A new approach to obtain EFMs using graph methods based on the shortest path between end nodes

José F. Hidalgo, Francisco Guil, and José M. García

Grupo de Arquitectura y Computación Paralela
Universidad de Murcia, Spain
{jhidalgo,fguil,jmgarcia}@um.es
http://www.um.es/gacop

Abstract. Genome-scale metabolic networks let us to understand the behavior of the metabolism in the cells of live organisms. The availability of great amounts of such data gives scientific community the opportunity to infer *in silico* new metabolic knowledge. Elementary Flux Modes (EFM) are minimal contained pathways or subsets of a metabolic network that are very useful to achieve the comprehension of a very specific metabolic function (as well as dis-functions), and to get the knowledge to develop new drugs. Metabolic networks can have large connectivity and, therefore, EFMs resolution faces a combinational explosion challenge to be solved. In this paper we propose a new approach to obtain EFMs based on graph methods and the shortest path between end nodes. Our method finds all the pathways in the metabolic network and it is able to prioritize the pathway search accounting the biological mean pursued. Our technique has two phases, the exploration one and the characterization one, and we show how it works in a well-known case study.

Keywords: Metabolic networks, graph theory, EFM, flux modes, pathways, systems biology

1 Motivation

Cellular metabolism is the set of biochemical enzyme-catalyzed reactions involved in the generation of nutrients and energy necessary for the cells in living organisms. Those reactions are equations of metabolites with stoichiometric coefficients. All the reactions and metabolites used to be grouped in a stoichiometric matrix. A metabolic pathway of a cell is a piece of the network, that is, a sequence of some of its reactions. Metabolic pathways have been found quite useful in different domains such as personalized medicine, drug discovery techniques or genomic feature discovery. Therefore, many efforts have been lately made to find pathways experimentally or by inferring them computationally.

Several mathematical methods modeling metabolism are emerging that are able to incorporate datasets provided by different *omics* technologies. Many of these methods are encompassed within constraint-based models, in which

a set of mathematical constraints are defined using a genome-scale metabolic network (GSMN) reconstruction as a starting point. Several curated GSMNs can be found in the literature [19]. However, being able to automatically characterize the biochemical reactions present in a particular metabolism through *omics* data truly constitutes a challenge [15].

The term constraint-based modeling (CBM) groups different approaches that analyze the metabolic behavior based on the stoichiometric relations between compounds participating in enzymatic reactions. CBM defines two constraints that pathways must fulfill. The first one is the steady-state condition that refers to the property of mass balance within the cell. In other words, the concentration of internal metabolites remains constant over the time. The second relevant constraint refers to thermodynamic feasibility, which restricts some fluxes from being non-negative (irreversibility constrain).

An elementary flux mode (EFM) [16] is a special type of metabolic pathway comprising a subset of reactions that meets the two aforementioned conditions plus the non-decomposability condition, that is, the pathway cannot be decomposed into smaller solutions (i.e., a subset of the pathway is not a feasible pathway as well). In other words, EFMs are solutions with the minimum support necessary to operate in stoichiometric steady-state balance with all reactions in the appropriate direction. EFMs are an effort to translate a complex network into a canonical expression of vector generators of a solution space.

In a typical metabolic network the number of reactions is higher than the number of metabolites, so that many possibilities can be found that are a solution to the system. As the metabolic network increases in size so do the amount of EFMs, which number explodes in a combinatorial fashion. Computing the full set of EFMs in large metabolic networks still constitutes a challenging issue.

Continuing with this effort, we have developed a new method to find systematically all the pathways from a metabolic network. In this paper we present our approach based on a novel strategy to find shortest pathways between end nodes in a graph representation of the network. Specifically, our approach exploits the well-known graph theory and tools to drive the search of EFMs prioritizing, if needed, the pathway search to account the biological mean quest. Our technique is composed of two phases, the exploration and the characterization one, and along the paper we describe the how the first phase works in a case study.

Unlike traditional Linear Programming (LP) approaches, our proposal avoids expensive floating-point calculations allowing us to speed-up the quest of all the available pathways in a certain metabolic network. Moreover, our approach is quite suitable to be developed in new commodity parallel architectures (such as multi- and many-cores and accelerators like GPUs), allowing shorter execution times and less energy consumption.

