CALCULATING THE *K*-SHORTEST ELEMENTARY FLUX MODES IN METABOLIC NETWORKS

L. F. de Figueiredo[1,2], A. Podhorski[3], A. Rubio[3], J. E. Beasley[4], S. Schuster[1] and F. J. Planes[3]
[1]Friedrich-Schiller-University Jena, Jena, Germany; [2] PhD Program in Computational Biology, Instituto Gulbenkian de Ciência, Oeiras, Portugal; [3]CEIT and TECNUN (University of Navarra), San Sebastián, Spain; [4]Brunel University, Uxbridge, United Kingdom

Corresponding Author: F. J. Planes, Paseo de Manuel de Lardizabal, 12, San Sebastian, TECNUN (University of Navarra), Spain Email: fplanes@tecnun.es

Abstract. In the post-genomic era elementary flux modes represent a key concept to analyze metabolic networks from a pathway-oriented perspective. In spite of considerable work in this field, the computation of the full set of elementary flux modes in large-sized metabolic networks still constitutes a challenging issue. In this paper we illustrate that the full set of elementary flux modes can be enumerated via integer linear programming. Technically, our approach produces elementary flux modes in increasing order of number of reactions by sequentially solving an optimization problem. Though our procedure is not particularly efficient in computing the full set of elementary flux modes for large-sized metabolic networks, it is very flexible. It can be applied to calculate the elementary flux modes satisfying a given criteria without having to calculate all the solutions first, in contrast to what is typically done by current methods. This greatly speeds up computation by focusing only on that part of the solution space that is of interest. Since computation time increases as the length increases, it is promising to start with the shortest, second shortest, etc., (overall called *K*-shortest) elementary modes. Detection of these is indeed of interest for several biological applications. Experimentally, it is difficult to insert and express a large number of heterologous genes and shorter pathways can carry higher fluxes. To illustrate the scope of our approach, we analyse a subset of elementary flux modes that produce L-lysine in *Escherichia coli*. Our analysis shows that our mathematical approach can be an effective tool to explore the capabilities of metabolic networks at the genome scale.

1 Introduction

Elementary flux modes (EFMs) represent a key concept for the analysis of metabolic networks in the post-genomic era. An EFM is defined as being a minimal subset of enzymes that operates in steady-state with all irreversible reactions used in the appropriate direction [17]. The relevance of EFMs in different fields has been broadly described in the literature [4], [5], [7], [13], [16].

However, EFMs analysis suffers from an important drawback. As the size of the metabolic network increases, the number of EFMs grows exponentially. More than two million EFMs have been reported for the metabolic network describing the central metabolism in *Escherichia coli*, which contains approximately 110 reactions [6], [25]. A number of attempts have been made to cope with the huge number of EFMs stemming even from relatively small networks. Stochastic optimization methods were employed to find a classification of external and internal compounds that minimised the number of EFMs [3]. Such an approach, although allowing a reduction in the number of EFMs, did not necessarily lead to a global minimum, as admitted by the authors. A similar approach was taken by Teusink *et al.*, where a reduction of the EFMs was obtained by carefully adjusting the status of metabolites between internal and external [22]. Partitioning of the whole network into smaller subnetworks was considered by Schuster et al., who created a tool called SEPARATOR that automatically performed the classification of internal vs. external metabolites and the corresponding decomposition of the network based on the connectivity of metabolites [18]. Recent approaches use binary tree search [6], [21] that allows pruning of children nodes that cannot contain a valid solution. They reduce memory requirements, decrease computation time and allow for parallelisation. The overall improvements allowed EFMs to be computed 4–5 times faster than previous methods and enabled computation of *E. coli* central metabolism in a reasonable time.

Although the number of EFMs that even small networks generate might be overwhelming, it is worthwhile to note that the number of EFMs is an intrinsic characteristic of the network, which as such can neither be avoided nor overcome. Rather, ways need to be found to cope with such complexity when inferring useful knowledge. One approach could be to focus only on those EFMs that are related to the particular problem of interest. This however is impeded by the fact that existing algorithms compute all EFMs before any further analysis is performed.

In this paper we present a novel approach that enables us to compute EFMs in increasing order of number of reactions by sequentially solving an optimization problem. Although our procedure is not particularly efficient to compute the full set of EFMs in large-sized metabolic networks, it allows the analysis of a selected part of the metabolic network without first having to compute all EFMs. This is performed by employing integer linear programming techniques, which have already proved useful for metabolic pathway recovery [2]. By imposing constraints on the EFMs to be computed only those that are of interest will be determined without having to enumerate all EFMs even if the metabolic network is large, as in genome-scale models.

2 Methods

Assume we have a metabolic network that comprises R reactions and C compounds. Note here that reversible reactions contribute two different reactions to the metabolic network. For this reason we can regard all fluxes as taking positive values. Let s_{cr} be the stoichiometric coefficient associated with compound c (c=1,…,C) in reaction r (r=1,…,R). As usual in the literature [14], [17] , input compounds have a negative stoichiometric coefficient, whilst output compounds have a positive stoichiometric coefficient. The matrix that contains these coefficients is typically named Stoichiometric Matrix.

Suppose we are concerned with finding the K-shortest EFMs in the metabolic network. We mean here by 1-shortest EFM, the EFM containing the minimum number of reactions; 2-shortest elementary flux mode the EFM containing the second minimum number of reactions; etc. Note here that we may have multiple EFMs containing the same minimum number of reactions.

2.1 Variables

Firstly, we need to decide the reactions involved in a particular EFM. Our model represents this situation by a zero-one (binary integer) variable, namely z_r=1 if reaction r (r=1,…,R) is active in the EFM, 0 otherwise. In addition, each reaction has also an associated non-negative (integer) flux t_r. Clearly, the reaction flux t_r=0 if the reaction is not active, i.e. z_r=0.

