BIOINFORMATICS ORIGINAL PAPER Vol. 30 no. 15 2014, pages 2197–2203
doi:10.1093/bioinformatics/btu193

Systems biology Advance Access publication April 10, 2014

Direct calculation of elementary flux modes satisfying several biological constraints in genome-scale metabolic networks

Jon Pey and Francisco J. Planes*
CEIT and TECNUN, University of Navarra, 20018 San Sebastian, Spain

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: The concept of Elementary Flux Mode (EFM) has been widely used for the past 20 years. However, its application to genome-scale metabolic networks (GSMNs) is still under development because of methodological limitations. Therefore, novel approaches are demanded to extend the application of EFMs. A novel family of methods based on optimization is emerging that provides us with a subset of EFMs. Because the calculation of the whole set of EFMs goes beyond our capacity, performing a selective search is a proper strategy.

Results: Here, we present a novel mathematical approach calculating EFMs fulfilling additional linear constraints. We validated our approach based on two metabolic networks in which all the EFMs can be obtained. Finally, we analyzed the performance of our methodology in the GSMN of the yeast *Saccharomyces cerevisiae* by calculating EFMs producing ethanol with a given minimum carbon yield. Overall, this new approach opens new avenues for the calculation of EFMs in GSMNs.

Availability and implementation: Matlab code is provided in the supplementary online materials

Contact: fplanes@ceit.es.

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

Received on January 23, 2014; revised on March 27, 2014; accepted on April 7, 2014

1 INTRODUCTION

From its inception, classical biology has relied on reductionist principles. However, the actual nature of the cell, in which all components interact, demanded switching the previous paradigm to a holistic molecular approach, extending the scope of the analysis. These are the foundations that motivated the development of systems biology in the second half of the twentieth century. In particular, systems biology aims at studying the biological processes in terms of the cellular components and their interactions from a holistic molecular perspective (Kitano, 2002).

Some of these components interact in the cellular metabolism, which comprises those biochemical reactions consuming and producing the smallest compounds of the cell, typically named metabolites. These reactions are close to be in thermodynamic equilibrium, requiring the presence of catalytic agents so as to achieve the appropriate rate of activity to sustain life. In most metabolic reactions, a set of proteins named enzymes act as catalysts. The conversion rate of the metabolic reactions is termed metabolic fluxes.

Metabolic reactions are organized into distinct functional modules, forming the so-called metabolic pathways. Some of these pathways have been experimentally reported and manually included in different databases (Kanehisa *et al*., 2012; Keseler *et al*., 2011), as well as in biochemistry books (Nelson and Cox, 2000), e.g. glycolysis, TCA cycle. However, metabolic pathways are not independent entities; for example, the same enzyme or metabolite may appear in different pathways. For this reason, a most general analysis of metabolism requires the simultaneous consideration of different metabolic pathways. In the most extreme scenario, when all the known metabolic pathways are considered, the resulting set of enzymes and metabolites is referred to as genome-scale metabolic networks (GSMNs). The outbreak of different high-throughput experimental techniques has allowed us to increase the accuracy and size of GSMNs, in terms of metabolites, reactions and gene regulation. Currently, GSMNs for several organisms are publicly available through different online repositories (Schellenberger *et al*., 2010).

The inherent complexity of GSMNs does not make it possible to perform a manual analysis per se. Different computational strategies have been developed (Planes and Beasley, 2008). Among them, a number of theoretical frameworks have been proposed to extend the concept of metabolic pathways from a network-oriented perspective (de Figueiredo *et al*., 2009; Pey *et al*., 2011, 2013). One of the most important pathway concepts is that of Elementary Flux Mode (EFM) (Schuster *et al*., 2000). EFMs are a minimum set of enzymes necessary to accomplish mass-balance and thermodynamic (irreversibility) conditions (Schuster *et al*., 2000). Although the mass-balance and thermodynamic constraints are directly imposed by means of two linear constraints, the condition that guarantees that only a minimum number of reactions is active, referred to as the non-decomposability condition (NDC), is more difficult and demands further mathematical considerations.

Efficient algebraic frameworks can be found in the literature for calculating the whole set of EFMs of a given metabolic network. These methods are based on an iterative process that has to be completed so as to guarantee that the obtained solutions are EFMs. However, the number of EFMs increases exponentially with the number of reactions constituting the network (Acuña *et al*., 2010). Consequently, applying these methodologies to GSMNs goes beyond their scope. Because it is not possible to calculate all the EFMs in GSMNs, different mathematical frameworks were later developed to provide a particular

*To whom correspondence should be addressed.

subset of them (de Figueiredo *et al*., 2009; Machado *et al*., 2012; Rezola *et al*., 2013), most of them based on Mixed Integer Linear Programming (MILP).

These optimization methods typically enumerate EFMs in an increasing number of reactions and have been proved effective for a number of applications (Rezola *et al*., 2013, 2014). However, they allow us to add only one biological constraint in the search procedure, e.g. computing the 100 shortest EFMs producing L-lysine, as stated by de Figueiredo *et al*. (2009). This limitation is due to NDC. In other words, currently in the literature, there is no general methodology for taking into account more than one biological constraint in the EFM computation without violating NDC.

