BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 4 2011, pages 534–540
doi:10.1093/bioinformatics/btq681

Systems biology

Advance Access publication December 10, 2010

Exploring metabolic pathways in genome-scale networks via generating flux modes

A. Rezola[1,†], L. F. de Figueiredo[2,3,†], M. Brock[4], J. Pey[1], A. Podhorski[1], C. Wittmann[5], S. Schuster[2], A. Bockmayr[6] and F. J. Planes[1,*]

[1]Biomedical Engineering, CEIT and TECNUN, University of Navarra, Manuel de Lardizabal 15, 20018 San Sebastian, Spain, [2]Department of Bioinformatics, Friedrich-Schiller-University Jena, 07743 Jena, Germany, [3]PhD Program in Computational Biology, Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal, [4]Junior Research Group of Microbial Biochemistry and Physiology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (HKI), 07745 Jena, [5]Institute of Biochemical Engineering, Technical University Braunschweig, 38106 Braunschweig and [6]DFG-Research Center Matheon, FB Mathematik und Informatik, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany

Associate Editor: Trey Ideker

ABSTRACT

Motivation: The reconstruction of metabolic networks at the genome scale has allowed the analysis of metabolic pathways at an unprecedented level of complexity. Elementary flux modes (EFMs) are an appropriate concept for such analysis. However, their number grows in a combinatorial fashion as the size of the metabolic network increases, which renders the application of EFMs approach to large metabolic networks difficult. Novel methods are expected to deal with such complexity.

Results: In this article, we present a novel optimization-based method for determining a minimal generating set of EFMs, i.e. a convex basis. We show that a subset of elements of this convex basis can be effectively computed even in large metabolic networks. Our method was applied to examine the structure of pathways producing lysine in *Escherichia coli*. We obtained a more varied and informative set of pathways in comparison with existing methods. In addition, an alternative pathway to produce lysine was identified using a detour via propionyl-CoA, which shows the predictive power of our novel approach.

Availability: The source code in C++ is available upon request.

Contact: fplanes@tecnun.es

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

Received on August 16, 2010; revised on November 12, 2010; accepted on December 7, 2010

1 INTRODUCTION

During the last decade, different approaches have been proposed to explore the structure of complex metabolic networks from a pathway-oriented perspective (Klamt and Stelling, 2003). As in other methods (Price *et al.*, 2004), these approaches are mainly built upon the steady state condition and irreversibility (thermodynamic) constraints, which define a polyhedral cone, typically referred to as the flux cone. In particular, pathway-based methods search for solutions in the flux cone that satisfy a simplicity condition, technically the non-decomposability condition. These properties were condensed in the definition of elementary flux mode (EFM) in the work of Schuster and Hilgetag (1994).

Since its introduction, the concept of EFM has received much attention, showing that a wide range of questions in bioengineering and bioinformatics can be addressed using such an approach (Schuster *et al.*, 2007; Trinh *et al.*, 2009). In particular, the prediction of novel biologically meaningful pathways constitutes one of the most important applications of EFM analysis. An example is the *in silico* prediction via EFM analysis of the PEP-glyoxylate cycle (Liao *et al.*, 1996; Schuster *et al.*, 1999), which was later confirmed experimentally (Fischer and Sauer, 2003).

However, due to combinatorial explosion (Klamt and Stelling, 2002), the computation of EFMs is not easy and, until recently, their analysis has been restricted to small metabolic networks. In this context, Terzer and Stelling (2008), have presented an improved method that expands the applicability of the EFMs approach to moderate-size metabolic networks; however, large metabolic networks remain beyond the capabilities of current algorithms. We have recently shown that the K-shortest EFMs (for small K values) can be calculated in large networks, even at the genome scale, using integer linear programming (de Figueiredo *et al.*, 2009). The use of optimization introduces more flexibility into EFM computation and, therefore, the direct exploration of a particular subset of EFMs of interest can be accomplished without having first to compute the full set of EFMs. This represents a clear advantage with respect to previous methods (Klamt *et al.*, 2005; Schilling *et al.*, 2000; Schuster *et al.*, 2000; Terzer and Stelling, 2008; Urbanczik and Wagner, 2005).

A concept related to EFMs is that of convex basis. First introduced in Pfeiffer *et al.* (1999), it was defined as a minimal set of EFMs that generate the solution space. This means that any vector in the flux cone, in particular any other EFM not included in this minimal set, can be written as a non-negative linear combination of the elements of the basis. For this reason, a convex basis is typically referred to as a minimal generating set and elements of this basis as generating flux modes, GFMs (Wagner and Urbanczik, 2005). However, the convex

*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.

basis is generally non-unique and does not necessarily involve all biochemical relevant pathways. Consequently, other mathematical descriptions of metabolism have been preferred over GFMs. In particular, special interest has been attracted by the extreme pathway approach (Schilling *et al.*, 2000), in which the network configuration is modified to overcome the non-uniqueness issue. Moreover, the number of GFMs and extreme pathways is lower than the number of EFMs.

The comparison between EFMs, extreme pathways and GFMs has been extensively addressed in the literature (Klamt and Stelling, 2003; Wagner and Urbanczik, 2005). This discussion is less relevant in large metabolic networks, since the full computation of EFMs and GFMs (and extreme pathways) is not possible due to combinatorial explosion. Instead, current research has been more devoted to identify novel methods to deal with such complexity and explore the EFM and GFM solution spaces (de Figueiredo *et al.*, 2009; Kaleta *et al.*, 2009). This is in fact the main purpose of this article, namely to present a new method for computing a subset of GFMs in large-scale metabolic networks.