2.2 Constraints

We need constraints relating the reaction variables: z_r and t_r. Equation (1) ensures that no flux traverses a reaction r if z_r=0. Equation (2) guarantees that t_r is non-zero if z_r=1. Note here that in the case a reaction r is active (z_r=*1*), its associated (integer) flux value t_r can take any value from the interval [1, M], M being a large constant value. Here we have scaled fluxes so that the maximum flux is M and the minimum (non-zero) flux is 1. This does not constitute an issue if we consider M sufficiently large. Computation time is directly related to this parameter. Clearly, computation time increases when M increases. In order to have a reasonable computation time, M was fixed to 10. We think that this value is appropriate, since, as we are calculating elementary (minimal) flux modes, flux values are not expected to be too high.

$$t_r \leq Mz_r \qquad r = 1,\ldots,R \tag{1}$$

$$z_r \leq t_r \qquad r = 1,\ldots,R \tag{2}$$

Defining the set B={(α,β)| reaction α and reaction β are the reverse of each other, α<β}, equation (3) ensures that a reaction and its reverse do not appear in an EFM.

$$z_\alpha + z_\beta \leq 1 \qquad \forall(\alpha,\beta) \in B \tag{3}$$

Equation (4) applies the steady state assumption to the set of internal compounds, I. This is a critical condition for EFMs. As opposed to internal compounds, external compounds are excluded from being balanced, because they are exchange metabolites between the outside and the system under study or they belong to metabolic pools whose concentration is assumed constant. They typically represent consumed substrates, excreted products and cofactors. We denote E the set of external compounds.

$$\sum_{r=1}^{R} s_{cr} t_r = 0 \qquad \forall c \in I \tag{4}$$

In order to avoid the trivial solution (z_r=t_r=0, r=1,..., R), we need to add Equation (5) to our formulation.

$$\sum_{r=1}^{R} z_r \geq 1 \tag{5}$$

2.3 Objective function

EFMs by definition cannot be decomposed into smaller entities without violating the steady state assumption, Equation (4). This is typically referred as to the non-decomposability condition [17]. In essence, this condition implies that no subset of reactions of an EFM can perform in steady state. For this reason EFMs are defined as being minimal set of reactions in steady state [16]. This constraint has not been explicitly considered above. However, by minimizing the number of active reactions involved in the solution pathway, as is done in Equation (6), we can be sure that the non-decomposability condition is satisfied. Clearly, the flux mode involving the minimum number of reactions will be elementary.

$$minimize \sum_{r=1}^{R} z_r \tag{6}$$

2.4 EFM elimination constraints

The mathematical optimization model given above (optimize (6) subject to equations (1)-(5)), once solved, allow us to obtain the shortest EFM. However, we are concerned here with calculating the *K*-shortest EFMs. In order to find the *K*-shortest EFM, we need to add further constraints to eliminate the (*K*-1)-shortest EFMs from the set of solutions. To illustrate this, suppose we are interested in finding the 2-shortest EFM. Let Z_r^1 be the zero-one binary (integer) solution associated with the shortest elementary flux mode. Note here that Z_r^1 takes value 1 when reaction r is active in the shortest EFM, 0 otherwise. We need to eliminate the shortest EFM from the set of solutions. To do this we add the following constraint to our previous formulation:

$$\sum_{r=1}^{R} Z_r^1 z_r \leq \left(\sum_{r=1}^{R} Z_r^1 \right) - 1 \tag{7}$$

This constraint ensures that, once we solve our model after equation (7) has been added to our formulation, the new solution found does not contain the shortest EFM. This also guarantees that the shortest EFM can never occur as a part of any other flux mode. In essence, we remove the shortest elementary flux mode from being a subset of any future solution.

In the general case, in order to find the *K*-shortest EFM, we need to include *k*-1 EFM elimination constraints related to the first (*K*-1) shortest EFMs, as shown in equation (8). Z_r^k represents the binary solution for the *k*-shortest EFM.

$$\sum_{r=1}^{R} Z_r^k z_r \leq \left(\sum_{r=1}^{R} Z_r^k \right) - 1 \qquad k = 1, \ldots, K-1 \tag{8}$$

Note here that the *K*-shortest EFMs described above are also elementary. For illustration, suppose that the *K*-shortest EFM (once solved) is not elementary, i.e. it contains a subset of reactions that would satisfy constraints (1)-(5), (8). The question is therefore, since we are constructing EFMs in increasing order of the number of reactions they contain, why did we not find this EFM before? Logically, as it involves fewer reactions, we must have encountered it before. But if we had encountered it before then we would have added a constraint, as described in equation (8), preventing it from ever appearing as a subset in future EFMs. So it cannot in that case ever be found as part of the *K*-shortest EFM. In other words, we have a contradiction here. For this reason, due to our construction procedure, every EFM we find must be elementary.

2.5 Biological constraints

In theory our procedure can be applied to enumerate all EFMs, namely by constructing them one by one. This is not particularly efficient for large-sized metabolic networks. However, the main advantage of our mathematical optimization model is that, by adding new constraints, special subsets of EFMs (of particular biomedical/biotechnological interest) can be found. This fact makes our algorithm more flexible than previous approaches [6], [16], [21]. We present below some of these constraints that can be easily added to our formulation.

Firstly, we may need to find EFMs that utilize a proper growth medium. We mean here by medium, *U*, the set of external compounds that can be taken up from the outside of the boundaries of the network. These compounds are assumed to have a high concentration outside the system, so that they can be easily taken up into the metabolic network. This uptake is typically carried out via transport reactions. In the case an external compound *c* is not included in the medium set, we need to deactivate the transport reactions associated with this compound. Equation (9) describes how this constraint is incorporated into our model.

$$t_r = 0 \;\; if \;\; \exists c \in [1, \ldots, C][c \notin I, c \notin U \; with \, s_{cr} \leq -1, r = 1, \ldots, R] \tag{9}$$

We may need to find the *K*-shortest EFMs that produce a particular external compound, μ. In order to do this, we need to add the following constraint:

$$\sum_{r=1}^{R} s_{\mu r} t_r \geq 1 \tag{10}$$

This can be easily reformulated if we want an external compound μ to be used as substrate, as observed in Equation (11).

$$\sum_{r=1}^{R} s_{\mu r} t_r \leq -1 \tag{11}$$

Note here that Equation (5) can be dropped from the formulation if we include Equation (10) or Equation (11).

