BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 23 2009, pages 3158–3165
doi:10.1093/bioinformatics/btp564

Systems biology

Computing the shortest elementary flux modes in genome-scale metabolic networks

Luis F. de Figueiredo[1,2], Adam Podhorski[3], Angel Rubio[3], Christoph Kaleta[1], John E. Beasley[4], Stefan Schuster[1] and Francisco J. Planes[3,*]

[1]Friedrich-Schiller-University Jena, 07743 Jena, Germany, [2]PhD Program in Computational Biology, Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal, [3]CEIT and TECNUN, University of Navarra, 20016 San Sebastián, Spain and [4]Brunel University, Uxbridge, UB8 3PH, UK

Received on May 11, 2009; revised on September 10, 2009; accepted on September 25, 2009

Advance Access publication September 30, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Elementary flux modes (EFMs) 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-scale metabolic networks still constitutes a challenging issue due to its underlying combinatorial complexity.

Results: In this article, we illustrate that the full set of EFMs can be enumerated in increasing order of number of reactions via integer linear programming. In this light, we present a novel procedure to efficiently determine the *K*-shortest EFMs in large-scale metabolic networks. Our method was applied to find the *K*-shortest EFMs that produce lysine in the genome-scale metabolic networks of *Escherichia coli* and *Corynebacterium glutamicum*. A detailed analysis of the biological significance of the *K*-shortest EFMs was conducted, finding that glucose catabolism, ammonium assimilation, lysine anabolism and cofactor balancing were correctly predicted. The work presented here represents an important step forward in the analysis and computation of EFMs for large-scale metabolic networks, where traditional methods fail for networks of even moderate size.

Contact: fplanes@tecnun.es

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

1 INTRODUCTION

In recent years, different approaches have been proposed to investigate the structure of complex metabolic networks (Price *et al.*, 2004). In particular, elementary flux modes (EFMs) have attracted increasing interest. An EFM is defined as a minimal set of enzymes that operates at steady state with all irreversible reactions used in the appropriate direction (Schuster and Hilgetag, 1994; Schuster *et al.*, 2000). An analogous concept in Petri net theory is provided by the minimal *T*-invariants (Koch *et al.*, 2005). The relevance of EFMs for various applications has been recently reviewed (Trinh *et al.*, 2009). EFM analysis has proved useful in elucidating novel metabolic pathways in addition to textbook knowledge, e.g. a new catabolic pathway that degrades glucose via the glyoxylate shunt (Fischer and Sauer, 2003; Liao *et al.*, 1996; Schuster *et al.*, 1999). Several software packages for computing EFMs have been developed, e.g. METATOOL (von Kamp and Schuster, 2006), CellNetAnalyzer (Klamt *et al.*, 2007), YANAsquare (Schwarz *et al.*, 2007) and efmtool (Terzer and Stelling, 2008). However, EFM analysis suffers from an important drawback: the number of EFMs grows exponentially with network size (Klamt and Stelling, 2002). For instance, more than two million EFMs have been reported for the metabolic network describing the central metabolism in *Escherichia coli*, which contains 110 reactions (Gagneur and Klamt, 2004). Despite a number of attempts to cope with such complexity (Dandekar *et al.*, 2003; Klamt *et al.*, 2005; Schuster *et al.*, 2002; Terzer and Stelling, 2008; Teusink *et al.*, 2006), computing the full set of EFMs in large metabolic networks still constitutes a challenging issue.

Based on the work of Beasley and Planes (2007), we show here that the full set of EFMs can be enumerated via integer linear programming. Technically, our approach produces EFMs in increasing order of number of reactions by solving a sequence of discrete optimization problems. Thus, it is promising to start with the shortest, second shortest, etc., overall called *K*-shortest EFMs. The '*K*-shortest' concept has been previously used in the context of graph theory and paths (see, for illustration, Planes and Beasley, 2009), but not in the context of EFMs. Acuña *et al.* (2009) have recently suggested that finding short EFMs should become interesting if size is considered a relevant criterion. Also, in Mavrouniotis *et al.* (1990), biochemical pathways (not EFMs) are obtained in increasing length order.

Detection of *K*-shortest EFMs is of interest for several biological applications. Experimentally, it is expensive and laborious to overexpress a large number of enzymes. On the other hand, since the highest increase in pathway flux is achieved if all enzymes (Kacser and Acerenza, 1993) or (at least) a considerable number of enzymes in a pathway (Fell and Thomas, 1995; Niederberger *et al.*, 1992) are overexpressed, shorter pathways are better suited as a target for genetic manipulation. Moreover, shorter pathways can carry higher fluxes (Meléndez-Hevia *et al.*, 1994; Pfeiffer and Bonhoeffer, 2004).

The use of integer linear optimization makes our procedure more flexible than previous approaches found in the literature

*To whom correspondence should be addressed.

(Schilling *et al.*, 2000; Schuster *et al.*, 2000), which require the computation of the full set of EFMs before any further analysis can be performed. Instead, our method allows us to directly explore the K-shortest EFMs related to a particular problem of interest, e.g. the K-shortest EFMs that consume/produce a particular metabolite.

