Development of an Integrated Framework for Minimal Cut Set Enumeration in Constraint-Based Models

Vítor Vieira, Paulo Maia, Isabel Rocha and Miguel Rocha

Abstract Under the realm of *in silico* Metabolic Engineering, pathway analysis approaches to strain optimization have shown a large potential as tools capable of providing an unbiased view over metabolic models. Most of these methods were difficult or impossible to use due to their heavy computational needs, since they are based in the calculation of elementary modes/minimal cut sets in large networks. However, a recent method (*MCSEnumerator*) has enabled the application of these approaches to genome-scale metabolic models. This work proposes a new software tool where this method is implemented in a novel Java library, that provides support for a plugin for the OptFlux metabolic engineering platform. Together, these tools implement the routines necessary for the calculation of minimal cut sets and their use to provide strain optimization methods. The aim is to provide an open-source software tool that includes an intuitive graphical user interface, thus facilitating its use by the community.

Keywords Metabolic pathway analysis · Constraint-based model · Strain optimization · Minimal cut sets · Metabolic engineering

1 Introduction

In recent years, genome-scale metabolic models (GSMMs) encompassing annotated whole genomes of living organisms have proved useful in predicting cell phenotypes through *in silico* methods. A particularly interesting application of GSMMs concerns the field of metabolic engineering (ME) which aims to design enhanced cell factories

V. Vieira(✉) · I. Rocha · M. Rocha
Centre of Biological Engineering, University of Minho, Braga, Portugal
e-mail: jose.vieira153@gmail.com

P. Maia
SilicoLife Lda., Braga, Portugal

M.S. Mohamad et al. (eds.), *10th International Conference on PACBB*,
Advances in Intelligent Systems and Computing 477,
DOI: 10.1007/978-3-319-40126-3_20

for products with added value in industrial biotechnology. The use of metabolic models allows for a rational design process, integrating vast amounts of data instead of trial-and-error methodologies [1, 2]. Most methods based on the use of GSMMs follow a constraint-based (CB) approach, considering various assumptions regarding cell metabolite balancing and discarding enzyme kinetics, as this information is only partially available. Various phenotype prediction, analysis and strain optimization methods have been developed using this approach [3]. Phenotype prediction methods are based on mathematical formulations that assume cell metabolism is driven towards certain goals. The methods developed so far include variants for wild-type strains, the most popular being Flux Balance Analysis (FBA) [4], and for mutant strains such as Minimization of Metabolic Adjustment (MOMA) [5] and Regulatory On/Off Minimization (ROOM) [6].

Computational strain optimization methods (CSOMs), with the purpose of finding genetic manipulation strategies able to overproduce selected compounds, have also been developed following the CB approach [7]. In this work, only optimization methods involving reaction deletions will be approached. These can be divided in two broad categories: bi-level constraint-based methods and pathway analysis (PA) methods. Bi-level approaches are nested optimization problems attempting to find engineering strategies that increase product yields or titers, as well as optimizing cellular objectives, through phenotype prediction methods, mostly FBA and related methods as MOMA or ROOM. While such assumptions allow faster computational time, assuming an objective may lead to bias which can result in less robust strategies. Some of these methods use deterministic methods, such as OptKnock [8], while others employ stochastic meta-heuristics, as Evolutionary Computation (e.g. OptGene [9]).

Pathway analysis approaches only consider the metabolite balance assumption, decreasing bias. Most are based on elementary modes (EMs), a concept representative of basic cell functions contained in a model. Complete enumeration of all EMs, and in some cases the related Minimal Cut Sets (MCSs), in a network is necessary for most PA methods, but incurs in heavy computational demands. However, EMs and MCSs [10] are important assets in strain optimization. Also, recent methods have allowed to extend the computation of EMs and MCSs to metabolic models at a genome-wide scale, through the MCSEnumerator framework [11].

In this scenario, given the potential of MCSs to guarantee robust production, regardless of the phenotype prediction method used, this work pursues the development of an open-source software tool capable of handling relevant tasks associated with their enumeration and applying those to strain optimization. Therefore, the main scientific/technological objectives of this work are:

- to implement a library containing the necessary routines for enumeration of MCSs in metabolic networks;
- to integrate MCS enumeration tasks in the OptFlux metabolic engineering platform, providing novel tools for strain optimization;
- to provide a simple and intuitive user-interface for the implemented routines.

