MCS^2: Minimal coordinated supports for fast enumeration of minimal cut sets in metabolic networks

by

Reza Miraskarshahi

B.Sc., Sharif University of Technology, 2017

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY
Fall 2019

Approval

Name:	**Reza Miraskarshahi**
Degree:	**Master of Science (Computing Science)**
Title:	**MCS^2: Minimal coordinated supports for fast enumeration of minimal cut sets in metabolic networks**
Examining Committee:	**Chair:** Nick Sumner Assistant Professor
	Leonid Chindelevitch Senior Supervisor Assistant Professor
	Tamon Stephen Supervisor Associate Professor Department of Mathematics
	Cedric Chauve Internal Examiner Professor Department of Mathematics
Date Defended:	**August 21, 2019**

Abstract

Motivation: Constraint-based modeling of metabolic networks helps researchers gain insight into the metabolic processes of many organisms, both prokaryotic and eukaryotic. Minimal Cut Sets (MCSs) are minimal sets of reactions whose inhibition blocks a target reaction in a metabolic network. Most approaches for finding the MCSs in constrained-based models require, either as an intermediate step or as a byproduct of the calculation, the computation of the set of elementary flux modes (EFMs), a convex basis for the valid flux vectors in the network. Recently, Ballerstein *et al.* [1] proposed a method for computing the MCSs of a network without first computing its EFMs, by creating a dual network whose EFMs are a superset of the MCSs of the original network. However, their dual network is always larger than the original network and depends on the target reaction.

Here we propose the construction of a different dual network, which is typically smaller than the original network and is independent of the target reaction, for the same purpose. We prove the correctness of our approach, MCS^2, and describe how it can be modified to compute the few smallest MCSs for a given target reaction.

Results: We compare MCS^2 to the method of Ballerstein *et al.* and two other existing methods. We show that MCS^2 succeeds in calculating the full set of MCSs in many models where other approaches cannot finish within a reasonable amount of time. Thus, in addition to its theoretical novelty, our approach provides a practical advantage over existing methods.

Availability: MCS^2 is freely available at https://github.com/RezaMash/MCS under the GNU 3.0 license.

Keywords: Metabolic network models; Minimal cut sets; Drug target identification; Linear programming duality; Mixed Integer Linear Programming

To Baba, Tahmineh, Bahar, and Maman

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Constraint-based modeling of metabolic networks has been a major subfield of systems biology thanks to its ability to identify key qualitative characteristics of networks for analyzing and extracting useful information [29, 3, 24]. A metabolic network is a collection of chemical reactions which comprise the metabolic activities (i.e., the biochemical transformation of molecules into other molecules for the purpose of maintenance and growth) of a specific organism.

One important application of metabolic network analysis is to find interventions that can block a reaction of interest, typically referred to as the *target reaction*, with applications in drug target identification [14, 40, 44, 18, 13] and metabolic engineering [27]. When this is achieved by disabling one or more other reactions, the disabled reactions are called a *cut set*. A cut set is called "minimal" if no proper subset of it can disable the target reaction. The concept of minimal cut sets (MCS) was introduced by Klamt and Gilles [21] and its applications are examined in detail in [20].

At the moment, the main approach used for enumerating the MCSs for a target reaction is to compute the elementary flux modes containing the target and then use a dualization procedure to produce the MCSs [11]. Here, *flux modes* are possible distributions of fluxes through the reactions, and those can be modelled as hyperedges on the vertex set of possible reactions.

Elementary flux modes (EFMs) are flux modes which are support-minimal, and it is known that any flux mode can be written as a non-negative linear combination of EFMs. Given the full set of elementary flux modes, minimal cut sets can be obtained through the dualization of the hypergraph they define [21, 15]. Two approaches to do this are Berge's algorithm [2] and Fredman and Khachiyan's dualization procedure [8]. However, both suffer from poor worst-case complexity and produce mixed results in practice.

A comparatively new approach [1] produces the MCSs without first computing the EFMs. It works by generating a dual network, which is larger than the original network and depends on the target

reaction, and then computing a subset of the EFMs of that network with a specific property, which guarantees that they are precisely the MCSs in the original network. We call this the *target-specific dual network method*. In this thesis, we develop a new method, MCS^2, which also generates a dual network (either explicitly or implicitly), but in a way that is independent of the target reaction from the original network, then computes the MCSs from those EFMs of the dual network that satisfy a certain property, which also guarantees that they are precisely the MCSs for the target reaction in the original network. MCS^2 is based on a generalization of some theoretical results by Chindelevtich [4].

We implement MCS^2 and find it to be effective on most instances we test it on. We compare it to three alternate methods for enumerating all the MCSs for a target set. The first two methods are to compute the EFMs, and then dualize them with either Berge's algorithm, or an optimized implementation of Fredman-Khachiyan dualization, respectively. For Berge's algorithm, we used the implementation in CellNetAnalayzer [22], containing the enhancements described in [15, 6]. For Fredman-Khachiyan dualization, we used the recent implementation of [35].

