BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 16 2011, pages 2256–2262
doi:10.1093/bioinformatics/btr367

Systems biology

Advance Access publication June 17, 2011

Decomposing flux distributions into elementary flux modes in genome-scale metabolic networks

Siu Hung Joshua Chan* and Ping Ji

Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Elementary flux mode (EFM) is a fundamental concept as well as a useful tool in metabolic pathway analysis. One important role of EFMs is that every flux distribution can be decomposed into a set of EFMs and a number of methods to study flux distributions originated from it. Yet finding such decompositions requires the complete set of EFMs, which is intractable in genome-scale metabolic networks due to combinatorial explosion.

Results: In this article, we proposed an algorithm to decompose flux distributions into EFMs in genome-scale networks. It is an iterative scheme of a mixed integer linear program. Unlike previous optimization models to find pathways, any feasible solutions can become EFMs in our algorithm. This advantage enables the algorithm to approximate the EFM of largest contribution to an objective reaction in a flux distribution. Our algorithm is able to find EFMs of flux distributions with complex structures, closer to the realistic case in which a cell is subject to various constraints. A case of *Escherichia coli* growth in the Lysogeny broth (LB) medium containing various carbon sources was studied. Essential metabolites and their syntheses were located. Information on the contribution of each carbon source not obvious from the apparent flux distribution was also revealed. Our work further confirms the utility of finding EFMs by optimization models in genome-scale metabolic networks.

Contact: joshua.chan@connect.polyu.hk

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

Received on February 28, 2011; revised on May 29, 2011; accepted on June 10, 2011

1 INTRODUCTION

Immediately after the success in genomics, the next step is to fill the gap between genotypes and phenotypes to have a full picture of biological systems. The genome-scale metabolic network has been playing this role and has been found useful in answering biological questions (Durot *et al.*, 2009). During the past decade, various methods have been proposed to analyze metabolic networks and the constraint-based modeling approach is exceptionally prominent (Durot *et al.*, 2009; Price *et al.*, 2004; Terzer *et al.*, 2009).

One object of particular importance and increasing interest in the approach is the elementary flux mode (EFM). EFM is an important theoretical concept as well as a practical tool in metabolic pathway analysis. Theoretically, it is the minimal operational unit at steady state in metabolic networks satisfying the thermodynamic constraint regarding the reversibility of each reaction (Schuster and Hilgetag, 1994). Practically, EFMs have been applied in a variety of studies: investigating network structures like network robustness (Stelling *et al.*, 2002), dynamic properties of pathways (Steuer *et al.*, 2007), pathway efficiency (Carlson, 2007) and modularity (Yoon *et al.*, 2007), exploring new pathways (de Figueiredo *et al.*, 2009b; Schuster *et al.*, 1999) and hence suggesting rational strain design (Klamt and Gilles, 2004; Trinh and Srienc, 2009; Trinh *et al.*, 2008), predicting mutants' behavior (Zhao and Kurata, 2009a), identifying interactions with other networks (Klamt *et al.*, 2006), etc. (reviewed in Durot *et al.*, 2009; Trinh *et al.*, 2009).

An important interpretation of EFMs is that every flux distribution $\mathbf{v}=[v_1 \ldots v_n]^T$ can be decomposed as a positive sum of a subset of EFMs $\{\mathbf{e}_j\}$ with the 'no cancelation' property that all $\mathbf{e}_j$ have a particular component being zero wherever $\mathbf{v}$ has that component being zero (Schuster *et al.*, 2002). That is,

$$\mathbf{v}=\sum_j w_j\mathbf{e}_j,\ w_j>0 \text{ and } Z(\mathbf{v})\subset Z(\mathbf{e}_j) \text{ for all } j$$

where $Z(\mathbf{v})=\{i|v_i=0\}$ is the index set for the zero components of $\mathbf{v}$. The 'no cancelation' property is unique compared to other generating sets of the flux space like extreme pathways (EPs) (Llaneras and Picó, 2010). In many studies, flux distributions are analyzed by similar decompositions. The first attempt is the α-spectrum to determine the range of attainable weights for each EP (Wiback *et al.*, 2003). This approach is also applicable to EFMs. Later, decomposing flux distributions with minimum sum of weights has been proposed (Schwartz and Kanehisa, 2005) and applied to study yeast glycolysis (Schwartz and Kanehisa, 2006). Decompositions with respect to other objectives have also been adopted to investigate particular cases, for instances, maximum number of active EFMs (Nookaew *et al.*, 2007), minimum relative error (Wang *et al.*, 2007), maximum yield rate (Song and Ramkrishna, 2009) and maximum entropy (Zhao and Kurata, 2009b). There are also other studies analyzing flux distributions by decompositions into EFMs (e.g. Carlson, 2009; Kurata *et al.*, 2007).

These approaches based on decomposing flux distributions, despite the insights they have provided, have their capabilities limited because the calculation always requires the complete set of EFMs *in prior* given a metabolic network, whose computation is notoriously hard due to the combinatorial explosion when the network size grows. This drawback was first studied in Klamt and Stelling (2002) and has been mentioned in many literatures (e.g. Durot *et al.*, 2009; Trinh *et al.*, 2009; Yeung *et al.*, 2007).

*To whom correspondence should be addressed.

