

Systems biology

Fast-SNP: a fast matrix pre-processing algorithm for efficient loopless flux optimization of metabolic models

Pedro A. Saa and Lars K. Nielsen*

Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, Corner College and Cooper Rds (Bldg 75), Australia

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on April 26, 2016; revised on July 15, 2016; accepted on August 21, 2016

Abstract

Motivation: Computation of steady-state flux solutions in large metabolic models is routinely performed using flux balance analysis based on a simple LP (Linear Programming) formulation. A minimal requirement for thermodynamic feasibility of the flux solution is the absence of internal loops, which are enforced using ‘loopless constraints’. The resulting loopless flux problem is a substantially harder MILP (Mixed Integer Linear Programming) problem, which is computationally expensive for large metabolic models.

Results: We developed a pre-processing algorithm that significantly reduces the size of the original loopless problem into an easier and equivalent MILP problem. The pre-processing step employs a fast matrix sparsification algorithm—Fast- sparse null-space pursuit (SNP)—inspired by recent results on SNP. By finding a reduced feasible ‘loop-law’ matrix subject to known directionalities, Fast-SNP considerably improves the computational efficiency in several metabolic models running different loopless optimization problems. Furthermore, analysis of the topology encoded in the reduced loop matrix enabled identification of key directional constraints for the potential permanent elimination of infeasible loops in the underlying model. Overall, Fast-SNP is an effective and simple algorithm for efficient formulation of loop-law constraints, making loopless flux optimization feasible and numerically tractable at large scale.

Availability and Implementation: Source code for MATLAB including examples is freely available for download at <http://www.aibn.uq.edu.au/cssb-resources> under Software. Optimization uses Gurobi, CPLEX or GLPK (the latter is included with the algorithm).

Contact: lars.nielsen@uq.edu.au

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Constrained-based methods are the most popular methods for exploring the capabilities of genome-scale metabolic models (GEMs) (Lewis *et al.*, 2012). GEMs have been reconstructed for many model organisms (Dal’Molin *et al.*, 2010a; Duarte *et al.*, 2007; Edwards and Palsson, 2000; Forster *et al.*, 2003), enabling comprehensive study of the genotype–phenotype relationship,

cellular physiology, metabolic capabilities, among others (Bordbar *et al.*, 2014a). More recently, these reconstructions have included larger and more complex networks, describing different cellular interactions (Bordbar *et al.*, 2010; Lewis *et al.*, 2010), microbial communities (Shoaie *et al.*, 2015; Stolyar *et al.*, 2007; Zhuang *et al.*, 2011) and multi-tissue organisms (Bordbar *et al.*, 2011; Dal’Molin *et al.*, 2010b, 2015). The increased complexity of current metabolic

models demands efficient constrained-based methods to compute possible network states, especially thermodynamically feasible ones. Computation of thermodynamically feasible states is computationally hard (Müller and Bockmayr, 2013), severely limiting its application to large-scale models. Efficient computational methods are thus needed to interrogate the consequences of thermodynamic feasibility on these models.

GEMs are mathematically represented by the stoichiometric matrix $S \in \mathcal{R}^{m \times n}$, which encodes the mass balances for m internal metabolites involved in n reactions. The capacity of each reaction is phenomenologically constrained by thermodynamics and enzyme kinetics through the use of appropriate lower \mathbf{lb} and upper \mathbf{ub} bounds on the vector of reaction fluxes $\mathbf{v} \in \mathcal{R}^n$ (Equation 1). Assuming vanishing accumulation of internal metabolites (Equation 2), the space of feasible steady-state fluxes is defined by the following set of constraints (hereafter referred as mass balance constraints),

$$\mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} \quad (1)$$

$$S \cdot \mathbf{v} = 0 \quad (2)$$

This typically highly undetermined space is readily explored using Flux Balance Analysis (FBA). Given an optimality criterion, FBA uses linear optimization to compute a steady-state flux solution that maximizes a defined objective such as growth rate, ATP production, among others (Orth *et al.*, 2010b). The flexibility of the network under (sub)optimal conditions can be further assessed using Flux Variability analysis (FVA) (Mahadevan and Schilling, 2003). FVA computes the minimum and maximum flux through each reaction capable of supporting a defined network state. Equations (3) and (4) describe the LP formulations of the above optimization problems,

$$\begin{aligned} \text{FBA : } & \max_{\mathbf{v}} \mathbf{c}^T \cdot \mathbf{v} \\ & \text{s.t. } \{\text{mass balance constraints}\} \end{aligned} \quad (3)$$

$$\begin{aligned} \text{FVA : } & \max_{\mathbf{v}} / \min_{\mathbf{v}} v_i, \quad i \in \{1, \dots, n\} \\ & \text{s.t. } \mathbf{c}^T \mathbf{v} \geq \alpha \cdot z_{\text{opt}} \\ & \quad \{\text{mass balance constraints}\} \end{aligned} \quad (4)$$

where \mathbf{c} represents the objective vector, z_{opt} denotes the optimal value found by FBA, and α is a parameter describing the degree of optimality w.r.t. FBA, i.e. for suboptimal analysis $0 \leq \alpha < 1$, whereas $\alpha = 1$ for alternative optima analysis. Although FBA and FVA provide feasible steady-state flux solutions, they are not guaranteed to be thermodynamically feasible. Additional constraints on the flux vector \mathbf{v} are needed to ensure this and a minimal criterion for thermodynamic feasibility is the absence of internal loops. Let us define the ‘loop-law’ matrix, $\mathbf{N}_{\text{int}} \in \mathcal{R}^{n_i \times l}$, the matrix containing a null space basis of the stoichiometric matrix of internal reactions $S_{\text{int}} \in \mathcal{R}^{m_i \times n_i}$, and, \mathbf{v}_{int} and $\Delta\boldsymbol{\mu}_{\text{int}}$ the internal reaction and the corresponding chemical potential vectors, respectively. A steady-state flux solution \mathbf{v}^* is thermodynamically feasible if, (i) the global potential energy of the network is balanced, i.e., $\mathbf{N}_{\text{int}}^T \cdot \Delta\boldsymbol{\mu}_{\text{int}}^* = 0$ (first law), and (ii) reactions proceed in the opposite direction of chemical potential change, i.e. $\text{diag}(\Delta\boldsymbol{\mu}_{\text{int}}^*) \cdot \mathbf{v}_{\text{int}}^* \leq 0$ (second law). These two conditions hold simultaneously true if and only if the net flux around all closed loops is equal to zero, i.e. \mathbf{v}^* is ‘loopless’ (Beard *et al.*, 2002). The loopless condition can be enforced by formulating mixed-integer linear constraints on \mathbf{v}_{int} (hereafter referred as loopless constraints) (Schellenberger *et al.*, 2011),

$$\begin{aligned} \mathbf{N}_{\text{int}}^T \cdot \Delta\boldsymbol{\mu}_{\text{int}} &= 0 \\ \mathbf{1} - \varepsilon_{\text{int}} \cdot (K + 1) &\leq \Delta\boldsymbol{\mu}_{\text{int}} \leq \mathbf{K} - \varepsilon_{\text{int}} \cdot (K + 1) \\ -K \cdot (\mathbf{1} - \varepsilon_{\text{int}}) &\leq \mathbf{v}_{\text{int}} \leq K \cdot \varepsilon_{\text{int}} \\ \varepsilon_{\text{int},i} &\in \{0, 1\} \quad i \in \{1, \dots, n_i\}, \quad \Delta\boldsymbol{\mu}_{\text{int}} \in \mathcal{R}^{n_i}, \quad K \text{ large (e.g., } 10^3) \end{aligned} \quad (5)$$

