**ABSTRACT**

Genome-scale metabolic networks let us understand the behaviour of the metabolism in the cells of living organisms. The availability of great amounts of such data gives the scientific community the opportunity to infer in silico new metabolic knowledge. Elementary Flux Modes (EFM) are minimal contained pathways or subsets of a metabolic network that are very useful to achieving the comprehension of a very specific metabolic function (as well as dysfunctions), and to get the knowledge to develop new drugs. Metabolic networks can have large connectivity and, therefore, EFM resolution faces a combinational explosion challenge to be solved. In this paper we propose a new approach to obtain EFM based on graph theory, the balanced graph concept and the shortest path between end nodes. Our proposal uses the shortest path between end nodes (input and output nodes) that finds all the pathways in the metabolic network and is able to prioritise the pathway search accounting the biological mean pursued. Our technique has two phases, the exploration phase and the characterisation one, and we show how it works in a well-known case study. We also demonstrate the relevance of the concept of balanced graph to achieve to the full list of EFMs.

**KEYWORDS**

metabolic networks; EFM; elementary modes; cellular metabolism; graph method

**INTRODUCTION**

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

Several mathematical methods modelling metabolism that are able to incorporate datasets provided by different omics technologies are emerging. Many of these methods are encompassed within constraint-based models, in which a set of mathematical constraints are defined using a genome-scale metabolic network (GSMN) reconstruction as a starting point. Several curated GSMNs can be found in the literature [1]. However, being able to automatically characterise the biochemical reactions present in a particular metabolism through omics data truly constitutes a challenge [2].

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

An elementary flux mode (EFM) [3] is a special type of metabolic pathway comprising a subset of reactions that meets the two aforementioned conditions plus the non-decomposability condition, that is, the pathway can not be decomposed into smaller solutions (i.e. a subset of the pathway is not a feasible pathway as well). In other words, EFM are solutions with the minimum support necessary to operate in stoichiometric steady-state balance with all reactions in the appropriate direction. EFM are an effort to translate a complex network into a canonical expression of vector generators of a solution space.

In a typical metabolic network the number of reactions is higher than the number of metabolites, so that many possibilities can be found that are a solution to the system. As the metabolic network increases in size so does the amount of EFM, whose number explodes in a combinatorial fashion [4]. Computing the full set of EFM in large metabolic networks still constitutes a challenge.

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

Unlike traditional Linear Programming approaches, our proposal avoids expensive floating-point calculations allowing us to speed-up the quest of all the available pathways in a certain metabolic network. Moreover, our approach is quite suitable to be developed in new commodity parallel architectures (such as multi- and many-core and GPU accelerators), allowing shorter execution times and less energy consumption. This article is an extension of the contribution presented at the 2015 IWBBIO conference [5].

The rest of the paper is structured as follows. In the Background Section, we comment briefly on the state of the art in the pathway extraction methods. Section Methods describes the constraint-based modelling, the artefacts needed and our strategy with detail. Section Discussion presents our contribution in relation to the balanced graph concept and a proposal method to build balanced graphs with minimum size to get EFMs. Section Results presents two case studies of our approach. The paper concludes by offering our conclusions and suggestions for future work.

BACKGROUND

The advantages of analysing metabolic networks based on EFMs have been shown in different works [6, 7]. However, the EFMs use has been limited because enumerating them is computationally demanding. Algorithms have been developed to enumerate all the EFMs in medium-size metabolic networks [8–10]. However, despite the development of novel methods using state of the art computational techniques expediting their application in larger networks [11], this family of algorithms fails on GSMNs using standard computers, because of the combinatorial explosion in the number of EFMs [4]. In this light, several methods have been recently proposed to determine a subset of EFMs in GSMNs [12–15].

Computational approaches to metabolic pathways can be classified into two groups: stoichiometric approaches and path-finding approaches [16]. Summarising, the first ones use the stoichiometric data to do calculations during the process. Linear Programming and Null-Space Algorithm [17] are some of the mathematical strategies applied to find pathways, mainly solving the system of linear equations proposed by the stoichiometric matrix. Stoichiometric approaches have the quality of impose biochemically meaningful stoichiometric constraints on the solutions but at the cost of intense floating point calculations. In general, these mathematical artefacts have to be assisted to lead to solutions with certain characteristics and to avoid others. Typically, the connectedness and minimality of solutions have to be treated complementarily during the process.