To overcome the difficulty, efforts have been put to improve the computational speed and memory demand for computing EFMs. The first algorithm was presented in Schuster *et al.* (1996). Improvements have then been made continuously by introducing the nullspace approach to reduce computational cost (Urbanczik and Wagner, 2005; Wagner, 2004); the binary approach to save a great deal of memory and allow bit operations (Gagneur and Klamt, 2004); the elementary testing by matrix rank to accelerate computations (Klamt *et al.*, 2005); the bit pattern trees to hasten elementary testing (Terzer and Stelling, 2006) and the latest recursive enumeration approach based on bit pattern trees (Terzer and Stelling, 2008). Software implementing these approaches has also been engineered, including *METATOOL* (Pfeiffer *et al.*, 1999; von Kamp and Schuster, 2006), *FluzAnalyzer* (Klamt *et al.*, 2003), *CellNetAnalyzer* (Klamt *et al.*, 2007), *EFMtool* (Terzer and Stelling, 2008). Despite these improvements, these algorithms still cannot cope with genome-scale metabolic networks reconstructed recently, usually consisting of at least a thousand of reactions and metabolites (e.g. Duarte *et al.*, 2007; Feist *et al.*, 2007).

One characteristic in common of the mentioned approaches to calculate EFMs is that they all compute the complete set of EFMs given the network stoichiometry and this is the major reason of the computational infeasibility in genome-scale networks, because the size of the full set grows extremely fast (Klamt and Stelling, 2002). As an alternative, recently there have been attempts to find specific metabolic pathways from stoichiometric information by optimization modeling. They are able to cope with the scale and furthermore to identify pathways with specific requirements. Beasley and Planes (2007) found pathways matched with literature by balancing ‘low presence’ compounds and minimizing the number of reactions involved and ATP consumed. Later by further considering connectivity of compounds, linear pathways were located (Planes and Beasley, 2009). Also, the K-shortest EFMs in genome-scale networks have been calculated successfully by an optimization model (de Figueiredo *et al.*, 2009a). More recently, the K-shortest generating flux modes, a subset of EFMs, have also been investigated by a similar model (Rezola *et al.*, 2010).

In this article, we present an algorithm to find a set of EFMs that decomposes a given flux distribution in a genome-scale metabolic network without finding the complete set of EFMs *in prior*. The algorithm solves optimization models recursively with data stored in a stack structure. Its implementation is highly dynamic. By editing the model being solved, a set of EFMs with certain properties can be found. In addition to the basic decomposition, the algorithm can approximate the EFM with largest contribution to a particular objective reaction in a given flux distribution.

2 METHODS

Given the network stoichiometric matrix $\mathbf{S}$, a flux distribution or equivalently a flux mode, is a vector $\mathbf{v}$ satisfying the steady-state assumption with its irreversible reactions having non-negative fluxes, i.e.

$$\mathbf{v}\in\left\{\mathbf{v}|\mathbf{Sv}=\mathbf{0}, v^{\mathrm{irr}}\geq 0\right\}.$$

An EFM by definition is a flux mode unable to be decomposed as the sum of two other flux modes, whose active reactions are proper subsets of those of that flux mode (Schuster and Hilgetag, 1994). This is called non-decomposability, elementarity or genetic independence. Mathematically, a flux mode $\mathbf{e}$ is an EFM if there are no flux modes $\mathbf{v}_1$, $\mathbf{v}_2$ such that

$$\mathbf{e}=\mathbf{v}_1+\mathbf{v}_2 \text{ with } Z(\mathbf{e})\underset{\neq}{\subset}Z(\mathbf{v}_1) \text{ and } Z(\mathbf{e})\underset{\neq}{\subset}Z(\mathbf{v}_2).$$

In this article, if for a flux mode $\mathbf{e}$, such $\mathbf{v}_1$, $\mathbf{v}_2$ exist, they are said to be bounded by $\mathbf{e}$. The algorithm proposed exploits this decomposability. In what follows, first, an optimization model to check the decomposability of a given flux mode is formulated, followed by the algorithm integrating the model to decompose flux distributions. Then, modifications of the algorithm to approximate EFMs of largest contributions are presented.

2.1 Decomposability check

Suppose there is a metabolic network consisting of m metabolites and n reactions with stoichiometric matrix $\mathbf{S}=[S_{ij}]$. All reactions are assumed irreversible. This can be done in reality by replacing a reversible reaction by two irreversible reactions with stoichiometry negative to each other.

The decomposability of a given flux mode $\mathbf{v}=[v_1\cdots v_n]^T$ is determined by examining whether a flux mode $\mathbf{p}=[p_1\cdots p_n]^T$ bounded by $\mathbf{v}$ can be found. First, $\mathbf{p}$ must satisfy the usual steady-state condition:

$$\sum_{j=1}^{n}S_{ij}p_j=0 \text{ for } i=1,\ldots,m \tag{1}$$

Each reaction is assigned with a binary integer variable a_j specifying the on/off condition by the following constraints:

$$a_j\leq p_j \text{ for } j=1,\ldots,n \tag{2}$$

$$p_j\leq M\delta_j a_j \text{ for } j=1,\ldots,n \tag{3}$$

where M is a large positive number and $\delta_j=1$ if $v_j>0$ and $\delta_j=0$ otherwise. If $a_j=0$, from (3), we have $p_j=0$ and thus reaction j is not involved, else if $a_j=1$, from (2) we have $p_j\geq 1$ and reaction j must has a positive flux. Constraints similar to (2) and (3) have been employed to find pathways before (Beasley and Planes, 2007; de Figueiredo *et al.*, 2009a; Rezola *et al.*, 2010). The difference of our approach lies in the introduction of δ_j. If $\delta_j=0$, a_j and p_j are forced to be zero by (2) and (3). In other words, if $v_j=0, p_j=0$. This implies $Z(\mathbf{v})\subset Z(\mathbf{p})$, necessary for $\mathbf{p}$ to be bounded by $\mathbf{v}$. For the large number M, it can be interpreted as the greatest stoichiometric ratio allowed within the flux mode $\mathbf{p}$, so it should be sufficiently large to allow all flux modes bounded by $\mathbf{v}$ to satisfy (2) and (3).