Addition of Equation (5) to Equations (3) to (4) yields two MILP formulations commonly known as ll-FBA and ll-FVA (‘ll’ stands for loopless). Inclusion of the loopless constraints not only increases the size of the problem (decision variables are now ε_{int} , \mathbf{v} , $\Delta\boldsymbol{\mu}_{\text{int}}$), but also yields a harder optimization problem. Computationally, this problem has recently been addressed by recasting ll-FVA into a series of LP problems, in which steady-state flux solutions are computed by conventional FBA and corrected by iteratively blocking all infeasible cycles (Desouki *et al.*, 2015; Müller and Bockmayr, 2013). Here, we proposed a different approach where loopless constraints are reduced to a sufficient set of feasible loop laws, enabling efficient computation of loopless flux optimization problems. In this way, once the sufficient set of loop laws has been found, any standard loopless optimization problem can be formulated and solved using a single optimization problem without post-processing.

In this work, we present a pre-processing algorithm—Fast-SNP—for efficient formulation of loopless constraints. We note that it is always possible to pre-process the loop-law matrix \mathbf{N}_{int} constrained to known directionalities, so that the reduced set of constraints yields an easier-to-solve, yet equivalent MILP problem. By using a fast sparsification algorithm, Fast-SNP generates a reduced feasible loop-law basis $\tilde{\mathbf{N}}_{\text{int}}$, accounting for these directionalities. Application of Fast-SNP considerably reduced computation times during loopless flux optimization compared with the traditional approach in different metabolic models. Furthermore, our approach identifies key directional constraints, which may enable elimination of infeasible cycles in the underlying GEM. Altogether, Fast-SNP enables efficient computation of thermodynamically feasible flux solutions in GEMs.

2 Methods

2.1 A useful matrix pre-processing algorithm for efficient loopless flux optimization

The motivation of our method stems from the observation that not all internal loop laws are feasible given the directionalities defined by Equation (1) (Saa and Nielsen, 2016). We previously exploited this fact to detect ‘potentially active or feasible loop-laws’ in metabolic models using the conventional representation of \mathbf{N}_{int} (e.g. reduced row echelon form of $\mathbf{N}_{\text{int}}^T$), enabling faster discovery of infeasible random flux samples (Saa and Nielsen, 2016). This reduced set of feasible loop-laws cannot immediately be used for loopless flux computation. However, if we are able to derive a reduced basis of \mathbf{N}_{int} containing a minimum set of feasible loop-laws, then the computed loopless flux solutions will be accurate (see Supplementary Fig. S1 for an illustrative example). In addition to minimizing the number of loop-laws, we decided to simplify the laws through sparsification. We developed a fast matrix pre-processing algorithm motivated by recent progress in the matrix sparsification field (Bian *et al.*, 2015). As opposed to sparsification algorithms based on MILP formulations (Bordbar *et al.*, 2014b), our approach relies on an LP formulation that is not guaranteed to find the ‘sparsest’ null space basis, i.e. least number of non-zero entries, but quickly generates a reduced feasible basis for efficient computation of thermodynamically feasible flux solutions.

The general workflow for efficient loopless flux optimization is illustrated with a toy model (Fig. 1). First, the network stoichiometry is formulated and the internal and exchange reaction matrices, S_{int} and S_{ext} , are defined (Fig. 1A). Next, an initial basis N_{int} is found using singular value decomposition (SVD). Conventional loopless optimization would use this matrix or a reduced row echelon form of N_{int}^T as constraint for loopless flux computation (Equation 5). In this example, out of the three possible independent loop laws found using these methods, there are only two feasible loop laws— L_1 and L_2 —given the directionalities of v_1 and v_2 , i.e. L_3 cannot form a loop (Fig. 1B). Thus, a reduced basis \tilde{N}_{int} can be constructed using just the first two loop laws. This basis is readily found using Fast-SNP by integrating the topology of S_{int} and directionality constraints. Implementation of this pre-processing step significantly improves computational performance in a diverse family of loopless optimization problems (Fig. 1C). In the following, we present our algorithm for pre-processing the loop-law matrix N_{int} .

2.2 A fast matrix sparsification for efficient formulation of loop-law constraints

The problem of finding the sparsest linear basis of the null space—or the Sparse Null-space Basis problem (SNB)—is motivated by its application to linear equality problems arising in constrained optimization problems (Gottlieb and Neylon, 2010). Coleman and Pothén (1986) demonstrated that a greedy algorithm must find the SNB of a matrix $A_{m \times n}$ in $r = n - \text{rank}(A)$ steps, provided that at

each step the subproblem of finding the sparsest null-space vector (SNV) can be solved. Since SNV is NP-complete, SNB is NP-hard (Coleman and Pothén, 1986), and the use of approximate algorithms to solve SNB is justified. Recently, Bian *et al.* (2015) proposed a convex-relaxation of the SNB—referred to as the Sparse Null-space Pursuit (SNP)—where a sparse basis is computed in $r \times n$ LP optimization runs. While this formulation is in principle attractive, it does not consider any directional constraints on the basis vectors, which is a key feature of our problem. Moreover, the n LP optimizations per basis vector are excessive.

