BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 20 2014, pages 2986–2988
doi:10.1093/bioinformatics/btu422

Systems biology

Advance Access publication July 3, 2014

FindPath: a Matlab solution for *in silico* design of synthetic metabolic pathways

Gilles Vieira[1,2,3], Marc Carnicer[1,2,3], Jean-Charles Portais[1,2,3] and Stéphanie Heux[1,2,3,*]

[1]Université de Toulouse; INSA, UPS, INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France [2]INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France and [3]CNRS, UMR5504, F-31400 Toulouse, France

Associate Editor: Igor Jurisica

ABSTRACT

Summary: Several methods and computational tools have been developed to design novel metabolic pathways. A major challenge is evaluating the metabolic efficiency of the designed pathways in the host organism. Here we present FindPath, a unified system to predict and rank possible pathways according to their metabolic efficiency in the cellular system. This tool uses a chemical reaction database to generate possible metabolic pathways and exploits constraint-based models (CBMs) to identify the most efficient synthetic pathway to achieve the desired metabolic function in a given host microorganism. FindPath can be used with common tools for CBM manipulation and uses the standard SBML format for both input and output files.

Availability and implementation: http://metasys.insa-toulouse.fr/software/findpath/.

Contact: heux@insa-toulouse.fr

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

Received on January 27, 2014; revised on June 3, 2014; accepted on June 25, 2014

1 INTRODUCTION

Broadening the spectrum of compounds that microorganisms are able to convert into biomass or other valuable products is attracting increasing interest in biotechnology and synthetic biology. Metabolic pathways are introduced in the organism to enable the conversion of the desired compound into host metabolites, which, in turn, can be further converted into biomass or other valuable products. Computational tools are already available to design such pathways, using the large repertoire of reactions and pathways found in the microbial world, or from novel—synthetic—reactions, or combinations of both. These tools use different strategies to identify all possible metabolic pathways able to support the desired metabolic function and to select the most suitable among the many possibilities. To search for all possible pathways, methods such as FMM (Chou *et al.*, 2009) and DESHARKY (Rodrigo *et al.*, 2008) rely on databases of known chemical reactions and compounds, whereas BNICE, RDM and others capture all biochemical transformations in the form of reaction rules extracted from known biochemical reactions (Finley *et al.*, 2009; Medema *et al.*, 2012; Oh *et al.*, 2007). Alternatively, Retropath (Carbonell *et al.*, 2011) uses a molecular signature-based approach. The best pathway(s) are then identified using algorithms allowing to rank all the pathways generated based on scoring criteria such as thermodynamic feasibility, pathway length, energetic costs, genetic loads, etc. (Medema *et al.*, 2012). However, the metabolic efficiency of the identified pathways in the host metabolism is rarely included in scoring criteria despite its critical importance (Medema *et al.*, 2012). Here we introduce FindPath, a standardized framework for the prediction of metabolic pathways enabling the conversion of one or more novel compounds into any host metabolite. The tool combines a database of all reactions associated with the metabolism of the novel compound, a constraint-based model (CBM) of the host organism, and tools to explore, prioritize and score metabolic pathways. Pathways are scored based on both topological feature (i.e. pathway length) and metabolic efficiency in the host organism, thus making it possible to tackle pathway design from a quantitative and functional point of view. FindPath is an easy-to-use Matlab package to streamline the design of novel synthetic pathways. Finally, FindPath predicts and ranks pathways in a single package that can be run in a couple of minutes.

2 FINDPATH WORKFLOW

2.1 Implementation

FindPath is implemented in Matlab, a programming language that is widely used in the metabolic modeling community, and requires minimal user programming skills. FindPath integrates existing tools for CBM manipulation, i.e. COBRA-Toolbox (Becker *et al.*, 2007) and EFMTool (Klamt and Gilles, 2004), as well as novel functions for the design of metabolic pathways. Such implementation allows to benefit from additional tools included in the COBRA-Toolbox and required by our method [i.e. flux balance analysis (FBA)]. All are then incorporated into a single platform. FindPath can use different file formats, including CSV, SBML (i.e. the standard format in system biology modeling) (Hucka *et al.*, 2003) and the COBRA-toolbox format. It also includes a function to convert CSV files into other file formats. Errors during import and export of data and files are reduced to a minimum, and exploitation of the workflow is facilitated for beginners. Finally, FindPath can be easily integrated into more complex and evolved workflows using other Matlab CBM-dedicated methods or tools that use the

*To whom correspondence should be addressed.

SBML format. A user manual is provided in Supplementary Materials.

A schematic diagram of the computational workflow of FindPath is given in Supplementary Figure S1 and is described step by step in the following sections.

2.2 Creation of the substrate-associated reaction database

FindPath requires a dedicated substrate-associated reaction (SAR) database including all the reactions and processes involved in the metabolic conversion of the compound of interest (see Supplementary Fig. S1A). This information can be extracted from metabolic databases such as MetaCyc (Karp *et al.*, 2002) or KEGG (Kanehisa *et al.*, 2008) or from the literature. The SAR database can be generated manually (See Supplementary Data). The reactions included in the database must be balanced, generic reactions must be instanced and reversibility of each reaction must be given. To ensure full compatibility between the different tools, the identifiers of metabolites and reactions in the SAR database and in the host organism CBM must be unified and must respect the COBRA-toolbox formalism.