Recently, Larhlimi and Bockmayr (2009) have presented a new constraint-based description of the flux cone. Based on the concept and properties of 'minimal metabolic behaviours' introduced in that paper, we adapted the K-shortest EFMs method presented in de Figueiredo *et al.* (2009), to calculate GFMs. Similarly to the K-shortest EFM approach, our method can be (in theory) applied to enumerate a full set of GFMs. However, it is particularly useful for large networks, where classical methods are not applicable (Urbanczik and Wagner, 2005). In these cases, our method allows the effective computation of a subset of GFMs by directly calculating only those GFMs that involve a particular reaction or metabolite of interest.

We present below the details of our mathematical optimization model for calculating GFMs. We used the standard network configuration and convex basis definition presented in Pfeiffer *et al.* (1999), as will become apparent in Section 2. Our approach, however, can be easily adapted to calculate extreme pathways just by changing the network configuration as defined in Schilling *et al.* (2000). We then examine the predictive power of our GFMs approach to find pathways producing the commercially important amino acid L-lysine in the genome-scale metabolic network of *Escherichia coli* (Feist *et al.*, 2007). The results are analyzed and compared with our previous method presented in de Figueiredo *et al.*, 2009. We show that our GFM approach presents a more descriptive and varied set of pathways than the K-shortest EFMs method. As a result of our analysis, we hypothesize a novel pathway to produce L-lysine in *E.coli* using propionyl-CoA (ppcoa).

2 METHODS

In this section, we present a novel mathematical method to calculate a subset (and, in principle, even a full set) of GFMs, i.e. a convex basis. Before describing this method, a brief introduction to metabolic network theory is carried out.

2.1 Definitions

Consider a metabolic network that comprises R reactions and C metabolites. Each reaction is associated with a flux variable $v_r, r=1,...,R$ where $v=[v_1,...,v_R]$ is the flux vector. Irr denotes the set of irreversible reactions, which satisfy the (qualitative) thermodynamic constraints, i.e. $v_r \geq 0$, $r_{\in} Irr$. Note that reversible reactions are not split into forward and backward reactions, as opposed to other approaches, e.g. Schilling *et al.*, 2000, where splitting into two irreversible steps is accomplished. The set of metabolites is divided into two subsets, namely external (E) and internal (I) metabolites. For internal metabolites, it is assumed that no accumulation or depletion is possible, therefore steady-state condition holds, i.e. $\sum_{r=1}^{R} s_{cr} v_r = 0$, $c_{\in} I$, where s_{cr} is the stoichiometric coefficient associated with compound c ($c=1,...,C$) in reaction r ($r=1,...,R$). As usual in the literature (Schilling *et al.*, 2000; Schuster *et al.*, 2000), substrates and products have negative and positive stoichiometric coefficients, respectively. External metabolites are typically considered as sources and sinks.

The set of flux vectors that satisfy the thermodynamic and steady state constraints defines a polyhedral cone, P, defined as

$$P = \left\{ v \middle| \sum_{r=1}^{R} s_{cr} v_r = 0,\ c_{\in} I;\ v_r \geq 0,\ r_{\in} Irr \right\} \tag{1}$$

Aside from being contained in the flux cone P, EFMs are solutions that satisfy a simplicity condition, namely the non-decomposability condition, which establishes that any subset of the reactions involved in an EFM cannot carry flux in steady state (Schuster *et al.*, 2000).

As in Pfeiffer *et al.*, 1999, a convex basis is defined here as a minimal set of EFMs that generates the solution space. Elements of a convex basis were termed above generating flux modes, GFMs. A convex basis defines two different subsets of GFMs, namely B reversible GFMs and G irreversible GFMs. While an irreversible GFM involves at least one irreversible reaction, a reversible GFM involves no irreversible reactions. Reversible and irreversible GFMs are denoted here in vectorial form, $b_j, j=1,...,B$ and $g_i, i=1,...,G$, respectively.

Any flux vector v in the polyhedral cone P can be generated as a linear combination with non-negative coefficients λ_i for irreversible GFMs and with unrestricted signed coefficients μ_j for reversible GFMs.

$$v = \sum_{i=1}^{G} \lambda_i g_i + \sum_{j=1}^{B} \mu_j b_j, \tag{2}$$

$$v_{\in} P,\ \lambda_i \geq 0,\ \ i=1,...,G;\ -\infty < \mu_j < \infty,\ j=1,...,B;$$

We describe below how reversible and irreversible GFMs can be separately determined.

2.2 Reversible GFMs

The determination of the reversible GFMs of a convex basis can be accomplished using the concept of reversible metabolic space (RMS), as introduced in Larhlimi and Bockmayr, 2009. RMS is defined as the linear subspace in which the irreversible reactions carry no flux:

$$\mathrm{RMS} = \{ v_{\in} P \mid v_r = 0,\ \forall r_{\in} Irr \} \tag{3}$$

Accordingly, the set of reversible GFMs ($b_j, j=1,...,B$) is a linear basis of RMS, namely:

$$v = \sum_{j=1}^{B} \mu_j b_j,\ v_{\in} \mathrm{RMS},\ -\infty < \mu_j < \infty,\ j=1,...,B; \tag{4}$$

Finding a linear basis is a simple task and it can be easily performed using classic linear algebra. Note that B is the number of reversible GFMs in a convex basis and it represents the dimension of RMS. Interestingly, if $B=0$, then the flux cone is pointed and the convex basis is unique up to multiplication by non-negative numbers (Pfeiffer *et al.*, 1999). However, genome-scale metabolic networks typically have $B \neq 0$. In the extreme pathway approach (Schilling *et al.*, 2000), for example, the network configuration is modified to have $B=0$.