The *target-specific dual network method* [1] first creates a dual network based on the given stoichiometric matrix and the given target reaction. It then proceeds to compute the EFMs of that dual network. Following some post-processing, the supports of these EFMs are reduced to give the required MCSs; because the MCSs correspond to only a subset of the vectors produced, this post-processing includes removing any supersets. Like MCS^2, the target-specific dual network method reports all the MCSs without first computing the EFMs or requiring them as an input. The authors of [1] did not provide a publicly available implementation of their method, so we did so ourselves, including all the enhancements mentioned in their supplementary materials. Most of these enhancements have improved the performance of the target-specific dual network method, for instance by reducing the size of the intermediate results. We remark that finding all the EFMs of this dual matrix gives a superset of the cut sets, which from which the minimal cut sets are distilled.

For the majority of the models we investigated, we find that MCS^2 is more efficient than these other methods, in terms of both running time and memory use. On the negative side, we show that our approach does not allow the enumeration of all MCSs through a given target reaction in incremental polynomial time, something that therefore remains a major open problem in the field.

Given the challenges in enumerating all MCSs (in part due to their large number, which can be exponential in the size of the network), some recent work [43, 42] uses Mixed Integer Linear Programming (MILP) formulations to enumerate a subset of the MCSs, in increasing order of size. In some practical applications, quickly obtaining a few MCSs of minimum size may be more desirable than enumerating the complete set. We therefore adapt the MCS^2 approach to use MILP formulations to address this task. We also implement this method, which we call MCS^2-MILP, and compare it to MCSEnumerator, the target-specific dual network approach adapted to MILP

[43]. The comparison shows that MCS^2-MILP performs at least as well as MCSEnumerator using a state-of-the-art MILP solver [16].

We conclude that MCS^2 is a promising approach for the computation of MCSs in metabolic networks, and expect it to be a beneficial addition to the analysis tools available for metabolic network models.

1.1 Preliminaries

Constraint based modeling is creating a mathematical model of a metabolic network which can be analyzed further using mathematical tools and algorithms. In this approach we convert the network to a matrix called the stoichiometry matrix which is filled based on contribution of each metabolite in each reaction. A constraint-based model contains other information, e.g., reversibility state of each reaction which is encoded to a single vector. It also contains other additional constraint on the reactions' flow rate.

Since most metabolites do not contribute in a given reaction, it is expected that stoichiometry matrix has many zero elements. The non-zero elements are typically small integers.

Using constraint-based approach we convert the metabolic network into a matrix with size $k \times q$ where q is the number of reactions and k is the number of metabolites. In almost every metabolic network we have $k < q < 2k$. However, there are a few exception to this rule. Calmodulin Allostery model is an example which its number of reactions is almost five times of its number of metabolites [38].

There is a core constraint on the fluxes of metabolic networks. It says the rate of production of each metabolites must be equal to the consumption rate of that metabolite. It is known as the steady-state constraint. If we set fluxes for each reaction and those fluxes satisfy the steady-state constraint and other constraints of the network (e.g., reversibility constraints), the vector of these fluxes is a flux mode of the network.

Let S be the stoichiometry matrix. Thus, $E = \{v | v = [f_1, f_2, \ldots, f_q], Sv = 0\}$ are all flux modes which satisfy the steady-state constraint. E forms a pointed cone in the q-dimensional space. Flux modes that have minimal number of active reactions are called elementary flux modes(EFM). EFMs are widely used in metabolic networks analysis and are essential in many applications of metabolic networks. Having this cone, we enumerate EFMs usually with Double Description method [9]. The set of active reactions for each EFM is unique and called the support of that EFM. We can find the EFM (up to a constant) from a given support using Linear Algebra.

1.2 Definitions

We now introduce the terminology we will be using throughout this thesis. When we speak of a metabolic network, it is understood that we are talking about a model in the constraint-based modeling formalism.

Definition 1 (Stoichiometric matrix)**.** The *stoichiometric matrix* S is an $m \times n$ matrix with each row representing a metabolite (indexed from 1 to m) and each column, a reaction (indexed from 1 to n). The entry S_{ij} indicates how many units of metabolite i are produced (if $S_{ij} > 0$) or consumed (if $S_{ij} < 0$) by reaction j. A vector v is *feasible* with respect to S if it is in the *null space* of S, i.e., if it satisfies $Sv = 0$.

Definition 2 (Reaction irreversibility)**.** The set $\mathcal{I}$ of *irreversible* reactions is a subset of the set of reactions constrained to have only non-negative fluxes. Its complement $\mathcal{I}^C$, the set of *reversible* reactions, is allowed to have fluxes of any sign. A vector v respects the reaction irreversibility constraints if $v_i \geq 0 \; \forall \; i \in \mathcal{I}$, also written as $v_{\mathcal{I}} \geq 0$.