In order to illustrate the applicability of our approach, we here analyse the K-shortest EFMs producing lysine in two different genome-scale metabolic networks, *E.coli* K-12 MG1655 (Feist *et al.*, 2007) and *Corynebacterium glutamicum* ATCC 13032 (Kjeldsen and Nielsen, 2009). Lysine is one of the essential amino acids in humans and is also used as supplement in animal feeds. The industrial production of lysine has a long history in biotechnology (Tosaka *et al.*, 1983; Wendisch *et al.*, 2006). Studying the production of lysine has been essential for the rational design of optimized strains. Nowadays, *C.glutamicum* is the organism of choice for lysine overproduction due to the higher yields obtained with it. The capability for producing lysine has been previously examined from a pathway oriented perspective (de Graaf, 2000; Mavrovouniotis *et al.*, 1990; Schuster *et al.*, 2007). However, these studies were not conducted at the genome-scale. Therefore, the results presented here extend these studies to a larger scale.

2 METHODS

The mathematical model proposed below formulates the task of finding EFMs as a sequence of optimization problems. Our method starts from the basis that the flux mode involving the minimum number of reactions must be elementary. We here refer to it as the shortest EFM. Accordingly, we first define the constraints and the function (objective) to be optimized that allows us the calculation of the shortest EFM. Based on this optimization model, we then show how to calculate the K-shortest EFMs. Finally, extensions of the K-shortest to other problems of interest are presented.

We mean here by 1-shortest EFM, the EFM containing the minimum number of reactions; 2-shortest EFM, the EFM containing the second minimum number of reactions, etc. We may have multiple EFMs containing the same minimum number of reactions. If this occurs, they are counted separately with different K values. The enumeration order of equally long EFMs depends on the actual implementation of the mathematical model and the solving procedure.

As noted above, EFMs are defined as minimal sets of enzymes in steady state (Schuster *et al.*, 2000). The meaning of 'minimal' in the definition of EFMs refers to the non-decomposability condition, i.e. the addition of an enzyme would turn the EFM into non-elementary. In contrast, we here refer the 1-shortest EFM as to the EFM that contains the (global) minimum number of enzymes.

2.1 Shortest EFM

Assume we have a metabolic network that comprises R reactions and C compounds. Here we decompose reversible reactions into two opposing reaction steps. Thus, we can regard all fluxes as taking positive values. Let s_{cr} be the stoichiometric coefficient associated with compound c ($c = 1, \ldots, C$) in reaction r ($r = 1, \ldots, R$). As usual in the literature (Schilling *et al.*, 2000; Schuster and Hilgetag, 1994), substrates and products have negative and positive stoichiometric coefficients, respectively. The matrix containing all these coefficients is called the stoichiometric matrix.

A zero-one (binary integer) variable is assigned to each reaction, namely $z_r = 1$ if reaction r ($r = 1, \ldots, R$) is active in the EFM, 0 otherwise. In addition, each reaction has an associated non-negative (integer) flux t_r. As we are studying structural properties of metabolic networks, it is appropriate to use integer fluxes. If the coefficients of the stoichiometric matrix (s_{cr}) take integer values, as it is assumed here and in many other approaches such as Petri net theory (Koch *et al.*, 2005), then the relative fluxes carried by EFMs can also be described using integer values. In addition, our computational experience reveals that the K-shortest method is more expensive when fluxes are allowed to be non-integer.

For the optimization model we need constraints relating the reaction variables z_r and t_r:

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

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

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. This does not constitute an issue if M is a sufficiently large value.

In our model, reversible reactions are decomposed into two irreversible reactions, and therefore, we define the set $B = \{(\alpha, \beta) |$ reaction α and reaction β are the reverse of each other, $\alpha < \beta\}$.

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

Equation (3) ensures that a reaction and its reverse do not appear in an EFM.

The steady-state condition is critical for the definition of EFMs and it is formulated as

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

where I is the set of internal compounds. 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 the set of external compounds by E.

In order to avoid the trivial solution ($z_r = t_r = 0$, $r = 1, \ldots, R$), we require that at least one reaction is active:

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

Equations (1–5) define the flux modes solution space for a particular metabolic network. In order to calculate the shortest EFM, we minimize the number of reactions:

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

As noted above, EFMs cannot be decomposed into smaller entities without violating the steady-state assumption, Equation (4). This is referred as to the non-decomposability (elementary) condition (Schuster and Hilgetag, 1994). In essence, this condition implies that no subset of reactions of an EFM can perform at steady state. We ensure that the non-decomposability condition is satisfied by minimizing the number of active reactions involved in the solution flux mode. Clearly, the flux mode involving the minimum number of reactions will be non-decomposable.

2.2 K-shortest EFMs

The mathematical optimization model given above [objective function (6) subject to Equations (1)–(5)], once solved, allows us to obtain the shortest EFM. 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 binary solution associated with the shortest EFM, where Z_r^1 equals to 1 if reaction r is active, 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:

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

The left-hand side of Equation (7) determines the number of reaction variables in the current solution that were active in the 1-shortest EFM solution. The right-hand side is the number of reactions that were active in the 1-shortest EFM less one. The inequality states that the number of active reactions repeating from the 1-shortest EFM should be less by at least one than the total number of active reactions in that EFM. This ensures that, once we solve our model, 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 EFM from the solution space. In the general case, the $K-1$ shortest EFM solution is eliminated before the K-th solution is computed and clearly the optimization problem for the K-th shortest EFM accumulates constraints from all $(1,\ldots,K-1)$ previous solutions, i.e. in order to find the K-shortest EFM, we need to include EFM elimination constraints related to the first $(K-1)$ shortest EFMs:

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

where Z_r^k is the binary solution for the k-shortest EFM.