We have developed a more efficient sparsification algorithm—Fast-SNP—inspired by the SNP formulation. Fast-SNP finds a minimal sparse representation of N_{int} in at most $2r$ LP optimization runs. Briefly, starting from an empty null space basis, the SNV is solved by finding the minimum l_1 -norm steady-state flux solution $v_{\text{int},k}$ that, (i) is consistent with the defined directionalities, and (ii) is contained in the orthogonal space of $N_{\text{int},k}$ derived from the previous ($k - 1$) iterations. The latter constraint ensures that the basis vector computed at iteration k is linearly independent from the previous ($k - 1$) vectors. This condition can be formulated as $P_{N_{\text{int},k}}^\perp \cdot v_{\text{int},k} \neq 0$, where $P_{N_{\text{int},k}}^\perp$ denotes the projection matrix onto $\text{null}(N_{\text{int},k})$. As this above equation defines neither a convex nor a compact region, we replace it with two equivalent constraints, namely:

$$w^T \cdot P_{N_{\text{int},k}}^\perp \cdot v_{\text{int},k} > \zeta \quad \vee \quad w^T \cdot P_{N_{\text{int},k}}^\perp \cdot v_{\text{int},k} < -\zeta \quad (6)$$

where w represents a vector of random weights and ζ is a small positive constant, e.g. 10^{-3} . Here we employed uniform random weights; however other choices can be also used yielding similar results (Supplementary Table S1). As w is non-zero, the above constraints are only violated if $P_{N_{\text{int},k}}^\perp \cdot v_{\text{int},k} \neq 0$, which is the condition to be avoided. The resulting SNV optimization problem can be then formulated as follows,

$$\begin{aligned} \text{SNV : } \min_{(x, v_{\text{int},k})} \quad & \sum_i x_i \\ \text{s.t. } \quad & S_{\text{int}} \cdot v_{\text{int},k} = 0 \\ & v_{\text{int},k} \leq x \\ & -v_{\text{int},k} \leq x \\ & lb_{\text{int}} \leq v_{\text{int},k} \leq ub_{\text{int}} \\ & w^T \cdot P_{N_{\text{int},k}}^\perp \cdot v_{\text{int},k} > \zeta \quad (\text{positive constraint}) \quad \vee \\ & w^T \cdot P_{N_{\text{int},k}}^\perp \cdot v_{\text{int},k} < -\zeta \quad (\text{negative constraint}) \\ & x \in R^m \end{aligned} \quad (7)$$

In this way, two LP optimizations are required to determine the SNV at each iteration, and thus $2r$ runs are needed to complete the null space basis. Notably, after each iteration, the sparsest (i.e. the minimum l_0 -norm) solution is kept for the next iteration. If at one iteration there are no feasible solutions for both SNV problems, then the algorithm terminates as there are no remaining non-zero basis vectors given the current $N_{\text{int},k}$ and the directionality constraints. The pseudocode of our Fast-SNP is shown in Algorithm 1.

2.3 Detection and removal of infeasible loops

Generation of a reduced loop-law matrix can accelerate other loopless flux optimization problems. Here, we evaluated two of such problems, namely: loop detection (LD) and loop removal (LR). The first problem determines whether a steady-state flux distribution is thermodynamically feasible (i.e. loopless), whereas the second computes the nearest feasible flux solution. Specifically, given a steady-

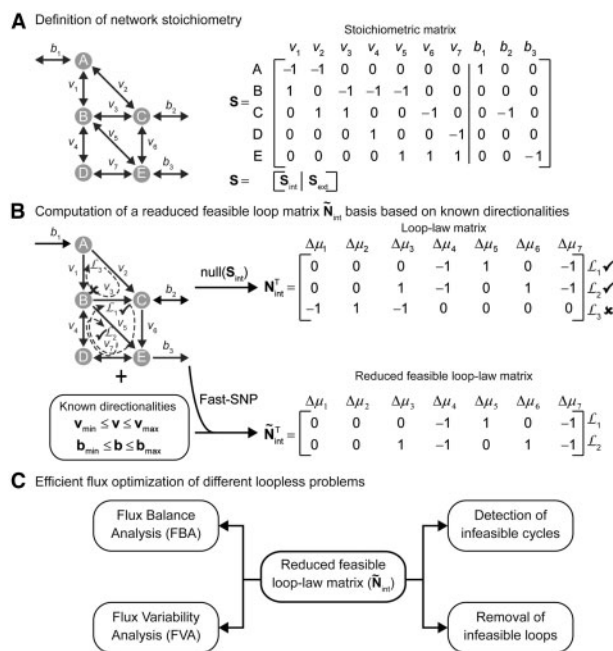


Fig. 1. Illustration of the optimization workflow using Fast-SNP. (A) Network stoichiometry definition. The toy model consist of seven internal reactions (v_i , $i = 1, \dots, 7$), three exchange reactions (b_j , $j = 1, 2, 3$) (nine reactions in total) and five metabolites. The stoichiometric matrix S is defined by two block matrices containing, respectively, the internal (S_{int}) and exchange (S_{ext}) reactions. **(B)** Pre-processing the loop-law matrix using known directionalities. An initial basis for the null space of internal reactions (N_{int}) is for example computed using SVD. This matrix contains $r = n_i - \text{rank}(S_{\text{int}}) = 3$ loop laws or basis vectors accounting only for the information in S_{int} . However, given the directional constraints on N_{int} , a minimal basis \tilde{N}_{int} can be found by including directionality information using Fast-SNP. The resulting basis \tilde{N}_{int} has $r^* = 2 \leq r$ feasible loop laws, enabling more efficient optimization of loopless problems. **(C)** Once a reduced basis for the loop-law matrix is found, a diversity of loopless optimization problems can be performed more efficiently

Algorithm 1. Fast SNP

Inputs: Stoichiometric matrix of internal reactions S_{int} , lower and upper bounds ($\mathbf{lb}_{\text{int}}, \mathbf{ub}_{\text{int}}$)

Outputs: Reduced feasible null space basis of the loop matrix N_{int}

Set $k \leftarrow 0$, $P_{N_{\text{int}},k} \leftarrow 0$, $N_{\text{int},k} \leftarrow \emptyset$

$\mathbf{w} \leftarrow \text{GenerateRandomWeights}$

while $k < r$ do

$P_{N_{\text{int}},k}^{\pm} \leftarrow I_{n_i \times n_i} - P_{N_{\text{int}},k}$

$\mathbf{v}_{\text{int},k}^+ \leftarrow \text{SolvePositiveConstraintSNV}(P_{N_{\text{int}},k}^{\pm}, \mathbf{w})$

$\mathbf{v}_{\text{int},k}^- \leftarrow \text{SolveNegativeConstraintSNV}(P_{N_{\text{int}},k}^{\pm}, \mathbf{w})$

Set $\mathbf{v}_{\text{int},k} \leftarrow \text{argmin}(\|\mathbf{v}_{\text{int},k}^-\|_0, \|\mathbf{v}_{\text{int},k}^+\|_0)$

if $\text{IsEmpty}(\mathbf{v}_{\text{int},k})$ then

break

else

$N_{\text{int},k+1} \leftarrow N_{\text{int},k} \oplus \mathbf{v}_{\text{int},k}$

$\hat{N}_{\text{int},k+1} \leftarrow \text{GenerateOrthogonalBasisUsingSVD}(N_{\text{int},k+1})$