The second ones translate the network into a directed graph to explore it. Path-finding approaches are considered to constitute some advance with respect to stoichiometry approaches mainly because they rest on the well-known graph theory [18] and allow the use of techniques based on distance metric, revealed as biologically relevant [19]. Shorter pathways are better suited to genetic manipulation [20, 21]. Moreover, shorter pathways can carry higher fluxes [22, 23], which are very interesting in metabolic engineering for compound-based production (e.g. biodiesel). Because of the combinatorial nature of the search, some proposals only find a subset of all feasible pathways, whereas other approaches get the full set of feasible pathways [24]. Another approach is the shortest EFM technique [12, 25], but the problem is that there is no guarantee that the picked start node and the end node are really participating in the shortest EFMs.

The major drawback of path-finding approaches is that the lack of use of stoichiometry during the exploration process can not offer assurance that the solution has biological meaning and meets all the constraints. Below, we show how the solutions that graph-building throws are up quite close to EFMs. Although with respect to that, an extra stage is needed to determine if a pathway found meets the constraints and constitutes an EFM.

There are many algorithms to compute EFMs in metabolic networks being the proposal of [15] considered the fastest one. Another very fast implementation is the Flux Mode Calculator [26], which enables large-scale EFM computation on ordinary desktop computers. Finally, other authors combine both approaches trying to build on the strengths of each and avoid respective drawbacks and computational expenses [27].

Our proposal uses graph methods and has the novelty to guarantee that the picked end nodes are really participating in the shortest EFMs, thanks to choosing them among the input and output reactions. Regarding the performance of our proposal, it does not have the intention of beating in speed former algorithms and implementations but to bring to the foreground some graph-based techniques for pathway extraction.

METHODS

Constraint-based modelling

Constraint-based modelling (CBM) starts with a stoichiometric matrix $S$ whose values are the stoichiometric coefficients for metabolites (rows) on each reaction (columns). Every reaction is characterised by the reaction rate (also known as flux rate) which numerically gives the rate at which the substrate metabolites are converted to the product metabolites. In addition, reactions can be input reactions (i.e., those that only produce metabolites), output reactions (i.e., those that only consume metabolites), and internal reactions (i.e., those that simultaneously consume and produce metabolites).

Metabolism involves fast reactions compared to events of gene regulation. Therefore, it is often assumed that the studied metabolite concentrations and reaction rates are in equilibrated, and thus quasi-steady state in the time-scale of study. A vector of flux rate $\vec{v}$ represents a pathway if it fulfils the steady-state and thermodynamic
conditions. The steady-state condition means that internal metabolites are balanced and concentration remains constant (Equation 3).

The thermodynamic feasibility constraint means that each irreversible reaction only participates with a positive rate (Equation 4) when it is part of the solution.

Let us refer to the equations 3 and 4 as primary conditions.

\[
S \cdot \vec{v} = \vec{0} \quad (3)
\]

\[
v_r \geq 0, \quad \forall r \in R 
\quad (4)
\]

Finally, \( \vec{v} \) represents an EFM if it is non-decomposable, that is, \( \vec{v} \) is not a linear combination with non-negative coefficients of other flux rate vectors corresponding to pathways.

Bipartite graph

A stoichiometric matrix \( S \) lets us build an adjacency matrix that corresponds to the graph \( G = (V, E) \), a non-weighted directed bipartite graph. The set \( V \) of vertices is the union of the sets of reactions and metabolites while the edges \( E \) indicate the existence or non-existence of any stoichiometric relationship between nodes. That is, there is an edge between two vertices that correspond to a reaction \( r \) and a metabolite \( m \) if \( m \) is one of the metabolites involved in the reaction \( r \) and the sense of that edge depends on the sign of the corresponding stoichiometric coefficients.

Matrix 1 corresponds to a toy metabolic network \( S \). Figure 1 shows the bipartite graph \( G \), built from \( S \). It can be observed that the edges go from reactions (rectangular nodes) to metabolites (circular nodes) when they are produced by those reactions. When the edge goes from a metabolite to a reaction, it means that the reaction consumes the metabolite.

