Computation of elementary flux modes by column generation

Homero Basso[1] and Rodolfo Dondo[1 2]

[1] Laboratorio de Fermentaciones, Facultad de Ciencias Biológicas y Bioquímica, U.N.L, Paraje El Pozo s/n, Santa Fe, 3000 Santa Fe, Argentina
[2] Instituto de Desarrollo Tecnológico para la Industria Química (U.N.L. - Conicet), Güemes 3450, 3000 Santa Fe, Argentina
rdondo@santafe-conicet.gov.ar

Abstract. Identification of elementary flux modes in genome-scale metabolic networks is difficult due to the combinatorial nature of the problem. However, in systems biology, constraint-based modeling utilizes only a few relevant elementary flux modes. So, their computation using optimization is emerging as a recent trend. We proposed an algorithm based on the column generation paradigm for computing elementary flux modes that allow for massive computational-time savings compared to previous approaches. The algorithm proved capable of efficiently generating elementary flux modes for genome-scale metabolic networks with thousands of intracellular reactions.

Keywords: column generation, elementary flux modes, metabolic networks.

Cómputo de modos elementales de flujo mediante generación de columnas

Resumen. La identificación de modos de flujo elementales en redes metabólicas a escala genómica es dificultosa debido a la naturaleza combinatoria del problema. Sin embargo, en biología de sistemas, el modelado basado en restricciones utiliza solo unos pocos modos de flujo elementales relevantes. Por ello, su cálculo mediante optimización se está convirtiendo en una tendencia reciente. Desarrollamos un algoritmo basado en el paradigma de generación de columnas para el cálculo de modos de flujo elementales que permite un ahorro considerable de tiempo computacional en comparación con enfoques previos. El algoritmo demostró ser capaz de generar eficientemente modos de flujo elementales para redes metabólicas a escala genómica con miles de reacciones intracelulares.

Palabras clave: generación de columnas, modos elementales de flujo, redes metabólicas.

1 Introduction

Elementary flux modes analysis analysis (EFMA) is a useful tool for metabolic engineering but the computation of all elementary flux modes (EFMs) from a metabolic network is very resources demanding (Zanghellini et al., 2013). Stoichiometric constraints that must be satisfied in a metabolic network at steady-state lead to the definition of the so called "flux cone" which comprises all possible steady-state flux distributions over the network. Given the steady-state assumption for flux distributions on a metabolic network, its metabolites can be classified as *internal* metabolites or *external* ones. Concentrations of internal metabolites are assumed to be in steady-state while the concentration external metabolites are not since they behave as sources or sinks of fluxes. If there is at least one EFM or a combination of EFMs that connects an external substrate and an external product, the bioproduction of such a product is possible and the stoichiometric efficiency of conversion is computable by following the chain of reactions of the EFM. There are several tools to compute EFMs that are implementations of the double description method (Fukuda and Prodon, 1996) which successively generates EFM-candidates by pairwise combinations of existing EFMs followed by verification that each EFM candidate has not been yet identified. These methods suffer from an important drawback because the number of EFMs grows exponentially with the network size. For instance, more than two million EFMs have been reported for the metabolic network describing the central *E. coli* metabolism, which contains just 110 reactions (Gagneur and Klamt, 2004). Therefore, mixed integer-linear programs (MILP) were used with the purpose of identifying a subset of useful EFMs. Later, decomposition methods aimed at alleviating the computational burden associated to the solution of MILPs were proposed. For instance, Oddsdottir et al. (2015) presented a decomposition method that dynamically generates a subset of EFMs by solving a master quadratic problem that fits measured external fluxes and a linear subproblem that generates an EFM. In this way, we develop a column-generation (CG) procedure aimed at computing a subset of EFMs.