To ensure $\mathbf{p}$ is non-trivial, i.e. has at least one non-zero flux, we have:

$$\sum_{j=1}^{n}a_j\geq 1 \tag{4}$$

For $\mathbf{p}$ to be properly bounded by $\mathbf{v}$, the number of non-zero fluxes in $\mathbf{p}$ must be less than the number of non-zero fluxes in $\mathbf{v}$, so we have:

$$\sum_{j=1}^{n}a_j\leq\sum_{j=1}^{n}\delta_j-1 \tag{5}$$

Finally, since all reactions are irreversible, all fluxes must be positive:

$$p_j\geq 0 \text{ for } j=1,\ldots,n \tag{6}$$

Constraints (1–6) suffice to define $\mathbf{p}$ as a flux mode bounded by $\mathbf{v}$. Hence, every objective can work. We simply choose maximizing zero:

$$\max 0 \tag{7}$$

Equations (1)–(7) form a mixed integer linear program, called Decomposability Check (DC). Given a flux mode $\mathbf{v}$, if any feasible solution can be found, then $\mathbf{v}$ is decomposable. Otherwise, we conclude that $\mathbf{v}$ is an EFM.

Different from de Figueiredo *et al.* (2009a), $\mathbf{p}$ is not required to be integers though this may lengthen computational time because in some of our investigations, fractional stoichiometric coefficients are involved, like the biomass composition, and turning these coefficients into integers will make them too large in magnitude and inconvenient for computation.

2.2 Decomposition of flux distributions

The algorithm to decompose flux distributions is an iterative scheme of DC. Flux modes involved are stored in a stack structure. The decomposability of the given flux distribution is first examined by DC. If it is decomposable, the flux mode bounded by it that is returned by DC will be stacked up and become the current flux mode. Else if it is non-decomposable, an EFM is reached and it leaves the stack. The EFM found is then used to update each intermediate flux mode by subtracting a scalar multiple of the EFM. After updating, intermediate flux modes unable to contribute to the first flux mode are removed. This procedure is repeated until all flux modes leave the stack. A set of EFMs decomposing the flux distribution is then obtained.

Let N be the number of flux modes in the stack, K be the number of EFMs found. The s-th flux mode in the stack is denoted as $\mathbf{fm}_s$ with the flux of the j-th reaction being fm_{sj}. $\mathbf{efm}_k$ and efm_{kj} are similarly defined for the k-th EFM. The steps of the algorithm can be summarized as follows:

Step 0. Initialize with $N=1$, $K=0$ and $\mathbf{fm}_1=\mathbf{v}$. Go to Step 1.
Step 1. Solve DC with $\mathbf{fm}_N$ as input. If there is a feasible solution $\mathbf{p}$, go to Step 2. Otherwise, go to Step 3.
Step 2. Update N by $N+1$. Set $\mathbf{fm}_N=\mathbf{p}$. Go to Step 1.
Step 3. Update K by $K+1$. Set $\mathbf{efm}_K=\mathbf{fm}_N$.
If $N=1$, terminate the algorithm, else go to Step 4.
Step 4. For $s=1,\ldots,N-1$, update $\mathbf{fm}_s$ by $\mathbf{fm}_s-r_s^K\times\mathbf{efm}_K$ where $r_s^K=\min\limits_j\left\{\mathrm{fm}_{sj}\Big/\mathrm{efm}_{Kj}\Big|\mathrm{efm}_{Kj}>0\right\}$. Go to Step 5.
Step 5. Remove $\mathbf{fm}_N$ from the stack. If $N>2$, for $s=2,\ldots,N-1$, check if $Z(\mathbf{fm}_1)\underset{\neq}{\subset}Z(\mathbf{fm}_s)$. If no, remove $\mathbf{fm}_s$ from the stack. Set N to be the current size of the stack. Go to Step 1.

The algorithm can be interpreted as decomposing $\mathbf{fm}_1$ repeatedly until $\mathbf{fm}_1$ becomes an EFM. Step 1 checks the decomposability of the current flux mode $\mathbf{fm}_N$. If a feasible solution exists, the flux mode is not an EFM and it is replaced by a new flux mode bounded by it which is found by DC as indicated in Step 2. The procedure is repeated until an EFM is reached.

Once an EFM is found, in Step 3, first we check whether it is the only flux mode in the stack. If it is, this means it is the last EFM and the algorithm is terminated. Otherwise, it will be used to update all flux modes in the stack as in Step 4. Note that after each updating step, $\mathbf{fm}_1$ is the flux mode remaining to be further decomposed. This updating process serves to eliminate the largest possible fluxes that are able to be contributed by the EFM found in each preceding flux mode. The number of non-zero entries in each flux mode is diminished. This step can accelerate computations because there are less non-zero entries needed to be dealt with in each flux mode. In practice, Step 4 can be efficiently performed by matrix multiplications.

After updating flux modes, some intermediate flux modes may have positive entries that have become zero in $\mathbf{fm}_1$. The 'no cancellation' property states that these flux modes cannot contribute to $\mathbf{fm}_1$ anymore, so they are removed from the stack in Step 5, which is carried out by bit operations. In fact, they can still contribute to $\mathbf{v}$, but it turns out that Steps 4 and 5, besides saving memory and computational cost, guarantee that the finally resulting set of EFMs has the following four nice properties.