In Figure 1 it can also be seen that there are two reversible reactions producing or consuming the same metabolite, given the appropriate biochemical conditions. A lot of reversible enzymatic reactions have been identified in the real world and they are included in cellular metabolic reconstructions. Usually CBM approaches, and specifically those that search EFMs, unfold the reversible reactions in two new irreversible reactions. The purpose of the unfolding is to get a set of reactions with one unique thermodynamic sense. This makes it easier to check the feasibility of one candidate pathway, validating as feasible only those flux rates \( \vec{v} \) that meet the primary conditions. The counterpart of the unfolding is the increase in the number of vertices to be explored. This can be seen in Graph 2 that corresponds to our example after unfolding its reversible reactions (or in its associated matrix 2).

Therefore, without loss of generality, we can consider \( S \) as a matrix with only irreversible reactions, because in any case this can be done by performing an additional step to the original one.

If we are interested in EFMs, we can also restrict ourselves to work with connected graphs, because non-decomposability of EFMs implies implicitly the connexity. Any approach to extract pathways must include strategies to guarantee the connexity of the solutions or some test to certify it.

**Representations of pathways**

A pathway can be expressed by giving the list of the reactions that appear with non-zero flux rate in that pathway. You can also include the implied metabolites in the list or consider that they are implicitly included.

Expression 5 shows an example of this equivalence.

\[
R_2, B, R_9, F, R_3 \equiv R_2, R_9, R_3 
\quad (5)
\]

Methods based on graphs used to employ graphs that contains only metabolite or reactions or a bipartite representation. For our purposes, the right election is a bipartite graph, but in some steps it is more effective to simplify the expression of the pathway by ignoring the metabolites.

The translation from the stoichiometric matrix to a graph is done to explore and visualise the relationships between pathways (or EFMs) and its graph representation. It is easy to see that for each pathway there exists a (full) sub-graph of \( G \), \( H = (V', E') \) with \( V' \subseteq V \) and \( E' \subseteq E \), but the contrary is not always true. A necessary but insufficient condition for a sub-graph to correspond to a pathway is to have all its metabolites balanced, that is, any metabolite in the sub-graph has
to be produced and consumed for some reactions that are also present in the sub-graph. The second graph in Figure 3 shows the sub-graph as correspondent to the expression 5. This sub-graph is clearly balanced. In the Discussion section, we discuss the relevance of balanced graphs while extracting pathways.

Associated with any (full) sub-graph $H$ of $G$, we can construct the sub-matrix $S'$ of $S$ composed only of the rows and columns of $S$ corresponding to the metabolites and reactions appearing in $H$. In Equation 6 we have the sub-matrix $S'$ associated with $H$.

$$S' = \begin{bmatrix} R2 & R3 & R9 \\ R1 & 1 & 0 & -1 \\ R2 & 0 & -1 & 1 \end{bmatrix}$$

Suppose that we have a pathway with associated sub-graph $H$ and sub-matrix $S'$. Let us call $R'$ the subset of reactions that appear with the non-zero rate in the pathway. Equations 3 and 4 can be reformulated more restrictively as shown in equations 7 and 8.

$$S' \cdot \vec{v} = \vec{0}$$

$$v_r > 0, \ \forall r \in R'$$

Equations 7 and 8 are different for each pathway because $S'$ and $R'$ are different for any pathway. If we want to check the accomplishment of the primary conditions, we can use this new simplified condition to reduce the calculations.

As it has been said, an EFM is a non-zero flux vector $\vec{\varphi}$, which is a solution for equation 7 and is also minimal. Minimality is the quality of a pathway to be non-decomposable into two or more different pathways.
Minimality can also be described from the point of view of graphs. Given a pathway represented by a sub-graph \( G_1 = (V_1, E_1) \), \( G_1 \) is an EFM if there exists no other pathway represented by a sub-graph \( G_2 = (V_2, E_2) \) with \( G_2 \neq G_1 \) such that \( \emptyset \neq V_2 \subseteq V_1 \). Under the appropriate conditions [28], it can be assured that for an elementary mode, there exists a bijection between the list of active reactions and the flux rate vector.