In this work, we present a novel approach based on MILP that is able to calculate a subset of EFMs fulfilling additional constraints. The framework is illustrated with a toy example and validated with two networks (Rezola *et al*., 2011; Schuster *et al*., 2000), where all EFMs can be obtained. Subsequently, the scalability of our approach is confirmed by calculating a subset of EFMs in the genome-scale metabolic network of *Saccharomyces cerevisiae* (Heavner *et al*., 2012). Here, we investigated EFMs simultaneously consuming glucose and producing ethanol, guaranteeing that a minimum yield is achieved. We also validated the performance of this new approach when more than two constraints are imposed, particularly by calculating 100 EFMs activating a random set of five reactions.

2 METHODS

Here, we introduce an MILP that allows us to calculate a subset of EFMs satisfying several biological constraints. At the end of the Section, the methodology is illustrated with a toy example.

As discussed earlier, an EFM is a minimum set of enzymes necessary to fulfill mass-balance and thermodynamic constraints. Although the last two conditions can be imposed by means of linear equations, the first one, referred to as the NDC, requires further considerations. In particular, a solution fulfills NDC if no subset of its active reactions can satisfy the mass-balance and thermodynamic constraints.

2.1 Mass-balance and thermodynamic constraints

Assume a metabolic network comprising *C* compounds and *R* reactions. For each reaction *r* ($r = 1, \ldots, R$), we assigned a continuous flux variable v_r representing its activity. These activities are included in the flux vector $v = [v_1, \ldots, v_R]$.

EFMs assume that the concentration of the so-called internal metabolites (*I*) remains constant over time (steady-state condition). This is represented by Equation (1), where S_{cr} is the stoichiometric coefficient associated with metabolite *c* ($c = 1, \ldots, C$) in reaction *r* ($r = 1, \ldots, R$). Note here that substrates have a negative stoichiometric coefficient, whereas products have a positive stoichiometric coefficient. These coefficients are grouped into the stoichiometric matrix (*S*). The steady-state condition provides as many linear constraints as internal metabolites present in the metabolic network under study.

$$\sum_{r=1}^{R} S_{cr} \cdot v_r = 0, \forall c \in I \tag{1}$$

In addition, EFMs must satisfy thermodynamic constraints associated with irreversible reactions, which can perform only in one direction. As typically done in the literature (Pey *et al*., 2011; Rezola *et al*., 2013), we count two irreversible steps for each reversible reaction. Therefore, all fluxes are non-negative:

$$v_r \geq 0, r = 1, \ldots, R \tag{2}$$

With this transformation, the solution space defined by Equation (1) and (2) becomes a pointed polyhedral cone, *P*.

2.2 NDC

As discussed elsewhere in the literature, e.g. Larhlimi and Bockmayr (2009), if spurious cycles are neglected, the extreme rays in *P* are precisely the EFMs of a network under consideration. Note that this claim holds when reversible reactions are divided into two irreversible steps (Rezola *et al*., 2011), as done precisely in Equation (2). For this reason, methods aiming to determine the full set of EFMs are based on algorithms to calculate extreme rays in a pointed polyhedral cone, such as the double description method (Terzer and Stelling, 2008). Our approach presented here is substantially different, as it relies on extreme points in a polytope and linear programming, as detailed later.

To convert *P* into a polytope, we include an additional linear constraint, as observed in Equation (3). Note that $b^T \in \mathbb{R}^R$ and $c \in \mathbb{R}$ are known. This generic linear inequality can be used, for example, to force the activation of at least a reaction in a given set *F*; see Equation (4).

$$b^T \cdot v \geq c \tag{3}$$

$$\sum_{r \in F} v_r \geq 1 \tag{4}$$

With the additional constraint in Equation (3), extreme rays in *P* are intersected and become extreme points of a polytope, as observed in Figure 1B. Therefore, EFMs now correspond to the extreme points in the feasible region defined by (1)–(3). Building a search procedure for EFMs based on extreme points is a powerful strategy and has been exploited previously (Kaleta *et al*., 2009), as highly developed algorithms for linear programming can be found in the literature, such as the Simplex algorithm (Dantzig *et al*., 1955), which provides an extreme point as an optimal solution. Our goal here is different, as we aim to search for EFMs that satisfy different biological constraints and, therefore, the theoretical background needs to be extended.