Note here that the K-shortest EFMs described above are also elementary. For an indirect proof, suppose that the K-shortest EFM (once solved) is not elementary, i.e. it contains a subset of reactions satisfying Equations (1–5) and (8). Since we are constructing EFMs in increasing order of the number of reactions they contain, we must have encountered the EFM corresponding to this subset before. However, 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, which contradicts the original assumption. Thus, every EFM we find must be elementary.

2.3 Extensions to *K*-shortest EFMs

Our procedure can be applied to enumerate all EFMs, namely by constructing them one by one. This is not particularly efficient for small-scale metabolic networks when compared with existing methods. The main advantage of our mathematical optimization model is that, by adding new constraints, special subsets of EFMs (of particular biomedical or biotechnological interest) can be found without having to first compute all EFMs as is the case in existing methods (Klamt *et al*., 2005; Schilling *et al*., 2000; Schuster *et al*., 2000; Terzer and Stelling, 2008). Below, we present some of these constraints that can be easily added to our formulation.

Genome-scale metabolic networks are typically compartmentalized models, in the simplest case containing the extracellular compartment and cytosol. We assume that metabolites in the extracellular compartment can be taken up or secreted as by-products, therefore these metabolites can be set to be external. We denote U the set of extracellular metabolites defining the growth medium. In the case an extracellular metabolite c is not included in the medium set, we need to avoid this compound to be consumed. Equation (9) describes how this constraint is incorporated into our model.

$$\sum_{r=1}^{R} s_{cr} t_r \geq 0 \qquad \forall c \in E, c \notin U \tag{9}$$

We may also need to find the K-shortest EFMs that produce a particular external compound, μ. To do so, 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 Equations (10) or (11), as both already require at least one compound to be produced or consumed, respectively, hence at least one reaction must be active. In addition, the non-decomposability condition is not guaranteed when more than one constraint based on Equations (10) or (11) is included in the formulation. For example, if we apply constraint (10) for metabolites μ_1 and μ_2, i.e. finding solutions to our model that produces μ_1 and μ_2, then we might obtain solutions containing two EFMs, namely one producing μ_1 and another producing μ_2. For this reason, in this article, we restrict our analysis to EFMs forced to produce/consume one metabolite. Equation (9) does not alter the non-decomposability condition.

2.4 Integer programming

Our mathematical optimization model given above for computing the K-shortest EFMs [objective function (6) subject to Equations (1–5) plus elimination constraints (8) and perhaps constraints (9–11)] is an integer linear program. Algorithmically such programs are solved by linear programming based tree search (Pardalos and Resende, 2002). Various free and commercial software tools are available to perform this task. We used ILOG CPLEX®.

3 RESULTS

We applied our method to three different metabolic networks. Firstly, we examined a well-known metabolic network that contains the tricarboxylic acid (TCA) cycle and some adjacent reactions (Schuster *et al*., 1999). Since this metabolic network is of moderate size, the full set of EFMs can be obtained using classic methods (Schuster *et al*., 1999). We used it as a benchmark to validate the capabilities of our method. Then, we applied our method to study the production of lysine in two different genome-scale metabolic networks, *E.coli* K-12 MG1655 (Feist *et al*., 2007) and *C.glutamicum* ATCC 13032 (Kjeldsen and Nielsen, 2009). Details of the three metabolic networks can be found in the Supplementary Material.

3.1 TCA cycle network

For the TCA cycle network, our method correctly enumerated, in increasing order of number of reactions, all 16 EFMs previously determined in Schuster *et al*. (1999). Details on the 16 EFMs are shown in Table 1. The shortest EFM contains two reactions, which are catalyzed by enzymes Pck and Ppc. The 2-shortest EFM also has two reactions. The 16-shortest EFM involves 13 reactions. These results confirm the applicability of our method.

We compared the computation time of our method with METATOOL (version 5.1) for this particular small network. Our method turned out to be less efficient than METATOOL, though both methods take <1 s (data not shown). However, as will be shown below, our method is particularly suitable for large-scale metabolic networks, where classical methods for EFMs computation are not applicable.

In addition, we extended the analysis by calculating the subset of EFMs that produces succinyl-CoA (SucCoAxt). This is done by incorporating a constraint based on Equation (10) for SucCoAxt into the K-shortest EFMs formulation. Our method directly enumerated the six EFMs producing SucCoAxt without having to first compute the full set of EFMs, as typically done by METATOOL and classic methods (Table 1).

Table 1. Full set of EFMs in the TCA cycle metabolic network

K	*L*	Enzyme set	SCA
1	2	Pck; Ppc	–
2	2	Pps; Pyk	–
3	5	AlaCon; Eno; Gdh; IlvE_AvtA; Pyk	–
4	5	AspC; AspCon; Eno; Gdh; Ppc	–
5	5	AspA; AspC; Fum; Gdh; Mdh	–
6	7	Eno; Ppc; SucCoAcon; -Fum; -Mdh; -Sdh; -SucCD	1
7	8	AspA; AspC; Eno; Gdh; Ppc; SucCoAcon; -Sdh; -SucCD	2
8	9	AceEF; Acn; 2 Eno; GltA; Icd; Ppc; Pyk; SucAB; SucCoAcon	3
9	9	AceEF; Acn; 2 Eno; Gdh; GltA; GluCon; Icd; Ppc; Pyk	–
10	10	2 AceEF; Acn; 2 Eno; GltA; Icl; Mas; Mdh; 2 Pyk; SucCoAcon; -SucCD	4
11	11	AceEF; Acn; Eno; Fum; GltA; Icd; Mdh; Pyk; Sdh; SucAB; SucCD	–
12	11	2 AceEF; Acn; Eno; Fum; GltA; Icl; Mas; 2 Mdh; Pck; 2 Pyk; Sdh	–
13	12	2 AceEF; Acn; 3 Eno; GltA; Icl; Mas; Ppc; 2 Pyk; 2 SucCoAcon; -Fum; -Sdh; -2 SucCD	5
14	13	3 AceEF; 2 Acn; 3 Eno; Fum; 2 GltA; Icd; Icl; Mas; 2 Mdh; 3 Pyk; Sdh; SucAB; SucCoAcon	6
15	13	3 AceEF; 2 Acn; 3 Eno; Fum; Gdh; 2 GltA; GluCon; Icd; Icl; Mas; 2 Mdh; 3 Pyk; Sdh	–
16	13	2 AceEF; Acn; AspC; AspCon; 2 Eno; Fum; Gdh; GltA; Icl; Mas; 2 Mdh; 2 Pyk; Sdh	–