Thanks to these considerations, the extraction of pathways using graphs can be done using the most suitable mathematical artefact at any time.

### Linear programming approaches

Linear programming is the most used technique for the state of the art EFM extraction methods. As linear programming is an optimisation tool, pathways search becomes in an optimisation problem restricted by the primary conditions. The EFM are more suitable objects of search using linear programming than pathways because the non-decomposability condition suggests clearly the objective function.

The posed linear program can be solved using the Simplex Algorithm. There are many references to detailed explanations of how to convert a stoichiometric matrix into a linear program [13] given the primary constraints. The direct translation of a stoichiometric matrix into a linear program defines a clean linear program (Equation 9).

\[
\begin{align*}
\text{Minimize} & \quad \sum_{i=1}^{n} v_i \\
\text{subject to} & \quad S \cdot \vec{v} = \vec{0} \\
& \quad v_i \geq 0 \quad \forall r_i \in R
\end{align*}
\]

A linear program has to be additionally constrained for obtaining different solutions.

Observe that the posed linear program does not guarantee that the solution is minimal in terms of amount of reactions because the function to optimise does not indicate the amount of reactions but the addition of the flux rates. Many linear programming based approaches complement linear programming with graph-oriented strategies in some manner.

### DISCUSSION

#### Balanced graphs

Recall that a pathway is a set of reactions with non-zero positive flux rate that meets the rest of primary conditions. A pathway is also an elementary mode if it is non decomposable, that is, it can not itself be expressed as an admissible convex combination of other elementary modes.

As we have said, we have a graph \( G \) associated to the stoichiometric matrix \( S \). Given a (full) sub-graph \( H \) of \( G \), we will call it a "balanced graph" if it fulfills the following two conditions:

1. All its metabolites can be produced and consumed by reactions included in \( H \)
2. If \( r \) is a reaction that corresponds to a vertex in \( H \), then all the metabolites involved in that reaction are also included in \( H \).

If \( m \) is a metabolite corresponding to a vertex in \( F \), the condition for the sub-graph \( H \) to be balanced in \( m \) can be expressed by the fact that \( H \) must contain at least one of the reactions that can produce \( m \) and another one that can consume it.

Observe that being balanced is a necessary but not sufficient condition to meet the steady-state condition. This is so because the stoichiometry does not play a role building balanced graphs, so a resulting balanced graph could be steady-state or not. Determining if the obtained solutions are pathways and, therefore, EFM or not is not possible ignoring the stoichiometry.

Figure 3 shows three balanced sub-graphs, with their respective vertices on gray. For example, the values \( x_2 = x_3 = x_9 = 1 \), \( x_1 = 0 \) correspond to the second balanced sub-graph. It is easy to check that not all these sub-graphs are EFM. The first graph in that figure contains another smaller pathway inside, the one constituted by the second graph. As the second graph is an EFM the first one can not be EFM as well. It is easy to prove, using stoichiometry, that the third graph is not steady-state.

Once we get a balanced candidate, the rank test or the resolution of the system of equations must be done to know if it is an EFM. These tests are expensive in terms of computations and the amount of tests to be done in relation to the size of the problem must be estimated. We have run over this sample network an exhaustive search of balanced graphs by testing all the possible subsets of reactions increasing the size on each step. That is not a part of our method but it lets us know the theoretical residual use of the rank test to see whether the balanced graphs were obtained in the right order. The results are listed in Table 1.

The balanced sub-graphs form a subset of all the possible combination of reactions. In Table 1, we show the results of our search (in the column corresponding to the number of possible combinations we have only taken into account those that do not contain a reversible reaction and its opposite one). The third column is the amount of possible balanced graphs classified by length.

Once we get the list of balanced graphs, if we certified that one of them is a pathway, in the next length balanced graph search we can avoid those combinations that contain this previously found pathway. The generalisation of this strategy can save the exploration of an increasing amount of sub-graphs as shown in the column Saved. The fifth column represents the amount of found balanced graphs once the saved combinations are discarded.