2 Definitions

The topology of a metabolic network is characterized by an $m \times n$ stoichiometric matrix, $\boldsymbol{S} = \{S_{ij}\}$, where m and n are the number of metabolites and reactions respectively. The value S_{ij} represents the stoichiometric coefficient of metabolite i in metabolic reaction j. For instance, S_{ij} is positive if metabolite i is produced in reaction j, negative if i is consumed in reaction j and zero if metabolite i does not participates into reaction j. A non-trivial flux vector $\boldsymbol{v} \in \mathrm{R}^n$ is an admissible flux mode if $\boldsymbol{v} \neq \mathbf{0}$ and it satisfies the steady-state condition $\boldsymbol{S}\,\boldsymbol{v} = \mathbf{0}$. Furthermore, all fluxes must be considered as nonnegative irreversible; i.e. $\boldsymbol{v} \geq \mathbf{0}$. A flux mode is called an EFM whenever the flux cannot operate in steady-state if any of its reactions is deleted. In such a case, the whole EFM would be inactive (Schuster et al., 2002). EFMs are non-decomposable steady-state pathways through a metabolic network and constitute minimal functional building blocks defined as vectors in such a network (Zanghellini et al., 2013). A flux vector $\boldsymbol{e_k}$ that satisfies the steady state constraint, thermodynamic

constraints and the non-decomposability condition is named an EFM. Let *I* and *J* the set of metabolites and metabolic reactions. A feasible steady-state flux distribution $\boldsymbol{v}$ associated to a stoichiometric matrix $\boldsymbol{S}$ defines the flux cone $\boldsymbol{P}$ as follows:

$$\boldsymbol{P} = \left\{ \begin{array}{ll} \sum_{j \epsilon J} S_{ij} v_j = 0, & \forall i \epsilon I \\ v_j \geq 0 & \forall j \epsilon J \end{array} \right\} \quad (1)$$

A convex basis **e** is a minimal set $\boldsymbol{K}$ of EFMs able to describe any flux distribution $\boldsymbol{v} \subset \boldsymbol{P}$ so that:

$$\boldsymbol{v} = \lambda\, \mathbf{e} \quad (2.a)$$

$$v_j = \sum_{\boldsymbol{k} \epsilon \boldsymbol{K}} \lambda_{\boldsymbol{k}} e_{kj} \qquad \forall j \epsilon J, k \epsilon K, \lambda_k \geq 0 \quad (2.b)$$

Flux balance analysis (FBA) is a mathematical approach, based on optimization, employed for estimating flows through a metabolic network (Orth et al., 2010). By FBA a linear program is employed to estimate the maximum possible flow $v_{j\ target}$ through a given reaction. FBA is formulated as follows:

$$Max \qquad v_{j\ target} \quad (3)$$

s.t.

$$\sum_{j \epsilon J} S_{ij} v_j = 0, \qquad \forall i \epsilon I \quad (4)$$

$$LB_j \leq v_j \leq UB_j \qquad \forall j \epsilon J \quad (5)$$

$$v_j \geq 0 \qquad \forall j \epsilon J$$

Usually the biomass flow $v_{biomass}$ is maximized but any flow can be maximized. Constraints of FBA defines a hyperplane $\boldsymbol{Q}$ in the solution space as illustrated by Fig. 1.a. Usually there are more EFMs than the necessary ones required to construct any steady-state flux distribution. The intersection of the flux cone $\boldsymbol{P}$ and the hyperplane defined by the constraints of the FBA leads to the polytope illustrated by Figure 1.c.

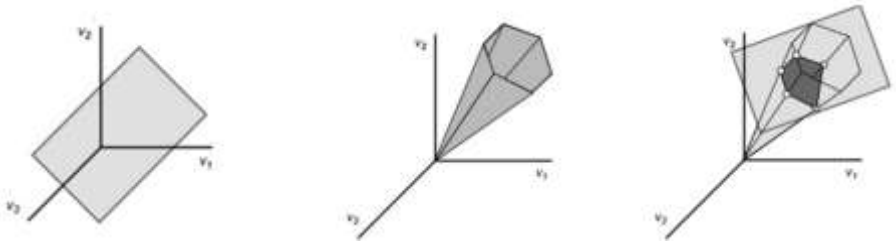

Figure 1: (a) Hyperplane defined by constraints of FBA; (b) flux cone of the metabolic network; (c) polytope defined by the intersection between the FBA-constraints and the flux-cone.

Metabolic processes are typically organized into biochemical networks. Pathways are defined as chains of consecutive reactions that convert source-metabolites(s) into sink-metabolite(s) as illustrated by Figure 2. A source reaction *j* is a transport flux moving an external source-metabolite *i* from the environment into the cell and have just one nonzero S_{ij} coefficient. Conversely, a sink *j* reaction will move sink-metabolite *i* outside the cell and will S_{ij} as a nonzero coefficient. EFMs are those pathways through a metabolic network that connect internal substrates to external metabolites.