$P_{N_{\text{int}},k+1} \leftarrow \hat{N}_{\text{int},k+1} \cdot \hat{N}_{\text{int},k+1}^T$

end if

Set $k \leftarrow k + 1$

end while

return $N_{\text{int},k}$

state flux solution \mathbf{v}_0 with $\mathbf{v}_{\text{int},0}$ denoting its internal reactions, LD tests for thermodynamic feasibility by directly applying the loopless constraints using a LP formulation with a blank objective (Schellenberger *et al.*, 2011),

$$\begin{aligned} \text{LD: } \min_{\Delta\boldsymbol{\mu}_{\text{int},0}} \quad & 0 \\ \text{s.t. } \quad & N_{\text{int}}^T \cdot \Delta\boldsymbol{\mu}_{\text{int},0} = 0 \\ & -K \cdot (1 + \text{sgn}(\mathbf{v}_{\text{int},0})) - \text{sgn}(\mathbf{v}_{\text{int},0}) \leq \Delta\boldsymbol{\mu}_{\text{int},0} \quad (8) \\ & \Delta\boldsymbol{\mu}_{\text{int},0} \leq K \cdot (1 - \text{sgn}(\mathbf{v}_{\text{int},0})) - \text{sgn}(\mathbf{v}_{\text{int},0}) \\ & K \text{ large (e.g., } 10^3) \end{aligned}$$

\mathbf{v}_0 is loopless if the previous LP problem admits a solution for $\Delta\boldsymbol{\mu}_{\text{int},0}$. If the flux solution is infeasible, LR finds the nearest feasible flux solution \mathbf{v}^* employing the following MIQP formulation (Schellenberger *et al.*, 2011),

$$\begin{aligned} \text{LR: } \min_{(\boldsymbol{\delta}_{\text{int}}, \mathbf{v}^*, \Delta\boldsymbol{\mu}_{\text{int}}^*)} \quad & \|\boldsymbol{\delta}\|^2 \\ \text{s.t. } \quad & \mathbf{v}^* - \mathbf{v}_0 = \boldsymbol{\delta} \\ & \{\text{mass balance constraints}\} \\ & \{\text{loopless constraints}\} \\ & \boldsymbol{\delta} \in R^n \end{aligned} \quad (9)$$

where $\boldsymbol{\delta}$ denotes the distance vector from \mathbf{v}_0 to \mathbf{v}^* . Thus, the closest feasible flux distribution to \mathbf{v}_0 is found by minimizing the Euclidean norm of $\boldsymbol{\delta}$.

2.4 Computational implementation

Fast-SNP and all the loopless flux formulations were coded and implemented in MATLAB 2015b (The MathWorks, Natick, MA). Our implementation relies on third-party solvers to solve LP, MILP and MIQP problems. Optimization runs were performed using Gurobi Optimizer 6.5 (Gurobi Optimization, Inc., Houston, TX) and IBM ILOG CPLEX 12.5.1 (IBM Corp., NY). All computations were run

Table 1. Characteristics of the metabolic models used in this study

Model	Organism	Mets	Rxns	Rev. rxns	Int. rev. rxns
Toy model	—	5	10	4	3
<i>E.coli</i> core	<i>E.coli</i>	68	87	15	13
iAF692	<i>Methanosarcina barkeri</i>	417	484	42	38
iNJ661	<i>Mycobacterium tuberculosis</i>	579	740	82	77
iYL1228	<i>Klebsiella pneumoniae</i>	830	1223	85	78
STM	<i>Salmonella Typhimurium</i> LT2	1086	1597	83	79
iJO1366	<i>E.coli</i>	1136	1679	115	110
Yeast 6	<i>Saccharomyces cerevisiae</i>	756	1018	78	76

on a Dell Studio OptiPlex 9020 workstation (Intel Core i7-4790 processor, 16 GB ram memory, 64-bit architecture) running on Windows 7.

3 Results

The computational performance of our strategy was compared with the traditional approach in different metabolic models covering a range of sizes. The models considered include the toy model shown in Figure 1, an *Escherichia coli* core model (Orth *et al.*, 2010a), and six GEMs, namely: iAF692 (Feist *et al.*, 2006), iJN661 (Jamshidi and Palsson, 2007), iYL1228 (Liao *et al.*, 2011), *Salmonella Typhimurium* Model (STM) (Thiele *et al.*, 2011), iJO1366 (Orth *et al.*, 2011) and Yeast 6 (Heavner *et al.*, 2013). We have assumed growth in minimal medium and growth rate maximization as the objective function for each model. In the case of the toy model, the upper bound of b_1 was set to 1 and reaction b_3 was maximized (see Fig. 1B). The characteristics of each model after removal of blocked reactions are shown in Table 1. The models used in this study can be downloaded from the BiGG Models database at <http://bigg.ucsd.edu> (King *et al.*, 2016).

3.1 Fast-SNP significantly reduces loop-law constraints enabling efficient ll-FBA computation

Fast computation of a reduced feasible loop-law matrix is critical for efficient loopless optimization. Even if a reduced basis accelerates optimization, if its computation is too expensive then the approach is simply not useful. Therefore, we first evaluated the computational performance of our sparsification algorithm in the studied models (Table 2). The computation of a reduced basis was achieved in <25 s for all models and Gurobi was ~60% faster than CPLEX. The speed depends on both the total number and the number of feasible loop laws. For example, Yeast 6 has considerably less internal reactions and therefore total loop laws than iJO1366 (224 versus 484), but it has almost 3-fold more feasible loop laws than iJO1366 (110 versus 34). In this case, the pre-processing time was similar in these two models using both solvers, highlighting the dependence on both factors. We further compared the performance of Fast-SNP against a previous method for generation of a minimal null space basis (Hay *et al.*, 2014). This method is based on the identification of reactions involved in infeasible loops through FVA and extreme pathway analysis. As such, this method would require $2n$ LP iterations + 1 basis computation for the generation of N_{int} . In contrast, Fast-SNP requires $2r$ LP iterations + r basis computations (in the worst case). Considering the greedy nature of Fast-SNP, the high efficiency of SVD algorithms for orthogonal basis generation and that $r < n$ as $r = n - \text{rank}(S_{\text{int}})$, Fast-