K: the order by which EFMs are computed; *L*: the number of reactions in each EFM; SCA—order by which EFMs producing SucCoAxt are computed. Reversible reactions active in the opposite direction have a minus sign before the flux value.

3.2 Genome-scale metabolic networks

We calculated the *K*-shortest EFMs that produce lysine in the genome-scale metabolic networks of *E.coli* and *C.glutamicum* with $K = 10$. These metabolic networks differ in the number of reactions and metabolites, as well as in the level of accuracy. During the computation of 10-shortest EFMs some errors in the *C.glutamicum* network were identified. In particular, an error in reaction dapB was responsible for a null lysine net synthesis. More details as to errors can be found in the Supplementary Material.

The *E.coli* network is larger than the *C.glutamicum* network. For this reason, the *E.coli* metabolic network represents a greater challenge in the computation of 10-shortest EFMs. Our method successfully computed them, though the difference in the computation time is significant (see Supplementary Material). We used glucose and ammonium as carbon and nitrogen sources, respectively, for both metabolic networks. See Supplementary Material for exact definition of the medium set, U. A sufficiently large M value is needed to ensure that no EFM information is lost. We conducted experimentation for different M values (see Supplementary Material) and selected $M = 10\,000$, since no change in the *K*-shortest EFMs solution was found with respect to smaller M values. This selected value is similar to that proposed in previous studies (Kjeldsen and Nielsen, 2009; Vallino and Stephanopoulos, 1993).

We first applied our mathematical model to the metabolic network of *E.coli*. Figure 1 shows a merged representation of the 10-shortest EFMs producing lysine in *E.coli*. The shortest

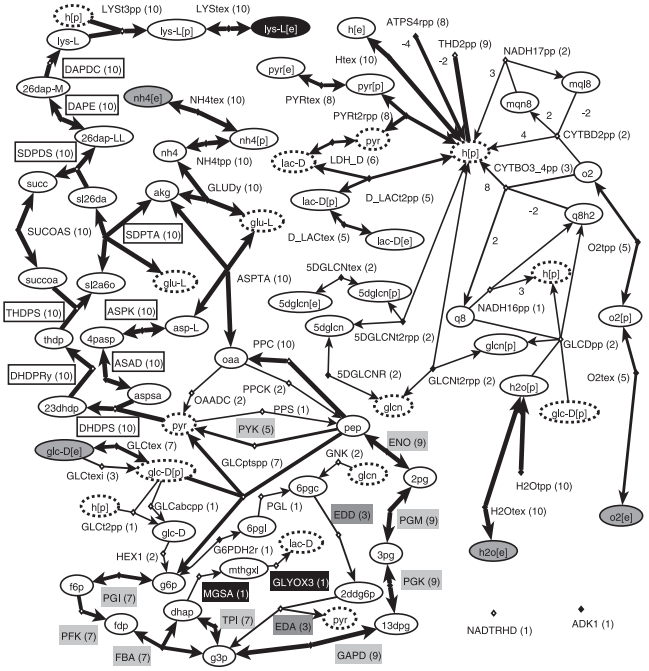

Fig. 1. Merged representation of the 10-shortest EFMs producing lysine in *E.coli* when cofactors are set as internal metabolites. Ellipses represent metabolites and arrows reactions. Stoichiometric coefficients higher than one are represented next to the edge linking the respective metabolite. Dashed ellipses are duplicated metabolite nodes, light grey ellipses are medium metabolites and the black ellipse is the target metabolite. Numbers in brackets after enzyme abbreviations correspond to the number of EFMs where these are present. Thickness of the arrows is proportional to this number. Boxed enzyme abbreviations represent the lysine biosynthetic pathway (Cohen and Saint-Girons 1987, Wittmann and Becker, 2007), enzyme abbreviations in light grey, in dark grey and black correspond to glycolysis, the Entner–Doudoroff pathway and the methylglyoxal bypass, respectively. The following metabolite nodes in the cytosolic compartment were removed from the representation for better visualization: atp, adp, amp, nad, nadh, nadp, nadph, h, coa, h2o, pi, co2. Note here that abbreviations are the same as in the original network (see Feist *et al*., 2007). Thus, reactions involving only these removed metabolites may seem disconnected from the sub-network when they are actually connected, e.g. NADTRHD.

EFMs are mainly fermentation modes and therefore, they require higher fluxes on glucose catabolism (see Supplementary Material for more information about the fluxes and the reaction sets). The combinatorial effect seen in EFM analysis can be immediately observed. This is particularly apparent for transport reactions. For example, there are two different reactions for the uptake of glucose (glc-D) from the extracellular compartment to the periplasm, specifically GLCtex and GLCtexi. Thus, there will be at least two EFMs among the 10-shortest EFMs that differ only in the use of one of these two reactions while the rest of the enzyme set remains the same. Such combinatorial features can also be found in the other *K*-shortest EFMs.