Definition 3 (Metabolic network)**.** A metabolic network $\mathcal{M}$ is a pair $(S, \mathcal{I})$, where $S \in \mathbb{Q}^{m \times n}$ is a stoichiometric matrix and $\mathcal{I} \subseteq [n]$ is the set of irreversible reactions. A vector v is a *flux mode* if it is feasible with respect to S and respects the irreversibility constraints, i.e., $Sv = 0$ and $v_{\mathcal{I}} \geq 0$. The set of all such vectors is called the network's *flux cone*.

Definition 4 (Reconfigured network)**.** Let $(S, \mathcal{I})$ be a metabolic network. We can *reconfigure* this network by replacing S with $S' = [S \; ; \; -S_{\mathcal{I}^C}]$ and $\mathcal{I}$ with $\emptyset$. This is equivalent to splitting each reversible reaction in the network into its forward reaction and reverse reaction.

Definition 5 (Null space matrix and network)**.** Let S be a matrix. A *null space matrix* of S is a matrix whose rows form a basis of the null space of S. The *null space network* of a metabolic network with stoichiometric matrix S is the fully reversible metabolic network (i.e., with $\mathcal{I} = \emptyset$) whose stoichiometric matrix is a null space matrix of S.

Definition 6 (Positive and negative support)**.** Let v be a vector. The *positive support* of v, $\mathcal{R}_+(v)$, is the set of positions i where v_i is positive: $\mathcal{R}_+(v) := \{i | v_i > 0\}$. The *negative support* of v, $\mathcal{R}_-(v)$, is the set of positions i where v_i is negative: $\mathcal{R}_-(v) := \{i | v_i > 0\}$. Their union $\mathcal{R}(v)$ is the *support* of v: $\mathcal{R}(v) := \mathcal{R}_+(v) \cup \mathcal{R}_-(v)$.

Definition 7 (Coordinated support)**.** Let v be a vector of size n and let $A \subseteq [n]$ be a set of coordinates. The *A-coordinated support* of v, $\mathcal{R}_A(v)$, is the union of its negative support on the positions in A and its support everywhere else: $\mathcal{R}_A(v) := (\mathcal{R}_-(v) \cap A) \cup (\mathcal{R}(v) \cap A^C)$.

Definition 8 (Elementary flux mode)**.** Let $\mathcal{M} = (S, \mathcal{I})$ be a metabolic network, and let v be a flux mode of $\mathcal{M}$. It is an *elementary flux mode* (EFM) if its support is minimal among all the flux modes of $\mathcal{M}$, i.e., $Sw = 0, w_{\mathcal{I}} \geq 0, \mathcal{R}(w) \subsetneq \mathcal{R}(v) \implies w = 0$ [34, 10].

Definition 9 (Minimal cut set)**.** Let $\mathcal{M} = (S, \mathcal{I})$ be a metabolic network, and let t be a reaction. C is a *cut set* for t if $Sv = 0, v_{\mathcal{I}} \geq 0, v_C = 0 \implies v_t = 0$. C is a *minimal cut set* (MCS) if it is inclusion-minimal: $D \subsetneq C \implies \exists v$ s.t. $Sv = 0,\ v_{\mathcal{I}} \geq 0, v_D = 0, v_t \neq 0$ [21].

Definition 10 (Canonical form of a network)**.** Let $\mathcal{M} = (S, \mathcal{I})$ be a metabolic network. We say that $\mathcal{M}$ is in *canonical form* if it satisfies:

1. No blocked reactions: for every reaction i, there exists a flux vector v with $v_i = 1$;
2. Proper directedness: for every reaction $i \in \mathcal{I}^C$, there exists a flux vector w with $w_i = -1$;
3. No enzyme subsets: no reaction pair $i \neq j$ satisfies $v_i = \kappa v_j$ with $\kappa \in \mathbb{Q}$ for all flux vectors v;
4. No redundant constraints: S has full row rank.

A metabolic network can be reduced to an equivalent one in canonical form (a.k.a. compressed form) in time polynomial in m and n [4].

Chapter 2

Previous Work

2.1 Hypergraph dualization methods

For reporting the Minimal Cut Set of a given target reaction $T = \{t_1, t_2, \ldots, t_k\}$ we will gather EFMs that at least one of reactions in T is in their support. If we somehow disable these EFMs, set of target reactions T gets blocked. The collection of supports of these EFMs will create a Sperner hypergraph. A Cut Set is a set which intersects with every EFMs in our collection a.k.a. edges of the hypergraph. These sets are also known as hitting sets. Minimal hitting sets are edges of another hypergraph known as *transversal hypergraph* or *dual hypergraph*. The traditional method for reporting the minimal cut sets is finding the EFMs containing the target reactions and create a hypergraph from these EFMs. Then computing the transversal hypergraph. For computing the transversal hypergraph we can use Berge's algorithm [2] or Fredman and Khachiyan(FK) algorithm [8].

