**Systems Biology**

**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, Manuel de Lardizabal 15, 20018 San Sebastian, Spain.

**ABSTRACT**

**Motivation:** The concept of Elementary Flux Mode has been widely used for the last 20 years. However, its application to genome-scale metabolic networks (GSMNs) is still under development due to methodological limitations. Therefore, novel approaches are demanded to extend the application of EFMs. A novel family of frameworks based on optimization is emerging which provide us a limited set of EFMs. Since the calculation of the whole set of EFMs goes beyond our capacity, performing a selective search is the proper strategy.

**Results:** We present here 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.

**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 et al., 2008), 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 thought different online repositories (Schellenberger et al., 2010).

The inherent complexity of GSMNs makes not 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; Pey et al., 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). Whilst 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. As it is not possible to calculate all the EFMs in GSMNs, different mathematical frameworks were later developed to provide a particular subset of them (Rezola et al., 2013; Machado et al., 2012; de Figueiredo et al., 2009), most of them based on Mixed Integer Linear Programming (MILP).

© The Author (2014). Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
These optimization methods typically enumerate EFMs in increasing number of reactions and have been proved effective for a number of applications (Rezola et al., 2013; Rezola et al., 2014). However, they only allow us to add only one biological constraint in the search procedure, e.g. computing the 100 shortest EFMs producing L-lysine, as in (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 able to calculate a subset of EFMs fulfilling additional constraints. The framework is illustrated with a toy example and validated with two networks (Schuster et al., 2000; Rezola et al., 2011), 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 5 reactions.

### 2 METHODS

Here we introduce a 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 above, an EFM is a minimum set of enzymes necessary to fulfill mass balance and thermodynamic constraints. While the last two conditions can be imposed by means of linear equations, the first one, referred to as the non-decomposability condition (NDC), requires further considerations. In particular, a solution fulfills NDC if no subset of its active reactions can satisfy the mass balance and thermodynamics constraints.

#### 2.1 Mass balance and thermodynamic constraints

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

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 $S_r$ is the stoichiometric coefficient associated with metabolite $c (c=1,...,C)$ in reaction $r (r=1,...,R)$. Note here that substrates have a negative stoichiometric coefficient, whilst 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.

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

In addition, EFMs must satisfy thermodynamic constraints associated with irreversible reactions, which can only perform 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:

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

With this transformation, the solution space defined by equation (1)-(2) becomes a pointed polyhedral cone, $P$.

#### 2.2 Non-decomposability condition

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 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 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 below.

![Fig 1. Solution space generated by A) equations (1)-(2) and B) equations (1)-(3).](http://bioinformatics.oxfordjournals.org/)

In order 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).

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

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

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 defined by (1)-(3). Building a search procedure for EFMs based on extreme points is a powerful strategy and have been exploited previously (Kaleta et al., 2009), as highly developed algorithms for linear programming can be found in the literature, such as Simplex algorithm (Dantzig et al., 1955), which provides an extreme point as 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 consider now 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:

$$S^* = \begin{bmatrix} S \\ b^T \end{bmatrix} \tag{5}$$
Without the abuse of notation, we introduce the operator $i$ acting in a given matrix, which represents the sub-matrix 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 $S_i^\top$ has zero degrees of freedom (dofs), as the solution $i$ is the geometric intersection of an extreme ray in $P$ and the hyperplane in (3). On the other hand, since removing (3) from $S_i^\top$ always reduces its rank in a unit, the null space of $S_i$ has one dof.

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 dof, and (ii) if $S_i$ has one dof, 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 the system of equations in (1)-(3). In analogy to (3), equation (6) represents a generic linear constraint, where $d^e \in \mathbb{R}^2$ and $e \in \mathbb{R}$ are known parameters.

$$d^T \cdot e \geq e$$  \hspace{1cm} (6)

To illustrate the effect of equation (6), Figure 2A and Figure 2B 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, sent the feasible region defined by (1)m(3) and by (1)m(3) and (6), respectively.

Similarly to equation (5), we introduce $S_i^{**}$ as the matrix grouping $S_i$ with $b^e$ and $d^e$, as illustrated by equation (7):

$$S_i^{**} = \begin{bmatrix} S_i & b^T \\ d^T \end{bmatrix}$$  \hspace{1cm} (7)

We now analyze the number of dof of the matrices $S_i^*$, $S_i^\top$ and $S_i$ 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 1 dof. 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.