The first two properties are the 'denseness' and 'uniqueness' of the solution. No EFM found is redundant and each EFM has a unique positive weight in the decomposition. The other two properties of greater theoretical interest are the linear independence and systemic independence of the solution set, i.e. the set of solution EFMs forms a linear basis as well as a convex basis for the flux distribution. These four properties, which follow from the fact that our algorithm decomposes a flux distribution by stepwise finding an EFM and reducing the flux mode, also hold in the original flux space in which a reversible reaction is not transformed into two irreversible reactions. They together give an exact role for each EFM in the solution. The cooperation between different pathways can be clearly revealed. This brings more exact biological interpretations. These properties are proved and explained with an example in detail in the Supplementary Material.

There are two major differences of the proposed algorithm compared with previous optimization models to find pathways. First, it is an iterative scheme to solve optimization models instead of solving a single optimization model. Second, any feasible solution can finally become an EFM and optimality is not necessarily needed. Every optimization objective works. In what follows, we propose one for a specific application.

2.3 Approximation of EFMs of largest contributions

In flux balance analysis (FBA), cellular metabolism is assumed to achieve an optimal state with respect to an objective like the maximum growth rate (Feist *et al.*, 2007) and ATP production (reviewed in Schuetz *et al.*, 2007).

Conversely, for a flux distribution (maybe experimentally measured, or simulated by methods other than FBA), if a biologically reasonable objective of the cell can be assumed, it will be insightful to decompose the flux distribution into EFMs that have considerable contributions to that objective. By 'contribution', we mean the flux of the objective reaction provided by an EFM in the flux distribution. Finding largest contributing EFMs can reveal principal operational modes in cells. We applied our algorithm to approximate such decompositions with only two modifications required. The first is the replacement of objective (7) by:

$$\max p_{j_0} \tag{7'}$$

where j_0 is the objective reaction that we are interested. The resulting solution flux mode then has the greatest flux at reaction j_0. Nonetheless, the flux mode can possibly contribute little to $\mathbf{fm}_1$. This comes to the second modification which replaces constraint (3) by:

$$p_j \le M'\mathrm{fm}_{1j}a_j \quad \text{for } j=1,\ldots,n \tag{3'}$$

Here δ_j is replaced by fm_{1j}. This takes the flux values of $\mathbf{fm}_1$ into account. Intuitively, the solution flux mode should have a better contribution because upper bounds for fluxes are not the same but proportional to the fluxes of $\mathbf{fm}_1$. In fact, the stepwise solution has the maximum contribution to the objective reaction flux in $\mathbf{fm}_1$ among all flux modes bounded by $\mathbf{fm}_1$. This also forms the rationale of applying the algorithm to approximate EFMs with largest contributions: in each step, the best flux mode bounded by the current flux mode is found until an EFM is reached. Hence, it is a greedy approach. For the large number M', it should be chosen large enough to properly scale $\mathbf{fm}_1$ to allow all feasible flux modes.

2.4 Implementations

Both versions of the optimization model DC [Version 1: objective (7) subject to constraints (1–6); Version 2: objective (7′) subject to constraints (1–2), (3′), (4–6)] were solved by ILOG CPLEX®. As for the algorithm, all the data and operations other than solving optimization models were processed in MATLAB®. For hardware, all computations were performed in a desktop computer with a 2.67 GHz CPU and 24 GB of RAM.

3 RESULTS

We applied our algorithm to three metabolic networks of different sizes. The first is the sample network used to demonstrate regulatory FBA (Covert *et al.*, 2001) and study α-spectrum (Wiback *et al.*, 2003). The second is the core metabolic network of *E.coli* K-12 MG1655 iAF1260 (Feist *et al.*, 2007). It is more realistic with a size considerably larger than the first one. The two networks were used as benchmarks to validate our algorithm.

To highlight the usefulness of our algorithm complementary to existing methods, a computational experiment was then conducted. Flux distributions of optimal growth rate subject to various substrate availability in the complete *E.coli* MG1655 iAF1260 metabolic network (Feist *et al.*, 2007) were used to compare the performance of our algorithm and *EFMtool* (Terzer and Stelling, 2008), which should be the most efficient algorithm to find full sets of EFMs up to the best of our understanding. Finally, a flux distribution simulating the growth of *E.coli* in the LB medium (Baev *et al.*, 2006a, b) was studied.

3.1 Benchmarks

For each of the two networks for benchmarks, we first randomly sampled 2000 flux modes (Almaas *et al.*, 2004), implemented with the COBRA Toolbox (Becker *et al.*, 2007) and decomposed them by our algorithm. We then checked whether the solutions were true EFMs able to be located by *EFMtool*. For the approximation of EFMs of largest contributions, we chose the biomass production or equivalently growth rate as the objective reaction. We checked the rankings of the EFMs found among the full set of EFMs. The detailed test procedure can be found in the Supplementary Material.

3.1.1 Sample metabolic network There are 20 reactions and 19 metabolites in the sample network. The number of EFMs found by *EFMtool* was 82, the same as in Wiback *et al*. (2003). In their paper, an algorithm for EPs was used, but since each reversible reaction was treated as two irreversible reactions, the result coincides with the set of EFMs (Llaneras and Picó, 2010).

In average, Version 1 of the algorithm took 5 s to decompose a flux distribution and Version 2 took 6 s. It was verified that all solutions belong to the set of EFMs computed by *EFMtool*.

For Version 2, in 98% of the 2000 samples, EFMs of maximum contributions to the growth rate in the flux distributions are the first EFMs found. The corresponding percentage for Version 1 is only ~15%. All other solutions by Version 2 except one contain EFMs of the second or third largest contributions.

3.1.2 Core E.coli K-12 MG1655 iAF1260 metabolic network The core *E.coli* metabolic network contains 95 reactions and 72 metabolites. By restricting substrates for uptake of glucose, phosphate, CO_2, H^+, H_2O, NH_4 and O_2 only, over 100 000 EFMs were found by *EFMtool*. The number is comparable to previous studies with similar configurations (Klamt and Stelling, 2002). One issue complicating the computation is the fractional stoichiometry in the biomass reaction estimated from molecular content (Feist *et al.*, 2007). Magnitudes of coefficients in the reaction highly vary. The ratio of the greatest to the smallest is 840.