We may also be interested in excluding from balancing certain cofactors or metabolites present in high concentration, such as ATP, NADH, H_2O, CO_2 etc. This is just done by removing Equation (4) for these biochemical compounds, including them in the set of external metabolites. In such case, cofactors can be freely consumed and produced.

2.6 Overview

Our mathematical optimization model given above for computing *K*-shortest EFMs (optimise (6) subject to (1)-(5) plus EFMs elimination constraints (8) and perhaps constraints (9)-(11)) is a linear integer program. Algorithmically such programs are solved by linear programming based tree search. Modern software packages to perform this task, such as ILOG CPLEX®, which we used, are well developed and highly sophisticated.

3 Results

3.1 Introduction

In this section we present results of application of our mathematical optimisation model to compute the *K*-shortest EFMs in two different metabolic networks.

Firstly, we considered the metabolic network of part of the monosaccharide metabolism [16], which comprises 19 reactions and 26 biochemical compounds. This small network includes the combined glycolysis/pentose phosphate pathway system. This network was chosen to show that our approach indeed produces EFMs.

Secondly, we examined the performance of our procedure in a large metabolic network, specifically a genome-scale metabolic network of *E. coli* K-12 (iJR904 GSM/GPR) [10]. Our approach was applied to analyse L-lysine biosynthesis pathways.

3.2 Network of Monosaccharide metabolism

Figure 1 shows the network of part of the monosaccharide metabolism presented in Schuster *et al.*, [16]. Aside from typical cofactors, ribose-5-phosphate (R5Pex), pyruvate (Pyr) and glucose-6-phosphate (G6P) were considered as external compounds, represented as grey nodes. Dashed nodes were duplicated for better visualization, e.g. glyceraldehyde-3-phosphate (GAP). Note that the same metabolite and reaction abbreviations adopted here are the same as in the original model [16].

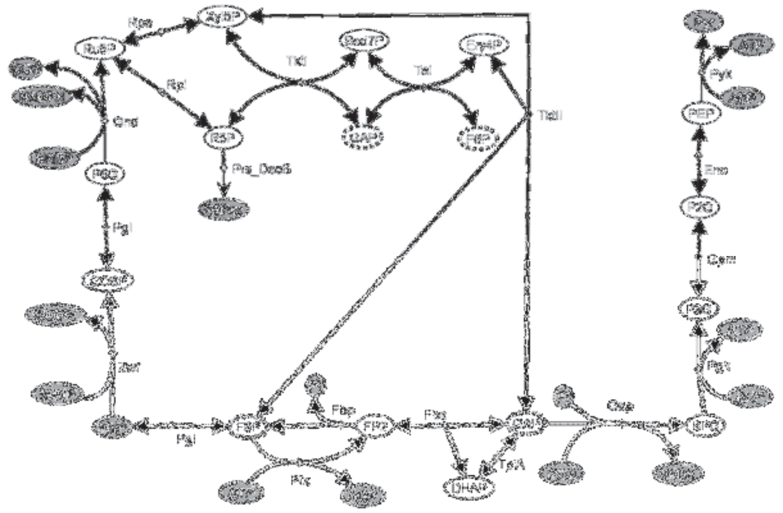

Figure 1: Reaction scheme of monosaccharide metabolism. Reversible reactions are indicated by bidirectional arrows. For the meaning of abbreviations, see [16].

As described above, optimising (6) subject to equations (1)-(5) plus EFM elimination constraints (8), produces EFMs in according to increasing number of reactions. Since the size of the network under consideration is small, our approach was applied to enumerate all EFMs.

Our optimization approach yielded exactly the same 7 EFMs that were reported in Schuster *et al.* [16], where METATOOL [8] had been used to compute the EFMs. Figure 2(a) shows the shortest EFM obtained from Figure

1. The shortest EFM involves just 2 reactions. Figure 2(b) shows the 2-shortest EFM, which contains 5 reactions. The 3-shortest, 4-shortest, 5-shortest and 6-shortest EFMs, though not shown here, contain 9, 10, 12 and 14 reactions, respectively. Figure 2(c) shows the 7-shortest EFM, which contains 16 reactions. This analysis shows that our approach indeed produces non-decomposable pathways at steady-state.

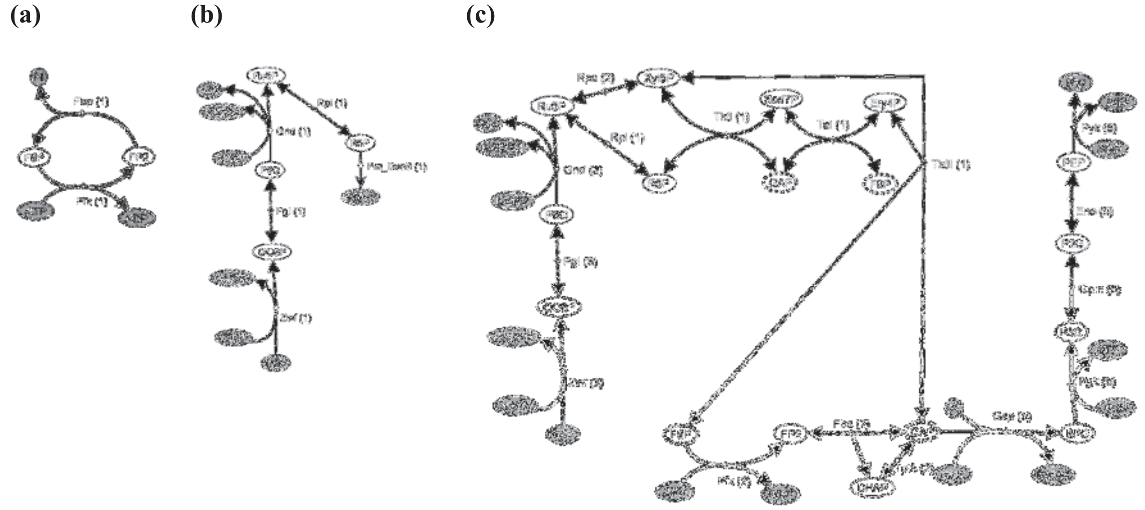

Figure 2: Graphical representation of the shortest (a), 2-shortest (b) and 7-shortest EFMs (c).