### Table 1. Dimension of the nullspace in each scenario

<table>
<thead>
<tr>
<th></th>
<th>(1)-(3)</th>
<th>(1)-(3) and (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is an EFM?</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$S_i^*$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$S_i^{**}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_i^\top$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Two considerations should be kept in mind. First, assuming that the solution $i$ is an extreme point of the feasible region defined by equations in (1)-(3) and (6), then $S_i^{**}$ has zero dof. 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 dof.

Overall, we aim to calculate extreme points shared by the feasible regions (1)-(3) and (1)-(3) and (6). In order to guarantee that the obtained solution is an EFM, $S_i^*$ and $S_i^{**}$ must have zero dof. In other words, removing $d^e$ from $S_i^{**}$ does not affect the number of dofs. This occurs if $d^e$ can be written as a linear combination of the rows of $S_i^\top$. 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 $S_i^\top$, not in $S_i^*$. Note also that this is not trivial, since the set of columns forming the target EFM are not known beforehand. To overcome this, we introduce here a Mixed-Integer Linear Programming framework, 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 equation (1)-(3), as observed in equation (8). Note that $S_i^\top$ is the transpose of the original stoichiometric matrix $S$. The dimensions of $W$ are $R \times (C+1)$ and, therefore, we have a row for each reaction denoted as $W_r$.

$$W = \begin{bmatrix} S_i^T \\ b \end{bmatrix}$$  \hspace{1cm} (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, $\varepsilon_r$ and $\delta_r$ ($r=1,..,R$), 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.

$$W_r \cdot x = d^e_r + \varepsilon_r - \delta_r , r = 1,..,R$$  \hspace{1cm} (9)

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

$$\varepsilon_r \geq z_r \cdot M, \quad M \cdot z_r \geq \varepsilon_r , \quad r = 1,..,R$$  \hspace{1cm} (10)

In addition, equation (11) allows us to impose that if $z_r=1$ then $\varepsilon_r=\delta_r=0$, 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:

$$M \cdot (I - z_r) \geq \varepsilon_r + \delta_r , \quad r = 1,..,R$$  \hspace{1cm} (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 equation (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 equation (3) and (6).

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

$$\min \sum z_r$$
\[
\min \sum_{r=1}^{R} z_r^k
\]

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 \(Z_j^k\) as the value of the \(z\)-variable in the \(k\)-th solution. With equation (13), we prevent previously calculated solutions from appearing again.

\[
\sum_{r=1}^{R} Z_j^k z_r^k \leq \sum_{r=1}^{R} Z_j^k - I
\]

Note here that this methodology, comprising equations (1)-(3), (6), (9)-(13), can be naturally extended so as 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 \(d^i\) and \(e^j\), \(j = 1, \ldots, J\), are input information.

\[
(d^i)^T \cdot v \geq e^j, \ j = 1, \ldots, J
\]

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

\[
W_r : x_r^j = d_r^j + e_r^j - \sum_{j=1}^{J} c_r^j, \ j = 1, \ldots, J, \ r = 1, \ldots, R
\]

\[
M(I - z_r) = \sum_{j=1}^{J} c_r^j - \sum_{j=1}^{J} \gamma_r^j, \ r = 1, \ldots, R
\]

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

### 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 1 metabolite and 4 reactions. Let us consider two different questions: (a) finding an EFM activating reaction 1 and 3 and (b) calculating an EFM with reactions 1 and 2.

![A)](image1)

\[
\begin{bmatrix}
1 & 1 \\
1 & 0 \\
1 & -1 \\
1 & -1
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 \\
\epsilon_1 - \gamma_1 & \epsilon_2 - \gamma_2 & \epsilon_3 - \gamma_3 & \epsilon_4 - \gamma_4
\end{bmatrix}
\]

In the first question we impose \(v_1 \neq 0\) and \(v_3 \neq 0\), leading to \(\gamma_1 = 1\). Therefore, based on the system of equations in Figure 3B, \(x_j = -1\) and \(x_j = 1\). 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 \(v_3\) and imposing \(\gamma_1 = 1\), the systems of equations in Figure 3C leads to \(x_j = 1\) and \(x_j = -1\) However, the mass-balance is not achieved activating \(v_1\) and \(v_3\). In particular, we need to consume the produced molecules of \(A\) by \(v_1\) and \(v_3\). Note that \(v_1\) and \(v_3\) can perform this task and, therefore, we have the following cases:

1. \(v_1 \neq 0\) and \(v_3 \neq 0\): the activation of \(v_1\), \(v_3\) and the subsequent inactivation of \(v_2\) leads to an incompatible system of equations in Figure 3C. Thus, this case does not fulfill the imposed systems of equations.
2. \(v_1 \neq 0\) and \(v_3 \neq 0\): As in the previous case, with the activation of \(v_1\), \(v_2\) and \(v_3\) and the subsequent inactivation of \(v_2\), the arising system of equations (Figure 3C) is not consistent.
3. \(v_1 \neq 0\) and \(v_3 \neq 0\): 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 \(v_3\). This second scenario constitutes one of the extreme points represented with the white crosses in Figure 2B.

### 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 if 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 in (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 (Von Kamp and Schuster, 2006; Terzer and Stelling, 2008; de Figueiredo et al., 2009; Kaleta et al., 2009). 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 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 in (Rezola et al., 2011), which was obtained after lumping all the reactions involved in the 100 shortest Generating Flux Modes producing Lysine in the GSNN reconstruction of *Escherichia Coli* presented in (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 efmmtool (Terzer and Stelling, 2008), obtaining 354225 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 if a MILP is feasible or not. Finding a feasible solution is sufficient so as 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 solution could be very 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 the 99% (21142) of the feasible pairs while discards 1% (214) of them, which represent strong outliers that distort the figure. In particular, out of 21356 feasible pairs, only 20 of them required more than 100 seconds, with the longest computational time being equal to 1400 seconds. The median and the mean values are equal to 0.2 seconds and 0.96 seconds, respectively. Note from Figure 4 that 5.34 seconds are sufficient for capturing the 99% of the feasible pairs.

In order to emphasize the progress brought by our approach, we explored the model presented in (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 Figure 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 if infeasible pairs lead to infeasible MILPs. Finding if 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, the 99% of the feasible pairs required less than 5.34 seconds to validate their feasibility. Based on this, we imposed an upper limit of 5.34 seconds 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 in this case, no feasible MILP can be obtained here. In order 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 21600 seconds (6 hours). 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 the 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 seconds. 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 *Saccharomyces 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 (Hevner et al., 2012; Förster et al., 2003).

In particular, much effort is being put into optimizing the yield of ethanol production. Improving this yield leads to reduce the cost per unit of produced ethanol and increase 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:

$$ v_{etoh} - 3v_{glucose} \geq 0 $$

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

$$ v_{glucose} \geq 1 $$

(18)

Note here that equations (17)-(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 Stelling, 2008) can be applied so as 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 seconds 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 5 reactions: (i) 6-phosphogluconolactonase, (ii) nucleoside triphosphatase, (iii) transketolase I, (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 minutes per EFM. Despite the expected time increment, this analysis reflects the capability of our approach to deal with multiple reactions.

### 4 CONCLUSIONS

Elementary Flux Modes (EFMs) constitute a well established approach developed in the last 20 years. The EFM approach has been applied to different questions in systems biology. However, its application to genome-scale metabolic networks is restricted due to the combinatorial explosion in the number of solutions. In the last years approaches to calculating a subset of EFMs in genome-scale metabolic networks have been developed. However, since 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 Mixed-Integer Linear Program (MILP) that enumerates several biological constraints. Although efficient algorithms and methods exist, solving MILPs 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 genome-scale metabolic networks was also confirmed studying the ethanol production by the yeast Saccharomyces cerevisiae. In this case study, we also showed different type of constraints that can be at the same time included in our model e.g. activate 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 to researchers novel efficient mechanisms so as to produce more efficiently ethanol.

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 have been previously shown in different works: analyzing omics data (Rezola et al.,...
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 source 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 co-expressed in a particular scenario, applying the methodology presented here will provide us the metabolic mechanisms connecting them. Analogously, this can be extended so as to integrate data from metabolomic experiments.

Overall, the contribution presented here will undoubtedly facilitate the application of EFM in genome-scale metabolic networks and overcome manifold of biological questions which were unreachable before.

ACKNOWLEDGEMENTS

The work of Jon Pey was supported by the Basque Government

REFERENCES