In average, Version 1 of the algorithm took 12 s to finish and Version 2 took 18 s. All solutions were verified to be true EFMs.

For Version 2, the greatest contributing EFMs were first found in 97% of the 2000 trials. Among all, 80% of the remaining solutions contain EFMs whose contributions rank top 10%. The success rate is significant in view of the number of different contributions by all EFMs in each sample, which is over 7000 in average.

The benchmark results for the two networks different in size show that our algorithm can decompose flux distributions into sets of EFMs. Furthermore, it is capable of approximating the best contributing EFM in a flux distribution with respect to a reaction.

3.2 *E.coli* K-12 MG1655 iAF1260 metabolic network

The complete *E.coli* K-12 MG1655 iAF1260 genome-scale metabolic network reconstructed by Feist *et al.* (2007) is a huge network with 1039 metabolites and 2382 reactions, of which 852 are reversible. After compartmentalization and transformation, there are 1668 metabolites and 3234 irreversible reactions in total.

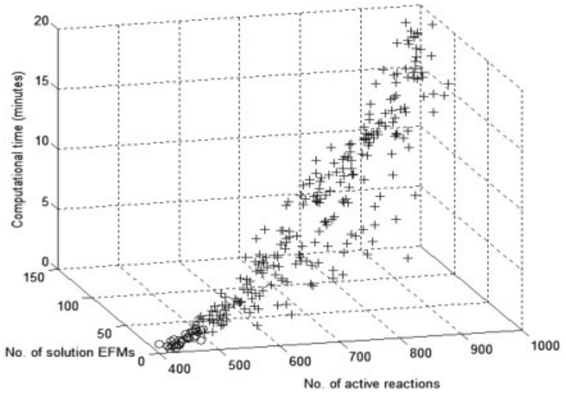

Fig. 1. The computational time of the proposed algorithm against the number of active reactions and the number of EFMs in the solution found. ‘o’ represents the case in which the set of EFMs of the subnetwork of active reactions can be calculated by *EFMtool*. ‘+’ represents the corresponding case in which the set cannot be calculated by *EFMtool*.

3.2.1 Computational experiment In cell culture experiments, minimal media containing one carbon source and necessary inorganic compounds only are often preferred due to the ease in analysis. Correspondingly, flux distributions of optimal growth rate simulating these cases consist of only few EFMs regardless of the network complexity. In this case, existing tools are able to find the set of EFMs that contributes to a flux mode by considering the subnetwork formed by the active reactions in the flux distribution (called subnetwork of a flux distribution). Realistic metabolic fluxes are, however, shaped by factors like gene regulation and in particular heterogeneous nutrients (Zamboni, 2010). Structures of flux distributions are thus more complicated and existing tools may not cope with them. To test the capability of our algorithm to analyze these cases, we performed a computational experiment.

In the *E.coli* MG1655 iAF1260 metabolic network, there are 299 possible uptake substrates. In each trial, in addition to inorganic compounds necessary for growth, a random set of carbon sources, whose uptake rates ≤ 1 mmol gDW^{-1} h^{-1}, was available and a flux distribution of optimal growth rate was determined. It was then decomposed by Version 2 of our algorithm with growth rate as the objective. Meanwhile, we tried to find the full set of EFMs of the subnetwork of the flux distribution by *EFMtool*.

In the computational experiment, the number of active reactions N_r, partially reflecting the complexity of a flux distribution, increases with the number of carbon sources consumed N_c linearly. In general, when $N_c \geq 30$, *EFMtool* was unable to find the full set of EFMs of the subnetwork of the flux distribution. In contrast, our algorithm succeeded to decompose every test flux distribution.

In certain cases with large N_r in the flux distributions but a small number of contributing EFMs in the solution found, N_{efm} (<50), *EFMtool* was unable to find the sets of EFMs of the subnetworks while our algorithm could find EFMs decomposing the flux distributions in relatively short time (Fig. 1, ‘+’s with small N_{efm}). This corresponds to the situation in which a flux distribution

Table 1. List of carbon sources available in the LB medium for uptake

Sugar	Sugar alcohol	Organic acid	Amino acid	Nucleotide
L-Arabinose (arab-L)	D-Mannitol (mnl)	D-lactate (lac-D)	L-Arginine (arg-L)	Cytidine monophosphate (cmp)
L-fucose (fuc-L)	Glycerol (glyc)	L-lactate (lac-L)	L-aspartate (asp-L)	Adenosine monophosphate (amp)
N-Acetyl-glucosamine (acgam)			L-Glutamine (gln-L)	Guanosine monophosphate (gmp)
D-Galactose (gal)			L-histidine (his-L)	Uridine monophosphate (ump)
D-Mannose (man)			Glycine (gly)	Inosine monophosphate (imp)
Melibiose (melib)			L-Methionine (met-L)	
D-glucosamine (gam)			L-serine (ser-L)	
L-Rhamnose (rmn)				
Trehalose (tre)				

Abbreviations adopted from Feist *et al.* (2007) are written in brackets.

having an enormous number of contributing EFMs can actually be represented as a convex sum of only a few of those EFMs. Our algorithm is very useful and efficient in this case. Moreover, the memory demand is not expensive as the most complicated cases in the computational experiment can also be solved by our algorithm in a computer with 2 GB memory only.