2.1.1 Berge

Let $E = \{e_1, e_2, \ldots, e_m\}$ be the set of edges of our hypergraph. Berge's algorithm finds the hitting sets in m iteration. Let $E_i = \{e_1, e_2, \ldots, e_i\}$. At iteration i Berge computes the hitting sets for E_i from the given hitting sets for E_{i-1}. Let's say we have the hitting sets for E_{i-1}. If for every vertex in e_i we add that vertex to all of the hitting sets of E_{i-1} and then remove non-minimal ones, we will have the hitting sets for E_i.

The speed of algorithm is highly depends on the set of intermediate hitting sets' size. It can be influenced by changing the order of edges in E. Therefore, naive implementation of Berge's algorithm typically have weak performance. In practical experiments, it has been seen that the following order has a better performance: ordering the rows by maximum number of zeros, then by lexicographical, then by absolute lexicographical, and then by fewest negative/positive pairs [39]. One of the most critical part of this algorithm is non-minimal sets removal. Since it is the time consuming part of the algorithm, a proper implementation of it can give a boost to the overall performance. Removing the super-set is also a critical part of the target-specific dual method which

we will talk about it in the chapter focused on our method, MCS^2 method. We use a simple approach for removing super-sets in the implementations of three method. This simple approach checks every pair of sets to see if one of them is a super-set of another, thus it needs $O(n^2)$ time. Regarding superset removal, exist algorithms that can do it in $O(n^2/\log n)$ or slightly better. Since it's not going to make a huge difference we sticked to the simple approach[30].

On the other hand, we used bitwise scripts to make sure each comparison only uses a single instruction and it can be done in one cycle of the CPU.

CellNetAnalyzer[22] have a proper implementation of Berge. It still removes the super-sets in $O(n^2)$ time but other parts are quite well-written.

Overall, Berge's algorithm is an efficient way for reporting full enumeration of minimal cut sets. It has been improved since first time it has been introduced but the main aspect of the algorithm remained the same. Almost every improvement happened either to finding a better order of edges or how to remove super-sets faster.

2.1.2 Fredman and Khachiyan

Finding the dual of a monotone Boolean function is well-known problem which is equivalent to finding the dual of a hypergraph. In our experiences, we used FK-B algorithm[35] which was first introduced by Fredman and Khachiyan[8].

Boolean functions are widely used in Theoretical computer science. In many practical usage of Boolean functions the Boolean function is monotone[12]. Thus we can look for their minimal true settings and maximal false settings. Let us describe what is a monotone Boolean function and what we mean by minimal true setting and maximal false setting.

Boolean function $f(x_1, x_2, ..., x_n)$ is monotone if by changing one of its inputs (that is $x_1, x_2, ..., x_n$) from 0 to 1 it remains 1 if it is already 1. In other words, consider two set of inputs $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ which for each $1 \leq i \leq n$ we have $x_i \leq y_i$ then f is monotone if $f(x_1, x_2, ..., x_n) \leq f(y_1, y_2, ..., y_n)$.

A minimal true setting $A = \alpha_1, \alpha_2, ...\alpha_n$ is a setting that while $f(A) = 1$ if we replace any of α_i with a smaller number then $f(A)$ is no longer equal to 1.

A minimal true setting $A = \alpha_1, \alpha_2, ...\alpha_n$ is a setting that while $f(A) = 0$ if we replace any of α_i with a larger number then $f(A)$ is no longer equal to 0.

Let M be metabolic network with reactions $r_1, r_2, \ldots, r_q$. Let $T = \{t_1, t_2, \ldots, t_k\}$ be target reactions. Define the Boolean function $f(x_1, x_2, \ldots, x_q)$ ($x_i \in \{0, 1\}$) to be 1 if $\{i|x_i = 1\}$ is a cut set for target reaction. This is a monotone function by definition. Finding a true minimal settings of this functions is equivalent to finding the minimal cut sets for T.

In this article we use an enhanced version of FK-B algorithm[35].