Let us now consider the mathematical properties of extreme points in the feasible region by (1)–(3) and EFMs. To that end, we first define S^* as the matrix grouping *S* with b^T, as shown below:

$$S^* = \begin{bmatrix} S \\ b^T \end{bmatrix} \tag{5}$$

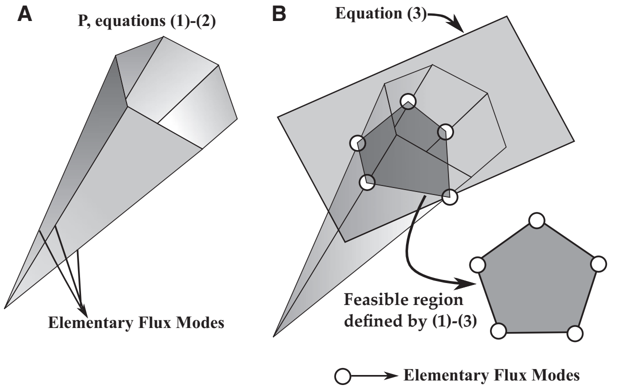

Fig. 1. Solution space generated by (**A**) Equations (1) and (2) and (**B**) Equations (1)–(3)

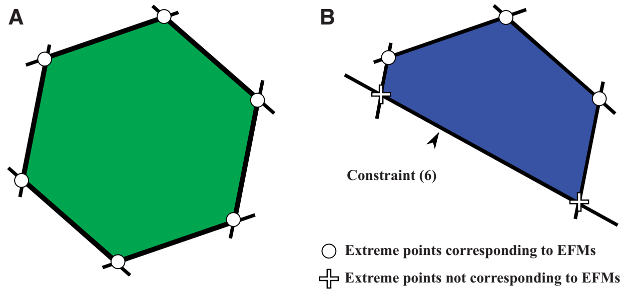
A
B
Constraint (6)
○ Extreme points corresponding to EFMs
✚ Extreme points not corresponding to EFMs

Fig. 2. Feasible regions defined by (**A**), (1)–(3) and (**B**), (1)–(3) and (6)

Without the abuse of notation, we introduce the operator i actuating in a given matrix, which represents the submatrix involving the columns corresponding to the active reactions in the solution i.

Considering that the solution i is an extreme point of (1)–(3) and therefore, an EFM, the null space of matrix S_i^* has zero degrees of freedom (*df*), as the solution i is the geometric intersection of an extreme ray in P and the hyperplane in (3). On the other hand, because removing (3) from S_i^* always reduces its rank in a unit, the null space of S_i has one *df*.

Two proofs can be found in the Supplementary Material reinforcing the relationship between extreme points and EFMs. In particular, we show that (i) if NDC is satisfied in solution i, S_i has one *df*, and (ii) if S_i has one *df*, NDC is satisfied and therefore, solution i is an EFM. These mathematical considerations constitute the starting point to understand the approach described below to include additional constraints.

2.3 Including additional constraints

We now introduce an additional constraint to Equations (1)–(3). In analogy to (3), Equation (6) represents a generic linear constraint, where $d^T \in \mathbb{R}^R$ and $e \in \mathbb{R}$ are known parameters.

$$d^T \cdot v \geq e \tag{6}$$

To illustrate the effect of Equation (6), Figure 2A and B represent the feasible region defined by (1)–(3) and by (1)–(3) and (6), respectively. With the addition of Equation (6), novel extreme points may arise, represented in Figure 2B by white crosses. The null space of matrix S_i for these extreme points has more than one *df* and, consequently, they are not EFMs. In particular, extreme points found in both feasible spaces in Figure 2 are precisely the solutions of interest, represented with a white dot. The mathematical properties of extreme points are considered next.

Similarly to Equation (5), we introduce S^{**} as the matrix grouping S with b^T and d^T, as illustrated by Equation (7):

$$S^{**} = \begin{bmatrix} S \\ b^T \\ d^T \end{bmatrix} \tag{7}$$

We now analyze the number of *df* of the matrices S^{**}, S^* and S so as to discuss when an extreme point from the feasible region defined by Equations (1)–(3) and (6) is an EFM. As shown in the Supplementary Material, a solution i is an EFM if its corresponding matrix S_i has one *df*. While this was always achieved for the extreme points in the feasible region defined by Equations (1)–(3), two possible scenarios arise when Equation (6) is included, as seen in Table 1.

Two considerations should be kept in mind. First, assuming that the solution i is an extreme point of the feasible region defined by Equations (1)–(3) and (6), then S_i^{**} has zero *df*. Second, if the solution i satisfies Equations (1)–(3) and (6) and is an EFM, then it will correspond to an extreme point in the system of Equations (1)–(3). Consequently, S_i^* has also zero *df*.

Table 1. Dimension of the nullspace in each scenario

	(1)–(3)	(1)–(3) and (6)	
Is an EFM?	✓	(I) ✓	(II) ✗
S_i	1	1	2
S_i^*	0	0	1
S_i^{**}	–	0	0

Overall, we aim to calculate extreme points shared by the feasible regions (1)–(3) and (1)–(3) and (6). To guarantee that the obtained solution is an EFM, S_i^* and S_i^{**} must have zero *df*. In other words, removing d^T from S_i^{**}does not affect the number of *df*. This occurs if d^T can be written as a linear combination of the rows of S_i^*. This is precisely the key point in our approach so as to guarantee that the obtained solution, which simultaneously satisfies (3) and (6), is certainly an EFM and, therefore, fulfills NDC. Note that the linear combination should be achieved by combining rows in S_i^*, not in S^*. This is not trivial because the set of columns forming the target EFM is not known beforehand. To overcome this, we introduce here a MILP which will allow us to directly calculate an EFM fulfilling at the same time (1)–(3) and (6).