3.3 Genome-scale metabolic network of *E. coli*

We used the metabolic network of *E. coli* K-12 (iJR904 GSM/GPR) [10]. This network distinguishes between two different compartments, namely the cytosol and extracellular compartments. The biochemical compounds present in the cytosol represent the set of internal compounds, whilst compounds included in the extracellular compartment represent the set of external compounds.

In order to illustrate the potential of our *K*-shortest EFMs method in large metabolic networks, we studied a biochemical problem related to the synthesis of L-lysine. Note here that Equation (10) was imposed for L-lysine, i.e. we add to our basic formulation (optimise (6) subject to (1)-(5) plus EFMs elimination constraints (8)) a constraint ensuring that L-lysine is released to the extracellular compartment at a positive rate. The addition of this constraint does not imply that the non-decomposability condition is lost. Rather it implies that we are only concerned with those EFMs producing L-lysine. The non-decomposability condition might be lost in the case where more than one constraint of this type is added to our basic formulation, e.g. EFMs that consume glucose and produce L-lysine. Further analysis would be required in such cases. However, this is beyond the scope of this paper and further research is needed, as noted in the Discussion section.

L-lysine is one of the essential amino acids in humans and is also used as supplement in animal feeds. The industrial production of this amino acid is very important and has a long history in the biotechnology field [23], [26].

Existing studies mainly focus on the regulation of L-lysine synthesis because, from the metabolic perspective, amino acids biosynthesis pathways are generally linear after a key metabolic precursor, such as pyruvate . Therefore, the regulation of such pathway plays an important role in microbial optimization [23].

Nevertheless, the improvement of cofactor regeneration, such as NADPH and ATP, is another optimization point [23]. Strains optimized for amino acid overproduction require higher regeneration rates of NADPH and ATP in order to cope with the high fluxes present in biosynthetic pathways. Consequently, it is of interest to know which pathways can supply (or reduce) cofactors required for amino acid synthesis. This problem can be studied in the framework of metabolic pathway analysis.

We considered three different cases concerning the synthesis of L-lysine in *E coli*. For each case, we defined a different medium set, *U*, and treatment for cofactors. Standard growth medium has been defined in experimental works [12]. Soybean hydrolysate was assumed to be composed of the amino acids present in Table 1.

In the first problem we used the basic formulation together with the amino acids shown in Table 1 as medium set. In the second problem the amino acids were removed from the medium set. In the last problem, the additional external metabolites set in Table 1 were excluded from the set of internal compounds, while keeping amino acids outside the medium set. Note here that, as described in Methods, those compounds in the extracellular compartment not included in the medium set cannot be consumed, but can be produced. Indeed, such compounds might represent by-products which are associated with a biosynthetic process.

Again, name nomenclature in the original model [10] was kept and therefore the reactions abbreviations are in upper case and the metabolites abbreviations are lower case. Also the reactions default directions are the same, c.f. [10].

Medium Metabolites		**Additional external metabolites set**
Basic Formulation	**Amino acids**	
pi[e]	glu-L[e]	co2
so4[e]	ser-L[e]	nadp
fe2[e]	asp-L[e]	adp
co2[e]	ile-L[e]	amp
h2o[e]	his-L[e]	h
na1[e]	pro-L[e]	nad
glc-D[e]	trp-L[e]	atp
k[e]	thr-L[e]	nadh
h[e]	phe-L[e]	pi
nh4[e]	ala-L[e]	nadph
o2[e]	lys-L[e]	h2o
	met-L[e]	coa
	arg-L[e]	
	gly[e]	
	leu-L[e]	
	val-L[e]	
	tyr-L[e]	

Table 1: Definition of medium metabolites set and cofactors set.

Table 2 shows the first 10 shortest EFMs that produce L-lysine (lys-L[e]) for the first problem under consideration. The computation time was approximately 4 minutes (Intel Core® Duo Processor T2400, 2GB RAM). For *K*=1 we have the shortest EFM; *K*=2 the 2-shortest EFM, i.e the second shortest EFM; etc. *L* represents the length of the EFM, i.e. the number of reactions involved in the EFM. The overall reactions equation given in that table is a representation of the complete transformation of substrates into products. Consequently, cofactors do not appear in this representation because they are constrained to being balanced and their net gain or consumption is null. For example the balancing of ATP is achieved by adding to the solution the reaction catalyzed by the ATP synthase (ATPS4r), which converts one molecule of ADP into one of ATP using the import of 4 protons from the extracellular compartment as driving force. From Table 2 it is possible to calculate the amount of ATP required for the conversion of the nutrients in the medium into L-lysine, which is approximately 2 to 3 molecules of ATP per molecule of L-lysine.

Figure 3 displays the first 3 shortest EFMs from Table 2. The number in brackets corresponds to the number of EFMs (among the first 3 shortest EFMs) in which the reaction is present. The thickness of the arrows is also proportional to the presence of the reactions in each EFM. The boxed reactions names represent the L-lysine biosynthesis pathway [9]. Dashed nodes were duplicated for better visualization. The following list of compounds was removed from the figure to simplify its visualization: NAD, NADH, ATP, ADP, NADP, NADPH, Pi, H2O, CoA and H. The visualization of the network used yEd® version 3.1.2.

It can be observed from Figure 3 that the main difference between the first 3 shortest EFMs relates to the transport reactions used for the uptake of L-aspartate (asp), namely ASPt2, ASPt2_2 and ASPt2_3. These three transport reactions only differ in the amount of protons released to the extracellular compartment. This fact illustrates the combinatorial complexity of the solution space and the hardness of calculating EFMs in large metabolic networks.

It can also be observed from Table 2 that the shortest EFMs consume amino acids from the medium instead of the main carbon source, glucose, and that compounds such as pyruvate (pyr) and 2-oxoglutarate (akg) are exported as by-products. These compounds are precursors of essential compounds for cell growth and therefore it is implausible they would be excreted.