The complexity of our algorithm was also examined. The computational time increases with N_{efm} as well as N_r (Fig. 1). Regression indicates a satisfactory linear relationship between the computational time and the product $N_{efm} \times N_r$ ($R^2 = 0.97$).

The computational experiment showed that our algorithm is more advantageous than existing methods in analyzing flux distributions with complex structures, which are closer to realistic cases in which a cell is subject to complicated factors. Details of the test and test results are provided in the Supplementary Material.

3.2.2 A case of growth in the LB medium The LB medium, or usually called Luria-Bertani medium, is one of the most common complex media used in bacterial growth. It is composed of tryptone, yeast extract and NaCl. In their studies, Baev *et al.* monitored the utilization of sugars, alcohols, organic acids (Baev *et al.*, 2006a) and amino acids, peptides, nucleotides (Baev *et al.*, 2006b) indirectly by transcriptional microarrays during the growth of *E.coli* MG1655 in the LB medium. Interestingly, during about 3–5 h of fermentation, simultaneous assimilation of a set of carbon sources was observed (Baev *et al.*, 2006a). Combining the result in Baev *et al.* (2006b), a large number of sugars, amino acid and other carbon sources were absorbed during the time period. We simulated a flux distribution describing the metabolism of *E.coli* under such conditions by optimizing the growth rate in FBA and allowing the list of carbon sources in Table 1 and some inorganic compounds for uptake. Maximum oxygen uptake rate (18.5 mmol gDW/h) and ATP maintenance (ATPM) cost (8.39 mmol gDW/h) under growth condition determined experimentally in Feist *et al.* (2007) were adopted. All maximum uptake rates of carbon sources were set to be 0.5 mmol gDW/h for simplicity because information on enzyme capacities was not given. We denote the simulated flux distribution by f_0 to avoid confusion in the following discussion.

f_0 consists of 493 reactions and consumes all 25 carbon sources. To understand the complexity and the difference of f_0 due to multiple sources, we found the maximum flux contribution of each individual carbon source to f_0 by maximizing the sum of fluxes, setting f_0 as the upper bound and allowing only that source for uptake (Supplementary Material for details). Surprisingly, only three non-zero flux modes consuming L-fucose, D-lactate and L-lactate, respectively, were resulted, each verified to be EFMs by *EFMtool*. They together provide the whole ATPM flux in f_0 and no single carbon source is independently consumed for growth.

Meanwhile, we tested whether each carbon source can be metabolized into biomass by maximizing the growth rate and allowing the uptake of one source only. All carbon sources except L-histidine and L-methionine can generate growth independently.

From these two results, we concluded that first, more efficient growth can be achieved by the simultaneous assimilation of the carbon sources in Table 1 if they are the only sources available. This may suggest a possible reason for the switch of the mode of assimilation from a sequential one to a simultaneous one as observed in Baev *et al.* (2006a). Second, f_0 cannot be simply decomposed into flux modes of individual carbon sources.

f_0 was then simplified into f_1 by subtracting the three EFMs found. We tried to find the set of EFMs of the subnetwork of f_1 by *EFMtool* but it was unsuccessful due to insufficient memory. We then decomposed f_1 by Version 2 of our algorithm. In all, 22 EFMs were found, all verified to be true EFMs by EFMtool. The first EFM found, also the one with the largest growth rate in the solution, accounts for 19% of the growth rate.

All EFMs of growth consume multiple sources simultaneously, at least 12. The EFMs consist of quite similar sets of reactions. In fact, among the 482 active reactions in f_1, 376 reactions were found to be shared by all EFMs, called the ‘backbone’ reactions (Fig. 2). They can be important reactions for optimal growth on the medium. The structures are complicated, as expected from the detailed biomass composition containing 63 metabolites. Various biochemical pathways are involved: glycolysis, pentose phosphate pathway, non-mevalonate pathway, cell envelope biosynthesis, glycerophospholipid metabolism, lipopolysaccharide biosynthesis, biosynthesis of different amino acids, cofactor and prosthetic group, etc. They are connected by branch point metabolites like pyruvate and chorismate. Interestingly, in the backbone, besides the uptake of eight extracellular carbon sources, some cytosolic metabolites are always synthesized in each EFM and then enter into the backbone acting as source nodes, including L-aspartate, L-serine, D-glucosamine 6-phosphate, glyceraldehyde 3-phosphate, D-glucose 6-phosphate, pyruvate, alpha-D-ribose 1-phosphate, uridine, D-xylulose 5-phosphate. Meanwhile, outside the backbone, no single extracellular carbon source is used by all EFMs.