2.3 Irreversible GFMs

The determination of irreversible GFMs is the actual challenge in computing a convex basis. To illustrate how to calculate such a set, we will use the concept of minimal metabolic behavior (MMB) introduced in Larhlimi and Bockmayr (2009).

A metabolic behavior is any non-empty set of irreversible reactions able to carry flux together in a steady state flux vector. A metabolic behavior D is minimal if there is no proper subset of D with this property. For a more rigorous definition of MMB, see Larhlimi and Bockmayr (2009).

According to Theorem 9 in Larhlimi and Bockmayr (2009), each irreversible GFM in a convex basis involves a different MMB and therefore the number of irreversible GFMs in a convex basis is equal to the number of MMBs. We can then define the set of irreversible GFMs in a convex basis as a subset of EFMs with the property that each EFM in this subset involves a different MMB.

Observe that in non-pointed flux cones we may have more than one EFM involving a particular MMB. Indeed, for this reason we can have a non-unique convex basis. However, in terms of determining a particular convex basis, we only need to select one such EFM for each MMB. We here selected the EFM with the minimum number of reversible reactions. The rest of EFMs involving the same MMB (but not included in the convex basis) can be generated by the EFM selected for the convex basis and the reversible GFMs. More details on MMBs and GFMs can be found in Larhlimi and Bockmayr (2009).

In this light, we propose a general optimization-based method to enumerate MMBs and their associated GFMs. Our method starts from the assumption that the flux mode involving the minimum number of irreversible reactions will involve the shortest MMB. By abuse of language, we here refer to an EFM involving the shortest MMB with the minimal use of reversible reactions as the shortest GFM. Accordingly, we first define the constraints and the objective function to be optimized that allows the calculation of the shortest GFM. Based on this optimization model, we then show how to calculate the K-shortest GFMs. Finally, extensions to other problems of interest will be presented.

We mean here by 1-shortest GFM, the EFM involving the minimum number of irreversible reactions, i.e. the shortest MMB, with the minimal use of reversible reactions; 2-shortest GFM, the EFM containing the second shortest MMB with the minimal use of reversible reactions; K-shortest GFM, the EFM containing the K-shortest MMB with the minimal use of reversible reactions. Note that we may have multiple GFMs containing the same number of irreversible reactions, i.e. MMBs with the same length. If this occurs, they are counted separately with different K values.

2.3.1 Shortest GFM Let z_r be a binary variable associated with each reaction r ($r=1,...,R$), where $z_r=1$ if the reaction is active in the shortest GFM, 0 otherwise. As noted above, v_r represents the flux variable for reaction r. Equation (5) introduces the constraints needed to relate the reaction variables z_r and v_r, where M is a large positive constant value, which represents the bounds for the reaction fluxes.

$$-Mz_r \leq v_r \leq Mz_r,\ r=1,...,R \tag{5}$$

Equations (6) and (7) define the thermodynamic and steady state constraints, as introduced above.

$$v_r \geq 0,\ r_{\in} Irr \tag{6}$$

$$\sum_{r=1}^{R} s_{cr} v_r = 0,\ c_{\in} I \tag{7}$$

In order to ensure that at least one irreversible reaction occurs in the shortest GFM, we need Equation (8), which avoids the trivial solution, ($v_r=0$, $r=1,...,R$).

$$\sum_{r=1,\ r_{\in} Irr}^{R} v_r \geq 1 \tag{8}$$

In order to calculate the shortest GFM, we pose the following objective function:

$$\text{Minimize} \qquad W \sum_{r=1,\ r_{\in} Irr}^{R} z_r + \sum_{r=1,\ r_{\notin} Irr}^{R} z_r \tag{9}$$

which involves two terms to be minimized. The first one is the number of active irreversible reactions multiplied by a large positive constant value, W; the second one is the number of active reversible reactions. We give W times more weight to the sum of active irreversible reactions to guarantee that the flux mode involves the minimum number of irreversible reactions and, therefore, the shortest MMB participates in the solution. The second term guarantees that (i) the flux mode containing the shortest MMB is elementary, i.e. non-decomposable, as defined for EFMs ; and (ii) the minimal use of reversible reactions as defined for the shortest GFM.