2.4 Mixed Integer Linear Programming framework

We introduce the W matrix that contains the transposed rows defined in Equations (1)–(3), as observed in Equation (8). Note that S^T refers to the transpose of the original stoichiometric matrix S. The dimensions of W are $R{\cdot}(C+1)$ and, therefore, we have a row for each reaction denoted as W_r.

$$W = \left[S^T b \right] \tag{8}$$

We need to write d vector in Equation (6) as a linear combination of columns in W. The introduction of non-negative slack variables, ε_r and δ_r ($r = 1, \ldots, R$), is required to allow a reaction to be neglected when imposing the linear combination, as it only applies to reactions activated in the final solution. This is shown in Equation (9). Note here that the vector $x \in \mathbb{R}^{c+1}$ stores the coefficients of the linear combination.

$$W_r \cdot x = d_r + \varepsilon_r - \delta_r, r = 1, \ldots, R \tag{9}$$

We now introduce the binary variable z_r, being $z_r = 1$ if $v_r > 0$ and $z_r = 0$ if $v_r = 0$. This is achieved with Equation (10). Note that M is a sufficiently big scalar.

$$v_r \geq z_r, \;\; M \cdot z_r \geq v_r, r = 1, \ldots, R \tag{10}$$

In addition, Equation (11) allows us to impose that if $z_r = 1$ then $\epsilon_r = \delta_r = 0$, namely, when reaction r is active, it is forcing its corresponding row in W to take part in the linear combination. This is imposed by the following equation:

$$M(1 - z_r) \geq \varepsilon_r + \delta_r, r = 1, \ldots, R \tag{11}$$

If a feasible solution is found in the set of constraints (1)–(3), (6), (9)–(11), we can be sure that there is, at least, one EFM satisfying Equations (3) and (6). The opposite also applies, namely, if the set of constraints (1)–(3), (6), (9)–(11) is infeasible; then no EFM can satisfy Equations (3) and (6).

Finally, minimizing the number of active reactions, we guarantee that the solution is an EFM:

$$\min \sum_{r=1}^{R} z_r \tag{12}$$

As done in previous works (de Figueiredo *et al.*, 2009; Rezola *et al.*, 2013), we can include constraints for enumerating a given number of EFMs. We introduce here Z_r^k as the value of the z_r variable in the *k*-th solution. With Equation (13), we prevent previously calculated solutions from appearing again.

$$\sum_{r=1}^{R} Z_r^k \cdot z_r \leq \sum_{r=1}^{R} Z_r^k - 1 \tag{13}$$

Note here that this methodology, comprising Equations (1)–(3), (6), (9)–(13), can be naturally extended to include additional constraints. Let us assume that, in analogy to Equation (6), *J* additional constraints must be satisfied, as observed in equation (14). Note that d^j and e^j, $j = 1, \ldots, J$, are input information.

$$\left(d^j\right)^T \cdot v \geq e^j, j = 1, \ldots, J \tag{14}$$

To guarantee that each d^j can be written as a linear combination of the rows of S and b, Equations (9) and (11) are redefined as follows:

$$W_r \cdot x^j = d_r^j + \varepsilon_r^j - \delta_r^j, j = 1, \ldots, J, r = 1, \ldots, R \tag{15}$$

$$M(1 - z_r) = \sum_{j=1}^{J} \varepsilon_r^j - \delta_r^j, r = 1, \ldots, R \tag{16}$$

As a result, for each new linear constraint *j*, we add up $(C + 1) + 2 \cdot R$ new (continuous) variables to the model (x^j, ε^j and δ^j), as well as *R* additional constraints. Because equation (15) does not involve any binary variable, including additional constraints does not dramatically increase the required computational time, as shown in the Section 3.

2.5 Illustrative example

The methodology presented above is now illustrated by means of a toy example. In particular, this example is based on the metabolic network in Figure 3A, which comprises one metabolite and four reactions. Let us consider two different questions: (i) finding an EFM activating reactions 1 and 3 and (ii) calculating an EFM with reactions 1 and 2.

In the first question, we impose $v_1 \neq 0$ and $v_3 \neq 0$, leading to $\varepsilon_1 = \gamma_1 = \varepsilon_3 = \gamma_3 = 0$. Therefore, based on the system of equations in Figure 3B, $x_1 = -1$ and $x_2 = 1$. In addition, the mass-balance constraint is satisfied by activating only v_1 and v_3, and, therefore, we can ensure that v_1 and v_3 can operate together in an EFM. Finally, as we minimize the number of active reactions, we have $v_1 = v_3 = 1$.