2 Methods

2.1 Constraint-Based Models and Pathway Analysis

Constraint-based models of metabolism comprise m intracellular metabolites and n reactions acting upon them. These reactions also include sinks for external metabolites, representing their uptake and/or production. The system is represented by a $m \times n$ matrix S, containing stoichiometric coefficients. In CB methods, metabolite concentrations are assumed to be time invariant, leading to a system of linear equations:

$$S \cdot v^T = 0 \tag{1}$$

with v as the column vector of fluxes (or rates) for each individual reaction. Additionally, thermodynamics assumptions and/or rate limits are added as additional constraints in the form:

$$\alpha \leq v \leq \beta \tag{2}$$

with α and β being respectively the vectors containing lower and upper limits for each element (flux) in v. Any irreversible reaction j must have a lower limit $\alpha_j = 0$. The system defined by Equations 1 and 2 can also be represented in space as a convex polyhedron hereby referred to as P, containing all feasible solutions to this system.

Considering this modeling framework, an **elementary mode** (EM) represents the smallest functional unit within it. Any elementary mode e equates to a flux distribution obeying three key properties [12]:

1. A flux distribution in e must comply with Equation 1;
2. Irreversible reactions must carry flux only through a single direction in any EM. These are specified in Equation 2;
3. Considering **supp**(e) as the reactions carrying flux in e, no subset of $supp(e)$ can yield a flux distribution obeying Equations 1 and 2.

Any point contained within P can be defined as a linear combination of EMs. It is possible to find desirable solutions to the metabolic model by finding points described by non-null combinations of EMs contained within a desired set of flux vectors D. Conversely, any set of undesired flux vectors T can be blocked by disabling EMs contained within that space. A set of reactions blocking all vectors in T is a cut set of T. If no reaction can be removed from the cut set without rendering it unable to block the vectors in T, it is considered a **minimal cut set** (MCS). However, MCSs do not necessarily guarantee the set of desired vectors D will not be blocked as well. An MCS M is considered a **constrained minimal cut set** (cMCS) [13] if it blocks all EMs describing the space in T, as well as ensuring points in D are feasible solutions to the system.

2.2 Enumerating Minimal Cut Sets

Most methods for the enumeration of MCSs involve prior knowledge of the full set of EMs in the network and are usually based on combinatorial algorithms. However, these are unsuitable for GSMMs due to the heavy computational demand of this task. Problem complexity and the number of total EMs rise exponentially with the size of the model, rendering it virtually impossible.

However, a recent approach, MCSEnumerator, has been proposed which is capable of enumerating MCSs and cMCSs in GSMMs, in some cases up to seven knockouts [11]. This algorithm involves a mixed-integer linear programming problem (MILP) that allows partial enumeration of MCSs with good time-efficiency. This formulation derives from a finding documented in [14] and describing the formulation of a dual system in which EMs correspond to MCSs in the original metabolic model. This is currently the most suitable approach for enumeration of MCSs in GSMMs, which prompted its use in this work.

A generic pipeline based on the original publication was assembled, covering all required steps for the enumeration task, as shown in the left panel of Figure 1. The model compression step in the pre-processing phase and subsequent MCS decompression in the enumeration phase are optional, but speed up computation times. The MILP framework developed in [11] as well as a basic algorithm for MCS enumeration are represented on the right side of the same figure. These constraints, along with the dual system, result in an EM enumeration problem (using the k-shortest EFM algorithm [15]) where EMs represent MCSs in the original network.

3 Development

This work has two main outcomes regarding the developed software. The first is a *Java* library implementing an entire pipeline for MCS enumeration using the algorithm in [11] and their use for strain optimization. The second outcome is a plugin developed for the OptFlux platform, providing a user interface for the library.

3.1 Enumeration Library

MCSEnumerator is currently only available as a part of the CellNetAnalyzer platform for MATLAB. One of the aims of this work was to build an independent library containing the necessary routines and algorithms for MCS enumeration using the MCSEnumerator [11] approach. This library was built using the *Java* programming language, allowing greater portability, as well as enabling the use of more advanced tools for the development of a graphical user interface (GUI). Currently, it requires