2.3.2 K-shortest GFMs Similarly to de Figueiredo *et al.* (2009), we need constraints to remove all the previous k-shortest GFMs ($k=1,...,K-1$) from the set of solutions in order to calculate the K-shortest GFM. We already know that the K-shortest GFM involves the K-shortest MMB. Since the convex basis needs only one EFM for each MMB, we need to guarantee that the K-shortest GFM does not involve the previous k-shortest MMBs ($k=1,...,K-1$), but only the K-shortest MMB. To do this, let Z_r^k be the binary solution associated with the k-shortest GFM ($k=1,...,K-1$), where $Z_r^k=1$ if reaction r is active in the k-shortest GFM, 0 otherwise, and we impose the following constraint:

$$\sum_{r=1,\ r_{\in} Irr}^{R} Z_r^k z_r \leq \left(\sum_{r=1,\ r_{\in} Irr}^{R} Z_r^k \right) - 1, \qquad k=1,...,K-1 \tag{10}$$

where the term in the left-hand side of Equation (10) determines the number of irreversible reactions in the new solution (z_r) that were active in the k-shortest GFM ($k=1,...,K-1$), and the right-hand side is the number of irreversible reactions that were active in the k-shortest GFM ($k=1,...,K-1$) less one. The inequality states that the number of irreversible reactions repeating from the k-shortest GFM ($k=1,...,K-1$) in the new solution should be strictly smaller than the total number of active irreversible reactions in the k-shortest GFM ($k=1,...,K-1$). In essence, Equation (10) ensures that, once we solve our model, the k-shortest MMBs are prevented from appearing again in the solution, for $k=1,...,K-1$. This guarantees that the solution will involve the K-shortest MMB, and therefore the K-shortest GFM is obtained.

2.3.3 Extensions to K-shortest GFMs The procedure described above can be applied (in theory) to enumerate a full set of irreversible GFMs ($g_i,\ i=1,...,G$) for a given metabolic network. However, our optimization method is not particularly efficient in small-scale metabolic networks when compared with classical methods. Indeed, the main advantage of our method is the enumeration of a subset of GFMs ($K=10-10000$) in large metabolic networks. In addition, the use of optimization allows the user to directly analyze those subsets of GFMs that may be of interest, as in de Figueiredo *et al.* (2009). This makes our approach a suitable tool for exploring EFMs (GFMs) in large metabolic networks.

To illustrate this, our method can be used to explore the K-shortest GFMs that involve a particular reaction ξ of interest, namely by imposing their flux v_ξ to be non-zero [Equation (11)]. Note here that Y is a scalar coefficient ($Y=-1$ or $Y=1$), whose value decides the direction of the flux of reaction ξ in the case the reaction is reversible.

$$Yv_\xi \geq 1, \quad Y=-1 \ or \ Y=1 \tag{11}$$

The application of the K-shortest GFM method here, Equations (5)–(10) plus Equation (11), requires to consider if reaction ξ is reversible or irreversible. If the reaction is irreversible, then our method will calculate a subset of irreversible GFMs containing this reaction ($Y=1$) . At the same time, if the reaction is reversible but it is not involved in any reversible GFM, i.e. a pseudo-irreversible reaction, as defined in Larhlimi and Bockmayr (2009), then our optimization method will produce GFMs containing this

reversible reaction. Note that Y can be -1 or $+1$ depending on the direction we want irreversible GFMs to occur. On the other hand, if the reaction is reversible and it participates in reversible GFMs, our current method might fail. To overcome this issue, a possible strategy would be to split this reversible reaction into two irreversible steps.

Note that we can only include one constraint based on Equation (11). For example, if we apply constraint (11) to reactions ξ_1 and ξ_2, i.e. finding solutions to our model that involve reactions ξ_1 and ξ_2, then we might obtain solutions containing two EFMs, namely one using ξ_1 and another using ξ_2. Therefore, the non-decomposability condition is not guaranteed. In addition, as in de Figueiredo *et al.* (2009), it is possible to define a growth medium. Extracellular metabolites not included in the growth medium can only be balanced/produced, but not consumed.

2.3.4 Implementation The mathematical optimization model given above for computing the K-shortest irreversible GFMs, consisting of objective function (9) subject to Equations (5)–(8), plus elimination constraints (10) and perhaps constraint (11), is a mixed-integer linear program. Algorithmically, such programs are solved via branch and bound and cutting planes methods (Pardalos and Resende, 2002). Various free and commercial software tools are available to perform this task. We used ILOG CPLEX®.

To reinforce the understanding of our method, Supplementary Material details the application of our method to the (toy) metabolic network presented in Larhlimi and Bockmayr (2009).

3 RESULTS

In this section, we present results corresponding to the application of our K-shortest GFMs to analyze the pathway structure of *E.coli* to produce lysine. We used the genome-scale metabolic network of *E.coli* K-12 MG1665 (Feist *et al.*, 2007), which involves 2082 reactions and 1668 metabolites. As noted above, our method is particularly useful for networks of this size, since the computation of a convex basis via traditional methods (Pfeiffer *et al.*, 1999; Terzer and Stelling, 2008; Wagner and Urbanczik, 2005) is not applicable in these networks.

A minimal medium based on glucose, ammonium and oxygen was utilized. We neglected constraints from proton and cofactor (ATP, NADH, NADPH, etc.) balances, fixing them as external metabolites. One may question the validity of this hypothesis for NADP and NADPH, since NADPH is required in stoichiometric amounts for lysine production and, as discussed in Wittmann and Becker (2007), its supply has substantial impact in metabolism. In this work, for the sake of simplicity, we focused on pathways for the supply of the carbon backbone of lysine. Moreover, extending the current work to consider the role of cofactors in lysine production can be accomplished without loss of generality, as shown in de Figueiredo *et al.* (2009). See Supplementary Material for the exact definition of the medium and cofactors set.