Regarding the second question, after activating v_1 and v_2 and imposing $\varepsilon_1 = \gamma_1 = \varepsilon_2 = \gamma_2 = 0$, the system of equations in Figure 3C leads to $x_1 = 1$ and $x_2 = -1$. However, the mass-balance is not achieved activating v_1 and v_2. In particular, we need to consume the produced molecules of *A* by v_1 and v_2. Note that v_3 and v_4 can perform this task and, therefore, we have the following cases:

(1) $v_3 \neq 0$: the activation of v_1, v_2 and v_3 and the subsequent inactivation of $\varepsilon_1 = \gamma_1 = \varepsilon_2 = \gamma_2 = \varepsilon_3 = \gamma_3 = 0$ leads to an incompatible system of equations in Figure 3C. Thus, this case does not fulfill the imposed systems of equations.

(2) $v_4 \neq 0$. As in the previous case, with the activation of v_1, v_2 and v_4 and the subsequent inactivation of $\varepsilon_1 = \gamma_1 = \varepsilon_2 = \gamma_2 = \varepsilon_4 = \gamma_4 = 0$, the arising system of equations (Fig. 3C) is not consistent.

(3) $v_3 \neq 0$ and $v_4 \neq 0$: As in the previous cases, the system of equations becomes infeasible.

Overall, there is no flux configuration satisfying Equations (1)–(3) and (6), as well as Equation (9). Therefore, no EFM can simultaneously activate v_1 and v_2. This second scenario constitutes one of the extreme points represented with the white crosses in Figure 2B.

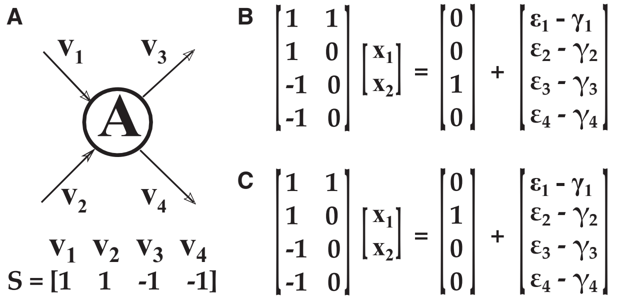

Fig. 3. (**A**) Example metabolic network and its corresponding stoichiometric matrix (*S*). (**B**) Equation (9) for scenario (i). (**C**) Equation (9) for scenario (ii)

3 RESULTS

3.1 Validation

In this section, an empirical validation of the methodology proposed above is conducted. With this analysis, we aim to show that our approach correctly decides whether there is an EFM activating two particular reactions and enumerates all their underlying solutions. To that end, we need to use a network where all the EFMs can be determined.

For that, let us first consider the metabolic network presented by Schuster *et al.* (2000), which merges the glycolysis, pentose phosphate pathway and part of the gluconeogenesis. This network involves seven EFMs. Several approaches can be applied to enumerate this set of EFMs (de Figueiredo *et al.*, 2009; Kaleta *et al.*, 2009; Terzer and Stelling, 2008; Von Kamp and Schuster, 2006). Using the approach described above, for each pair of reaction involved in at least one EFM, we successfully enumerated all its underlying solutions. We extended this analysis for the rest of the pairs of reactions, and, as expected, an infeasible MILP was obtained.

Next, we studied the computational performance of our methodology in the light of a more complex metabolic network. In particular, we considered the metabolic network presented by Rezola *et al.* (2011), which was obtained after lumping all the reactions involved in the 100 shortest Generating Flux Modes producing lysine in the GSMN reconstruction of *Escherichia coli* presented by Feist *et al.*(2007). Overall, this metabolic network comprises 99 metabolites and 195 reactions, after splitting reversible reactions into two irreversible steps. We computed the full set of EFMs with efmtool (Terzer and Stelling, 2008), obtaining 354 225 solutions. For convenience, we define here a feasible pair as a couple of reactions being simultaneously activated in, at least, one of these EFMs. Similarly, an infeasible pair of reactions is defined as a subset of two reactions that do not appear together in any EFM. The validation here is divided in two parts with the objective of confirming that (i) a feasible pair leads to a feasible MILP and (ii) an infeasible pair always produces an infeasible MILP. Therefore, here we are not concerned with finding an EFM, but proving whether an MILP is feasible. Finding a feasible solution is sufficient to guarantee that at least one EFM exists involving a pair of reactions and, therefore, the objective function is not needed.