The rest of the paper is structured as follows. Section 2 gives some background on the constraint-based mathematical modeling. In Section 3 we show the method we have followed to design our technique. Section 4 presents a case study of our approach, and the paper concludes giving some related work in Section 5, and offering our conclusions and future work in Section 6.

2 Background

Constraint based modeling (CBM) starts with a stoichiometric matrix S where the values are the stoichiometric coefficients for metabolites (rows) on each reaction (columns). Every reaction is characterized by the reaction rate (also known as flux rate) which numerically gives the rate at which the substrate metabolites are converted to the product metabolites.

Be $\overrightarrow{r}$ a vector of flux rate that represents a pathway, therefore fulfilling the steady-state and the thermodynamic feasibility constrains. The steady-state condition means that internal metabolites are balanced and concentration remains constant ($S \cdot \overrightarrow{r} = \overrightarrow{0}$), and the feasibility constraint means that each irreversible reaction only participates with a positive rate ($\forall i, r_i \geq 0$) when it is part of the solution. Finally, $\overrightarrow{r}$ represents an EFM if the non-decomposability condition is met ($\overrightarrow{r}$ is not a lineal combination of other flux rate vectors).

The stoichiometric matrix S let us build an adjacency matrix that corresponds to the graph $G = (V, E)$, a non-weighted directed bipartite graph where V are both reactions and metabolites, and the edges E are directed attending the sign of the stoichiometric coefficients. A pathway is a sub-graph of G, $G' = (V', E')$, which is equivalent to $\overrightarrow{r}$ and vice versa.

A known drawback of graph exploration methods is that the flux rate vector is missing at the final of the process. In order to verify which ones of the pathways founds are EFMs, it is needed to do a final verification test using stoichiometry.

3 The shortest path technique to find EFMs

We propose a new CBM approach based on path-finding techniques. Our method consists of two phases, the exploration phase and the characterization one. The exploration phase consists of 3 stages for traversing the graph and finding the feasible pathways. In the first stage, we use the pathway distance metric approach (that is, the amount of reactions participants at the pathway) and take advantage of the fact that it should be biologically meaningful [1]. Therefore, the quest starts the graph exploration by building an axis between a source and a target of the network applying the Dijkstra's shortest path algorithm [2]. The choice of the path end nodes (source, target or both) comes from the biological problem we are dealing with.

At the end of this stage, an axis has been built using the Dijkstras algorithm that traverses the graph through metabolites and reactions from the source node to the target one using the shortest distance. Some of the reactions included in the axis can need metabolites that have not been included yet. We name this kind of metabolites as orphan metabolites.

The second stage goes back from the target to the source (bottom-up approach) to solve the orphan metabolite problem. This process traverses the inverted graph and it is done in a recursive way.

The third stage consists of the simulation of all the reactions that should occur due to the presence of the required metabolites produced by other reactions.

The third stage ends when the end nodes are connected by a complete graph of reactions without orphans nor non-consumed internal metabolites.

After the third stage, our approach has found systematically all the pathways in the axis formed by those end nodes in a metabolic network. This process should be iterative by every pair of interesting end nodes.

The characterization phase needs to check all the pathways produced to determine which of them are EFMs. The final pathways obtained seem minimal because none of the elements can be eliminated without sacrificing consistency. Moreover, pathways fulfill the necessary conditions to have the steady-state balance. However, it cannot be assured the steady-state condition because the stoichiometry is not playing a role during the run of our approach. Without stoichiometry, and depending on the network structure, it can be got a lot of false positives but also some other real positives. In terms of feasibility, the pathways are built fulfilling the necessary conditions to be feasible, that is, respecting the positive direction of every reaction, but the feasibility constrain is only met conditioned to the steady-state consistency.

This second phase is needed as the steady-state constraint has not been granted during the exploration phase of the graph and, therefore, it must be checked afterwards. Currently, we are developing some heuristics to properly select EFMs from the full set of feasible pathways produced.