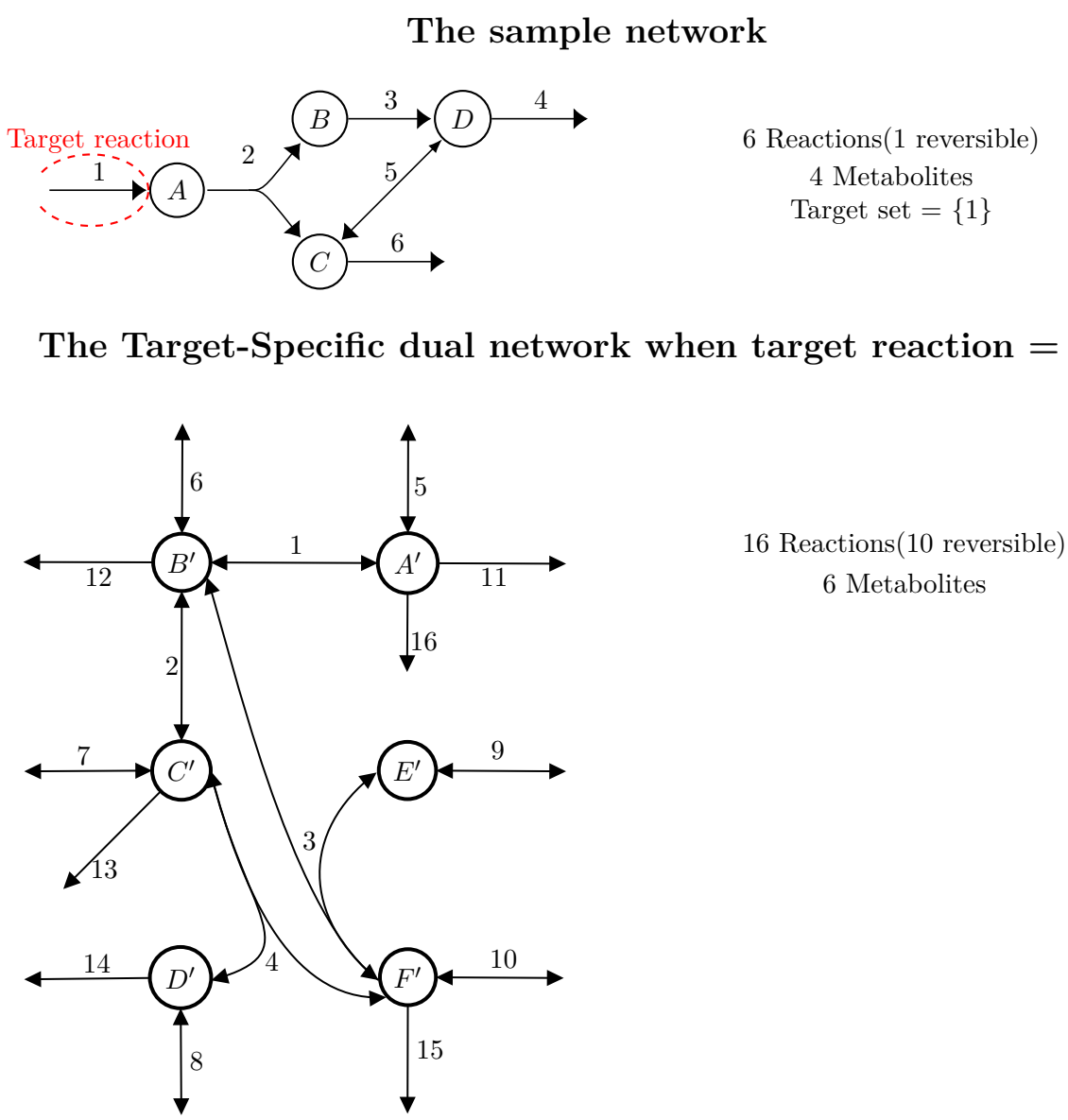


Figure 2.1: Target-specific dual network of a toy example when the target reaction has been set to the first reaction.

2.2 Enhanced Target-specific dual method

Working with any target set is the advantage of this method comparing to our method. Creating the dual network and finding the EFMs of that every time for different target set is the downside of this method. Unlike their method, we can save the time by extracting the EFMs of transposed null-space once and chose the ones that satisfies our constraints.

In this approach, this four matrices get concatenated to each other making the result matrix which is a dual network.

$$S^{\text{dual}} = \Big[S^t \Big| I \Big| - I_{\mathcal{I}} \Big| - t\Big]$$

S is the given stoichiometry matrix. I is the Identity matrix with size q which is the number of reactions or in other words number of columns in S. $I_{\mathcal{I}}$ is a $q \times |\mathcal{I}|$ matrix. It's a Identity matrix which columns corresponded to reversible reactions are removed. t is a $\{0,1\}^q$ vector with ones in the position of target reactions. In result, S^{dual} is a $q \times (m + q + |\text{irrev}| + 1)$ matrix.

The rest of the approach is simple. Find the EFMs of this network. The non-zero elements of an EFM in position between $m+1$ to $m+q$ which are correspond to reactions created by the identity matrix in this dual network is a cut-set for the defined target set. Every minimal cut-set appear at least in an EFM in this network. For the proof, see the section 2.2 in (Ballerstein, K, and et al)[1].