K	*L*	**Overall reaction equations**
1	21	3 ala-L[e] + 1 asp-L[e] + 7 glu-L[e] + 9 h[e] → 7 4abut[e] + 8 co2[e] + 1 lys-L[e] + 2 nh4[e] + 2 pyr[e]
2	21	3 ala-L[e] + 1 asp-L[e] + 8 glu-L[e] + 10 h[e] → 8 4abut[e] + 9 co2[e] + 1 lys-L[e] + 2 nh4[e] + 2 pyr[e]
3	21	3 ala-L[e] + 1 asp-L[e] + 9 glu-L[e] + 11 h[e] → 9 4abut[e] + 10 co2[e] + 1 lys-L[e] + 2 nh4[e] + 2 pyr[e]
4	22	1 asp-L[e] + 10 glu-L[e] + 9 h[e] + 1 ser-L[e] → 7 4abut[e] + 3 akg[e] + 8 co2[e] + 1 lys-L[e] + 3 nh4[e]
5	22	1 ala-L[e] + 1 asp-L[e] + 9 glu-L[e] + 9 h[e] → 7 4abut[e] + 2 akg[e] + 8 co2[e] + 1 lys-L[e] + 2 nh4[e]
6	22	1 ala-L[e] + 1 asp-L[e] + 10 glu-L[e] + 10 h[e] → 8 4abut[e] + 2 akg[e] + 9 co2[e] + 1 lys-L[e] + 2 nh4[e]
7	22	4 asp-L[e] + 4 glu-L[e] + 9 h[e] → 4 4abut[e] + 8 co2[e] + 1 lys-L[e] + 2 nh4[e] + 2 pyr[e]
8	22	8 asp-L[e] + 9 h[e] → 4 ala-L[e] + 8 co2[e] + 1 lys-L[e] + 2 nh4[e] + 2 pyr[e]
9	22	1 asp-L[e] + 11 glu-L[e] + 10 h[e] + 1 ser-L[e] → 8 4abut[e] + 3 akg[e] + 9 co2[e] + 1 lys-L[e] + 3 nh4[e]
10	22	1 ala-L[e] + 1 asp-L[e] + 11 glu-L[e] + 11 h[e] → 9 4abut[e] + 2 akg[e] + 10 co2[e] + 1 lys-L[e] + 2 nh4[e]

Table 2: 10 shortest EFMs using a standard growth medium.

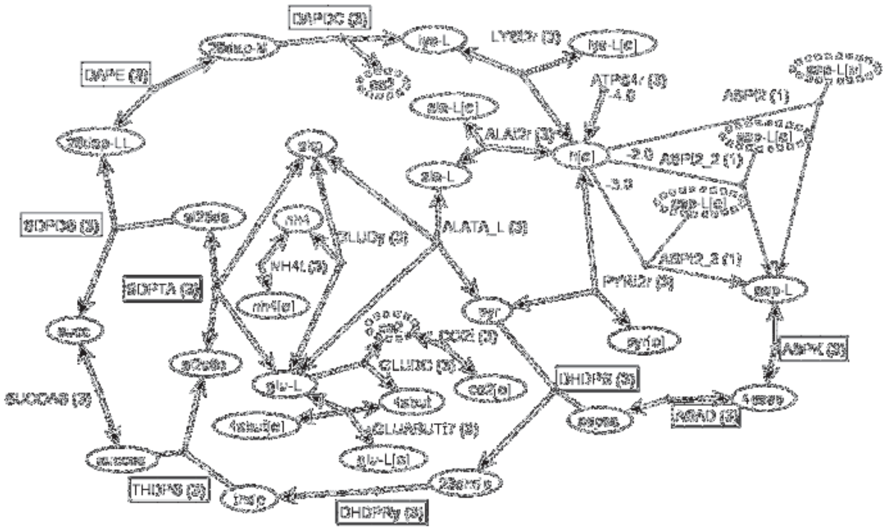

Figure 3: Superposition of the first 3 shortest EFMs calculated for conditions of a standard medium.

In the second problem we reformulated the question so as to find the shortest EFMs that produce L-lysine using glucose as carbon source. In order to do this, we removed the amino acids from the medium set. This implies that amino acids can only be produced, but not consumed. Table 3 shows the first 10 shortest EFMs for this second problem. The computation time was 10 minutes.

From Table 3 it can be observed that the length of the EFMs has considerably increased in comparison with the previous problem. In addition, it is interesting to note that the 5-shortest EFM consumes oxygen. However, there are still a large number of compounds that are excreted as by-products.

K	*L*	**Overall reaction equations**
1	30	4 glc-D[e] + 2 nh4[e] → 4 h2o[e] + 7 h[e] + 4 lac-D[e] + 1 lys-L[e] + 2 pyr[e]
2	31	4 glc-D[e] + 2 nh4[e] → 4 h2o[e] + 7 h[e] + 4 lac-D[e] + 1 lys-L[e] + 2 pyr[e]
3	31	4 glc-D[e] + 2 nh4[e] → 4 h2o[e] + 7 h[e] + 4 lac-D[e] + 1 lys-L[e] + 2 pyr[e]
4	31	4 glc-D[e] + 2 nh4[e] → 4 h2o[e] + 7 h[e] + 4 lac-D[e] + 1 lys-L[e] + 2 pyr[e]
5	31	8 glc-D[e] + 2 nh4[e] + 2 o2[e] → 4 dha[e] + 4 glcn[e] + 4 h2o[e] + 7 h[e] + 1 lys-L[e] + 2 pyr[e]
6	31	8 glc-D[e] + 6 nh4[e] → 4 ala-L[e] + 8 dha[e] + 8 h2o[e] + 7 h[e] + 1 lys-L[e] + 2 pyr[e]
7	31	6 glc-D[e] + 2 nh4[e] → 6 dha[e] + 4 h2o[e] + 5 h[e] + 2 lac-D[e] + 1 lys-L[e] + 2 pyr[e]
8	31	4 glc-D[e] + 2 nh4[e] → 4 h2o[e] + 7 h[e] + 4 lac-D[e] + 1 lys-L[e] + 2 pyr[e]
9	31	3 glc-D[e] + 4 nh4[e] → 2 ala-L[e] + 6 h2o[e] + 5 h[e] + 1 lys-L[e] + 2 pyr[e]
10	31	5 glc-D[e] + 4 nh4[e] → 8 h2o[e] + 8 h[e] + 2 lac-D[e] + 2 lys-L[e] + 4 pyr[e]

Table 3: 10 shortest EFMs using a glucose-based medium.