4 Case study

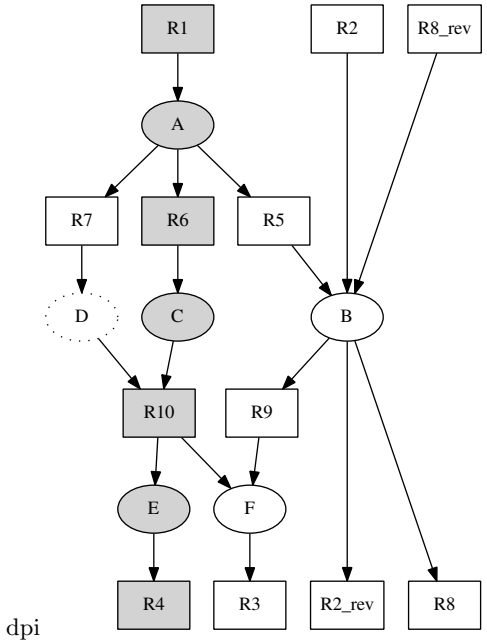
As mentioned before, our approach produces all possible pathways and, for certain cases, EFMs can inferred from those pathways. In simple networks like the EFMtool example published in [4] (6 metabolites and 12 reactions), once all the pathways has been found, the characterization phase has got the full list of EFMs easily discarding decomposable pathways. In addition, for this small and not complex network, the flux rate vector has easily been calculated for any found EFM.

Let us consider as an example the aforementioned network represented by the stoichiometric matrix S shown in the matrix 1. Note that the reactions $R2$ and $R8$ are reversible reactions. For the rest of the process these reactions need to be unfolded in $R2$, $R2\_rev$, $R8$ and $R8\_rev$ automatically. Unfolded reactions must be included in the matrix with individual columns in the new extended stoichiometric matrix that it is shown next. Therefore, all the reactions are from now irreversible.

$$S = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccccccc} R1 & R2^r & R3 & R4 & R5 & R6 & R7 & R8^r & R9 & R10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \quad (1)$$

Based on the extended S we build the graph $G = (V, E)$ represented graphically in the figure 1, where vertices V are both metabolites and reactions, and edges E are the incidence arcs following the direction of the reactions.

$$S = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccccccccc} R1 & R2 & R2\_rev & R3 & R4 & R5 & R6 & R7 & R8 & R8\_rev & R9 & R10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$



dpi

Fig. 1. Graph obtained after the first stage

Our technique starts in the exploration phase, which has three stages. In its first stage, the Dijkstra's shortest algorithm is run to build an axis for the foreseeable pathway. A shortest path for this example is shown in the Figure 1 with the participating nodes in gray. This shortest path is a route between $R1$ as input extreme of our metabolic network and $R4$ as output extreme. Obviously every pair of extreme points can be considered. Many times, the paths obtained in this stage could have the orphan metabolite problem. In the example we are

considering, $R10$ needs that the metabolite D (dotted in the Figure) is also available in the cell to be part of the pathway.

The second stage has the objective to fix this inconvenience. Following with the example, this stage try to include the metabolite D in the axis $\{R1,\ A,\ R6,\ C,\ R10,\ E,\ R4\}$ to form the axis $\{R1,\ A,\ R6,\ C,\ D,\ R10,\ E,\ R4\}$. Many solutions with different complexity can be developed for each found shortest path. In our case, this stage incorporates the reaction $R7$ to the pathway to supply D.

The third stage is responsible to assure that every metabolite produced by the pathway has consumer reactions, that is, it should be consumed inside the pathway. In our example, $R10$ produces the metabolite F but there is no consumer reaction for it. This stage looks for what reactions could occur with the metabolite F in order to be consumed. In this example, there is only one possibility ($R3$ reaction), and it will be incorporated to the pathway. After this three stages we have the pathway $\{R1,\ R6,\ R10,\ R4,\ R7,\ R3\}$ with the metabolites $\{A,\ C,\ D,\ E,\ F\}$ involved in it. The reactions have been shown in the same order they were obtained. The pathway is shown in the Figure 2.

