessing step for FASTGAPFILL. This improves the fidelity of FASTGAPFILL as demonstrated in Figure 2. Mode 3: In this mode, we include the flux constraints (used in the standard constraint-based gap filling formulation) in step C described above. This enables BoostGAPFILL to be used for growth inconsistency reconciliation like tools such as SMILEY.

Running BoostGAPFILL in mode 1 is preferred for initial screening of a large reactions database, with mode 2 and mode 3 preferred for more biologically realistic predictions. Mode 2 is best for pure gap filling while mode 3 can be used for growth data reconciliation and predicting reactions to unblock metabolites in turn. The limitations and technical implementation detials are discussed in the supplementary file.

3 Results

We test the performance of BoostGAPFILL on seven different metabolic network reconstructions downloaded from the BiGG database (King *et al.*, 2016). Figure 2 presents the comparison of the performance of BoostGAPFILL and FASTGAPFILL on the *E. coli* model iAF1260. BoostGAPFILL automatically fixes gaps (also see Supplementary Fig. S2). It also appears to perform well even when a large number of reactions are missing. The algorithm was able to predict several new reactions added in iJO1366 (the latest *E.coli* model) from an earlier version (iAF1260) including new content (15 gap filling reactions and 4 new content reactions), as shown in Supplementary Figure S3 in the supplementary file. While tools like FASTGAPFILL (Thiele *et al.*, 2014) and SMILEY (Reed *et al.*, 2006) perform well in predicting reactions that close as many gaps as possible (Fig. 2C), BoostGAPFILL outperforms them in terms of preserving the network topology (Fig. 2). This illustrates the fact that constraint-based techniques can sometimes fail to capture the embedded patterns in metabolic networks and thus their predictive fidelity is compromised. BoostGAPFILL provides that missing functionality and easily integrates with the existing gap filling tools. Similar performance was observed in other metabolic network reconstructions as seen in Supplementary Figures S4 and S6. BoostGAPFILL can also make predictions of reactions contaning metabolites not in the original network (see supplementary file for dicussion and Supplementary Fig. S5).

4 Conclusions

A unique methodology, integrating topology-based and constraint-based approaches to refining metabolic network reconstructions has been presented. The performance of BoostGAPFILL has been rigorously tested on different metabolic reconstructions. Approaches that combine machine learning models and pure mechanistic models to describe biological phenomena will prove useful in decoding complex interactions that exist in living systems. Integrating pattern-based methods with constraint-based techniques can potentially enhance their predictive fidelity in computational strain design for metabolic engineering.

Funding

This work was supported by the National Science Foundation (MCB 1616619 and DBI1356669).

Conflict of Interest: none declared.

References

Bordbar,A. *et al.* (2014) Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.*, **15**, p107–120.

Ganter,M. *et al.* (2014) Predicting network functions with nested patterns. *Nat. Commun.*, **5**, 3006.

King,Z.A. *et al.* (2016) BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, **44**, p D515–D522.

Koren,Y. *et al.* (2009) Matrix factorization techniques for recommender systems. *Computer*, **42**, 42–49.

Latendresse,M. *et al*. (2012) Construction and completion of flux balance models from pathway databases. *Bioinformatics*, **28**, 388–396.

Maranas,C.D. and Zomorrodi,A.R. (2016) *Optimization Methods in Metabolic Networks*, John Wiley & Sons, Inc., Hoboken, NJ.

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

Palsson,B. (2015) *Systems Biology: Constraint-Based Reconstruction and Analysis*, Cambridge University Press, Cambridge, UK.

Reed,J.L. *et al*. (2006) Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U. S. A*., **103**, p17480–17484.

Rendle,S. (2012) Factorization machines with libfm. *ACM Trans. Intell. Syst. Technol. (TIST)*, **3**, 57.

Thiele,I. *et al*. (2014) FASTGAPFILL: efficient gap filling in metabolic networks. *Bioinformatics*, **30**, 2529–2531.

Zhang, Muhan *et al*. (2016) Recovering Metabolic Networks using A Novel Hyperlink Prediction Method. arXiv preprint arXiv:1610.06941.