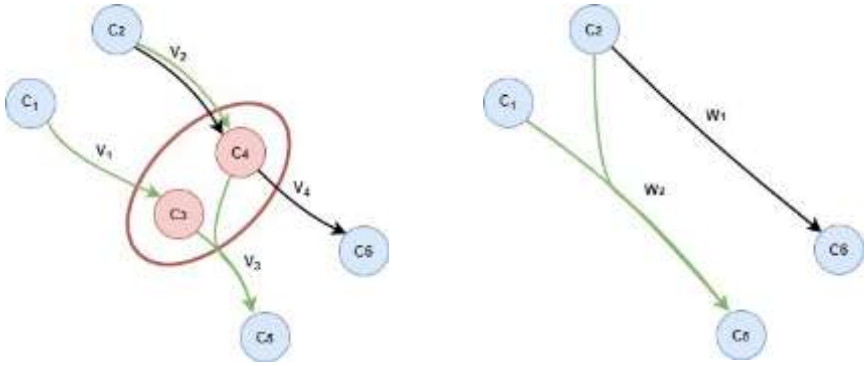

Figure 2: (a) conversion of source-metabolites (substrates) into sink-metabolites (products) by a metabolic network through a biochemical network; (b) macroscopic representation.

3 Numerical Methods

Metabolic pathways do not act as individual reactions but work simultaneously trough highly connected and complex reaction-networks. Reconstruction of genome-scale metabolic networks has allowed the scientific community to study metabolism by using mathematical modelling techniques and several mathematical approaches have been developed under the paradigm of constraint-based modelling. These methods are based on the use of the so called stoichiometric matrix $\boldsymbol{S}$ (Orth et al., 2010) as defined above. Constraint based methods combine flux analysis and metabolic pathway analysis and allow to estimate the fluxes over every metabolic pathway. A flux cone $\boldsymbol{P}$ is defined by assuming that all reactions are irreversible; i.e. $\boldsymbol{v} \geq \boldsymbol{0}$, but FBA deals with both reversible and irreversible reactions. EFM from networks with reversible reactions can be computed with a pre-processing step which consists on splitting a reversible reaction into two irreversible reactions (Gagneur and Klamt, 2004). This translates the network into an extended network where columns of the matrix associated to reversible columns are split into two columns associated respectively to "forward" and "backward" irreversible reactions:

$$\boldsymbol{S} = [\boldsymbol{S_{irr}} \quad \boldsymbol{S_{rev}}] \rightarrow \boldsymbol{S'} = [\boldsymbol{S_{irr}} \quad \boldsymbol{S^{+}_{rev}} \quad \boldsymbol{S^{-}_{rev}}] \tag{6.a}$$

$$J = J_{irr} \cup J_{rev} \rightarrow J' = J_{irr} \cup J^{+} \cup J^{-} \tag{6.b}$$

There would be a longer flux vector $\boldsymbol{v'}$ that contains just irreversible reactions. From them, fluxes in the original $\boldsymbol{P}$ space can be calculated as follows:

$$v_j = v'_j \qquad \forall j \in J_{irr} \tag{7.a}$$

$$v_j = v'^{+}_j - v'^{-}_j \qquad \forall j \in J_{rev} \tag{7.b}$$

3.1. Column generation.

Given *all* EFMs from a metabolic network, an EFM selection problem can be formulated as a MILP. Eq. (8) minimizes the cost of selected modes while eq. (9) forces to select EFM associated to active reactions v_j. The term "cost" will be later discussed. In eq. (9) the parameter ε_{kj} indicates the stoichiometric contribution of the k-EFM into reaction j. Eq. (10) is an on/off activation constraint. If $X_k = 1$, then EFM-k would be switched-on on the description of the metabolism and λ_k will be its weight on the selected reactions.