Table 2. Pre-processing results for each metabolic model

Model	Feasible loop laws/ total loop laws	No. non-zero entries Fast-SNPN _{int} /no. non-zero entries N _{int} ^a	Pre-processing time ^b (s)	
			Gurobi	CPLEX
Toy model	2/3	7/10	$3.6 \cdot 10^{-2} \pm 9.0 \cdot 10^{-3}$	$4.7 \cdot 10^{-2} \pm 1.6 \cdot 10^{-3}$
<i>E. coli</i> core	1/13	2/80	$4.2 \cdot 10^{-2} \pm 9.0 \cdot 10^{-3}$	$4.2 \cdot 10^{-2} \pm 9.0 \cdot 10^{-3}$
iAF692	12/64	38/409	$5.5 \cdot 10^{-1} \pm 9.7 \cdot 10^{-2}$	$8.5 \cdot 10^{-1} \pm 1.3 \cdot 10^{-1}$
iNJ661	21/149	63/1183	$1.6 \pm 1.7 \cdot 10^{-1}$	$2.6 \pm 1.1 \cdot 10^{-1}$
iYL1228	23/368	76/4635	$4.3 \pm 8.9 \cdot 10^{-2}$	$8.9 \pm 2.4 \cdot 10^{-1}$
STM	23/446	59/5453	$6.5 \pm 3.8 \cdot 10^{-1}$	$1.4 \pm 4.6 \cdot 10^{-1}$
iJO1366	34/484	96/6164	$12.0 \pm 1.2 \cdot 10^{-1}$	$21.6 \pm 3.3 \cdot 10^{-1}$
Yeast 6	110/224	1802/3825	$15.14 \pm 4.8 \cdot 10^{-1}$	$23.63 \pm 3.1 \cdot 10^{-1}$

^aIn order to ensure a fair comparison between the two matrices, the reduced row echelon form of N_{int}^T was employed.

^bRuntimes are shown as mean \pm standard deviation calculated from three independent runs.

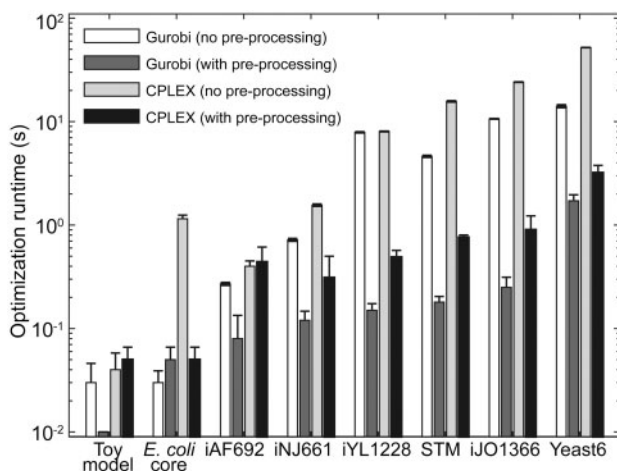


Fig. 2. Optimization performance with and without pre-processing the loop-law matrix. Comparison of ll-FBA runtimes with and without pre-processing using distinct solvers in different models. The error bars represent the standard deviation of three optimization runs. Pre-processing of the loop-law matrix substantially reduces the computation time in large models using both solvers

SNP produces a sparse null basis substantially faster (~ 10 times) than the latter method (Supplementary Fig. S2).

Fast-SNP also yielded substantially reduced loop-law matrices with fewer feasible loop laws in all GEMs (Table 2). Reduction of GEMs loop laws varied from 50.9% for Yeast 6 to 94.8% for STM, underscoring the impact of known directionalities on the number of feasible loop laws. In addition to shrinking the loop matrices, the algorithm created much sparser matrices. For instance, the computed loop-law matrix for iJO1366 contained 98.5% less non-zero entries compared with a reduced column echelon of the latter. The above results are encouraging for the potential increase in algorithmic efficiency of loopless optimization algorithms. Whether they actually translate into substantial speedups is evaluated next.

Pre-processing the loop-law matrix significantly accelerated ll-FBA optimization (Fig. 2). The runtime reductions increased with the size of the model and Gurobi was ~ 3 -fold faster than CPLEX (Fig. 2). For the largest GEMs (STM, iJO1366 and Yeast 6) a runtime reduction of around 20-fold was observed. Importantly, in all the studied models there was a perfect agreement on the computed optimal values (Supplementary Fig. S3).

3.2 Reduced loop-law constraints improve computational performance of ll-FVA and related loopless optimization problems

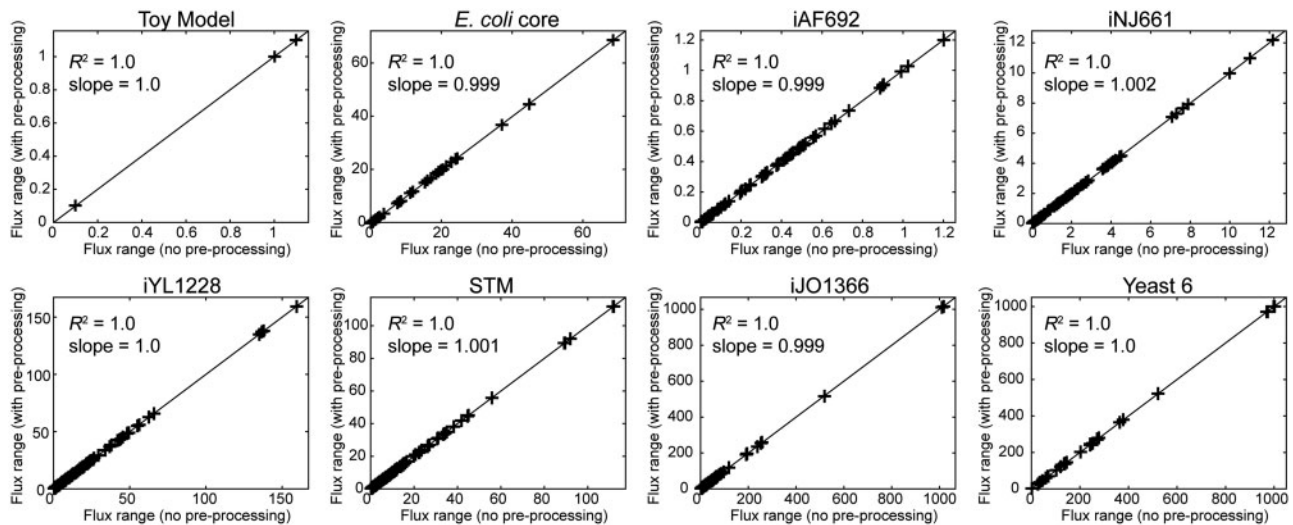
The computational improvement of Fast-SNP was further evaluated in ll-FVA setting $\alpha = 0.9$, i.e. 90% of the optimal growth rate calculated previously. As with ll-FBA, our strategy displays considerably lower runtimes using either solver compared with the conventional approach (Table 3). In general, the greater the reduction in the loop-law constraints, the greater the relative acceleration. For large GEMs, reductions in computation time varied from 7-fold in Yeast 6 using Gurobi (13-fold improvement with CPLEX) up to 50-fold in STM (31-fold improvement with CPLEX). Very promising speedups were also achieved in iJO1366; 41- and 47-fold improvements were obtained using CPLEX and Gurobi, respectively. In the case of medium-scale GEMs, speedups were moderate ranging from 2-fold runtime reduction for iAF692 using CPLEX, to ~ 20 -fold improvement for iYL1228 using Gurobi. Notably, flux ranges computed with Fast-SNP yielded almost identical results as in conventional ll-FVA with no statistical significant difference (Fig. 3 and Supplementary Table S4). We further compared the computational improvement of Fast-SNP against recent approaches for efficient loopless flux optimization, namely: fast-tFVA (Müller and Bockmayr, 2013) and CycleFreeFlux (Desouki et al., 2015). Overall, Fast-SNP yielded more modest improvements compared with the other approaches though magnitude varied dramatically between problems and solvers used (Supplementary Tables S2 and S3).