Figure 4 displays the first 10 EFMs for this second problem. As in Figure 3, most frequent reactions (among the 10 shortest EFMs) have thicker arrows and numbers in brackets represent the number of EFMs in which a reaction participate. Reactions with the name abbreviation in white boxes belong to glycolysis, whilst those reactions in grey boxes correspond to the Entner-Doudoroff pathway. Compounds removed in Figure 3 were also removed here for the sake of simplicity.

In Figure 4, it can be observed that there are three main pathways conducting the conversion of glucose-6-phosphate (g6p) into pyruvate: Entner-Doudoroff, glycolysis and a pathway whose existence has been recently hypothesized. The latter differs from glycolysis due to the conversion of fructose-6-phosphate (f6p) into glyceraldehyde-3-phosphate (g3p) using dihydroxyacetone (dha) as intermediate [15], [24]. Nevertheless, the toxicity of dha and the rapid conversion of this compound into methylglyoxylate, which is also toxic to the cell, make the catalysis of glucose through this alternative pathway highly improbable [20]. Note here that our approach could easily be modified to make dha never appear in the solution.

From Figure 3 and Figure 4, it is clear that shortest EFMs include specific reactions to regenerate cofactors, e.g. reactions consuming extracellular protons (h[e]) in Figure 3 or producing AMP (amp) in Figure 4. This fact makes the solution more difficult to analyze and includes information that is already known by biochemists, namely the concept of moiety conservation [11]. In order to overcome this issue we removed the balancing constraints of some cofactors because they often belong to such moieties. The list of additional external metabolites, including some cofactors, was defined in Table 1.

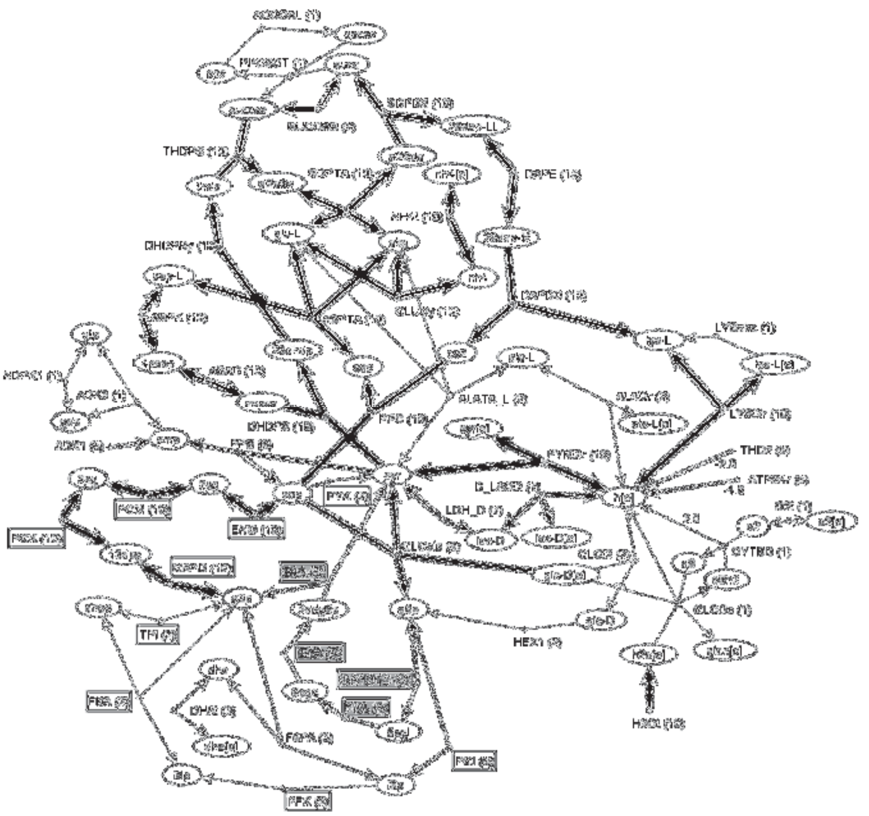

Figure 4: Representation of 10 shortest EFMs in a glucose-based medium.

Table 4 shows the 10 shortest EFMs for our last problem, in which we used glucose as main carbon and an additional set of external metabolites which include some cofactors. EFMs are now more comprehensive. One of the main improvements places is the small number of EFMs containing by-products. This means that the carbon source is completely converted into L-lysine and consequently the yield is higher. The second shortest EFM has a 1:1 ratio of molecules of L-lysine produced per glucose consumed. Indeed, 7 of the EFMs presented here have this 1:1 ratio.

K	*L*	**Overall Equation**
1	24	1 atp + 2 glc-D[e] + 2 nad + 4 nadph + 2 nh4[e] → 1 amp + 2 dha[e] + 2 h2o + 1 h[e] + 1 lys-L[e] + 2 nadh + 4 nadp + 2 pi
2	24	1 atp + 1 glc-D[e] + 1 h + 2 nad + 4 nadph + 2 nh4[e] → 1 adp + 3 h2o + 1 h[e] + 1 lys-L[e] + 2 nadh + 4 nadp + 1 pi
3	25	2 atp + 1 glc-D[e] + 1 nad + 3 nadph + 2 nh4[e] → 2 adp + 2 h2o + 1 h + 1 lys-L[e] + 1 nadh + 3 nadp + 2 pi
4	25	2 atp + 1 glc-D[e] + 1 nad + 3 nadph + 2 nh4[e] → 1 adp + 1 amp + 1 h2o + 1 h + 1 h[e] + 1 lys-L[e] + 1 nadh + 3 nadp + 3 pi
5	25	1 atp + 2 glc-D[e] + 1 h[e] + 2 nad + 4 nadph + 2 nh4[e] → 1 adp + 2 dha[e] + 3 h2o + 1 h + 1 lys-L[e] + 2 nadh + 4 nadp + 1 pi
6	25	1 atp + 1 glc-D[e] + 2 nad + 4 nadph + 2 nh4[e] → 1 amp + 2 h2o + 1 h[e] + 1 lys-L[e] + 2 nadh + 4 nadp + 2 pi
7	25	5 atp + 1 glc-D[e] + 3 h2o + 2 nad + 4 nadph + 2 nh4[e] → 3 adp + 2 amp + 5 h + 1 h[e] + 1 lys-L[e] + 2 nadh + 4 nadp + 7 pi
8	25	5 atp + 2 glc-D[e] + 4 h2o + 2 nad + 4 nadph + 2 nh4[e] → 2 adp + 3 amp + 2 dha[e] + 6 h + 1 h[e] + 1 lys-L[e] + 2 nadh + 4 nadp + 8 pi
9	25	4 atp + 1 glc-D[e] + 2 h2o + 2 nadph + 2 nh4[e] → 2 adp + 2 amp + 4 h + 1 h[e] + 1 lys-L[e] + 2 nadp + 6 pi
10	25	5 atp + 1 glc-D[e] + 4 h2o + 2 nad + 4 nadph + 2 nh4[e] → 2 adp + 3 amp + 6 h + 1 h[e] + 1 lys-L[e] + 2 nadh + 4 nadp + 8 pi