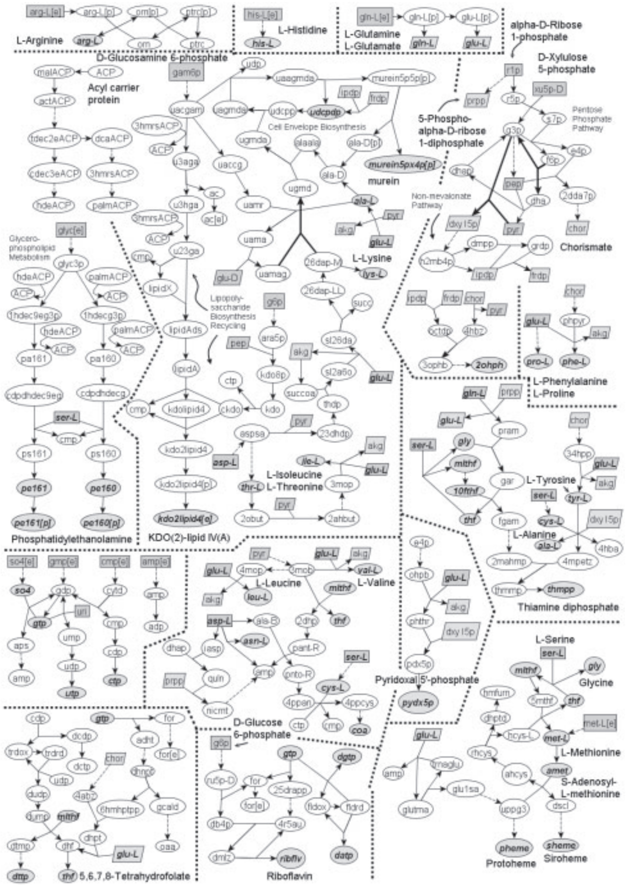

Fig. 2. 'Backbone' reactions for f_1. Sources entering the backbone are in squares. Metabolites involved in many pathways are in parallelograms. Metabolites comprising biomass are in bold and italic text. Independent intersecting edges are bolded for clarity. Small molecules, cofactors and their reactions are not shown for simplicity. Abbreviations are the same as in Feist *et al.* (2007).

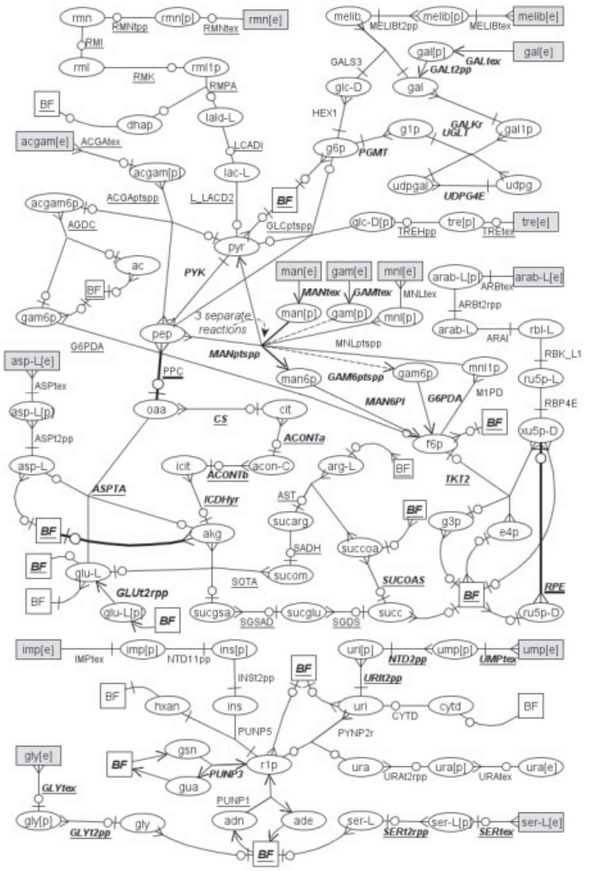

Fig. 3. Reactions outside the backbone of the top five EFMs contributing to growth. Extracellular sources are in squares. 'BF' stands for 'backbone fluxes'. Legend for EFMs (source to sink is from left to right): first: connecting lines vertically; second: connecting lines with open circle; third: connecting lines with arrow; fourth: reaction name bold and italic; fifth: reaction name underscored. Independent intersecting edges are bolded for clarity. Small molecules, cofactors and their reactions are not shown for simplicity. Abbreviations are the same as in Feist *et al.* (2007).

This means that these cytosolic sources of the backbone must be first synthesized regardless of the carbon sources. Hence, the ability to synthesize them may reflect the efficiency of the cell's growth. Of the five EFMs with the largest growth rate (accounting for 80% of the growth rate in total), the reactions outside the backbone are shown in Figure 3, from which different pathways to synthesize those necessary cytosolic metabolites entering the backbone can be seen. All EFMs are in general different arrangements of these pathways to strike delicate balance to produce biomass.

We further looked into the contribution of each carbon source. The 'marginal relative contribution' of each carbon source was calculated by summing up the relative contribution by that source in each EFM (Supplementary Material). The carbon source with largest relative contribution to biomass is *N*-acetyl-glucosamine with over 5.4% and L-histidine has the lowest contribution with only 0.9%. The information cannot be obtained from f_0 or f_1 alone in which nearly all uptake fluxes of carbon sources are equal. The value of the marginal relative contribution, however, seems to change with the non-unique decomposition by EFMs. A further analysis will be needed to formally address this issue.

It is remarked that the above analysis focuses on f_0 and f_1. In-depth studies are required to prove the claims from it. For example, alternative optimal flux distributions and different maximum uptake rates chosen may alter the results on the dependence of carbon sources and the necessity of certain metabolites. Still, the resolution brought by the present decomposition has revealed some information not observed from the apparent flux distribution.

4 CONCLUSION

In this article, we presented an algorithm which solves MILPs iteratively to decompose a flux distribution into EFMs in genome-scale metabolic networks. One advantage of the algorithm is the elementarity of solutions independent of the optimality of MILPs. The objective function can then be used for other purposes and we suggested approximating the largest contributing EFMs in flux distributions.

We demonstrated the ability of the algorithm to decompose flux distributions into EFMs and approximate EFMs with largest contributions to flux distributions in genome-scale metabolic networks. In particular, complementary to existing methods, it can find EFMs of flux distributions with complex structures. The case of growth of *E.coli* on the LB medium gives an exemplary flux distribution with a complex structure in which a simultaneous mode of assimilation of carbon sources is seen. By our algorithm, essential reactions and metabolites, contribution of carbon sources of the flux distribution were studied under the resolution of the decomposition by EFMs. We conclude that the algorithm can facilitate metabolic pathway analysis in genome-scale metabolic networks. It provides an analytic method that prepares for the future breakthrough in experimental techniques to measure *in vivo* fluxes in a huge scale.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful comments made by the two anonymous reviewers.