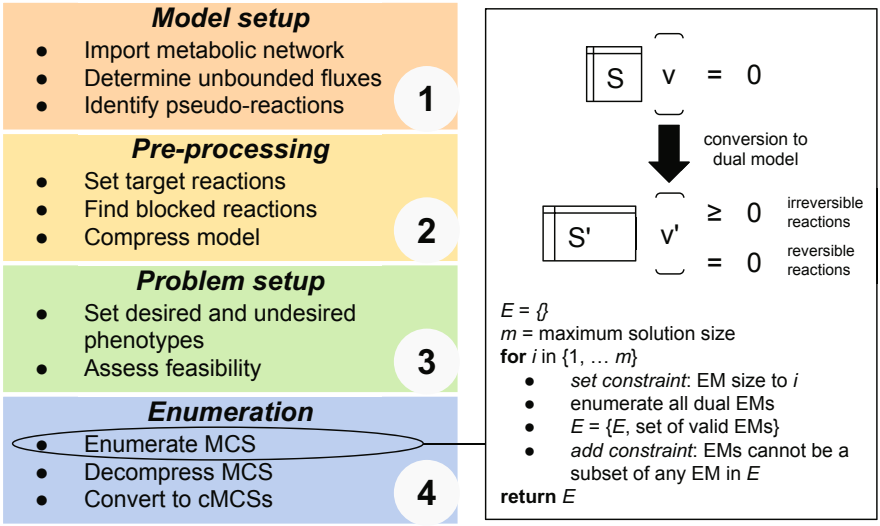

Fig. 1 **Left:** Representation of the pipeline used in this work. Step **(1)** concerns model setup which consists mainly on determining reaction reversibility and pseudo-reactions that will not be a part of any solution; Step **(2)** aims at reducing the size of the problem, mostly through removal of blocked reactions and network compression by lumping correlated reactions. Step **(3)** assembles the enumeration problem and validates it, so that in Step **(4)** the proper formulation is built and solved. MCSs that are not feasible in the desired space are discarded, leaving only cMCSs. **Right:** Brief overview of the MILP formulation. S' and v' are derived from the dual model formulation in [14] which already includes the undesired phenotypes.

the usage of the IBM ILOG CPLEX Optimization Studio[1] for solving the MILP problem described in [11].

This library contains three newly developed packages:

1. **Enumeration**: Contains methods needed to implement the MCSEnumerator MILP formulation, given a suitable problem.
2. **Metabolic**: Provides a framework upon which constraint-based metabolic models can be defined, as well as constraints for typical linear programming problems such as yield or capacity constraints on the reactions.
3. **Utilities**: Includes methods that execute the entire pipeline given a set of parameters for the optimization. Functions to run the algorithms in a command-line environment are also provided, capable of reading parameters contained within text files.

The libraries provide routines capable of executing these tasks in a command-line environment using only a metabolic model in Systems Biology Markup Language (SBML) format and a file containing the parameters of the MCS enumeration problem, which may include undesired or desired limits for fluxes or yields, exclusion of target reactions, among other constraints.

[1] http://www-03.ibm.com/software/products/en/ibmilogcpleoptistud

3.2 OptFlux Plugin

This work also aimed at providing a simple and clear user interface (GUI) as a plugin within OptFlux [16]. Currently, this framework includes important tools used in CB approaches, including phenotype simulation, analysis methods, and strain optimization algorithms developed in-house (OptGene[9] and derivatives [17]).

As far as this work is concerned, OptFlux provides the necessary methods to read and write metabolic models, serving as inputs for our algorithms. The developed plugin provides a simple GUI, shown in Figure 2 for the MCSEnumerator approach requiring minimal user input and providing a useful abstraction for the concepts discussed in the previous section. The user is only required to specify the maximum number of knockouts, which reactions correspond to biomass, product synthesis and substrate uptake, the desired thresholds for production and growth, and whether the production threshold is a yield or a rate constraint. Additionally, environmental conditions can be added, and knockout targets can be discarded from the search either by supplying a list of critical reactions or a gene ID corresponding to spontaneous reactions, should the model represent those as being associated with a placeholder pseudo-gene. The solutions are displayed using OptFlux's GUIs, using the format of previously available optimization algorithms. So, these solutions can be processed and simulated afterwards using other tools from OptFlux. From OptFlux 3.3 onwards, this plugin is available in the software's plugin repository.

4 Results

4.1 Library Validation

The set of case studies was defined with the aim of ensuring that the outputs provided by the developed software match the ones from MCSEnumerator's original implementation. As such, the iAF1260 *Escherichia coli* GSMM [18] was used with different enumeration problems for which the cMCSs were previously determined in the original publication [11]. The results from all case studies were accurately replicated and are highlighted in the Table 1.