2.3 Prediction of possible pathways

FindPath predicts possible pathways capable of converting the novel compound into any host metabolites by elementary flux mode (EFM) analysis (Klamt and Stelling, 2003). To this end, the information contained in the SAR database is converted into an SAR model (see Supplementary Fig. S1B and Supplementary Materials). The resulting SAR model is then converted by FindPath into the EFMTool format (Klamt and Gilles, 2004) by linking all equivalent fields (stoichiometric matrix, list of reactions list of reactions list, reversibility, etc.). The input metabolite is the novel compound to be converted. The output metabolites are all host metabolites, but a subset of host metabolites can be specified as output metabolites when appropriate (e.g. to exclude cofactors or undesired intermediates). All the resulting EFMs enabling the conversion of the novel compounds into a host metabolite are considered as possible solutions. Using this process, both naturally occurring pathways and pathways resulting from novel combinations of reactions—i.e. synthetic pathways—can be predicted.

2.4 Prioritization of pathways

Depending on the number of reactions in the SAR database, the number of pathways predicted by FindPath may increase dramatically. FindPath makes it possible to prioritize the different pathways based on two criteria: the length of the pathway and the metabolite composition of the pathway (see Supplementary Fig. S1C). Short pathways reduce both the genetic effort needed to construct the strain and the energetic costs related to the biosynthesis and maintenance of enzymes (Medema *et al.*, 2012). The second criterion helps prioritize pathways according to metabolic considerations, e.g. selecting pathways containing metabolites of interest or discarding pathways containing toxic compounds. The pathways selected during this step are kept for the next step.

2.5 Ranking of selected pathways

For each pathway selected, FindPath creates a specific pathway/organism CBM by incorporating the pathway reactions into the genome-scale model of the selected host. Constraints on these newly added reactions (i.e. upper and lower bound of the flux) can be set by the user or an arbitrary range can be automatically applied by the COBRA-Toolbox functions. For each model, FBA is performed after an objective function corresponding to the desired metabolic capability (e.g. production of biomass or product of interest, etc.) has been defined in the model. To determine the best trade-off between genetic effort and metabolic efficiency, pathways are ranked based on a weighted score that accounts for both pathway length (i.e. data obtained during the previous step) and efficiency (i.e. results of the FBA) (see Supplementary Fig. S1D and Supplementary Materials). The weight of each parameter can be defined by the user. This allows long pathways with high productivity to be selected over short pathways with low productivity. Results are exported to SBML file to ensure the data produced are fully compatible when incorporated in a larger workflow and in a log file. The number of pathways to be exported can be specified by the user.

3 TEST CASE

As proof of concept, FindPath was used to select the best pathways to enable growth of *Saccharomyces cerevisiae* on D-xylose. Although this sugar is one of the main monomers present in lignocellulosic hydrolysates, *S.cerevisiae* is unable to grow on or consume it (Kim *et al.*, 2013). To achieve economically viable ethanol production from lignocellulosic feedstocks, it is essential to generate a pentose using yeast, and this is the focus of many current studies. Starting with an SAR database of 29 reactions and a genome-scale model of 1412 reactions, i.e. iMM904 (Zomorrodi and Maranas, 2010), FindPath was able to identify 15 potential pathways in <1 min in an XEON E5410 (8 cores) and 16 GB of RAM. Interestingly, the two pathways with the highest scores matched the pathways (i.e. xylose isomerase and xylulose kinase (XK) or xylose reductase, xylitol dehydrogenase and XK) that were successfully used to obtain a xylose-fermenting *S.cerevisiae* strain (Kim *et al.*, 2013). More details can be found in Supplementary Materials.

4 CONCLUSION

FindPath is a simple unified workflow for the *in silico* design of pathways to convert novel compounds into any host metabolites. FindPath rapidly identifies the best solution but also alternative solutions with topological and efficiency interest. It provides additional features to the COBRA-Toolbox and can also be used in addition—or in parallel—to other computational tools that accept SBML formats. FindPath is a flexible tool that can be used for many purposes, including: (i) prediction of pathways involved in the conversion of a carbon source that is not naturally consumed by the host organism; (ii) design of degradation pathways for the detoxification of toxic compounds; (iii) curation of CBMs, by providing candidate pathways when experimental evidence points to metabolic conversion, but the pathway has not been identified; and (iv) choice among a large set of

compounds, the most suitable starting compound (e.g. carbon source) to achieve the high-yield synthesis of the desired product. This new tool should speed up the development of efficient microbial cell factories.

Funding: PROMYSE, 7th FWP, KBBE.2011.3.6-04 Applying Synthetic Biology principles towards the cell factory notion in biotechnology.

Conflict of interest: none declared.

REFERENCES

Becker,S.A. *et al.* (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protoc.*, **2**, 727–738.

Carbonell,P. *et al.* (2011) A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst. Biol.*, **5**, 122.

Chou,C.H. *et al.* (2009) FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res.*, **37**, W129–W134.

Finley,S.D. *et al.* (2009) Computational framework for predictive biodegradation. *Biotechnol. Bioeng.*, **104**, 1086–1097.

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Karp,P.D. *et al.* (2002) The MetaCyc database. *Nucleic Acids Res.*, **30**, 59–61.

Kim,S.R. *et al.* (2013) Strain engineering of *Saccharomyces cerevisiae* for enhanced xylose metabolism. *Biotechnol. Adv.*, **31**, 851–861.

Klamt,S. and Gilles,E.D. (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics*, **20**, 226–234.

Klamt,S. and Stelling,J. (2003) Two approaches for metabolic pathway analysis? *Trends Biotechnol.*, **21**, 64–69.

Medema,M.H. *et al.* (2012) Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol.*, **10**, 191–202.

Oh,M. *et al.* (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Inf. Model.*, **47**, 1702–1712.

Rodrigo,G. *et al.* (2008) DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, **24**, 2554–2556.

Zomorrodi,A.R. and Maranas,C.D. (2010) Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data. *BMC Syst. Biol.*, **4**, 178.