Funding: The Hong Kong Polytechnic University(project G-RPCC).

Conflict of Interest: none declared.

REFERENCES

Almaas,E. *et al.* (2004) Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, **427**, 839–843.

Baev,M.V. *et al.* (2006a) Growth of *Escherichia coli* MG1655 on LB medium: monitoring utilization of sugars, alcohols, and organic acids with transcriptional microarrays. *Appl. Microbiol. Biotechnol.*, **71**, 310–316.

Baev,M.V. *et al.* (2006b) Growth of *Escherichia coli* MG1655 on LB medium: monitoring utilization of amino acids, peptides, and nucleotides with transcriptional microarrays. *Appl. Microbiol. Biotechnol.*, **71**, 317–322.

Becker,S.A. *et al.* (2007) Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols*, **2**, 727–738.

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

Carlson,R.P. (2007) Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics*, **23**, 1258–1264.

Carlson,R.P. (2009) Decomposition of complex microbial behaviors into resource-based stress responses. *Bioinformatics*, **25**, 90–97.

Covert,M.W. *et al.* (2001) Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.*, **213**, 73–88.

de Figueiredo,L.F. *et al.* (2009a) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, **25**, 3158–3165.

de Figueiredo,L.F. *et al.* (2009b) Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, **25**, 152–158.

Duarte,N.C. *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA*, **104**, 1777–1782.

Durot,M. *et al.* (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.*, **33**, 164–190.

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.

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

Klamt,S. and Gilles,E.D. (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics*, **20**, 226–234.

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

Klamt,S. *et al.* (2003) FluxAnalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics*, **19**, 261–269.

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

Klamt,S. *et al.* (2006) A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, **7**, 56.

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

Kurata,H. *et al.* (2007) Integration of enzyme activities into metabolic flux distributions by elementary mode analysis. *BMC Syst. Biol.*, **1**, 31.

Llaneras,F. and Picó,J. (2010) Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J. Biomed. Biotechnol.*, **2010**, 753904.

Nookaew,I. *et al.* (2007) Identification of flux regulation coefficients from elementary flux modes: a systems biology tool for analysis of metabolic networks. *Biotechnol. Bioeng.*, **97**, 1535–1549.

Pfeiffer,T. *et al.* (1999) METATOOL: for studying metabolic networks. *Bioinformatics*, **15**, 251–257.

Planes,F.J. and Beasley,J.E. (2009) An optimization model for metabolic pathways. *Bioinformatics*, **25**, 2723–2729.

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

Rezola,A. *et al.* (2010) Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, **27**, 534–540.

Schuetz,R. *et al.* (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.*, **3**, 119.

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.* (1996) Elementary modes of functioning in biochemical networks. In Cuthbertson,R., Holcombe,M. and Paton,R. (eds) *Computation in Cellular and Molecular Biological Systems*. World Scientific, Singapore, pp. 151–165.

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

Schwartz,J.M. and Kanehisa,M. (2005) A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics*, **21** (Suppl. 2), ii204–ii205.

Schwartz,J.M. and Kanehisa,M. (2006) Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics*, **7**, 186.

Song,H.S. and Ramkrishna,D. (2009) Reduction of a set of elementary modes using yield analysis. *Biotechnol. Bioeng.*, **102**, 554–568.

Stelling,J. *et al.* (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Steuer,R. *et al.* (2007) From structure to dynamics of metabolic pathways: application to the plant mitochondrial TCA cycle. *Bioinformatics*, **23**, 1378–1385.

Terzer,M. and Stelling,J. (2006) Accelerating the computation of elementary modes using pattern trees. In Bucher,P. and Moret,B.M.E. (eds) *WABI*, Vol. 4175 of *Lecture Notes in Computer Science*. Springer, Heidelberg, pp. 333–343.

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

Terzer,M. *et al.* (2009) Genome-scale metabolic network. *Wiley Interdis. Rev. Syst. Biol. Med.*, **1**, 285–297.

Trinh,C.T. and Srienc,F. (2009) Metabolic engineering of *Escherichia coli* for efficient conversion of glycerol to ethanol. *Appl. Environ. Microbiol.*, **75**, 6696–6705.

Trinh,C.T. *et al.* (2008) Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Appl. Environ. Microbiol.*, **74**, 3634–3643.

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.

Urbanczik,R. and Wagner,C. (2005) An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, **21**, 1203–1210.

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

Wagner,C. (2004) Nullspace approach to determine the elementary modes of chemical reaction systems. *J. Phys. Chem. B*, **108**, 2425–2431.

Wang,Q. *et al.* (2007) Metabolic network properties help assign weights to elementary modes to understand physiological flux distributions. *Bioinformatics*, **23**, 1049–1052.

Wiback,S.J. *et al.* (2003) Reconstructing metabolic flux vectors from extreme pathways: defining theα-spectrum. *J. Theor. Biol.*, **224**, 313–324.

Yeung,M. *et al.* (2007) Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics*, **8**, 363.

Yoon,J. *et al.* (2007) Modular decomposition of metabolic reaction networks based on flux analysis and pathway projection. *Bioinformatics*, **23**, 2433–2440.

Zamboni,N. (2010) 13C metabolic flux analysis in complex systems. *Curr. Opin. Biotechnol.*, **22**, 103–108.

Zhao,Q. and Kurata,H. (2009a) Genetic modification of flux for flux prediction of mutants. *Bioinformatics*, **25**, 1702–1708.

Zhao,Q. and Kurata,H. (2009b) Maximum entropy decomposition of flux distribution at steady state to elementary modes. *J. Biosci. Bioeng.*, **107**, 84–89.