The reaction transporting lysine from the periplasm compartment to the external compartment (LYStex, see Feist *et al.*, 2007) is pseudo-irreversible, i.e. no reversible GFMs exist that produce lysine. Thus, we need only to focus on irreversible GFMs producing lysine. Our K-shortest GFMs method was then applied with $K=100$ and constraint (11) for lysine. We also computed the K-shortest EFMs, as described in de Figueiredo *et al.*, 2009, in order to compare the results. We selected $M=100$ to guarantee that no pathway information is missed, similarly to de Figueiredo *et al.* (2009). In a 64 Bits, 2.00 GHz, 12 GB RAM PC, CPLEX 12.1 took approximately 30 and 100 min to compute the 100 shortest EFMs and GFMs, respectively. Therefore, though computation time is acceptable, our method is less efficient than the K-shortest EFMs method.

Note that, by setting cofactors as external metabolites, we may find solutions that use some cofactors as carbon source, which are irrelevant for the biosynthesis of lysine using glucose as carbon source. As it can be observed in the Supplementary Material, we found this issue in longer pathways using AMP as carbon source. To prevent these solutions from appearing, we constrained the adenine pool formed by ATP, ADP and AMP to be in balance by adding the following constraint to our previous formulation: $\sum_{r=1}^{R} s_{ATPr} v_r + \sum_{r=1}^{R} s_{ADPr} v_r + \sum_{r=1}^{R} s_{AMPr} v_r = 0$, where s_{ATPr}, s_{ADPr} and s_{AMPr} are the stoichiometric coefficients of ATP, ADP and AMP, respectively, for reaction r.

3.1 Comparison between K-shortest EFMs and GFMs

Figure 1 is an overview representation of the 100-shortest EFMs and GFMs. It can be seen that the 100-shortest GFMs present a more diverse and varied set of pathways than the 100-shortest EFMs. This hypothesis is further confirmed in Table 1. First, the 100 shortest EFMs and 100 shortest GFMs involve 54 and 132 reactions, respectively. In addition, while the maximum length for 100 shortest EFMs is 26, this is 37 for the 100 shortest GFMs, which just indicates that the K-shortest GFMs method produce a set of longer pathways. We also used the Hamming distance to evaluate the differences between each pair of solutions of the 100 shortest EFMs (GFMs), finding that the average Hamming distance is higher for GFMs, which just indicates that the shortest EFMs approach determine a less diverse set of solutions.

Furthermore, it can be observed from Table 1 that from 54 reactions that participate in the 100 shortest EFMs, 49 are included in the 100 shortest GFMs. This implies that most information in the shortest EFMs is included in the shortest GFMs and, therefore, the GFMs approach is more informative. To validate this, we created a subnetwork with the reactions forming the 100 shortest EFMs and computed the full set of EFMs for this subnetwork via efmtool (Terzer and Stelling, 2008). As observed in Table 1, we obtained 1665 EFMs, 1636 of which produce L-lysine. We did the same for the 100 shortest GFMs, finding 354 238 EFMs in total, 25 007 of which produce L-lysine. The difference in the number of pathways is clearly significant. If the RMS is included in the subnetworks designed, the difference is much more important, though efmtool was not capable of calculating the number of EFMs for the subnetwork created with the reactions participating in the 100 shortest GFMs plus RMS.

From the biochemical point of view, though difficult to observe in Figure 1 (see Supplementary Figs S2, S4 and S5 for a clearer picture), different insights can be gained. The interpretation of the network graphs in the Supplementary Material becomes easier when we focus on pyruvate (pyr) and oxolacetate (oaa), since the variability in the set of EFMs and GFMs is mainly related to the production and use of those metabolites, mainly due to the fact that they are the precursors of lysine.

Among the 100 shortest EFMs (see Supplementary Figs S2 and S5), the main difference is the degradation of glucose into pyruvate. As we are producing shorter pathways, we obtain the shortest routes to do this function, namely glycolysis, the Entner–Doudoroff (ED) pathway, the methylglyoxal bypass and some variations of these pathways such as a detour via dihydroxyacetone (dha). Minor differences relate to the transportation of glucose into