Table 4: First 10 shortest EFMs using a glucose-based medium and excluding cofactors from the set of internal compounds.

Figure 5 depicts the shortest EFM that has a better yield. Note here that grey compounds are external compounds. It is possible to identify glycolysis as the glucose catabolic pathway. Glycolysis is one of the most efficient pathways converting glucose to pyruvate. The latter is then used in the biosynthesis of L-lysine starting from L-aspartate. L-lysine contains 2 nitrogen atoms and, from Figure 5, it can be identified that the import of ammonia indeed occurs, requiring the use of 2 NADPH. The function of akg and L-glutamate (glu-L) carrying the import of nitrogen is also visible in Figure 5. Note here that curved lines linking grey nodes are associated with different moieties, such as ATP and ADP. This was also used in Figure 1. In the case of reactions SDPTA and ASPTA, curved lines help to understand which metabolites are substrates or products, e.g. akg and sl26da are substrates in the reaction catalyzed by SDPTA when the forward direction is taken, being glu-L and sl2a6o products.

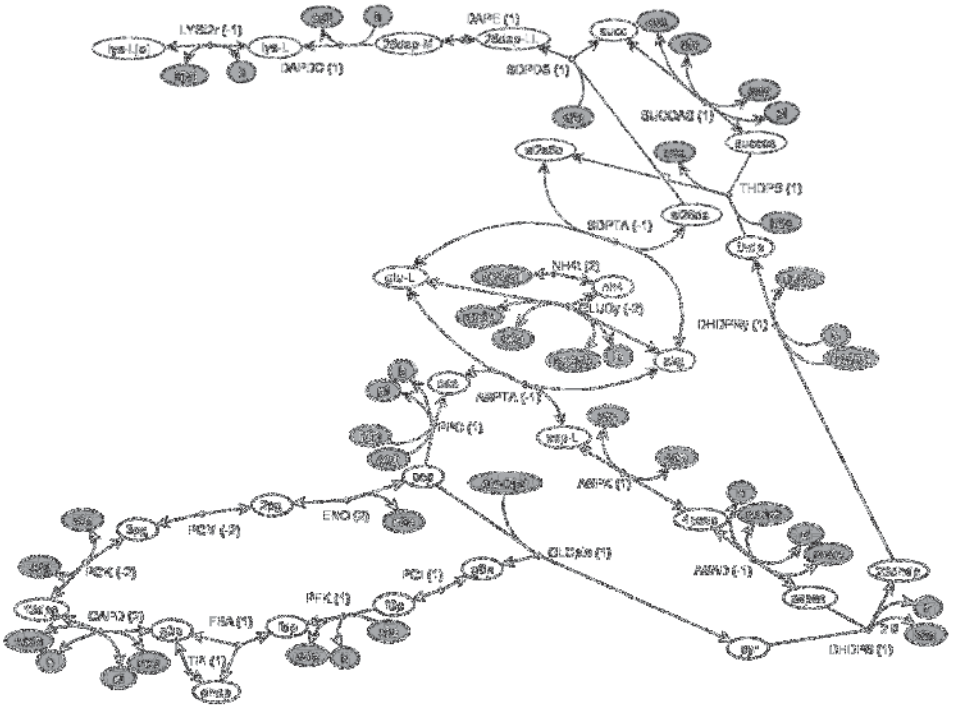

Figure 5: Shortest EFM with better lysine per glucose ratio.

4 Discussion

In this paper we have introduced a novel optimization method to compute the *K*-shortest EFMs. We have illustrated that our approach can be applied to compute the full the set of EFMs. Our procedure is not particular efficient for large-sized metabolic networks. However, the underlying flexibility of linear optimization and feasibility in large systems enables us to explore particular subsets of EFMs. For illustration, we applied our approach to analyse monosaccharide metabolism and L-lysine biosynthesis pathways.

As opposed to other approaches [14], [16], we constrained fluxes (t_r) to have integer values. An alternative formulation of the *K*-shortest EFMs problem using continuous fluxes was also tested. However, computation time for this case far exceeds the formulation presented here. Since most biochemical reactions do have integer stoichiometric coefficients, our integrality condition is suitable. Note that the biomass reaction presented in the genome-scale model of *E. coli* used here [10] was excluded from the reactions set.

As a future direction, we plan to expand our current formulation to include multiple constraints, similar to Equation (10) for L-lysine, such that the non-decomposability condition is satisfied. As noted in the Results section, in this paper we only focus on EFMs producing L-lysine. By minimising the number of reactions, our formulation clearly satisfies the non-decomposability condition. However, were we meant to calculate EFMs that consume glucose (Equation 11) and produce L-lysine, the non-decomposabilty may not be satisfied. Indeed, our model may obtain two different EFMs, namely one consuming glucose and other producing L-lysine. Further research is needed to clarify this issue.