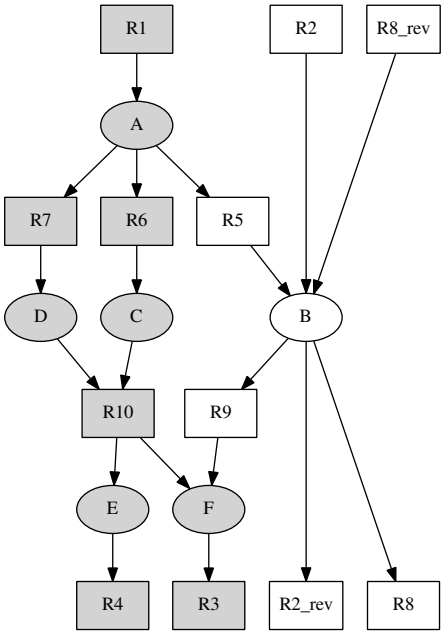


Fig. 2. Final pathway

Following with the example, in [4] the set of EFMs correspondent to the metabolic network are available. One of those EFMs is given by the *flux rate* $\overrightarrow{r} = (2, 0, 1, 1, 0, 1, 1, 0, 0, 1)$ and it corresponds with our pathway $\{R1,\ R6,$

R10, R4, R7, R3} (same non-zero and positive coefficients). Therefore, as there is not possible to have other EFMs with the same non-zero and positive rates, our approach has obtained the same EFM that `efmtool`.

In the end, $\overrightarrow{r}$ can be translated into the stoichiometric sub-matrix S' by maintaining columns and rows correspondent to those nodes of G'. It is important to note that G', S' and $\overrightarrow{r}$ are fully equivalent. It can be proven that if $\overrightarrow{r}$ is an EFM, then $S' \cdot \overrightarrow{r} = \overrightarrow{0}$.

$$S' = \begin{array}{c} \\ A \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} R1 & R3 & R4 & R6 & R7 & R10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Finally, our approach is currently being checked against other small network like the network of *E. coli core model* [3] with 95 reactions and 72 metabolites, and it is obtaining some promising results.

5 Related work

The advantages of analyzing metabolic networks based on EFMs have been shown in different works [5] [14]. However, their use has been limited because enumerating them is computationally demanding. Algorithms have been developed to enumerate all the EFMs in medium-size metabolic networks [20][21] [17]. However, despite the development of novel methods using state of the art computational techniques expediting their application in larger networks [7], this family of algorithms fails on GSMNs using standard computers, because of the combinatorial explosion in the number of EFMs [9]. In this light, several methods have been recently proposed to determine a subset of EFMs in GSMNs [6] [12] [13].

Computational approaches to metabolic pathways can be classified in two groups: stoichiometric approaches and path-finding approaches [10]. Summarizing, the first ones use the stoichiometric data to do calculations during the process. Linear Programming and Null-Space Algorithm [8] are some of the mathematical strategies applied to find pathways, mainly solving the system of linear equations propose by the stoichiometric matrix. Stoichiometric approaches have the quality of impose biochemically meaningful stoichiometric constraints to the solutions but at the cost of intense floating point calculations.

The second ones translate the network into a directed graph to explore it. Path-finding approaches are considered to constitute some advance with respect to stoichiometry approaches mainly because they rest on the well-known graph theory and let the use of techniques based on distance metric, revealed as biologically relevant [1]. Because of the combinatorial nature of the search, some proposals only find a subset of all feasible pathways, whereas other approaches get the full set of feasible pathways [18]. The major drawback of path-finding

approaches is that the lack of use of stoichiometry during the exploration process cannot assure that the solution has biological meaning and meets all the constraints. Therefore, an extra stage is needed to determine if a found pathway meets the constraints and it constitutes an EFM.

Finally, some other authors combine both approaches trying to build on strengths of each and avoid respective drawbacks and computational expenses [11].

6 Conclusions and future work

In this paper we propose a new approach to obtain EFMs based on graph methods and the shortest path between end nodes. The novel approach we have presented here constitutes an advance with respect to previous approaches as it relies on a three-stage method based on the Dijkstra's shortest path algorithm, and an extra heuristic and mathematical phase that can produce systematically candidates to EFM.

Our method finds all the pathways in the metabolic network and it is able to prioritize the pathway search accounting the biological mean pursued. Our technique has two phases, the exploration one and the characterization one, and we show how it works in a well-known case study.