We also evaluated the performance of Fast-SNP in the related loopless flux optimization problems, LD and LR (see ‘Methods’ section). These optimization problems have been respectively proposed for detection and correction of infeasible flux distributions generated using random sampling (Schellenberger et al., 2011). Given the large size of samples required for the analysis of GEMs (typically 10^5), efficient detection and removal of infeasible cycles is fundamental (Saa and Nielsen, 2016). As for ll-FBA and ll-FVA, pre-processing the loop-law matrix improves runtimes for detection and removal of loops in iJO1366 (Fig. 4). Comparing the performance of LD optimizations with and without pre-processing of iJO1366, runtimes decreased 1.3- and 7.4-fold using Gurobi and CPLEX, respectively. Interestingly, on this occasion CPLEX was faster than Gurobi. The results are even more impressive for the case of LR. Neither Gurobi nor CPLEX were capable of finding a feasible solution within 6 h, when the loop-law matrix was not processed using Fast-SNP (Fig. 4). With Fast-SNP, a solution was found within minutes using either Gurobi or CPLEX. This result highlights that an

Table 3. Runtime comparison of II-FVA with and without reduction of loopless constraints

Model	II-FVA without pre-processing ^a		II-FVA with pre-processing ^a		Average runtime fold-change ($t_{\text{conv II-FVA}}/t_{\text{Fast-SNP II-FVA}}$)	
	CPLEX	Gurobi	CPLEX	Gurobi	CPLEX	Gurobi
Toy model	$4.7 \cdot 10^{-1} \pm 3.0 \cdot 10^{-2}$	$1.1 \cdot 10^{-1} \pm 2.4 \cdot 10^{-2}$	$3.1 \cdot 10^{-1} \pm 1.0 \cdot 10^{-2}$	$1.4 \cdot 10^{-1} \pm 1.6 \cdot 10^{-2}$	1.7	0.8
<i>E.coli</i> core	15.1 ± 1.9	3.3 ± 0.2	$2.6 \pm 3.3 \cdot 10^{-1}$	$1.9 \pm 1.3 \cdot 10^{-1}$	5.8	1.8
iAF692	$4.4 \cdot 10^2 \pm 4.7$	$2.9 \cdot 10^2 \pm 6.3$	$2.1 \cdot 10 \pm 7.8$	69.0 ± 5.6	2.1	4.2
iNJ661	$2.9 \cdot 10^3 \pm 24.5$	$1.6 \cdot 10^3 \pm 40.1$	$6.4 \cdot 10^2 \pm 30.2$	$3.1 \cdot 10^2 \pm 5.6$	4.5	5.1
iYL1228	$2.9 \cdot 10^4 \pm 61.5$	$1.3 \cdot 10^4 \pm 40.6$	$1.7 \cdot 10^3 \pm 36.6$	$6.5 \cdot 10^2 \pm 36.1$	17.5	19.3
STM	$5.9 \cdot 10^4 \pm 4.9 \cdot 10^2$	$2.8 \cdot 10^4 \pm 5.6 \cdot 10^2$	$1.9 \cdot 10^3 \pm 41.9$	$5.7 \cdot 10^2 \pm 15.1$	31.1	49.7
iJO1366	$8.3 \cdot 10^4 \pm 2.3 \cdot 10^2$	$4.2 \cdot 10^4 \pm 42.3$	$2.0 \cdot 10^3 \pm 7.65$	$9.0 \cdot 10^2 \pm 19.8$	40.7	46.8
Yeast 6	$9.2 \cdot 10^4 \pm 2.8 \cdot 10^2$	$2.9 \cdot 10^4 \pm 91.0$	$7.2 \cdot 10^3 \pm 22.3$	$4.4 \cdot 10^3 \pm 1.9 \cdot 10^2$	12.7	6.5

^aRuntimes are shown as mean \pm SD calculated from three independent runs.

**Fig. 3.** Flux ranges computed using the reduced loop-law matrix are consistent with the ranges obtained from conventional II-FVA in the studied models

efficient formulation of the loop matrix might not only be a convenient strategy, but an absolute requirement to solve complex loopless formulations.

3.3 Analysis of the reduced loop-law matrix reveals key directional constraints required to avoid infeasible loops

A reduced feasible loop-law matrix can also serve as a valuable source of information for unravelling infeasible cycles in genome-scale models. The set of vectors encoded in the rows of \tilde{N}_{int} describes a minimal basis of feasible loop laws allowed by the directionalities in the model, and thus, they can yield useful insights about potential infeasibilities in the model. In the following, we analyzed the topology of these laws in iJO1366 after performing II-FVA with $\alpha = 0.9$ (see previous section). Application of II-FVA yielded 13 blocked reactions and changed the directionality of 728 reactions in iJO1366. Importantly, the resulting model has 18 loop laws out of the initial 34, easing its topological analysis (Fig. 5).

Figure 5A depicts the complete set of loop laws found in iJO1366 after definition of growth conditions (minimal medium), optimality criterion (at least 90% of optimal growth rate) and thermodynamic feasibility (loopless condition). The set of feasible loop laws covers 45 reactions (11 reversible) involved in ion and metabolite inter-

compartment transport, alternate carbon, acetate and glycogen metabolism (Supplementary Table S1). Reactions with the highest participation in loop laws are: Nat3pp (5), PPKr (5), CA2t3pp (4) and Cat6pp (4) (Supplementary Fig. S4). Notably, the latter reaction is also flux-forcing, i.e. it has a strictly non-zero flux, and thus it cannot be removed without rendering the model infeasible.