Let us walk through a toy example. Figure 2.1 is showing a sample metabolic network which its stoichiometry matrix and reversibility vector are:

$$\text{Stoichiometric matrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & -1 & 0 \end{bmatrix}$$

$$\text{Reversibility vector} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Having the stoichiometry matrix and reversibility vector the method will create the following stoichiometry matrix:

Stoichiometric matrix =

$$\left[\begin{array}{cccc|cccccc|cccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \end{array}\right]$$

Reversibility Vector =

$$\left[\begin{array}{cccc|cccccc|cccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

Now the stoichiometry matrix and reversibility vector of the dual is created we can rebuild the dual network(Figure 2.1).

Now every valid Elementary Flux Modes of this network which the last reaction is active in it, can be use to extract a cut set from.

For instance, consider the following EFM:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

This EFM can be used to find a cut set for reaction 1. Consider the green area. Support of this area is a cut set for reaction 1 which is $\{5, 6\}$. Not every EFM which contain last reaction produces a minimal cut set. Even we different EFMs can produces the same cut set. These limitations of this approach does not mentioned clearly in the [1].

There are some enhancements mentioned in the supplementary material of the paper [1]. In the implementation we applied these enhancements. Enhancements have favorable influence on the time or space by making the main approach faster, or saving time by make the result smaller which means fewer super-sets of MCSs or repeated MCSs.

Chapter 3

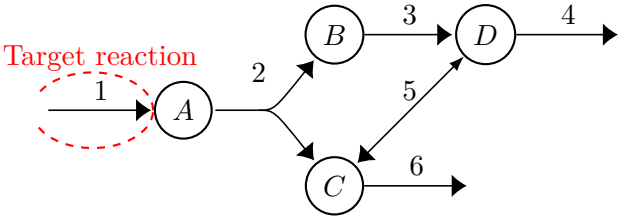
The MCS^2 Method

Let S_i be the i-th row of the stoichiometric matrix S. Then S_{ir} represents the amount of metabolite i consumed or produced by reaction r (in these cases, $S_{ir} < 0$ and $S_{ir} > 0$, respectively). Assume that reaction r produces metabolite i if it has a positive flux. Then, in a steady state where no reaction consuming metabolite i is active, reaction r must be inactive in the forward direction. If reaction r happens to be reversible, it must be consuming metabolite i, and its flux must be negative. This shows that reaction r is blocked in the forward direction if we disable every reaction that can consume metabolite i, i.e., every irreversible reaction with a negative value in row i and every reversible reaction with a non-zero value in row i. The set of such reactions is then a cut set for the forward direction of reaction r. Every row gives us some, not necessarily minimal, cut set in this manner.

We can apply the same reasoning to linear combinations of the metabolites. Consider a new *virtual* metabolite x, which represents a linear combination of rows S_i and S_j corresponding to metabolites i and j respectively, say for example $u_x = 2S_i - S_j$.

Since the fluxes producing and consuming each metabolite are balanced in any admissible vector, so are the fluxes of their linear combinations, so the virtual metabolite x must also be balanced. If we pick a reaction with a positive value in u_x, it produces a virtual metabolite x when it has a positive flux. It will therefore be blocked if we cut all irreversible reactions with negative values in v_x and all reversible reactions with non-zero values in v_x. Thus, we can obtain cut sets from the vector u_x, which is a member of the row space of S, as we did with S_i and S_j. A proposal for finding cut sets by analyzing the row space of the stoichiometric matrix was introduced in the Ph.D. thesis of Chindelevitch [4]. The intuition we described shows how vectors in the row space can generate cut sets. However, the lemmas proven in [4] only work for the fully reversible networks (i.e., with $\mathcal{I} = \emptyset$) or fully reversible networks (i.e., with $\mathcal{I} = [n]$) . We generalize them here to networks with both irreversible reactions and reversible reactions.

The sample network

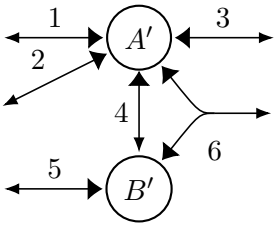


6 Reactions(1 reversible)
4 Metabolites
Target set = {1}

$$\text{Stoichiometric matrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & -1 & 0 \end{bmatrix}$$

$$\text{Reversibility vector} = \begin{bmatrix} 0, 0, 0, 0, 1, 0 \end{bmatrix}$$

The MCS^2 dual network



6 Reactions(6 reversible)
2 Metabolites

$$\text{The nullspace network's stoichiometric matrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 & 1 \end{bmatrix}$$

$$\text{Reversibility vector} = \begin{bmatrix} 1, 1, 1, 1, 1, 1 \end{bmatrix}$$

Every Minimal Cut sets for reaction 1

$$\left\{ \begin{array}{cc} \{4,6\} & \xleftarrow{\mathcal{I}\text{-Coordinated support}} [1, 0, 0, \frac{-1}{2}, 0, \frac{-1}{2}] \\ \{3\} & \xleftarrow{\mathcal{I}\text{-Coordinated support}} [1, 0, -1, 0, 0, 0] \\ \{5,6\} & \xleftarrow{\mathcal{I}\text{-Coordinated support}} [1, 0, 0, 0, 1, -1] \\ \{4,5\} & \xleftarrow{\mathcal{I}\text{-Coordinated support}} [1, 0, 0, -1, -1, 0] \\ \{2\} & \xleftarrow{\mathcal{I}\text{-Coordinated support}} [1, -1, 0, 0, 0, 0] \end{array} \right\} \text{Every EFMs in nullspace network}$$

Figure 3.1: Example of a metabolic network and an associated row-space network. In this example no redundant Cut Set(non-minimal or repetitive Cut Sets) is produced.