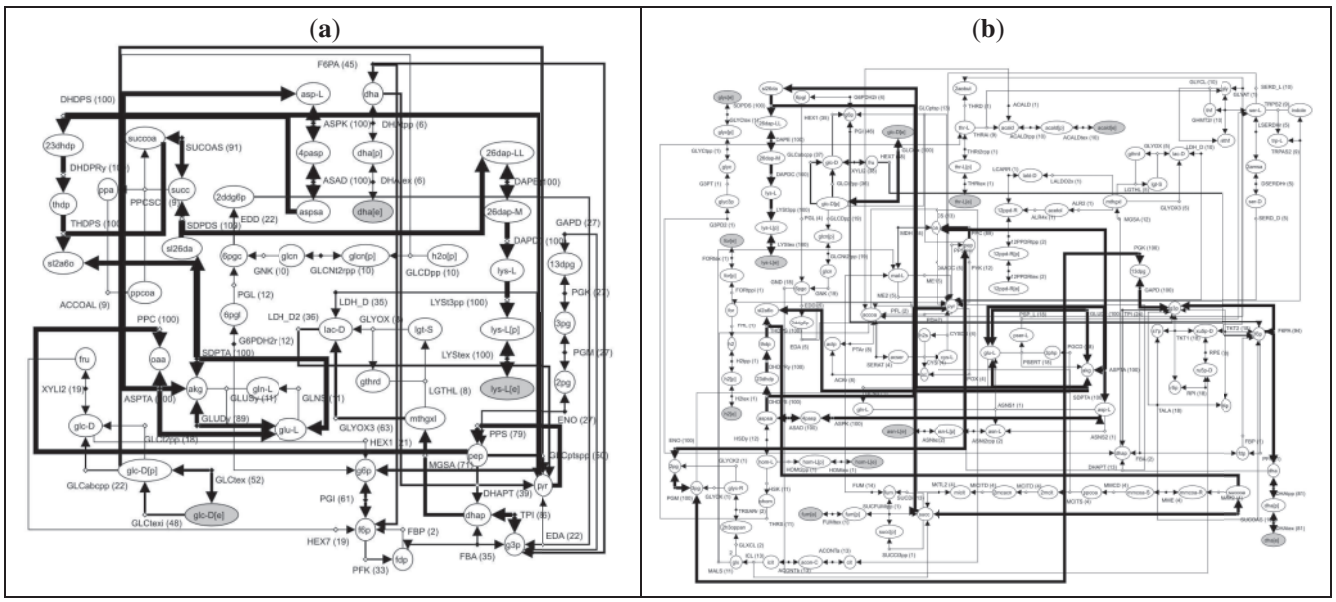

Fig. 1. Overview of 100 shortest (**a**) EFMs and (**b**) GFMs. Ellipses and arrows represent metabolites and reactions, respectively. Gray ellipses are sources and sinks. In brackets, the corresponding compartment for metabolites is shown, namely [e], extracellular; [p], periplasm; nothing, cytosol. Cofactors metabolites were removed from the picture for a better visualization. Thickness of the arrows is proportional to the number of times a reaction appears in the 100 shortest EFMs/GFMs.

Table 1. Information about the 100 shortest GFMs/EFMs

	NoR	LI	OvR	AHD	No EFMs	No EFMs(R)
100 shortest EFMs	54	25–26	90.74% (49/54)	12.79	1665 Lys: 1636	5398 Lys: 3960
100 shortest GFMs	132	25–37	37.12% (49/132)	16.21	354238 Lys: 25007	$5 \times 10^5 - 5 \times 10^{10}$ Lys: $*$

NoR, number of reactions; LI, length interval; OvR, overlapping percentage of reactions of the 100 shortest EFMs (GFMs) in the shortest GFMs (EFMs); AHD, average hamming distance; No EFMs, number of EFMs emerging from the network formed by the 100 shortest EFMs (GFMs); No EFMs (R), number of EFMs emerging from the network formed by the 100 shortest EFMs (GFMs) and the reversible metabolic space (RMS); Lys, L-lysine.

cytosol, nitrogen metabolism, the balancing of succinate (succ) and succinyl-CoA (succoa) and the production of dha.

With respect to the 100 shortest GFMs, we have longer routes than the 100 shortest EFMs, though they are still short (maximum 37 reactions). The set of pathways is now much more varied (see Supplementary Figs S4 and S5). For instance, the pentose phosphate (PP) pathway and the glyoxylate shunt are now recovered. Moreover, we obtain several routes around the switch node per-pyr-oaa, which is known to be responsible for the distribution of the carbon flux among catabolism, anabolism and energy supply in several bacteria (Sauer and Eikmanns, 2005). Overall, most of previous biochemical knowledge on the production of lysine is recovered in the 100 shortest GFMs.

A relevant mechanism appearing in the 100 shortest GFMs is the metabolic cycle shown in Figure 2a, which suggests a novel pathway to produce lysine. This metabolic cycle consumes one molecule of oaa and produces one molecule of pyr and co2 each, making a detour through propionyl-CoA (ppcoa). This cycle is an alternative mechanism to the OAADC reaction: oaa → pyr + co2. However, it is less efficient than the OAADC reaction, since it involves six reactions and consumes one molecule of ATP to carry out the conversion. We refer to it as the ppcoa cycle. To the best of our knowledge, this cycle has not been previously described. Note here that other routes for the conversion of oaa to pyr have been previously described in the literature (Wittmann and Becker, 2007).

The importance of the ppcoa cycle in the production of lysine is however questionable. The decarboxylation of oaa is generally counterproductive, since the overexpression of genes associated with enzymes producing oaa as well as the deletion of genes encoding enzymes that withdraw oaa are well-known strategies to increase the yield of lysine (Wittmann and Becker, 2007). A counterexample here is the decarboxylation of oaa via malate (reactions MDH and ME2) that converts NADH into NADPH, which, as mentioned above, is important for lysine production. On the other hand, it is known from studies in *Aspergillus nidulans* that ppcoa inhibits the activity of pyruvate dehydrogenase and thus, its presence decreases the flux through the tricarboxylic acid (TCA) cycle (Brock and Buckel, 2004). This may be of interest for lysine production, since the increase in the yield of lysine is linked to a decrease in the flux through TCA (Kiss and Stephanopoulos, 1992). The overall effect requires experimental validation. We discuss below the functional feasibility of this cycle in *E.coli*.