Next, our interest is to analyze the impact of blocking or fixing the directionalities of reactions involved in infeasible cycles. We have selected three illustrative cases to show different strategies for rational analysis and removal of loops (Fig. 5B). Case 1 illustrates a simple loop comprised of two duplicated reactions: ASPt2pp (irreversible) and ASPt2rpp (reversible). In this case, the most sensible action is to remove the irreversible reaction (ASPt2pp) and preserve the reversible one (ASPt2rpp). The resulting model has the same capabilities as the original without the artificial loop L_2 . Cases 2 and 3 are more subjective and illustrate different modelling rationales. In Case 2, if we are not interested in pyruvate alternate metabolism, reaction HPYRI can be removed from the model effectively eliminating L_1 . Case 3 illustrates that loops can be also eliminated by fixing the directionalities of reversible reactions rather than deleting a reaction. L_7 consists of a loop where the succinyl-CoA synthetase operates in reverse. However, under fast-growing glucose-limited aerobic conditions, this reaction proceeds in the forward direction in *E.coli* (Ishii *et al.*, 2007). By fixing its direction to forward, L_7 is eliminated as no

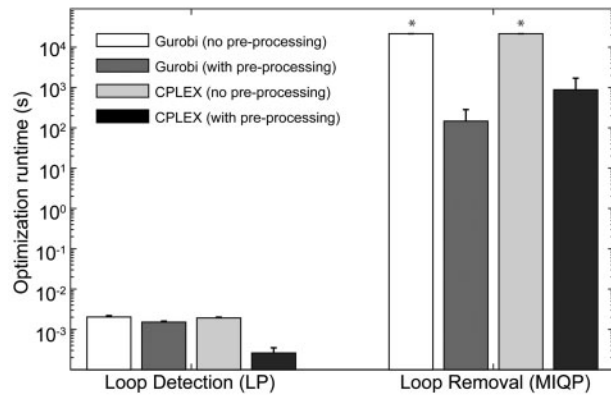


Fig. 4. Reduced loopless constraints generated from Fast-SNP reduce optimization runtimes of other loopless problems in iJO1366. Computational performance in LD and LR problems using flux samples from iJO1366. Error bars represent the standard deviation of five optimization runs. Bars shown with (*) denote cases where it was not possible to compute a solution in $<6\text{ h}>$

other reactions in this cycle can be reversed (Fig. 5B). In summary, the topology encoded in the reduced loop-law matrix facilitates rational infeasibility analysis of large metabolic models.

4 Discussion

Computation of loopless steady-state flux solutions is an NP-hard MILP problem (Müller and Bockmayr, 2013). Recent approaches such as fast-tFVA (Müller and Bockmayr, 2013) and CycleFreeFlux (Desouki *et al.*, 2015) have exploited particular features of the problem and used LP formulations following alternative approaches to solve different loopless flux optimization problems. Although these approaches enabled fast computation of feasible fluxes (acceleration up to two orders of magnitude), they rely on post-processing methods for blocking infeasible cycles after steady-state flux computation. In the current study, we developed a pre-processing strategy where the conventional MILP problem is reformulated using a minimal set of loop-law constraints, enabling efficient computation of loopless flux solutions as well as topological analysis of infeasible cycles in large metabolic models.

All modern MILP solvers use pre-processing to recast MILP formulations into easier-to-solve, yet equivalent representations (Savelsbergh, 1994). It is often possible to develop superior pre-processing strategies using insights into the specific MILP problem. We have developed a fast pre-processing algorithm, Fast-SNP, which finds a minimal set of feasible loop laws taking into consideration the directionalities of the model. Once the reduced loop-law matrix is found, a smaller loopless MILP problem can be formulated and solved efficiently. Speedups up to 50-fold were achieved with pre-processing across diverse types of loopless flux optimization problems (Tables 2 and 3, Figs 2 and 4). Although the computational improvement of Fast-SNP is more modest than previous approaches (Supplementary Tables S2 and S3), it is the only strategy capable of efficiently solving loopless related problems (i.e. Il-FBA, Il-FVA, LD and LR) using a single optimization problem based on suitable LP, MILP or MIQP formulations subject to the loopless constraints (Equation 5). Importantly, such constraints have recently been algebraically proven to always yield thermodynamically feasible flux solutions (Noor *et al.*, 2012).

Finally, we highlight additional and distinctive features of Fast-SNP for the analysis of metabolic models. As shown in the previous section, this approach enables a priori assessment of model complexity in the ‘loop law sense’, as it explicitly computes a minimal

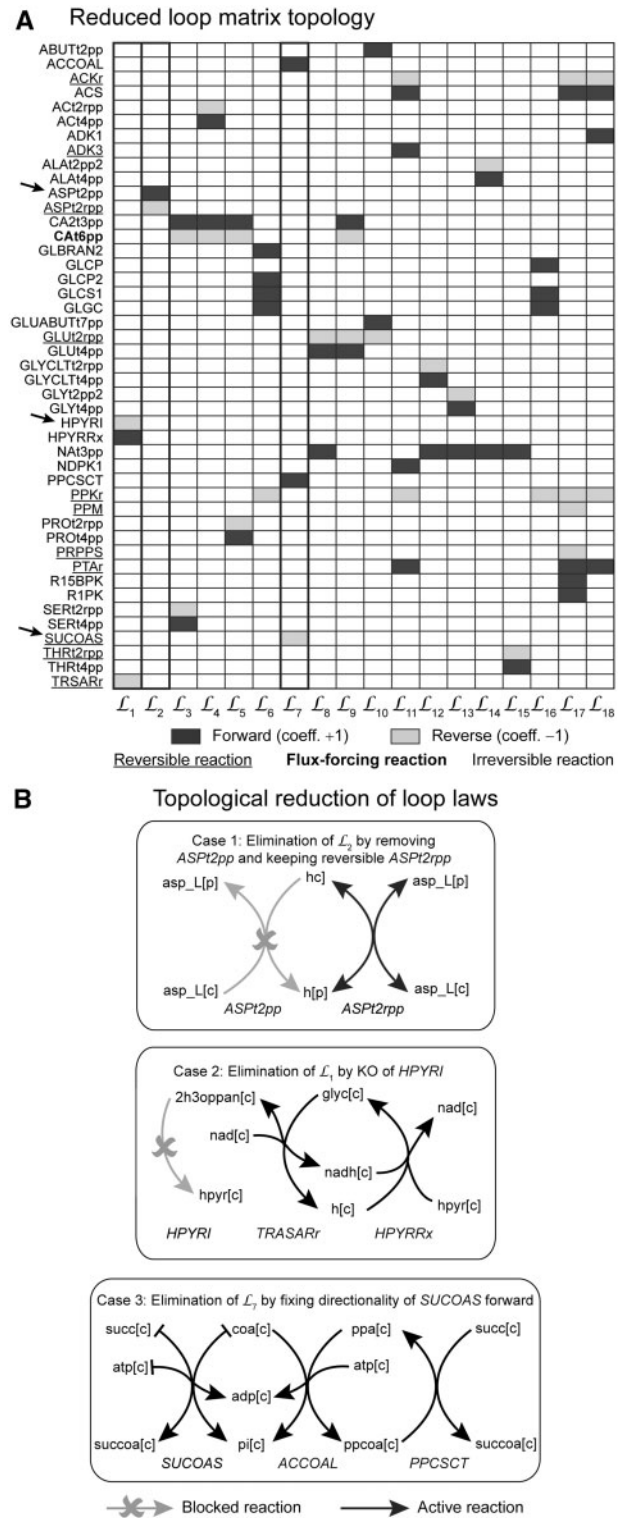


Fig. 5. Topological exploration of the reduced loop-law matrix obtained from Fast-SNP enables rational elimination of infeasible loops in iJO1366. (A) Topological analysis of the reduced loop-law matrix. Reactions indicated with arrows can be either blocked or their directionalities fixed to eliminate infeasibilities. (B) Examples of different approaches to remove infeasible loops