4.2 Plugin Operation

This section shows in more detail the plugin's mode of operation using one of the case studies described above (anaerobic ethanol production in *Escherichia coli* using glucose as carbon source). To run this case study:

Table 1 Overview of the validation case studies. Y represents product/substrate yield and Glc represents glucose uptake ($mmol \cdot gDW^{-1} \cdot h^{-1}$). Note that aerobic conditions were allowed only for fumarate and serine production. Computation times were determined in a single run using 12 cores (from two Intel® Xeon® E5-2650 CPUs) and 30GB of RAM.

Objective	Scenario	#MCS/#cMCS	Computation time (h)	Maximum size
Synthetic lethals	-	1018 / -	17	4
Anaerobic ethanol production	Glc $\leq$ 10 Y $\geq$ 1.4	185302 / 8342	7.5	7
	Glc $\leq$ 10 Y $\geq$ 1.8	153338 / 1987	9.1	7
	Glc $\leq$ 18.5 Y $\geq$ 1.4	156477 / 8819	12.7	7
	Glc $\leq$ 18.5 Y $\geq$ 1.8	138675 / 4618	2	7
Fumarate production	Y $\geq$ 0.5	17338 / 30	12.4	7
Serine production	Glc $\leq$ 20	18449 / 140	1	6

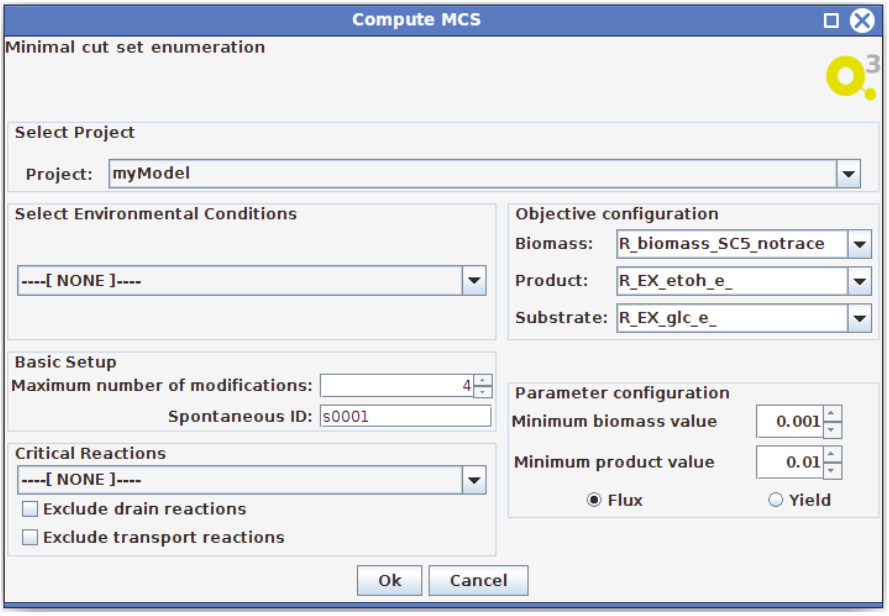

Fig. 2 Graphical interface provided by the plugin to formulate an enumeration problem.

1. Start a new OptFlux project using the **New project wizard** option, click on **OptFlux model repository** and select the iAF1260 *Escherichia coli* model. Assume default options in the process.
2. Create an environmental condition using the **New...** menu, option **Create...** and then click **Environmental condition**. Add a constraint for reaction *R_EX_glc_e* with lower bound as -20 and upper bound as 999999 (definition of glucose uptake

rate), and another for *R_EX_o2_e_* with 0 as lower bound and 999999 as upper bound (definition of anaerobic conditions).

3. Access the **Optimization** tab, and click on **Minimal cut sets**.
   a. Select the environmental condition that was created in the previous step.
   b. Allow at most 3 modifications and set the spontaneous ID for s0001.
   c. Configure the objectives as follows: Biomass as *R_Ec_biomass_core_59p81M*, substrate as *R_EX_glc_e_* and product as *R_EX_etoh_e_*.
   d. Set the biomass value to 0.1, choose yield and set the minimum product value to 0.2.