Efficient algorithms solving MILPs can be found in the literature. However, the computation time required to obtain a

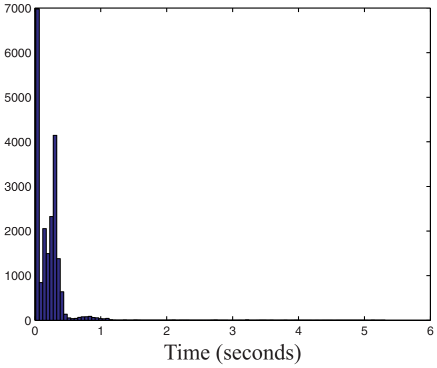

Fig. 4. Computation times found in the analysis of the network presented by Rezola *et al.* (2011)

solution could be high. The causes that increase the computation time are not always tied to the complexity of the problem. In particular, trivial effects, such as the order of the constraints, as well as the selected solver, may severely affect the computation time. Aiming at minimizing these trivial effects, we detail in the Supplementary Material an alternative formulation. It should be highlighted here that in essence both formulations represent the same problem. A priori, we cannot conclude which formulation is more efficient for a particular pair of reactions. Therefore, we solved both problems and selected the one involving the minimum computation time.

We successfully found a solution for each feasible pair. Figure 4 shows the computation time histogram corresponding to 99% (21 142) of the feasible pairs, while discarding 1% (214) of them, which represent strong outliers that distort the figure. In particular, out of 21 356 feasible pairs, only 20 of them required >100 s, with the longest computational time being equal to 1400 s. The median and the mean values are equal to 0.2 and 0.96 s, respectively. Note from Figure 4 that 5.34 s are sufficient for capturing 99% of the feasible pairs.

To emphasize the progress brought by our approach, we explored the model presented by de Figueiredo *et al.* (2009) when additional constraints are directly included. For that, as this model does not prevent non-elementary solutions (those represented with a cross in Fig. 2B) from appearing, we calculated the number of them required to obtain an EFM per each feasible pair. In particular, we enumerated an average of 1197 solutions before obtaining an EFM, which clearly highlights the necessity of the methodology presented here.

We consider now the second part of this validation, namely, whether infeasible pairs lead to infeasible MILPs. Finding whether a given MILP is infeasible can be extremely hard. To overcome this, we impose a time limit for the solving procedure. In other words, if a particular MILP exceeds this time limit and no solution has been found, we assume that it is infeasible. As discussed above, 99% of the feasible pairs required <5.34 s to validate their feasibility. Based on this, we imposed an upper limit of 5.34 s when solving each MILP corresponding to infeasible pairs. After performing this analysis, we found two scenarios: (i) the MILP was identified as infeasible or (ii) the time limit was reached. It is important to emphasize that no feasible MILP can be obtained here. To reinforce this, we randomly selected 10 infeasible pairs for which the time limit was reached and repeated the analysis increasing the time limit to 21 600 s (6 h). After performing this task, no feasible solution was found, which supports here again the correct behavior of our approach.

Finally, we validated the performance of our approach when more than two constraints are imposed. For that, we generated 10 000 random triplets of reactions and evaluated, in each case, the solution obtained when the reactions in the triplet are forced as active. On one hand, we checked that an elementary solution was directly obtained for each feasible triplet. On the other hand, we confirmed that when no EFM exists involving a given triplet of reactions, an infeasible model arises or no solution is found before reaching the time limit, namely, 5.34 s. This also generalizes our approach to consider multiple constraints.

3.2 Case study

The microbial-based biofuel production constitutes a relevant topic in the field of bioengineering. In this light, the yeast *S.cerevisiae* is one of the most widely used cell factories (Caspeta and Nielsen, 2013). In addition, the metabolism of this yeast has been extensively studied for years, and many genome-scale metabolic reconstructions can be found in the literature (Förster *et al.*, 2003; Heavner *et al.*, 2012).

In particular, much effort is being put into optimizing the yield of ethanol production. Improving this yield leads to a reduction of the cost per unit of produced ethanol and an increase in the economic efficiency of the process. In this section, we will calculate EFMs with a 100% yield from glucose. Including this information in the model can be easily achieved by means of the following linear equation:

$$v_{etoh} - 3 \cdot v_{glucose} \geq 0 \qquad (17)$$

where v_{etoh} represents the flux of the reaction producing ethanol and $v_{glucose}$ the activity of the reaction consuming glucose. The coefficient of 3 multiplying $v_{glucose}$ imposes that the proportion of consumed glucose and produced ethanol is that corresponding to a yield of 100%.

In addition, because solutions deactivating both v_{etoh} and $v_{glucose}$ lack from interest, we impose a lower bound for the glucose consumption, as shown below:

$$v_{glucose} \geq 1 \qquad (18)$$

Note here that Equations (17) and (18) can be included in the model by substituting Equations (3) and (6). Therefore, we will obtain EFMs activating both v_{etoh} and $v_{glucose}$ and guaranteeing the minimum yield imposed by Equation (17). In addition to glucose, we also simulated a free supply of glycine, which has a positive effect in the ethanol concentration (Thomas *et al.*, 1994). Therefore, we may obtain EFMs consuming both glucose and glycine, leading to a yield above the imposed in (17).