3.1 Enumerating the full set of MCSs

In the MCS^2 method, the dual network is the null space network of the original network, i.e., a fully reversibile network whose stoichiometric matrix is the null space matrix of the original stoichiometric matrix. The EFMs in this dual network map to minimal cut sets of the original network, though, as we will see, the mapping can be many to one. The dual network has the

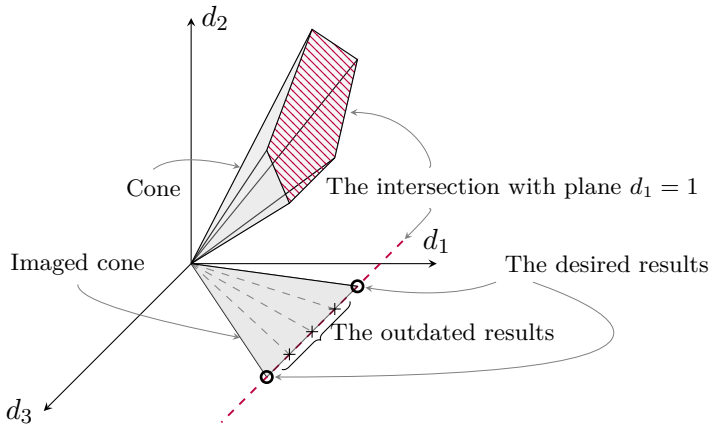


Figure 3.2: Each extreme ray of the projected cone is an image of an extreme ray in the original cone, while some extreme rays of the original cone do not project to extreme rays. It is also possible that two or more extreme rays in the original cone project onto the same one. Our desired projections lie in the plane where the value in the target position is 1.

same number of reactions, n, as the original one, but it typically has fewer metabolites; if the original network has m metabolites and the stoichiometric matrix is full rank, the dual has $n - m$ metabolites.

Lemma 1 (MCSs for an irreversible reaction)**.** Let $\mathcal{M} = (S, \mathcal{I})$ be a metabolic network. Let $t \in \mathcal{I}$ be an irreversible target reaction. Then C is a cut set for t if and only if there exists a vector u in the row space of S, $u \in Row(S)$, such that $u_t = 1$ and $\mathcal{R}_{\mathcal{I}}(u) \subseteq C$.

Proof. This lemma is an extension of Lemma 3 of [4]. We observe that C being a cut set for irreversible reaction t is equivalent to:

$$S_{-C}v = 0 \text{ and } \forall i \in \mathcal{I} - C \quad v_i \geq 0 \implies v_t = 0 \tag{3.1}$$

Based on Farkas' Lemma and the irreversibility of t, we only need to find a constraint that implies $v_t \leq 0$. Thus, there exists a y such that:

$$y^T S_{-C} = e_t + \sum_{i \in \mathcal{I} - \{C \cup t\}} \alpha_i e_i, \text{ s.t. } \forall i \in \mathcal{I} - \{C \cup t\} \; \alpha_i \geq 0 \tag{3.2}$$

Here, e_i is a vector with a 1 in the i-th position and 0 elsewhere. Thus:

$$y^T S = u, u_t = 1, \forall i \in \mathcal{I} - C \; u_i \geq 0 \; , \; \forall j \in \mathcal{I}^C - C \; u_j = 0 \tag{3.3}$$

Therefore $\mathcal{R}_-(u) \cap \mathcal{I} \subseteq C \cap \mathcal{I}$ and $\mathcal{R}(u) \cap \mathcal{I}^C \subseteq C \cap \mathcal{I}^C$, and so $\mathcal{R}_{\mathcal{I}}(u) \subseteq C$.

$$\left[\begin{array}{c|c} S & -S \\ \hline I & I \end{array}\right]$$

Figure 3.3: This $(m+n) \times 2n$ matrix is the nullspace of the reconfigured nullspace of stoichiometry matrix S. The double description method begins on this space and finds extreme rays with length $2n$.