A detailed analysis of Figure 1 reveals that there are three major pathways for glucose catabolism: glycolysis, the Entner–Doudoroff (ED) pathway and the methylglyoxal bypass. Glycolysis provides higher quantities of ATP but does not produce any NADPH and therefore the periplasmic NAD(P) transhydrogenase, THD2pp,

is required to reduce NADP by oxidizing NADH. The ED pathway can use two different precursors of 6-phospho-D-gluconate (6pgc), namely, 6-phospho-D-glucono-1,5-lactone (6pgl) and D-gluconate (glcn). In case 6pgl is used as precursor the oxidative part of the pentose phosphate (PP) pathway produces NADPH and therefore, the THD2pp is not required in this mode, in contrast with the rest of EFMs. When the methylglyoxal bypass is used there is a very low ATP yield from glucose catabolism and therefore, the ATP synthase, ATPS4rpp, has a higher flux when compared with the other modes. It should be noted that this pathway, though possible, is very unlikely to be the main catabolic route of glucose due to the toxicity of methylglyoxal (Subedi *et al*., 2008).

In *E.coli*, ammonium assimilation can be carried out via the glutamine synthetase/glutamate synthase (GLNS/GLUSy) cycle or exclusively using glutamate dehydrogenase (GLUDy). The GLNS/GLUSy cycle constitutes the main ammonium assimilation route even for growth conditions with high extracellular ammonium content (Yuan *et al*., 2006). In the 10-shortest EFMs, the assimilation of ammonium is however conducted by GLUDy, which involves fewer steps and consumes less ATP. The other route would appear for EFMs containing 40 reactions.

In addition, it is well-known that *E.coli* has only one pathway for lysine biosynthesis using aspartate and pyruvate as precursors (Cohen and Saint-Girons, 1987; Wittmann and Becker, 2007). This is also observed in the left upper corner in Figure 1, where the thickness of the involved arrows is maximal, i.e. they appear in all 10-shortest EFMs.

On the right-hand side of Figure 1, there are many reactions around the periplasmatic proton node, h[p]. These reactions are mainly involved in the establishment of a proton gradient so that ATP and NADPH can be produced. We assumed that cofactors are buffered in the metabolic network and set them as external metabolites. We repeated our K-shortest procedure ($K = 10$) and found that the shortest EFM involves 27 reactions, as opposed to the case described above where the shortest EFMs involved 38 reactions.

In Figure 2, there are no EFMs producing by-products such as lactate or pyruvate. The main reason is that there is no need of fermentative modes or other modes producing cofactors in small reaction steps and with high fluxes, since cofactors are now external metabolites. The catabolism of glucose in Figure 2 is again accomplished by the same three pathways: glycolysis, the ED pathway and the methylglyoxal bypass. Combinations of these three pathways are also found in the 10-shortest EFMs, e.g. in the 7-shortest EFM, the ED pathway is combined with the triose phosphate part of glycolysis, while in the 10-shortest EFM the ED pathway is combined with the methylglyoxal bypass. There is a detour to the classical glycolysis described in textbooks, via dihydroxyacetone (dha). This detour has been recently hypothesized by van Winden *et al.* (2003). However, the use of dha as intermediate is questionable due to its toxicity and possible conversion to methylglyoxal (Molin *et al.*, 2003; Subedi *et al*., 2008).

The results also show that, with glycolysis as single catabolic pathway, it is possible to produce one mole of lysine per mole of glucose consumed, consuming four moles of NADPH and one mole of ATP and producing two moles of NADH. Thus, from a molecule containing six carbon atoms, glucose, it is possible to produce another six-carbon molecule, lysine, requiring two NADPH for ammonium assimilation, plus two NADPH and one ATP for

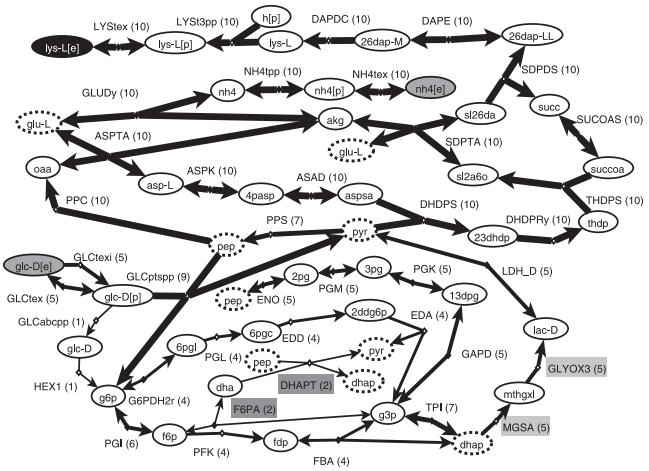

Fig. 2. Merged representation of the 10-shortest EFMs producing lysine in *E.coli* when cofactors are set as external metabolites. Enzyme abbreviations in light grey and dark grey represent the methylglyoxal bypass and a detour of the classical glycolysis over dha, respectively. The following metabolite nodes in the cytosolic compartment were removed from the representation for better visualization: atp, adp, amp, nad, nadh, nadp, nadph, h, coa, h2o, pi, co2. Note here that abbreviations are the same as in the original network (see Feist *et al*., 2007).

the intermediate metabolites inter-conversion. However, due to the carboxylation and decarboxylation reactions, this 1:1 conversion cannot be deduced directly from the number of carbons.