Unlike traditional Linear Programming (LP) approaches, our proposal avoids expensive floating-point calculations allowing us to speed-up the quest of all the available pathways in a certain metabolic network. We realize that the fact of the combinatorial explosion while exploration of the graph is a common problem to path-finding approaches (loops and the increasing size of the networks worsen the problem), so we foresee that the parallelization of this process could give us a lot of benefits. Our approach is quite suitable to be developed in new commodity parallel architectures (such as multi- and many-cores and accelerators like GPUs), allowing shorter execution times and less energy consumption.

As for future work, the characterization phase of the EFMs from the set of pathways obtained is still immature and more work should be done in relation with it, as developing some heuristics from artificial intelligence techniques like ants colony. Another direction of future work is the parallelization of all of the stages of our method using HPC commodity architectures, as multicore processors and accelerators (like GPUs or Xeon Phi).

Acknowledgments

This work was jointly supported by the Fundación Séneca (Agencia Regional de Ciencia y Tecnología, Región de Murcia) under grant 15290/PI/2010 and the Spanish MEC and European Commission FEDER under grant TIN2012-31345.

References

1. Croes D, Couche F, Wodak SJ, et al. Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res 2005;33;W326-330

2. Dijkstra E W. A note on two problems in connexion with Graphs. Numerische Mathematik 1959; 1, 269-271
3. Roman MT Flemming, B.O. Palsson. Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. 2010, Chapter 10.2.1 in Escherichia coli and Salmonella: Cellular and Molecular Biology, Washington, DC
4. Elementary Flux Mode Tool, `http://www.csb.ethz.ch/tools/efmtool`
5. De Figueiredo,L F et al. (2008) Can sugars be produced from fatty acids? A test case for pathway analysis tools. Bioinformatics, 24 , 2615-2621.
6. De Figueiredo,L F et al. (2009) Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics, 25, 3158-3165.
7. Hunt,K.A. et al. (2014) Complete enumeration of elementary flux modes through scalable, demand-based subnetwork definition. Bioinformatics, in press.
8. Dimitrije Jevremovic, Daniel Boley, Carlos P. Sosa. Divide-and-Conquer Approach to the Parallel Computation of Elementary Flux Modes in Metabolic Networks. IPDPS, 2011 IEEE International Symposium;50-511.
9. Klamt S, Stelling J. Combinatorial complexity of pathway analysis in metabolic networks. Mol Biol Rep 2002;29(1-2):233-6.
10. F.J. Planes, J.E. Beasly. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. Briefings in Bioinformatics, 2008,;9;422-436
11. Pey J, Prada J, Beasley JE, Planes FJ., Path finding methods accounting for stoichiometry in metabolic networks. Genome Biol. 2011;12(5):R49
12. Pey,J. and Planes, F.J. (2014). Direct calculation of Elementary Flux Modes satisfying several biological constraints in genome-scale metabolic networks. Bioinformatics, in press.
13. Rezola,A. et al. (2013) Selection of human tissue-specific elementary flux modes using gene expression data. Bioinformatics, 29, 2009-2016.
14. Rezola,A. et al. (2014) Advances in network-based metabolic pathway analysis and gene expression data integration. Brief. Bioinform In press.
15. Schmidt, B.J., et al. (2013) GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. Bioinformatics 29, 2900-2908.
16. Schuster S, Hilgetag C. (1994) On elementary flux modes in biochemical reaction systems at steady state. J Biol Syst 2;165-182.
17. von Kamp A and Schuster S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics 22(15):1930-1931.
18. H. Seo, D.-Y. Lee, S. Park, L.T. Fan,S. Shafie, B. Bertk, F. Friedler. (2001) Graph-theoretical identification of pathways for biochemical reaction. Biotechnology Letters, V23; 1551-1557
19. Thiele,I. and Palsson,B. . (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc, 5, 93-121.
20. Terzer,M. and Stelling, J. (2008) Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics, 24 ,2229-2235.
21. Urbanczik,R. and Wagner,C. (2005) An improved algorithm for stoichiometric network analysis: theory and applications. Bioinformatics, 21, 1203-10.