We based our analysis on the recently published genome-scale reconstruction Yeast 5 (Heavner *et al.*, 2012). Aiming at reducing the computation time, the solving procedure was slightly modified. In particular, the MILP solver provides the first feasible integer solution (FFI) obtained. We guarantee that within the reactions active in FFI at least one EFM fulfilling the imposed conditions is involved. Efficient algebraic techniques (Terzer and

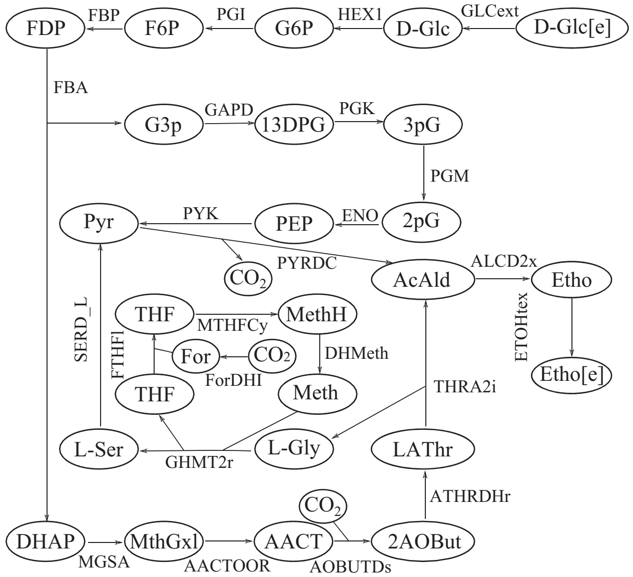

Fig. 5. An EFM producing ethanol from glucose with a carbon yield of 100%. For the sake of simplicity, cofactors, despite being balanced in the obtained solution, are not included in the figure. Reactions: AACTOOR, Aminoacetone: oxygen oxidoreductase (deaminating) (flavin-containing); ALCD2x, alcohol dehydrogenase; AOBUTDs, L-2-amino-3-oxobutanoate decarboxylation; ATHRDHr, L-allo-threonine dehydrogenase; DHMeth, methylenetetrahydrofolate dehydrogenase (NAD); ENO, enolase; ETOHtex, ethanol transport; FBA, fructose-bisphosphate aldolase; FBP, fructose-bisphosphatase; FTHFl, formate-tetrahydrofolate ligase; ForDHI, formate dehydrogenase; GAPD glyceraldehyde-3-phosphate dehydrogenase; GHMT2r, glycine hydroxymethyltransferase; GLCext, glucose transport; HEX1, hexokinase (D-glucose:ATP); MGSA, methylglyoxal synthase; MTHFCy, methenyltetrahydrofolate cyclohydrolase; PGI, glucose-6-phosphate isomerase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; PYK, pyruvate kinase; PYRDC, pyruvate decarboxylase; SERD_L, L-serine deaminase; THRA2i, L-allo-threonine aldolase. Metabolites: 10fthf, 10-formyl-THF; 13DPG, 1,3-bisphospho-D-glycerate; 2AObut, L-2-amino-3-oxobutanoate; 2pG, 2-phospho-D-glyceric acid; 3pG, 3-phosphoglycerate; AcAld, acetaldehyde; CO_2, carbon dioxide; D-Glc, D-glucose; D-Glc[e], D-glucose[extracellular]; DHAP, dihydroxyacetone phosphate; Etho, ethanol; Etho[e], ethanol[extracellular]; F6P, D-fructose 6-phosphate; FDP, D-fructose 1,6-bisphosphate; For, formate; G3p, glyceraldehyde 3-phosphate; G6P, D-glucose 6-phosphate; L-Gly, L-glycine; L-Gly, L-glycine; L-Ser, L-serine; Meth, 5,10-methylenetetrahydrofolate(2-); MethH, 5,10-methenyl-THF; MthGxl, methylglyoxal; PEP, phosphoenolpyruvate; Pyr, pyruvate; THF, 5,6,7,8-Tetrahydrofolate; aact, aminoacetone; athr-L, L-allothreonine

Stelling, 2008) can be applied to calculate such EFM from the FFI. This procedure reduces the computation time from few minutes to seconds.

Overall, we calculated 100 EFMs in an average time of 60 s per solution. Next, we validate that the obtained solutions are, indeed, EFMs satisfying (17)–(18). We present in Figure 5, one of these solutions where all the carbons of the glucose are diverted to ethanol production. The list of calculated EFMs can be found in the Supplementary Material.

For completeness, we evaluated our approach when more than two additional constraints are imposed. In particular, we calculated 100 EFMs activating a given random set of five reactions: (i) *6-phosphogluconolactonase*, (ii) *nucleoside triphosphatase*, (iii) *transketolase 1*, (iv) *glyoxylate transport* and (v) *malate dehydrogenase*. We were able to enumerate 100 EFMs using a regular desktop computer in an average time of 20 min per EFM. Despite the expected time increment, this analysis reflects the capability of our approach to deal with multiple reactions.

4 CONCLUSIONS

EFMs constitute a well-established approach developed in the past 20 years. The EFM approach has been applied to different questions in systems biology. However, its application to GSMNs is restricted because of the combinatorial explosion in the number of solutions. In the past years, approaches to calculating a subset of EFMs in GSMNs have been developed. However, because the number of EFMs emerging from large metabolic networks goes beyond the scope of any approach, alternative strategies are required.