A similar analysis was conducted for *C.glutamicum*. We found that the shortest EFM contains 33 reactions when cofactors are set to internal. The shortest EFMs for *C.glutamicum* are not fermentative (Fig. 3) in contrast to *E.coli* (Fig. 1) and the main route for glucose catabolism is the PP pathway. A reasonable question that can be posed is why there is no fermentative mode in the shortest EFMs for *C.glutamicum*. This is due to the fact that the reaction catalyzed by lactate dehydrogenase, which reduces pyruvate to lactate, is not present in the metabolic network, nor any other pathway linking pyruvate to lactate. Note, however, that such reaction is present in the genome annotation of this organism and there is experimental data on lactate dehydrogenase mutants (Inui *et al*., 2004).

In Figure 3, it is also apparent that the main variability in the 10-shortest EFMs is in the balancing of cofactors and there are no alternative pathways for glucose catabolism in comparison to the 10-shortest EFMs of *E.coli* (Fig. 1). This fact can be attributed to the differences in the metabolic networks caused by evolution. While in *E.coli* the ED pathway and the methylglyoxal bypass are present, to date they have not been identified in *C.glutamicum* (Eggeling and Bott, 2005). Moreover, there are differences in some anaplerotic reactions. Nevertheless, there is also an evident difference in the accuracy of both networks, since the number of reactions in the metabolic network of *E.coli* is almost 5-fold higher while the size of the genome and the number of predicted proteins for both organisms is of the same order of magnitude (Blattner *et al.,* 1997; Kalinowski *et al.,* 2003).

As mentioned above, the PP pathway is the only glucose catabolic pathway present in the EFMs, which is due to the requirement of redox anabolic power. An alternative pathway would have been the TCA cycle or anaplerotic reactions between oxaloacetate, malate,

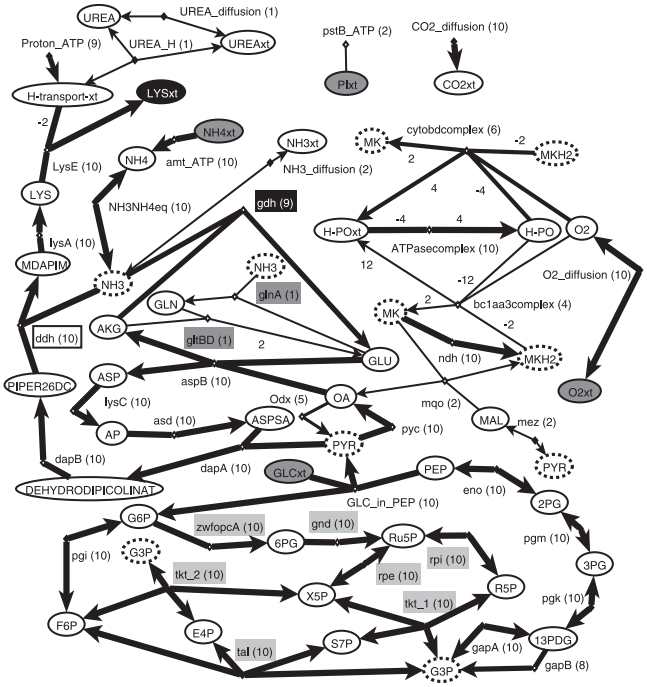

Fig. 3. Merged representation of the 10-shortest EFMs in *C.glutamicum* producing lysine and with cofactors as internal metabolites. Boxed enzyme abbreviation is characteristic for *C.glutamicum* (Eggeling, 1994; Wittmann and Becker, 2007), enzyme abbreviations in light grey, dark grey and black represent the PP pathway, the longest and the shortest pathways for ammonium assimilation, respectively. The following metabolite nodes, in the cytosolic compartment, were removed from the representation for better visualization: ATP, ADP, NAD, NADH, NADP, NADPH, H-transport, COA, PI, CO_2.

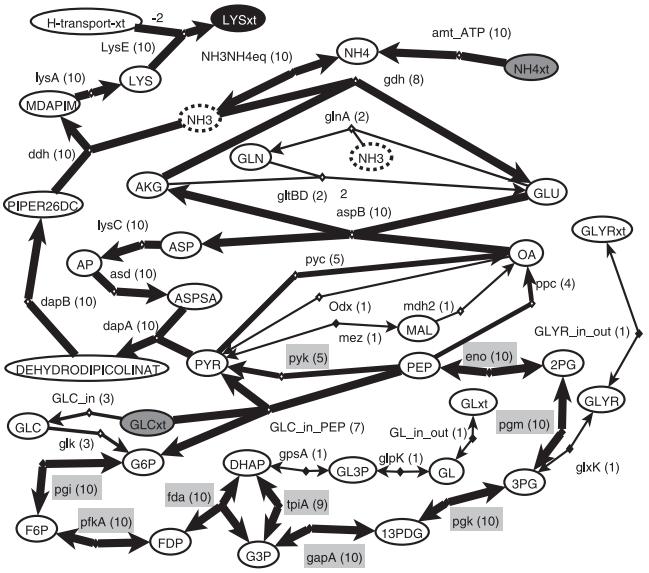

Fig. 4. Merged representation of the 10-shortest EFMs producing lysine in *C.glutamicum* and with cofactors as external metabolites. Enzymes with abbreviations in light grey represent glycolysis. The following metabolite nodes, in the cytosolic compartment, were removed from the representation for better visualization: ATP, ADP, NAD, NADH, NADP, NADPH, H-transport, COA, PI, CO_2.

phosphoenolpyruvate and pyruvate. However, the presence of the complete TCA cycle requires more enzymes to reduce NADP using glucose. Experimentally, the PP pathway also has a more important role in NADPH synthesis than the TCA cycle. Indeed, metabolic flux analyses have shown that ∼70% of the NADPH is generated by the PP pathway and the remaining 30% by isocitrate dehydrogenase of the TCA cycle (Eggeling and Bott, 2005).