The rank test must only be done to determine if a balanced graph is an EFM. Note that we also save rank tests and, therefore, floating-point calculations. It is remarkable that if the search starts from the smallest to the biggest combinations and we saved those that contain a previous EFM, the rest of the strategy guarantees that all the pathways founds are also EFM. This strategy is similar to others based on the rising size of the elementary modes [12].
The shortest path technique to find EFMs

We propose a new CBM approach based on path-finding techniques. We have seen that finding balanced graphs seem an interesting way to achieve a list of EFMs so we propose a novel approach to generate balanced sub-graphs from the full network starting from the smallest to the biggest.

Our method consists of two phases: exploration and characterisation. The exploration phase consists of three stages for traversing the graph and finding balanced sub-graphs and the characterisation phase identifies which ones correspond to EFMs.

In the first stage, we use the pathway distance metric approach (that is, the amount of reactions of participants at the pathway) and take advantage of the fact that it should be biologically meaningful [19]. Therefore, the quest starts the graph exploration by building an axis between an input reaction (source node) and an output reaction (target node) of the network, applying Dijkstra’s shortest path algorithm [29]. The choice of the path end nodes (source, target or both) could be influenced by the biological problem we are dealing with. At the end of this stage, an axis has been built using the Dijkstra algorithm that traverses the graph through metabolites and reactions from an input reaction to an output one using the shortest distance.

Of course, these paths are usually not balanced. The first problem we can find is that some of the reactions included in the axis can need metabolites that have not been included yet. We name this kind of metabolites as orphan metabolites.

In the second stage, the exploration goes back from the target to the source (bottom-up approach) to solve the orphan metabolites problem. This process traverses the inverted graph and it is done in a recursive way. The inverted graph is obtained by transposing the adjacency matrix used for the original graph G. During this stage, sets of reactions are proposed for each orphan metabolite. The objective is that all of the metabolites needed for the axis are at least produced by a reaction present in the partial solution and, therefore, in the final pathway.

The third stage consists of the simulation of all the reactions that should occur due to the presence of the required metabolites produced by other reactions. The third stage ends when the end nodes are connected by a complete graph of reactions without orphans or non-consumed internal metabolites.

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

An additional step, the characterisation phase, has to be done to evaluate if the balanced graph can be considered a pathway. The obtained sub-graphs seem minimal because none of the elements can be eliminated without sacrificing consistency. Moreover, sub-graphs fulfil the necessary conditions to have the steady-state balance. However, it can not be assured the steady-state condition because the stoichiometry is not playing a role during the run of our approach. Without stoichiometry, and depending on the network structure, a lot of false positives can be got but also some real positives. In terms of feasibility, the candidates to the pathway are built fulfilling the necessary conditions to be feasible, that is, respecting the positive direction of every reaction, but the feasibility constraint is only met based on the steady-state consistency.

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

<table>
<thead>
<tr>
<th>Length</th>
<th>Combinations</th>
<th>Bal1</th>
<th>Saved</th>
<th>Bal2</th>
<th>EFMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>406</td>
<td>2</td>
<td>16</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>560</td>
<td>4</td>
<td>65</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>532</td>
<td>4</td>
<td>127</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>344</td>
<td>4</td>
<td>145</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>145</td>
<td>11</td>
<td>96</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>36</td>
<td>12</td>
<td>32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1791</td>
<td>43</td>
<td>485</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 1: Balanced graphs and EFMs analysis in the toy network. The network has 10 reactions (2 reversible reactions) and 6 metabolites. Bal1 is the total amount of balanced graphs and Bal2 is the amount of balanced graphs obtained after the savings.

In the Results section, we study the proportion between balanced graphs and candidates to be a pathway in real networks.