In this article, instead of calculating a large number of EFMs so as to properly represent any metabolic phenotype, we define a novel framework providing those EFMs of interest. This question is not straightforward and requires additional mathematical developments, as presented here. Based on them, we define a MILP that enumerates EFMs fulfilling several biological constraints. Although efficient algorithms and methods exist to solve MILPs, they may take an unacceptable amount of time. Aiming at reducing the computation time, redefining the formulation preventing the use of integer variables constitutes a future line of research. For situations with several biological constraints, this is particularly relevant, as the number of variables and constraints further increases and, therefore, the arising MILP will be even more difficult to solve.

The methodology was validated in two metabolic networks for which the full set of EFMs can be calculated. Its application to GSMNs was also confirmed studying the ethanol production by the yeast *S.cerevisiae*. In this case study, we also showed different types of constraints that can be included in our model, e.g. activation of a set of reactions, imposing a minimum yield etc. We emphasize here that, based on the metabolic reconstruction under consideration, different EFMs with a carbon yield of 100% were obtained. The thermodynamic feasibility of these solutions requires further consideration. Nevertheless, these EFMs may provide researchers with novel efficient mechanisms to produce ethanol more efficiently.

Our approach allows us to more effectively compute a selective subset of EFMs and, therefore, it opens new possibilities to explore metabolic pathways without computing the complete set of EFMs. The value of computing a subset of EFMs has been previously shown in different works: analyzing omics data (Rezola *et al*., 2013), elucidating novel metabolic pathways in human metabolism (Kaleta *et al*., 2011), revealing potential strategies for metabolic engineering (Ip *et al*., 2011) or understanding the role of different carbon sources in the growth of *E.coli* (Chan and Ji, 2011). Our approach can certainly complement and extend these previous analyses.

In particular, the inherent potential of this approach can be strengthened with the integration of omics data. For instance, if we observe that a given set of enzymes are coexpressed in a particular scenario, applying the methodology presented here will provide us the metabolic mechanisms connecting them. Analogously, this can be extended to integrate data from metabolomic experiments.

Overall, the contribution presented here will undoubtedly facilitate the application of EFMs in GSMNs and overcome the manifold biological questions that were unreachable before.

Funding: The work of Jon Pey was supported by the Basque Government.

Conflict of Interest: none declared.

REFERENCES

Acuña,V. *et al.* (2010) A note on the complexity of finding and enumerating elementary modes. *Biosystems*, **99**, 210–214.

Caspeta,L. and Nielsen,J. (2013) Economic and environmental impacts of microbial biodiesel. *Nat. Biotechnol.*, **31**, 789–793.

Chan,S.J. and Ji,P. (2011) Decomposing flux distributions into elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, **27**, 2256–2262.

Dantzig,G.B. *et al.* (1955) The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pac. J. Math.*, **5**, 183–195.

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.

Förster,J. *et al.* (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.*, **13**, 244–253.

Heavner,B.D. *et al.* (2012) Yeast 5–an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Syst. Biol.*, **6**, 55.

Ip,K. *et al.* (2011) Analysis of complex metabolic behavior through pathway decomposition. *BMC Syst. Biol.*, **5**, 91.

Kaleta,C. *et al.* (2009) EFMEvolver: computing elementary flux modes in genome-scale metabolic networks. *Lect. Notes Inform.*, **P-157**, 179–189.

Kaleta,C. *et al.* (2011) In silico evidence for gluconeogenesis from fatty acids in humans. *PLoS Comput. Biol.*, **7**, e1002116.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Keseler,I.M. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

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

Machado,D. *et al.* (2012) Random sampling of elementary flux modes in large-scale metabolic networks. *Bioinformatics*, **28**, 515–521.

Nelson,D.L. and Cox,M.M. (2000) *Lehninger Principles of Biochemistry*. 3rd edn. W.H. Freeman, New York, NY.

Pey,J. *et al.* (2011) Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biol.*, **12**, R49.

Pey,J. *et al.* (2013) Refining carbon flux paths using atomic trace data. *Bioinformatics*, **30**, 975–980.

Planes,F.J. and Beasley,J.E. (2008) A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Brief. Bioinform.*, **9**, 422–436.

Rezola,A. *et al.* (2011) Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, **27**, 534–540.

Rezola,A. *et al.* (2013) Selection of human tissue-specific elementary flux modes using gene expression data. *Bioinformatics*, **29**, 2009–2016.

Rezola,A. *et al.* (2014) Advances in network-based metabolic pathway analysis and gene expression data integration. *Brief. Bioinform.*

Schellenberger,J. *et al.* (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, **11**, 213.

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

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

Thomas,K.C. *et al.* (1994) Effects of particulate materials and osmoprotectants on very-high-gravity ethanolic fermentation by *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.*, **60**, 1519–1524.

Von Kamp,A. and Schuster,S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22**, 1930–1931.