3.2 Physiological feasibility of ppcoa cycle in *E.coli*

As observed in Figure 2a, the ppcoa cycle comprises a route linking succinyl-CoA (succoa) to propionyl-CoA (ppcoa), which is part of

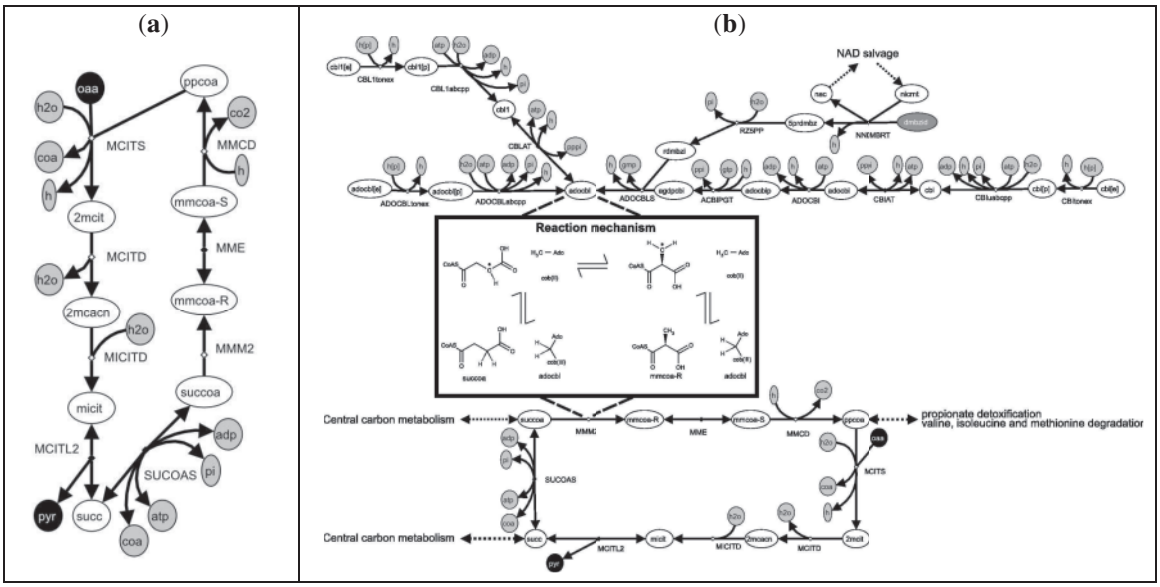

Fig. 2. (**a**) Novel ppcoa cycle detected in the GFMs; (**b**) Interdependency of 5-deoxyadenosylcobalamin (adocbl) synthesis and the ppcoa cycle. Schematic representation of the reaction mechanism associated to methylmalonyl-Coa mutase (MMM2)–cobalamin dependent (adapted from Banerjee, 1997) which is not present in the genome-scale network of *E.coli*. Ellipses represent metabolites and arrows reactions. Light gray ellipses are external metabolites, dark gray are dead-end metabolites and black are source and sink metabolites.

the methylmalonyl-CoA pathway, and a route linking propionyl-CoA to pyruvate accompanied by a reduction of oxaloacetate to succinate (succ), as a part of the 2-methylcitric acid cycle. These two routes are joined by succinyl-CoA synthetase, which belongs to the TCA cycle. We describe below the feasibility of the coexistence of the methylmalonyl-CoA pathway and the 2-methylcitric acid cycle in *E.coli*.

The unidirectional 2-methylcitric acid cycle has been shown to be essential for propionyl-CoA degradation in several bacteria and fungi and an impairment of the cycle prohibits growth on propionate, odd chain fatty acids and propionyl-CoA generating amino acids (Brock and Buckel, 2004).

An alternative pathway for propionyl-CoA degradation is the coenzyme B12-dependent methylmalonyl-CoA pathway, which is also operative in the direction of propionyl-CoA formation (Buckel *et al.*, 2005; Michenfelder *et al.*, 1987). Despite a functional 2-methylcitric acid cycle in *E.coli*, genes encoding proteins of the methylmalonyl-CoA pathway were recently discovered in this organism (Froese *et al.*, 2009; Haller *et al.*, 2000). Although the contribution of the latter pathway to propionyl-CoA degradation is unknown, all genes encode functional enzymes as shown by heterologous integration of the pathway for secondary metabolite production in *Salmonella enterica* serovar thyphimurium (Aldor *et al.*, 2002) and polypetide production in *E.coli* (Aldor *et al.*, 2002).

Assuming a coexistence of both pathways, they can be integrated into the central carbon metabolism. Thus, we conducted additional simulations to examine this hypothesis, which resulted in novel potential pathways linking glycolysis and TCA cycle. Results are given in the Supplementary Material.

Besides the general functionality of proteins involved in the methylmalonyl-CoA pathway, an important aspect of the methylmalonyl-CoA pathway not reflected in the genome-scale network of *E.coli* refers to the fact that the methylmalonyl-CoA mutase (MMM2) requires 5-deoxyadenosylcobalamin (adocbl) as prosthetic group (cf. (Banerjee, 1997; Ludwig and Matthews, 1997). Detailed reaction mechanisms, as in Figure 2b for MMM2, are not usually included in genome-scale networks. In particular, when a cofactor is essential for cell growth, which is not the case for adobcl, constraint-based methods, such as flux balanced analysis (FBA), include its consumption in the biomass equation to simulate its requirement (Feist *et al.*, 2007, 2009). In contrast, in metabolic pathway analysis the biomass equation is not typically included and, consequently, these interactions are not taken into account. To deal with this issue in metabolic pathway analysis, a possible strategy is to create an additional layer of logical interactions containing the interdependency between prosthetic groups and enzymes, reflecting that if a given cofactor cannot be *de novo* synthesized (which can be tested with EFM analysis or FBA) and it is not supplied in the growth medium, then the reaction requiring such cofactor cannot take place.