Algorithm 1: EFM search using Dijkstra’s shortest algorithm

Data: Matrix S, set of reactions R
Result: Set EFMs of EFMs

\[ \text{EFMs} \leftarrow \emptyset; \]
\[ A \leftarrow \text{FromStoiToAdjacencyMatrix}(S); \]
for \( i \in \text{inputs}(R) \) do
  for \( o \in \text{outputs}(R) \) do
    \[ \text{DijkstrasList} \leftarrow f_{\text{Dijkstra}}(A, i, o, \text{knockouts} = \text{true}); \]
    for \( d \in \text{DijkstrasList} \) do
      \[ \text{OrphanFreeList} \leftarrow f_{\text{CompleteOrphanProducers}}(A, d); \]
      for \( f \in \text{OrphanFreeList} \) do
        \[ \text{BalancedList} \leftarrow f_{\text{BalanceGraph}}(A, f); \]
        for \( b \in \text{BalancedList} \) do
          if \( f_{\text{isEFM}}(S, b) \) then
            \[ \text{EFMs} \leftarrow \text{EFMs} + [b]; \]
        end
      end
    end
  end
end
Cycles and repetitions during the exploration

The second and third stages have a lot of similarities. In fact, they are identical but with slightly different objectives and run in inverted graphs. Both stages also have the same problems like cycle formation and the occurrence of repeated partial solutions.

The degree of connections for any node in the graph and the evolutionary redundancy inside the metabolic network of a cell produce cycles along the graph that should be avoided. The main reason to avoid cycles is so as not to visit the same part of the graph more than once for each iteration. In term of programming, the cycles can not be produced if the representation chosen to do for each iteration. In term of programming, the cycles can not be produced if the representation chosen to do so are avoided. The main reason to avoid cycles is so as not to visit the same part of the graph more than once for each iteration. In term of programming, the cycles can not be produced if the representation chosen to do for each iteration.

Avoiding repeated sub-graphs is mandatory for any approach, mainly because getting repeated solutions is a waste of time and resources. This occurs because dealing with a combinatorial problem in many cycles can constantly produce repetitions in any exploration strategy. This repetition occurs when we get a previously obtained sub-graph although the start point is different. Repeated solutions can be prevented by keeping a list of the obtained sub-graphs and considering them as sets. At this point, it must be recalled that the order between reactions and metabolites inside a pathway is not relevant.

There is a kind of possible pathway that accomplishes all the conditions but without input or output reactions. These kinds of pathways are based on internal cycles to meet the steady-state condition. Our method does not consider any of those pathways as a feasible solution. However, it is possible to do a slight modification to include them by allowing the two initial vertices to be internal (even equal, in this case the shortest path between them is of zero length).

Characterisation phase

After each iteration of the exploration phase we get a balanced sub-graph but, as we have pointed out, we still need to check if Equations 7 and 8 have a non-trivial solution. This checking can be done using, for example, linear programming. Here we propose a different approach that uses properties of pathways and EFM to speed up the process.

If the sub-graph is an EFM, then it corresponds to a unique vector of flux rate \( \bar{v} \). This is a consequence of compliance with Equations 7 and 8 and the minimality condition. The uniqueness of the solution requires that the rank of \( S' \) accomplishes the equation 10, where \( |R'| \) is the cardinality of \( R' \).

\[
\text{rank}(S') = |R'| - 1 \tag{10}
\]

If this equality occurs, the process of calculating the rank of \( S' \) can also be used to calculate the unique solution (up to scalar multiples) and to check if all its entries have the same sign, and so inform us if we have found a pathway or not.

Many candidates to be checked with the test 10 can be previously filtered using some tests that are less computationally expensive. It also allows us to reduce as much as possible the floating-point calculations. We want to propose the following two tests for filtering.

- In relation to the rank test, it must be observed that the rank of \( S' \) depends on the number of rows as well. Some candidate sub-graphs do not have enough metabolites (linear equations) such that \( S' \) can accomplish in 10. Therefore, the first test proposed is expressed in equation 11.

\[
|C'| \geq |R'| - 1 \tag{11}
\]

- On the other hand, it must be observed that a necessary but not sufficient condition for a pathway to be an EFM is that its associated sub-graph has all its nodes connected (clearly, only a balanced sub-graph with all its nodes connected can be minimal). Thus, minimality requires the undirected graph associated to our sub-graph to be connected.

As explained before, after applying the above two filters to our obtained balanced sub-graphs, we finally characterise them as EFM or not by calculating the rank of the associated sub-matrix \( S' \). We emphasise again that we only use floating-point calculations in this last step.