$$Min \quad \sum_{k \in K} c_k X_k \tag{8}$$

$$s.t. \tag{9}$$

$$\sum_{k \in K} \varepsilon_{kj} \lambda_k = v_j \qquad \forall j \in J_{active} \tag{10}$$

$$0 \leq \lambda_k \leq X_k$$

$$X_k \in \{0, 1\}$$

This formulation assumes that all EFMs are known but feasible-EFMs may run into trillions for large metabolic networks and it is not possible to realistically generate all of them. The column generation (CG) approach handles this combinatory complexity by implicitly considering all EFMs trough the solution of the linear relaxation of this selection problem. In this way, the selection variable is relaxed so that $X_k = \lambda_k$ and $0 \leq X_k \leq 1$. EFMs are scaled to have maximum weights λ_k not larger 1. A simple way for scaling is to multiply each mode by $\min_j \{v_j / e_{jk} > 0\}$. According to the CG technique, a subset of feasible EFMs must be first enumerated and the resulting linear relaxation with this partial set must be solved. Two problems, the master problem (MP) and the pricing subproblem (SP), are later iteratively solved. The iterative procedure involves obtaining the optimal vector of prices from the MP to pass it to the SP. The SP solution either finds a new EFM that improves the objective of the MP or indicates that the solution of the MP cannot be improved by the addition of more EFMs. The prices (dual variable values) are used to determine if there is a new EFM not included in the subset of enumerated modes that can reduce the objective function value of the primal problem. Using the value of the optimal dual vector $\boldsymbol{\pi} = \{\pi_j\}$ with respect to the partial EFM set, new EFMs are generated and incorporated. The so called pricing problem or EFM-generation problem can be defined as follows:

$$\underset{}{\textit{Min}} \quad C_k - \sum_{j \in J} \pi_j \varepsilon_j \tag{11}$$

$$\textit{s.t.} \quad \sum_{j \in J} S_{ij} \varepsilon_j = 0 \qquad \forall i \in I \tag{12}$$

$$\varepsilon_j \leq E_j \qquad \forall j \in J \tag{13}$$

$$\varepsilon_j \in R; E_j \in \{0, 1\}$$

Eq. (11) aims at identifying the EFM with minimum reduced cost while eq. (12) ensures that the identified mode fulfills the steady-state metabolites balance. Eq. (13) is and activation constraint indicating that just the selected reaction belongings to the minimum reduced cost EFM can have a non-zero flux. While binary variable E_j is employed to activate reaction j in the new EFM, the continuous variable ε_j will be uploaded to the RMP as the parameter ε_{kj} for the newly generated k-EFM.

The pricing subproblem would locate an extreme point corresponding to the intersection of an extreme ray of the flux cone $\boldsymbol{P}$ with the hyperplane defining contraints of the FBA..

3.2. Objective functions

Not all EFMs have biological meaning. The non-uniqueness of generated EFMs and its dependence on the choice of the mathematical program and the objective functions lead to an ambiguity that weakens the biological significance of EFMs (Zanghellini et al., 2013). A key concept for finding biologically significant EFMs is that short metabolic pathways have better biological properties because they can carry higher fluxes (de Figueiredo et al., 2009) and canonical pathways typically use a minimal number of step-reactions to achieve their metabolic objective. An alternative objective function would be the minimization of the aggregated reaction fluxes according to the proposal by Rezola et al., (2014). This leads to the relaxation of integrality of variables E_j thus defining a linear pricing problem. So, within this CG approach, the following EFM "costs" can be employed to capture such meanings:

$$C_k = \sum_{j \in J} E_j \tag{14.a}$$

$$C_k = \sum_{j \in J} \varepsilon_j \tag{14.b}$$

The selection of the objective function would depend on the specific case and on the aim of the metabolic engineer but this issue is out of the scope of this work.

3.3. Algorithmic implementation and numerical issues

CG is a standard decomposition technique that must incorporate some specific modules for computing EFMs. The proposed algorithm is summarized by the pseudocode shown in Figure 3. First, in the pre-processing **Stage 0**, FBA and flux-variability-analysis –FVA- (Gudmundsson et al., 2010) are solved to determine fluxes compatible with available experimental rates of external metabolites. Fluxes intervals, in turn, allow classifying reactions into reversible, forward-irreversible and backward-irreversible reactions. Since CG must start from an initial feasible solution, an important issue is the initialization stage. Detecting feasible EFMs for large genomic scale networks may be quite hard and therefore, in this application, the initialization phase comprises two sub-stages. In the first one (**Stage 1.A**) artificial EFMs comprising just a single active reaction and a large cost are generated to conform the initial bank of EFMs. In this way, for any $j \in J$ a fake EFM $\boldsymbol{e_j}$ with $e_j = 1$ and $e_{j'} = 0$ for al $j' \neq j$ will be created. A large M-cost will be allocated to any fake-EFM. From them the initial and biologically meaningful RMP can be computed. Then, feasible EFMs are generated by successively activating all reactions of the network. Each new and biologically feasible EFM generated in this stage identifies, in addiction to active reaction j, some other active reactions $j' \neq j$. After completing **Stage 1.B**, all artificial