This example allows determination of up to 3 knockouts guaranteeing a production yield of at least 0.2 with a growth rate of $0.1h^{-1}$. The results can be browsed and sorted and also saved to disk as a text file. Specific solutions (deletion sets) can be saved to the clipboard, and simulated or analyzed through other OptFlux tools.

5 Conclusions and Further Work

The availability of PA-based strain design methods is scarce when considering GSMMs. The new library proposed in this work presents a useful resource for the metabolic engineering community, allowing for the enumeration of MCSs, in a way that is standardized fit for most problems with generic CB models, while also allowing flexibility regarding problem setup. The proposed OptFlux plugin facilitates an abstraction from complex concepts surrounding cMCS enumeration, improving ease of use and extending the already wide variety of optimization algorithms within OptFlux, maintaining a coherent overall computational interface. Also, the provided software is all made available to the community as open source allowing for third party contributions in the future. Despite all efforts, the library is still dependent on a commercial solver, but this provides a free academic license.

Computation times for larger sets of deletions, even when using a state-of-the-art optimizer, can be time consuming for some enumeration problems, and others still remain out of reach. Heuristic methods or alternative formulations may help in achieving solutions for larger sizes and this will be a line of future work.

Acknowledgments The authors thank the project "DeYeastLibrary - Designer yeast strain library optimized for metabolic engineering applications", Ref. ERA-IB-2/0003/2013, funded by national funds through FCT/MCTES.

References

1. Stephanopoulos, G.: Metabolic Fluxes and Metabolic Engineering 11 (1999)
2. Patil, K.R., Åkesson, M., Nielsen, J.: Use of genome-scale microbial models for metabolic engineering. Curr. Opin. Biotechnol. **15**(1), 64–69 (2004)

3. Szallasi, Z., Stelling, J., Periwal, V.: System Modeling in Cell Biology (2010)
4. Varma, A., Palsson, B.O., Arbor, A., Varma, A.: Stoichiometric Flux Balance Models Quantitatively Predict. Appl. Environ. Microbiol. **60**(10), 3724–3731 (1994)
5. Segrè, D., Vitkup, D., Church, G.M.: Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. U. S. A. **99**(23), 15112–15117 (2002)
6. Shlomi, T., Berkman, O., Ruppin, E.: Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proc. Natl. Acad. Sci. U. S. A. **102**(21), 7695–7700 (2005)
7. Maia, P., Rocha, M., Rocha, I.: In silico constraint-based strain optimization methods: the quest for optimal cell factories. Microbiology and Molecular Biology Reviews **80**(1), 45–67 (2016)
8. Burgard, A.P., Pharkya, P., Maranas, C.D.: Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. **84**(6), 647–657 (2003)
9. Patil, K.R., Rocha, I., Förster, J., Nielsen, J.: Evolutionary programming as a platform for in silico metabolic engineering. BMC Bioinformatics **6**(1), 308 (2005)
10. Klamt, S., Gilles, E.D.: Minimal cut sets in biochemical reaction networks. Bioinformatics **20**(2), 226–234 (2004)
11. von Kamp, A., Klamt, S.: Enumeration of Smallest Intervention Strategies in Genome-Scale Metabolic Networks. PLoS Comput. Biol. **10**(1), e1003378 (2014)
12. Schuster, S., Hilgetag, C.: On Elementary Flux Modes in Biochemical Reaction Systems At Steady State. J. Biol. Syst. **02**(02), 165–182 (1994)
13. Hädicke, O., Klamt, S.: Computing complex metabolic intervention strategies using constrained minimal cut sets. Metab. Eng. **13**(2), 204–213 (2011)
14. Ballerstein, K., von Kamp, A., Klamt, S., Haus, U.U.: Minimal cut sets in a metabolic network are elementary modes in a dual network. Bioinformatics **28**(3), 381–387 (2012)
15. de Figueiredo, L.F., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J.E., Schuster, S., Planes, F.J.: Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics **25**(23), 3158–3165 (2009)
16. Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J.P., Nielsen, J., Patil, K.R., Ferreira, E.C., Rocha, M.: OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst. Biol. **4**(1), 45 (2010)
17. Rocha, M., Maia, P., Mendes, R., Pinto, J.P., Ferreira, E.C., Nielsen, J., Patil, K., Rocha, I.: Natural computation meta-heuristics for the in silico optimization of microbial strains. BMC Bioinformatics **9**(1), 499 (2008)
18. Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., Palsson, B.Ø.: A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. **3**(121), 1–18 (2007)