RESULTS

Case study: a toy network

Let us consider as an example the aforementioned network represented by the stoichiometric matrix \( S \) shown in the matrix 1. Note that the reactions \( R_2 \) and \( R_8 \) are reversible. For the rest of the process, these reactions need to be unfolded in \( R_2 \), \( -R_2 \), \( R_8 \) and \( -R_8 \) automatically. Unfolded reactions must be included in the matrix with individual columns in the new unfolded stoichiometric matrix (matrix 2). Therefore, all the reactions are not now irreversible.

Based on the unfolded \( S \), we build the graph \( G = (V, E) \) represented graphically in Figure 2. Vertices \( V \) are both metabolites and reactions, and edges \( E \) are the incidence arcs following the direction of the reactions.

Our technique starts in the exploration phase, which has three stages. In its first stage, the Dijkstra's shortest algorithm is run to build an axis for the foreseeable pathway. The shortest path for this example is shown in Figure 4 with the participating nodes in gray. This shortest path is a route between \( R_1 \) as input extreme of our metabolic network and \( R_4 \) as output extreme. Obviously every pair of extreme points can be considered. Many times, the obtained paths in this stage could have the orphan metabolite problem. In the example we are considering, \( R_{10} \) needs that the metabolite \( D \) (dotted in Figure 4) is also a part of the pathway.

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

The third stage is responsible for assuring that every metabolite produced by the pathway has consumer reactions, that is, it should be consumed inside the pathway. In our example, \( R10 \) produces the metabolite \( F \) but there is no consumer reaction to it. This stage looks for what reactions could occur with the metabolite \( F \) in order to be consumed. In this example, there is only one possibility (reaction \( R3 \)), and it will be incorporated to the pathway.

After these three stages we have the balanced graph \{R1, R6, R10, R4, R7, R3\} with the metabolites \{A, C, D, E, F\} involved in it. It is the candidate to be a pathway. The reactions have been shown in the same order they were obtained and the corresponding sub-graph is shown in Figure 5.

Once we get the balanced graph, it is mandatory to evaluate if it is an EFM. This is done with the bundle of tests of the characterisation phase. The first test a candidate sub-graph must pass is a connectivity test. Our approach guarantees that all nodes are connected because of the process of construction of the sub-graph, so it can be overlooked. Other approaches not based on graphs should do this test.

The second test is the rank test. Before building the matrix and calculating the rank, we can try the lighter rank test proposed in Equation 11. Being \(|R'| = 6\) and \(|C'| = 5\), the test is positive, so the candidate passes the test. This test is done by counting the elements of the set \( V' \). Each time the result of the test is negative, we get bad news, but we saved time and programming resources using this light version of the rank test.

Matrix 15 shows the extended system of equations.
The flux rate of $R_{10}$ is fixed to $1$. 

\[
\begin{pmatrix}
1 & 0 & 0 & -1 & -1 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 \\
0 & 0 & 0 & 0 & 1 & -1 \\
0 & 0 & -1 & 0 & 0 & 1 \\
0 & -1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 \\
\end{pmatrix} \cdot \begin{pmatrix} v'_1 \\ \vdots \\ v'_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}
\]

(15)

Solving this system of equations, if all the rates are positive, it can be concluded that the balanced sub-graph is at least a pathway. In this case, all of them are strictly positive and we know that $\text{rank}(S') = 6$. Therefore, the unique solution to system (15) is $\vec{v} = (2, 1, 1, 1, 1, 1)$, and it can be concluded that it is an EFM. The flux rate $\vec{v}$ with $|R|$ elements is $(2, 0, 1, 1, 0, 1, 0, 1)$. 

To get the full list of EFMs (Table 2), 13 Dijkstra’s algorithm runs that produce 14 orphans-free graphs and finally 8 balanced graphs have been needed. Those 8 candidates passed the rank test and, therefore, they are EFMs.

Table 3: Balanced graphs and EFMs analysis in the TCA cycle and some adjacent reactions. The network has 24 reactions (9 reversible reactions) and 16 metabolites.

The results are listed in Table 3. It must be observed that the saved combinations make up almost 30% of the combinations and all the obtained balanced graphs are EFMs. 