and meaningless EFMs from Stage 1.A would have been rendered redundant. Then the RMP will contain, in addition to meaningless EFM $\boldsymbol{e_k}$: ord(k) $\leq |J|$ biologically feasible EFM generated in **stage 1.B**. Later, in **Stage 2**, the proper CG procedure starts. In this phase, for any couple of source and sink fluxes, the CG generates iteratively EFMs until no more EFMs with reduced cost can be identified. Finally, in **Stage 3**, fluxes intervals computed by FBA/FVA are employed to solve the integer EFM selection problem given by eqs. (8)-(10) that allows obtaining a macroscopic representation of a the metabolism.

Stage 0: Data input
Solve FBA
Solve FVA
S; E = [ ]; c = [];
$J' = J^{irr} \cup J^{+} \cup J^{-}$
Stage 1.A: Pre-initialization
for j = 1,…J'
$\boldsymbol{e_j} = [e_{jj} = 1; e_{jj'} = 0\ \forall j' \in J: j' \neq j]$; c_j = M;
E ← [E $\boldsymbol{e_j}$]; c ← [c c_j]
end for
Stage 1.B: Initialization
for j = 1,…, J'
$\boldsymbol{v} = [v_j = 1; v_{j'} = 0\ \forall j' \in J: j' \neq j]$
while c^r <0
Solve RMP
Solve Pricing problem
E ← [E $\boldsymbol{e}^{new}$]; c ← [c c^{new}]
end while
end for
Stage 2: EFM-generation
$\boldsymbol{v}$ = [*0,…, 0*]
for j' = 1,…, J^{sink}
$v_{j'} = 1$;
for j'' = 1,…, J^{source}
$v_{j''} = 1$;
while c^r <0
Solve RMP
Solve Pricing problem
E ← [E $\boldsymbol{e}^{new}$]; c ← [c c^{new}]
end while
end for
end for
Stage 3: EFM-selection
Solve integer EFM-selection problem

Figure 3: Pseudo-code of the CG-based EFMs generation algorithm

4 Results

The algorithm summarized by the pseudo-code of Fig. 3 was built as an .m by employing the COBRA toolbox (Heirendt et al., 2019) to read .xbml files downloaded from the BIGG database (http://bigg.ucsd.edu/models) and to compute FBA and FVA. The CG algorithm was coded in GAMS 35.0 and the transfer of information between both languages was performed by the GAMS-MATLAB inter-phase developed by Dorfner (2018). The .m and .gms codes can be provided to the interested reader for reproduction of the results. We applied the CG based algorithm to twenty biochemical and genomic-scale networks on a PC with an Intel CPU 2.9 GHz processor and 64 GB RAM. Tables 1 report results indicating the number of generated EFMs and the CPU time consumed in each stage of the algorithm. Table 1.a summarizes the results found with an integer pricing problem considering eq. (14.a) as the cost of any generated EFM while Table 1.b. presents results for the linear pricing problem considering eq. (14.b) as its objective function. The procedures was successfully employed jointly with networks ranging from the 72 x 95 *E coli* core model to the 2172 x 4483 *Ph tricornutum* CCAP 1055 iLB1027_lipid model. The size of networks in term of cardinality of sets I, J and J' and the parameter DOF (degrees of freedom), defined as ($|J'| - |I|$) are also reported. Note that total CPU time exceeds the sum of CPU time consumed by both problems of the CG procedure because of the computation of FBA, FVA and because of the information transfer between modules of the algorithm.

The last step of the algorithm involves the generation of a macroscopic representation of the metabolism from the EFMs previously obtained. This corresponds to the stage 3 of the pseudocode depicted in Figure 3. This stage is here illustrated with EFMs generated from the *E coli* core network. The maximum specific growth rate determined by FBA (with fluxes bounds unchanged from the ones downloaded from BIGG) is $v_{biomass} = 0.874$. First FVA ran to compute fluxes-bounds compatible with $v_{biomass}$ and with these bounds, the integer problem (8)-(10) was solved. To avoid infeasibilities due to numerical issues, eq. (9) was replaced by the two corresponding "$\leq$" and "$\geq$" constraints. Also a small ε-value was respectively added and subtracted to their right hand side of these equations. The problem selected the six EFMs, detailed in Table 2. In the first column of the table the macroscopic representation involving just external metabolites transported by sink and source flows is reported. The second column reports the weights vector ($\lambda * 100$) associated to each selected mode.