Possible NADPH regenerating cycles, involving anaplerotic reactions, which are often mentioned in the literature (cf. Wittmann and Becker, 2007), are not found with this function. Instead, they can only convert NADPH into NADH because in the genome-scale network the reactions mdh and mqo are set to irreversible forcing these cycles to be irreversible. The existence of two glyceraldehyde-3-phosphate dehydrogenases, gapA and gapB, also allows the conversion of NADPH into NADH, but not the reverse. If the reaction catalysed by lactate dehydrogenase is included in the metabolic network, the fermentative pathways are still not the shortest because there is no alternative to the PP pathway for NADPH synthesis, and therefore, the EFMs with this pathway are the shortest (data not shown).

Regarding the ammonium assimilation, it can be seen that a larger number of EFMs uses glutamate dehydrogenase (gdh) and only two EFMs use the glutamine synthase/glutamate synthase (glnA/gltBD) pathway. The appearance of a longer route is due to the fact that the 10-shortest EFMs in *C.glutamicum* have more widely distributed lengths than the 10-shortest EFMs in *E.coli*. Nevertheless, for *C.glutamicum*, the shorter pathway is more relevant at high ammonium concentrations (Eggeling and Bott, 2005).

If cofactors are set external, the PP pathway, the cycles converting NADPH to NADH and enzymes from the respiratory chain do not appear in the 10-shortest EFMs. Instead, glycolysis is the main route for glucose catabolism (Fig. 4). This pathway is indeed the shortest catabolic pathway in this network, as the ED pathway and the glyoxylate bypass are not present. The main variability in these EFMs is found in the synthesis of by-products such as glycerate and glycine and in the interconnection of the catabolic and anabolic part of the EFMs. The latter is evident by the detour made through malate (Fig. 4).

From Figures 3 and 4, it can be observed that the 10-shortest EFMs involve the shortest lysine biosynthetic pathway described in the literature (Wittmann and Becker, 2007). An alternative longer route does exist in *C.glutamicum*, which differs in three reactions and requires one additional reaction to balance succinate and succinyl-CoA, as shown in the 10-shortest EFMs of *E.coli* (Figs 1 and 2). This means that EFMs with higher length are needed so as to obtain the alternative pathway for lysine synthesis.

4 CONCLUSION

The computation of EFMs in genome-scale metabolic networks has been very difficult if not impossible so far. In order to explore the metabolic capabilities of a given organism via EFMs, often smaller sub-networks are delimited. However, the analysis of small sub-networks can be misleading (Kaleta *et al*. 2009; Terzer and Stelling, 2008) and therefore, the computation of EFMs in genome-scale networks is essential for a more comprehensive analysis of

the metabolic capabilities of an organism. In such large networks, detecting short EFMs is of interest from the biological viewpoint. Experimentally, it is expensive to overexpress a large number of enzymes, so that shorter pathways are better suited for genetic manipulation. Moreover, shorter pathways usually carry higher fluxes.

In this article we showed that the full set of EFMs can be theoretically enumerated via discrete optimization. This is a promising development in EFM computation and it might serve as a basis for building new methods to explore the structure of large metabolic networks. We presented an effective method to compute the shortest EFMs even in genome-scale networks, as opposed to classic approaches, where EFM analysis cannot be accomplished. A clear advantage of our method in comparison to the classic approaches for EFMs computation is its inherent flexibility. Certainly, the use of optimization enables one to directly search for EFMs that produce/consume a certain metabolite or involve a particular reaction. For this reason the *K*-shortest EFMs is a suitable concept when exploration of a specific subset of EFMs is of interest.

It is beyond the scope of this article to analyse the run-time complexity of the algorithm. Interesting results in that direction have been presented by Acuña *et al.* (2009). Here we have shown by numerical examples that even for genome-scale networks, the *K*-shortest EFMs can be computed in reasonable time.

Our procedure was applied to find the 10-shortest EFMs that produce lysine in the genome-scale metabolic networks of *E.coli* and *C.glutamicum*. The computation of the 10-shortest EFMs in *C.glutamicum* was faster than in *E.coli*, mainly due to the difference in network complexity. The sets of reactions in the computed EFMs can be divided into four parts: catabolism of glucose; anabolism of lysine; ammonium assimilation and a subset responsible for cofactor balancing, when cofactors are set internal metabolites. This classification is in agreement with the presentation in many biochemical textbooks.

The catabolic subset converts glucose into aspartate and pyruvate, precursors of lysine, and plays an important role in cofactor supply, in particular of NADPH. In the genome-scale network of *E.coli*, a variety of pathway combinations exists for glucose catabolism because NADPH can be obtained via a NAD(P) transhydrogenase, whereas in the network of *C.glutamicum* the PP pathway is preponderant for NADPH supply. The cofactor balancing subset is more influenced by the catabolic subset than by the anabolic subset. The latter partially overlaps in the solutions of both organisms and does not change in the 10-shortest EFMs. Shorter routes are clearly favored by the *K*-shortest EFMs method and this fact is evident in the anabolic subset and ammonium assimilation subsets. When cofactors are removed from the balancing constraints, pathways with 100% yield are obtained, hence highlighting the impact of cofactors consumption/supply in lysine synthesis.