basis describing all the potentially infeasible cycles in the studied model. Formally, these loop laws represent elementary flux modes that do not exchange mass with the surroundings (Schuster *et al.*, 2000). These internal cycles have previously been studied in the

context of oriented matroid theory (Beard *et al.*, 2004; Oliveira *et al.*, 2001) and more recently, infeasible cycles have been enumerated in GEMs using sampling approaches (De Martino *et al.*, 2013). However, efficient assessment of the impact of directionality constraints as well as topological analysis of these loop laws for their potential removal, has been lacking in the literature. Fast-SNP aids comprehensive assessment, unravelling simple strategies for removing infeasible loops based on thermodynamic or practical considerations (see Fig. 5). Additionally, we note that the proposed matrix sparsification algorithm in Fast-SNP can also be employed for other useful analyses. For example, this algorithm can readily be modified to explore alternative solutions by incorporating suitable additional constraints (e.g., $v_i \geq z^*$) using an appropriate objective (e.g. l_1 -norm minimization). The resulting greedy algorithm would generate a set of linearly independent solutions or pathways (i.e. a basis) consistent with the previous constraints by solving only LP problems. This avoids the need of computing expensive MILP formulations. In conclusion, Fast-SNP complements existing methods in the constrained-based modelling toolbox for exploration and topological analysis of infeasible cycles and metabolic pathways.

Funding

P.S. was supported by Becas-Chile, UQ Centennial and IPRS Scholarship Programmes.

Conflict of Interest: none declared.

References

- Beard, D.A. *et al.* (2004) Thermodynamic constraints for biochemical networks. *J. Theor. Biol.*, **228**, 327–333.
- Beard, D.A. *et al.* (2002) Energy balance for analysis of complex metabolic networks. *Biophys. J.*, **83**, 79–86.
- Bian, X. *et al.* (2015) Sparse null space basis pursuit and analysis dictionary learning for high-dimensional data analysis. *40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia, pp. 3781–3785.
- Bordbar, A. *et al.* (2011) A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *Bmc Syst. Biol.*, **5**, 180.
- Bordbar, A. *et al.* (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.*, **6**, 422.
- Bordbar, A. *et al.* (2014a) Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.*, **15**, 107–120.
- Bordbar, A. *et al.* (2014b) Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Mol. Syst. Biol.*, **10**, 737.
- Coleman, T.F. and Pothén, A. (1986) The null space problem. 1. Complexity. *Siam J. Algebra Discr.*, **7**, 527–537.
- Dal'Molin, C.G.D. *et al.* (2010a) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol.*, **152**, 579–589.
- Dal'Molin, C.G.D. *et al.* (2010b) C4GEM, a genome-scale metabolic model to study C-4 plant metabolism. *Plant Physiol.*, **154**, 1871–1885.
- Dal'Molin, C.G.D. *et al.* (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Front. Plant Sci.*, **6**, 4.
- De Martino, D. *et al.* (2013) Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks. *Metabolites*, **3**, 946–966.
- Desouki, A.A. *et al.* (2015) CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. *Bioinformatics*, **31**, 2159–2165.
- 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.
- Edwards, J.S. and Palsson, B.O. (2000) The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*, **97**, 5528–5533.
- Feist, A.M. *et al.* (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol. Syst. Biol.*, **2**, 2006.0004.
- Forster, J. *et al.* (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.*, **13**, 244–253.
- Gottlieb, L.A. and Neylon, T. (2010) Matrix sparsification and the sparse null space problem. *Approx. Random. Comb.*, **6302**, 205–218.
- Hay, J.O. *et al.* (2014) Integration of a constraint-based metabolic model of Brassica napus developing seeds with C-13-metabolic flux analysis. *Front. Plant Sci.*, **5**, 724.
- Heavner, B.D. *et al.* (2013) Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database-Oxford*, **2013**, bat059.
- Ishii, N. *et al.* (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science*, **316**, 593–597.
- Jamshidi, N. and Palsson, B.O. (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *Bmc Syst. Biol.*, **1**, 26.
- King, Z.A. *et al.* (2016) BiGG models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, **44**, D515–D522.
- Lewis, N.E. *et al.* (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.*, **10**, 291–305.
- Lewis, N.E. *et al.* (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.*, **28**, 1279–1291.
- Liao, Y.C. *et al.* (2011) An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *J. Bacteriol.*, **193**, 1710–1717.
- Mahadevan, R. and Schilling, C.H. (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, **5**, 264–276.
- Müller, A.C. and Bockmayr, A. (2013) Fast thermodynamically constrained flux variability analysis. *Bioinformatics*, **29**, 903–909.
- Noor, E. *et al.* (2012) A proof for loop-law constraints in stoichiometric metabolic networks. *Bmc Syst. Biol.*, **6**, 140.
- Oliveira, J.S. *et al.* (2001) An algebraic-combinatorial model for the identification and mapping of biochemical pathways. *B. Math. Biol.*, **63**, 1163–1196.
- Orth, J.D. *et al.* (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism - 2011. *Mol. Syst. Biol.*, **7**, 535.
- Orth, J.D. *et al.* (2010a) Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. *EcoSal plus*, **4**, 1.
- Orth, J.D. *et al.* (2010b) What is flux balance analysis? *Nat Biotechnol*, **28**, 245–248.
- Saa, P.A. and Nielsen, L.K. (2016) Il-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks. *Bioinformatics*, doi:10.1093/bioinformatics/btw132.
- Savelsbergh, M.W.P. (1994) Preprocessing and probing techniques for mixed integer programming problems. *ORSA J. Comput.*, **6**, 445–454.
- Schellenberger, J. *et al.* (2011) Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys J.*, **100**, 544–553.
- 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.
- Shoaei, S. *et al.* (2015) Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell Metab.*, **22**, 320–331.
- Stolyar, S. *et al.* (2007) Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.*, **3**, 92.
- Thiele, I. *et al.* (2011) A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *Bmc Syst. Biol.*, **5**, 8.
- Zhuang, K. *et al.* (2011) Genome-scale dynamic modeling of the competition between Rhodospirillum rubrum and Geobacter in anoxic subsurface environments. *ISME J.*, **5**, 305–316.