With respect to the analysis of the results obtained, it is clear that the definition of the problem is essential so as to obtain a meaningful solution. We started by formulating a very general problem. We found that these results did not fulfil our expectations in terms of how the main carbon source, glucose, is converted into the desired product, L-lysine. Instead, other carbon sources were used. Results were improved by imposing a medium without amino acids in the second problem. However, balancing equations for cofactors still made the solution difficult to interpret.

A correct solution was achieved in the last problem formulation by including additional information to the network, specifically the concept of moiety conservation. The moiety conservation for a particular set of compounds implies that their concentration remains constant due to the high number of reactions where they can be interconverted. The classical example is the phosphate moiety represented by AMP, ADP, ATP, orthophosphate and pyrophosphate. In order to reflect this concept, some of these metabolites were excluded from being balanced. Nevertheless, the definition of external metabolites will always depend on the problem that is going to be studied. Problems may arise if the set of external metabolites is too large, since we increase the degrees of freedom of the system.

In summary, we showed that our approach partially overcomes the combinatorial explosion associated with the calculation of EFMs, by selecting specific EFMs, the shortests, from the solutions space. Thus our approach is a useful tool to calculate specific EFMs in larger metabolic networks, which was previously thought to be very hard [1].

5 Acknowledgments

The authors gratefully acknowledge the helpful discussions with Christoph Kaleta and Rolf Backofen.

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

6 References

[1] Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, MF. and Stougie, L.: *Modes and cuts in metabolic networks: Complexity and algorithms*. Biosystems, 95 (2008), 51-60.

[2] Beasley, J.E. and Planes, F.J.: *Recovering metabolic pathways via optimization*. Bioinformatics, 23(2007), 92–98.

[3] Dandekar, T., Moldenhauer, F., Bulik, S., Bertram, H. and Schuster, S.: *A method for classifying metabolites in topological pathway analyses based on minimization of pathway number*. Biosystems, 70(2003), 255–270.

[4] de Figueiredo, L.F., Schuster, S., Kaleta, C. and Fell, D. A.: *Can sugars be produced from fatty acids? A test case for pathway analysis tools*. Bioinformatics, 24(2008), 2615-2621. (correct version republished in 25(2009), 152-158)

[5] Förster, J., Gombert, A.K. and Nielsen, J.: *A functional genomics approach using metabolomics and in silico pathway analysis*. Biotechnol. Bioeng., 79(2002), 703–712.

[6] Klamt, S., Gagneur, J. and von Kamp, A.: *Algorithmic approaches for computing elementary modes in large biochemical reaction networks*. IEE Proc. Syst. Biol., 152(2005), 249–255.

[7] Liao, J.C., Hou, S.Y. and Chao, Y.P.: *Pathway analysis, engineering and physiological considerations for redirecting central metabolism.* Biotechnol. Bioeng., 52 (1996), 129–140

[8] Pfeiffer, T., Sánchez-Valdenebro, I., Nuño, J., Montero, F. and Schuster, S.: *METATOOL: for studying metabolic networks*. Bioinformatics, 15(1999):251–257.

[9] Michal, G.: *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Spektrum Akademischer Verlag, Heidelberg, 1999.

[10] Reed, J. L., Vo, T. D., Schilling, C. H. and Palsson, B. Ø.: *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)*. Genome. Biol., 4 (2003), R54.

[11] Reich, J. G. and Selkov, E. E.: *Energy metabolism of the cell: a theoretical treatise.* Academic Press, New York, 1981.

[12] Reverend, B., Boitel, M., Deschamps, A. M., Lebeault, J.-M., Sano, K., Takinami, K. and Patte, J.-C.: *Improvement of Escherichia coli strains overproducing lysine using recombinant DNA techniques*. Appl. Microbiol. Biotechnol. 15(1982) 227-231.

[13] Rohwer, J.M. and Botha, F.C.: *Analysis of sucrose accumulation in the sugar cane culm on the basis of in vitro kinetic data*. Biochem. J., 358 (2001), 437–445.

[14] Schilling, C. H., Letscher, D. and Palsson, B. Ø.: *Theory for the Systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented prespective*. J. theor. Biol, 203 (2000), 229-248.

[15] Schürmann, M. and Sprenger, G. A.: *Fructose-6-phosphate aldolase is a novel class I aldolase from Escherichia coli and is related to a novel group of bacterial transaldolases*. J. Biol. Chem., 276 (2001), 11055-11061.

[16] Schuster, S., Fell, D. A. and Dandekar, T.: *A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks*. Nat. Biotechnol., 18 (2000), 326–332.

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

[18] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. and Dandekar, T.: *Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae*. Bioinformatics, 18 (2002), 351–361.

[19] Schuster, S., von Kamp, A. and Pachkov, M.: *Understanding the Roadmap of Metabolism by Pathway Analysis.* Methods Mol. Biol., 358(2007), 199-226.

[20] Subedi, K. P., Kim, I., Kim, J., Min, B. and Park, C.: *Role of GldA in dihydroxyacetone and methylglyoxal metabolism of Escherichia coli K12*. FEMS Microbiol. Lett., 279 (2003), 180-187.

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

[22] Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W. M., Siezen, R. J. and Smid, E. J.: *Analysis of Growth of Lactobacillus plantarum WCFS1 on a Complex Medium Using a Genome-scale Metabolic Model.* J. Biol. Chem., 281(2006), 40041–40048.

[23] Tosaka, O., Enei, H. and Hirose, Y.: *The production of L-lysine by fermentation*. Trends in Biotechnology, 1(1983), 70-74.

[24] van Winden, W. A., van Gulik, W. M., Schipper, D., Verheijen, P. J. T., Krabben, P., Vinke, J. L., and Heijnen, J. J.: *Metabolic flux and metabolic network analysis of Penicillium chrysogenum using 2D [^{13}C, ^{1}H] COSY NMR measurements and cumulative bondomer simulation*. Biotechnol. Bioeng., 83(2003), 75-92.

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

[26] Wendisch, V. F., Bott, M. and Eikmanns, B. J.: *Metabolic engineering of Escherichia coli and Corynebacterium glutamicum for biotechnological production of organic acids and amino acids*. Curr. Opin. Microbiol., 9(2006)268-274.