Finally, contrary to the widely held belief that the computation of EFMs in large-scale metabolic networks is impossible, the work presented here represents an important step forward.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful comments made by three anonymous reviewers. The groups in which L.F.F., S.S. and C.K. work belong to the Jena Centre for Bioinformatics (JCB).

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.).

Conflict of Interest: none declared.

REFERENCES

Acuña,V. *et al*. (2009) Modes and cuts in metabolic networks: complexity and algorithms. *Biosystems*, **95**, 51–60.

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

Blattner,F.R. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Cohen,G.N. and Saint-Girons,I. (1987) Biosynthesis of threonine, lysine, and methionine. In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella typhimurium—Cellular and Molecular Biology.* Vol. 1, 1st edn, American Society for Microbiology, Washington, pp. 429–444.

Dandekar,T. *et al*. (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems*, **70**, 255–270.

de Graaf,A.A. (2000) Metabolic flux analysis of *Corynebacterium glutamicum*. In Schügerl,K.B. and Bellgardt,K.H. (eds) *Bioreaction Engineering, Modelling and Control.* Springer, New York, pp. 506–555.

Eggeling,L. (1994) Biology of L-lysine overproduction by *Corynebacterium glutamicum*. *Amino Acids*, **6**, 261–272.

Eggeling,L. and Bott,M (2005) Handbook of *Corynebacterium glutamicum*. CRC Press, Boca Raton.

Feist,A.M. *et al*. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Fell,D.A. and Thomas,S. (1995) Physiological control of metabolic flux: The requirement for multisite modulation. *Biochem. J.*, **311**(Pt 1), 35–39.

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

Gagneur,J. and Klamt,S. (2004) Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, **5**, 175.

Inui,M. *et al*. (2004) Metabolic analysis of Corynebacterium glutamicum during lactate and succinate productions under oxygen deprivation conditions. *J. Mol. Microbiol. Biotechnol.*, **7**, 182–196.

Kacser,H. and Acerenza,L. (1993) A universal method for achieving increases in metabolite production. *Eur. J. Biochem.*, **216**, 361–367.

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

Kalinowski,J. *et al*. (2003) The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J. Biotechnol.*, **104**, 5–25.

Kjeldsen,K.R. and Nielsen,J.(2009) *In silico* genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.*, **102**, 583–597.

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

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

Klamt,S. *et al*. (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol.*, **1**, 2.

Koch,I. *et al*. (2005) Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, **12**, 1219–1226.

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

Mavrovouniotis,M.L. *et al*. (1990) Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.*, **36**, 1119–1132.

Meléndez-Hevia,E. *et al*. (1994) Optimization of metabolism: the evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *J. Theor. Biol.*, **166**, 201–220.

Molin,M. *et al*. (2003) Dihydroxyacetone kinases in *Saccharomyces cerevisiae* are involved in detoxification of dihydroxyacetone. *J. Biol. Chem.*, **278**, 1415–1423.

Niederberger,P. *et al*. (1992) A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. *Biochem. J.*, **287**(Pt 2), 473–479.

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

Pfeiffer,T. and Bonhoeffer,S. (2004) Evolution of cross-feeding in microbial populations. *Am. Nat.*, **163**, E126–E135.

Planes,F.J. and Beasley,J.E. (2009) Path finding approaches and metabolic pathways. *Disc. Appl. Math.*, **157**, 2244–2256.

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

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

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

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

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

Schuster,S. *et al*. (2002) Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J. Math. Biol.*, **45**, 153–181.

Schuster,S. *et al*. (2007) Understanding the roadmap of metabolism by pathway analysis. In Weckwerth,W. (ed.), *Metabolomics – Methods and Protocols*, Vol. 358, Human Press, Totowa, New Jersey, pp. 199–226.

Schwarz,R. *et al*. (2007) Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics*, **8**, 313.

Subedi,K.P. *et al*. (2008) Role of GldA in dihydroxyacetone and methylglyoxal metabolism of *Escherichia coli* K12. *FEMS Microbiol. Lett.*, **279**, 180–187.

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

Teusink,B. *et al*. (2006) Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *J. Biol. Chem.*, **281**, 40041–40048.

Tosaka,O. *et al*. (1983) The production of L-lysine by fermentation. *Trends Biotechnol.*, **1**, 70–74.

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

Vallino,J. and Stephanopoulos,G. (1993) Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. *Biotechnol. Bioeng*., **41**, 633–646.

van Winden,W.A. *et al*. (2003) 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**, 75–92.

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

Wendisch,V.F. *et al*. (2006) Metabolic engineering of *Escherichia coli* and *Corynebacterium glutamicum* for biotechnological production of organic acids and amino acids. *Curr. Opin. Microbiol.*, **9**, 268–274.

Wittmann,C. and Becker,J. (2007) The L-lysine story: from metabolic pathways to industrial production. In Wendisch,V.F. (ed.) *Amino acid biosynthesis – Pathways, Regulation and Metabolic Engineering.* Springer, Heidelberg, pp. 39–70.

Yuan,J. *et al*. (2006) Kinetic flux profiling of nitrogen assimilation in *Escherichia coli*. *Nat. Chem. Biol.*, **2**, 529–530.