Table 1.a: Results on biochemical and genomic scale networks with shortest pathway (eq. 14.a) as objective function of the SP

Network	Organism	FVA μ^{max}	|I|	|J|	|J'|	DOF	Generation stage				Selection stage		Total CPU time
							Generated-EFM	CPU RMP	CPU Pricing	CPU total	Selected-EFM	CPU	
Core model	E coli	0.8739	72	95	95	23	102	1.06	1.01	2.07	2	0.36	29.40
iAB_RBC_283	H.Sapinens	2.9356	342	469	521	179	1391	45.78	56.02	101,8	56	0.61	830.2
iLJ478	Th maritima MSB8	0.2284	570	652	668	98	171	6.19	10.71	16.90	18	0.88	196.74
iAF692	M str. Fusaro		628	690	706	78	176	7.42	13.16	20,58	18	0.89	220.3
iSB619	S aureus N315	0.1581	655	743	759	104	501	13.47	20.42	33,89	Not found	-	352.90
iCN718	A baumannii AYE	1.3136	888	1015	1085	197	1990	106.13	92.75	198.88	79	0.22	2208.07
iJN746	P putida KT2440	1.400	907	1054	1145	238	645	19.76	38.43	58,19	92	0.02	647.51
iAM_Pb448	Plasmodium berghei	0.6587	903	1067	1183	280	1661	93.25	93.06	186.31	121	0.47	2054.09
iEK1008	M tuberculosis H37Rv	0.0582	998	1226	1245	247	1334	25.32	47.34	72,66	21	0.44	822.56
iND750	S cerevisiae S288C	0.094	1059	1266	1325	266	782	16.85	33.45	50,3	60	0.02	657.81
iYL1228	K pneumoniae MGH 78578	1.0426	1658	2262	2312	654	1250	64.92	123.57	188,49	53	0.05	2427.26
iAF1260	E coli K-12. MG1655	0.7367	1668	2382	2434	766	2163	126.48	211.99	338,47	54	0.06	4180.15
iRC1080	C reinhardtii	0.000	1706	2191	2461	755	2794	189.44	178.91	368,35	282	0.17	4705.98
iSDY_1059	Shigella dysenteri-ae Sd197	0.9379	1888	2539	2593	705	1890	153.83	269.26	423.89	57	0.09	6189.77
STM_v1_0	Salmonella str. LT2	0.4778	1802	2545	2596	794	2502	206.43	271.57	478.00	54	0.26	5949.54
iS-SON_1240	Shigella sonnei Ss046	0.9826	1936	2693	2761	825	2436	207.37	312.82	520,19	70	0.14	6284.21
iUMN146_1321	E coli UM146	0.9825	1942	2735	2806	864	2555	198.42	308.36	506.78	73	0.10	7620.42
iEcDH1_1363	E coli DH1	0.9825	1949	2750	2822	873	2628	206.57	310.04	516,61	74	0.09	6525.23
iYS1720	Salmonella pan-reactome	0.4885	2436	3357	3414	978	5596	671.73	515.59	1187.32	59	0.47	14135.96
iLB1027_lipid	Ph tricornutum CCAP 1055	0.3596	2172	4456	4483	2311	5862	194.26	595.99	790.25	128	0.08	22200.72

Table 1.b: Results on biochemical and genomic scale networks with aggregated fluxes (eq. 14.b) as objective function of the SP