4 CONCLUSION

In this article, we have presented a novel theoretical method to compute a convex basis based on the recent work of de Figueiredo *et al.* (2009) and Larhlimi and Bockmayr (2009). Our method was effectively applied to determine a subset of GFMs in large metabolic networks, where classical approaches fail. Though the computation of K-shortest GFMs is more expensive than the K-shortest EFMs method presented in de Figueiredo *et al.* (2009), we here obtained a more diverse and informative set of pathways.

In the analysis of pathways producing lysine obtained by the K-shortest GFMs method, we identified a new cycle via propionyl-CoA that produces one molecule of pyruvate and consumes one

molecule of oxaloacetate. This cycle integrates part of two metabolic pathways that have an important biotechnological application. Experimental validation of the cycle is needed to examine its functional feasibility and biotechnological impact, but it shows the potential of our *K*-shortest GFMs method in predicting novel metabolic pathways and generating hypotheses.

ACKNOWLEDGEMENT

The authors would like to acknowledge the helpful comments made by the two anonymous reviewers.

Spanish entities Funding: Asociación de Amigos de la Universidad de Navarra (PhD grant to A.R.) and Basque Government (PhD grant to J.P.).

Portuguese entities Funding: Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (PhD grant SFRH/BD/32961/2006 to L.F.F.).

Conflict of Interest: none declared.

REFERENCES

Aldor,I.S. *et al.* (2002) Metabolic engineering of a novel propionate-independent pathway for the production of poly(3-hydroxybutyrate-co-3-hydroxyvalerate) in recombinant *Salmonella enterica* serovar Typhimurium. *Appl. Environ. Microbiol.*, **68**, 3848–3854.

Banerjee,R. (1997) The yin-yang of cobalamin biochemistry. *Chem. Biol.*, **4**, 175–186.

Brock,M. and Buckel,W. (2004) On the mechanism of action of the antifungal agent propionate. *Eur. J. Biochem.*, **271**, 3227–3241.

Buckel,W. *et al.* (2005) Stabilisation of methylene radicals by cob(II)alamin in coenzyme B12 dependent mutases. *Chemistry*, **12**, 352–362.

de Figueiredo,L.F. *et al.* (2009) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, **25**, 3158–3165.

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.

Feist,A.M. *et al.* (2009) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.*, **7**, 129–143.

Fischer,E. and Sauer,U. (2003) A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*. *J. Biol. Chem*., **278**, 46446–46451.

Froese,D.S. *et al.* (2009) Sleeping beauty mutase (*sbm*) is expressed and interacts with *ygfd* in *Escherichia coli*. *Microbiol. Res.*, **164**, 1–8.

Haller,T. *et al.* (2000) Discovering new enzymes and metabolic pathways: conversion of succinate to propionate by *Escherichia coli*. *Biochemistry*, **39**, 4622–4629.

Kaleta,C. *et al.* (2009) Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.*, **19**, 1872–1883.

Kiss,R.D. and Stephanopoulos,G. (1992) Metabolic characterization of a L-lysine-producing strain by continuous culture. *Biotechnol. Bioeng.*, **39**, 565–574.

Klamt,S. and Stelling,J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep*., **29**, 233–236.

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

Klamt,S. *et al.* (2005) Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *Syst. Biol.*, **152**, 249–255.

Larhlimi,A. and Bockmayr,A. (2009) A new constraint-based description of the steady-state flux cone of metabolic networks. *Discr. Appl. Math.*, **157**, 2257–2266.

Liao,J.C. *et al.* (1996) Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.*, **52**, 129–140.

Ludwig,M.L. and Matthews,R.G. (1997) Structure-based perspectives on B12-dependent enzymes. *Annu. Rev. Biochem.*, **66**, 269–313.

Michenfelder,M. *et al.* (1987) Quantitative measurement of the error in the cryptic stereospecificity of methylmalonyl-CoA mutase. *Eur. J. Biochem.*, **168**, 659–667.

Pardalos,P. and Resende,M. (2002) *Handbook of Applied Optimization*. Oxford University Press, New York, USA.

Pfeiffer,T. *et al.* (1999) Metatool: for studying metabolic networks. *Bioinformatics*, **15**, 251–257.

Price,N.D. *et al.* (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, **2**, 886–897.

Sauer,U. and Eikmanns,B.J. (2005) The pep-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria. *FEMS Microbiol. Rev.*, **29**, 765–794.

Schilling,C.H. *et al.* (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol*., **203**, 229–248.

Schuster,S. and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, **2**, 165–182.

Schuster,S. *et al.* (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*., **17**, 53–60.

Schuster,S. *et al.* (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Schuster,S. *et al.* (2007) Understanding the roadmap of metabolism by pathway analysis. *Methods Mol. Biol.*, **358**, 199–226.

Terzer,M. and Stelling,J. (2008) Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, **24**, 2229–2235.

Trinh,C.T. *et al.* (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl. Microbiol. Biotechnol.*, **81**, 813–826.

Urbanczik,R. and Wagner,C. (2005) An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, **21**, 1203–1210.

Wagner,C. and Urbanczik,R. (2005) The geometry of the flux cone of a metabolic network. *Biophys. J.*, **89**, 3837–3845.

Wittmann,C. and Becker,J. (2007) The L-lysine story: from metabolic pathways to industrial production. In Wendisch,V.F. (ed.) *Amino Acid Biosynthesis - Pathways, Regulation and Metabolic Engineering*. Springer, Heidelberg, Germany, p. **5**.