Table 3: Balanced graphs and EFMs analysis in the TCA cycle and some adjacent reactions. The network has 24 reactions (9 reversible reactions) and 16 metabolites.

Table 4: The EFMs obtained from the TCA cycle reconstruction. The first column shows the order in which the k-shortest approach finds them [12].

Case study: TCA cycle

To study a real network, we use the well-known real metabolic network that contains the tricarboxylic acid (TCA) cycle as a benchmark. As metabolite $PG$ is considered as an external reaction, $Eno$ is an input reaction.

For visualisation purposes, Figure 6 shows the bipartite graph for this metabolic model. Bipartite form and the split of reversible reactions prevent the seeing of the expected cycled figure.

Identically, as we have done with the sample network, before running our method, we have done the exhaustive search of balanced graphs by testing all the possible subsets of reactions increasing the size on each step. The results are listed in Table 3. It must be observed that the saved combinations make up almost 30% of the combinations and all the obtained balanced graphs are EFMs.

Table 3 shows the amount of the possible sub-graphs that are balanced, pathway and, finally, EFMs. To build this table it has purged the model quitting the blocked reactions proposed by [30]. Reaction $P_{ps}$ is one of those blocked reactions that do not appear except in one EFM with its respective reversible pair. Our method also sacrifices this EFM from the list of solutions.

Although this network is considered small, it is a real example. The most important conclusion is that the results suggest that filtering the balanced graphs from all possible solutions is a good approach to find EFMs if it can be done efficiently. 

Table 4 shows the 11 EFMs obtained out of the 16 possible. To get these results, it was necessary to run 49 Dijkstra’s (4 clean plus 45 knock-outs), 128 partial orphan-free runs, 139 balanced graph operations (where only 129 are unique) and 129 rank tests.

As it can be seen, our approach missed five of the possible EFMs. The reason is that our method always starts building the shortest path from an input reaction to an output reaction. Therefore, all those EFMs that only have internal reactions (i.e., forming internal loops) can not be found because they do not cross the network from an input to an output reaction. Notice that reactions in any model can be classified
as input reactions (i.e., those that only produce metabolites), output reactions (i.e., those that only consume metabolites), and internal reactions (i.e., those that simultaneously consume and produce metabolites).

This way, our approach produces all possible balanced graphs between a pair of end nodes (input and output nodes). Then, all EFMs can be inferred from those balanced graphs.

Finally, our approach has been checked against larger networks such as the *E. coli core model* [31] with 95 reactions and 72 metabolites. As an example of provisional results that we have obtained, a few minutes of execution was necessary to obtain 512 balanced graphs from the first 100 shortest paths from inputs to outputs. Among them, we get 32 EFMs. Those promising results encourage us to further research on our approach.

**CONCLUSIONS AND FUTURE WORK**

In this paper we propose a new approach to obtain EFMs based on graph theory and the shortest path between end nodes. Our novel approach has two phases, the exploration and the characterisation phase. The exploration phase relies on a three-stage method, namely Dijkstra's shortest path algorithm, orphan metabolites problem solving and the construction of the balanced graph. The characterisation phase is based on heuristic and mathematical artefacts.
Our method finds all the pathways in the metabolic network that crosses from input to output and it is able to prioritize the pathway search, accounting the biological mean pursued. Along the paper, we showed how it works using a well-known case study.

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

As for future work, the characterisation phase to filter, which is EFM from the set of obtained balanced sub-graphs, is still immature and more work should be done in relation to it, such as developing some heuristics using artificial intelligence techniques. The restriction of the approach derived from the fact that Dijkstra’s algorithm is designed to work only between two end nodes must also be improved. Another direction of future work is the parallelisation of all of the stages of our method using High Performance Computing commodity architectures.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTIONS

All authors participated in the design of the method and the preparation of the manuscript. José F. Hidalgo has implemented the method and carried out the tests.

CONFICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution figure files are available at Genomics and Computational Biology online.

ABBREVIATIONS

CBM: Constraint-Based Modelling
EFM: Elementary Flux Mode
GSMN: Genome-Scale Metabolic Network
TCA: Tricarboxylic Acid

REFERENCES


NOTES