Network	Organism	FVA μ^{max}	|I|	|J|	|J'|	DOF	Generation stage				Selection stage		Total CPU time
							Generated-EFM	CPU RMP	CPU Pricing	CPU total	Selected-EFM	CPU	
Core model	E coli	0.8739	72	95	95	23	52	1.71	9.75	11.47	3	0.92	34.04
iAB_RBC_283	H.Sapinens	2.9356	342	469	521	179	896	28.03	88.89	116.91	56	0.81	¿?
iLJ478	Th maritima MSB8	0.2284	570	652	668	98	119	11.40	46.75	58.15	18	0.88	233.54
iAF692	M str. Fusaro	0.0268	628	690	706	78	109	12.70	47.48	60.18	18	0.94	255.73
iSB619	S aureus N315	0.1581	655	743	759	104	241	14.28	79.30	93.58	Not found	-	334.34
iCN718	A baumannii AYE	1.3136	888	1015	1085	197	1106	67.74	9940.7	10008.44	79	0.22	11460.91
iJN746	P putida KT2440	1.400	907	1054	1145	238	532	26.34	837.21	863.55	92	0.91	1421.04
iAM_Pb448	Plasmodium berghei	0.6587	903	1067	1183	280	1235	66.76	1689.3	1756.1	121	0.78	3086.1
iEK1008	M tuberculosis H37Rv	0.0582	998	1226	1245	247	1297	31.94	846.72	878.66	24	0.89	1699.8
iND750	S cerevisiae S288C	0.094	1059	1266	1325	266	518	18.18	927.18	945.36	60	0.91	1459.1
iYL1228	K pneumoniae MGH 78578	1.0426	1658	2262	2312	654	1092	46.84	2147.5	2194.3	54	0.89	4084.4
iAF1260	E coli K-12. MG1655	0.7367	1668	2382	2434	766	1534	96.06	5149.4	5245.4	54	0.72	8718.2
iRC1080	C reinhardtii	0.000	1706	2191	2461	755	Crashed						
iSDY_1059	Shigella dysenteri-ae Sd197	0.9379	1888	2539	2593	705	1529	136.6	5953.1	5089.6	59	0.89	10845.9
STM_v1_0	Salmonella str. LT2	0.4778	1802	2545	2596	794	1395	109.6	7464.6	7784.1	54	0.90	11762.5
iS-SON_1240	Shigella sonnei Ss046	0.9826	1936	2693	2761	825	2021	183.1	10553	10736	73	0.77	16811.3
iUMN146_1321	E coli UM146	0.9825	1942	2735	2806	864	2199	181.9	9075.1	9257.0	73	0.10	15701.2
iEcDH1_1363	E coli DH1	0.9825	1949	2750	2822	873	1650	144.5	8344.0	8488.5	77	0.86	13596.9
iYS1720	Salmonella pan-reactome	0.4885	2436	3357	3414	978	2330	217.6	15619	15387	60	0.66	24819.1
iLB1027_lipid	Ph tricornutum CCAP 1055	0.3596	2172	4456	4483	2311	62435	103.1	7955.1	8054.2	128	0.80	17311.8

Table 2: Macroscopic representation of the *E coli* metabolism

Macroscopic stoichiometry	λ
$\begin{bmatrix} 0 & 0.296 & 0 & 0.148 & 0.074 & 0.296 & 0.148 & 0 & 0.370 & 0 & 0 \\ 0 & 0.338 & 0 & 0 & 0.056 & 0.338 & 0 & 0 & 0.338 & 0 & 0 \\ 0 & 0.211 & 0 & 0.421 & 0.105 & 0.211 & 0.421 & 0 & 0.421 & 0 & 0 \\ 0.025 & 0.977 & 0.684 & 0 & 0.570 & 0.477 & 0.506 & 0.138 & 0.264 & 0.093 & 0 \\ 0.020 & 0.497 & 0 & 0 & 0.223 & 0.641 & 0.398 & 0.108 & 0.474 & 0.073 & 0 \\ 0.023 & 0.443 & 0 & 0 & 0.349 & 0.778 & 0.797 & 0.126 & 0.500 & 0.085 & 0.168 \end{bmatrix} \begin{bmatrix} Ex\ Biomasa \\ Ex\ CO_2 \\ Ex\ EtOH \\ Ex\ For \\ Ex\ Gluc \\ Ex\ H_2O \\ Ex\ H \\ Ex\ NH_3 \\ Ex\ O_2 \\ Ex\ Pi \\ Ex\ Succ \end{bmatrix} = [0]$	$\begin{bmatrix} 12.83 \\ 4.76 \\ 5.56 \\ 5.72 \\ 17.43 \\ 16.69 \end{bmatrix}$

Substrates transported by source fluxes into the cell (e.g. *Glc, O_2, NH_4*) would have negative flux values while metabolic products fluxed out of the cell; like biomass, CO_2, H_2O, etc; would have positive flux values. Note that three different EFMs for producing biomass were selected. As stated by Oddsdottir et al. (2015), the subset of obtained EFMs and therefore, the macroscopic representation of the metabolism may not be necessary unique.

5 Conclusions

A CG-based method to generate EFMs compatible with FBA and FVA has been presented. CG is a well-known technique which has been employed in many fields but was not employed in metabolic engineering. Our proposal generates EFMs connecting each source-flux with each sink-flux of a metabolic network after computing feasible flux intervals for every network-reaction. These intervals are employed in a post-generation stage to select EFMs previously computed by CG. This procedure allows getting a set of EFMs in short CPU time for large genomic-scale networks and also allows to rapidly generating macroscopic representations of these networks. It must be noted that the subset of EFMs computed by CG may not be unique. The macroscopic representation arises as a selection of generated EFMs compatible with feasible fluxes-intervals. This framework allows finding solutions with a minimal number of EFMs and showed to be able to identify and select EFMs in moderate CPU times even for large genomic-scale metabolic networks with more than four thousand irreversible reactions.

References

de Figueiredo, L; , Podhorski, A.; Rubio, A.; Kaleta, C.; Beasley, J.; Schuster, S; Planes, F. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics 25 (23) 2009, 3158–3165. https://doi:10.1093/bioinformatics/btp564

Dorfner, J., 2018. GAMS-MATLAB Documentation. Release 0.1. https://gams-matlab.readthedocs.io/en/latest/index.html

Fukuda, K., Prodon, A., Double description method revisited, in: Deza, M., Euler, R., Manoussakis, I., (Eds.), Combinatorics and Computer Science, Volume 1120 of Lecture Notes in Computer Science, Springer, 1996, pp. 91–111. https://doi.org/10.1007/3-540-61576-8_77

Gagneur, J.; Klamt, S. Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics. 5, 175 (2004). https://doi.org/10.1186/1471-2105-5-175

Gudmundsson, S., Thiele, I. Computationally efficient flux variability analysis. BMC Bioinformatics. 11, 489 (2010). https://doi.org/10.1186/1471-2105-11-489

Heirendt, L; S. Arreckx, T. Pfau, S. Mendoza, A. Richelle, A. Heinken, H. Haraldsdottir, J. Wachowiak, S. Keating, V. Vlasov, S. Magnusdottir, C. Ng, G. Preciat, A. Zagare, S. Chan, M. Aurich, C. Clancy, J. Modamio, J. Sauls, A. Noronha, A. Bordbar, B. Cousins, D. El Assal, L. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, Ni. Sompairac, H. Le, D. Ma, Y. Sun, L. Wang, J. Yurkovich, M. Oliveira, P. Vuong, L. El Assal, I. Kuperstein, A. Zinovyev, H. Hinton, W. Bryant, F. Aragon Artacho, F. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M. Saunders, C. Maranas, N. Lewis, T. Sauter, BØ. Palsson, I. Thiele, R. Fleming. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0, Nature Protocols, volume 14, pages 639–702, (2019) https://doi.org/10.1038/s41596-018-0098-2

Oddsdóttir H.; Hagrot E.; Chotteau V; Forsgren A. On dynamically generating relevant elementary flux modes in a metabolic network using optimization. J Math Biol. (2015) 71(4):903-20. https://doi:10.1007/s00285-014-0844-1

Orth, J.; Thiele, I.; Palsson, BØ. What is flux balance analysis? Nature Biotechnology (2010) 28, 245–248. https://doi.org/10.1038/nbt.1614

Rezola, A.; Pey, J.; Tobalina, L.; Rubio, A.; Beasley, J; Planes, F. Advances in network-based metabolic pathway analysis and gene expression data integration. Briefings in bioinformatics. Vol 16. No 2. 265-279 (2014) https://doi:10.1093/bib/bbu009

Schuster S, Hilgetag C, Woods JH 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. https://doi.org/10.1007/s002850200143

Zanghellini, J.; Ruckerbauer, D.; Hanscho, M.; Jungreuthmayer, C.Elementary flux modes in a nutshell: Properties, calculation and applications. Biotech Journal (2013) 8, 1009-1016. https://DOI:10.1002/biot.201200269