For the other direction, suppose that u satisfies equation (3.3) for some C. Then the union of the indices i in $\mathcal{I}$ for which $u_i < 0$ and the indices i in $\mathcal{I}^C$ for which $u_i \neq 0$ is a subset of C, i.e., $\mathcal{R}_{\mathcal{I}}(u) \subseteq C$. Then equality (3.2) holds and, by Farkas' lemma, so does condition (3.1). □

Lemma 2 (MCSs for one direction of a reversible reaction)**.** Let $\mathcal{M} = (S, \mathcal{I})$ be a metabolic network. Let t be a reversible target reaction. Then C is a cut set for the forward (reverse) direction of t if and only if there exists a vector $u \in Row(S)$ such that $\mathcal{R}_{\mathcal{I}}(u) \subseteq C$ and $u_t = 1$ ($u_t = -1$), respectively.

Proof. If we assume that t is irreversible for a moment, the first part already follows from the previous lemma. For the second part, replace t with $-t$ in S to create S'. Reaction t is blocked in the forward direction in S' if and only if reaction t is blocked in the reverse direction in S, and there is a bijection between the vectors in $Row(S)$ and those in $Row(S')$ via the mapping that multiplies the t-th coordinate by -1. □

With these lemmas, Algorithm 1 below can be used to find the minimal cut sets for a set of target reactions $T = \{t_1, t_2, ..., t_k\}$ in an arbitrary metabolic network $\mathcal{M} = (S, \mathcal{I})$, where T has separate elements for the opposite directions of a reversible reaction. We call this method MCS^2 because it computes MCSs as the minimal coordinated supports (MCSs) of the elementary flux modes in the dual network.

Algorithm 1 MCS enumeration via the MCS^2 method

Input: A metabolic network $\mathcal{M} = (S, \mathcal{I})$, and a target set $T = \{t_1, t_2, \ldots, t_k\}$
Output: Minimal cut sets of target reactions T.

1: **function** $MCS_$ENUMERATION$(S, \mathcal{I}, T)$
2: Reduce S to its canonical form.
3: Compute the null space matrix N of S.
4: Compute all elementary flux modes $\mathcal{F}$ of N.
5: **for all** $1 \leq i \leq k$ **do** Compute $\mathcal{F}_i$, the set of all elements of $\mathcal{F}$ involving target t_i.
6: **for all** $1 \leq i \leq k$ **do** Let $\mathcal{C}_i$ be the set of minimal $\mathcal{I}$-coordinated supports of the elements of $\mathcal{F}_i$.
7: Let $\mathcal{C} = \{x = x_1 \cup x_2 \cup ... \cup x_k | x_i \in \mathcal{C}_i \; \forall \; i\}$.
8: Let $\mathcal{C}'$ be the result of pruning $\mathcal{C}$ to remove any supersets.
9: Return $\mathcal{C}'$.

The MCS^2 method computes the null space network of the original network. Then, by applying coordinated support to the EFMs in which each target reaction is active in turn, it generates the cut sets related for this target. All the MCSs are among these cut sets, and are obtained by pruning. The null space network is fully reversible, and since the null space of the null space is the original space, the dual network of the dual network is equivalent to the original network all of whose reactions have become reversible. An example is shown in Figure 3.1, where the dual network appears alongside the original one. As shown in the example, flux modes with an active target reaction in the dual network map to cut sets for the target reaction in the original network.

Flux modes finders such as FluxModeCalculator [41] reconfigure the network before applying the double description method. The double description method is an algorithmic approach for finding the extreme rays of a pointed cone described by linear constraints. The reconfigured network is $N' = [N\ ;\ -N]$, where N is the null space network of S. Figure 3.3 shows the null space matrix of N', which is the starting point of double description method. The double description method begins by using elementary row operations to put the matrix in the form suggested in [45] which contains an identity matrix of size $m+n$. When the double description method finishes, it outputs the extreme rays describing a cone in $2n$-dimensional space [39]. These extreme rays are the non-zero vectors in the flux cone with minimal support. On the other hand, $\mathcal{I}$-coordinated support does not count non-zero values in some dimensions, namely, those that correspond to positive values in irreversible reactions. If we ignore these dimensions, we project the cone into a lower-dimensional subspace. While the image of a pointed cone remains a pointed cone, the extreme rays of the new cone are those in the original flux cone with minimal support in the remaining dimensions. Figure 3.2 shows why all the minimal coordinated supports are among the minimal supports, while arguing that there may be some redundant results among them as well. The next theorem formalizes this:

Theorem 1 (Correctness of the method)**.** Algorithm 1 returns precisely the set of minimal cut sets of the network $\mathcal{M}$ for a single target reaction.

Proof. Let t be the target reaction. We prove the inclusion in both directions. First, let $C \in \mathcal{C}'$ be one of the sets returned by the method above. Then C is a cut set for t in the reconfigured network, by Lemma 1 and by construction. Indeed, $\mathcal{F}$ contains flux modes of N involving t, which are precisely the vectors in the row space of S involving t, and $\mathcal{C}$ (as well as $\mathcal{C}'$) contains the $\mathcal{I}$-coordinated supports of these vectors.

Now, let C be a minimal cut set for t in $\mathcal{M}$. We will show that $C \in \mathcal{C}'$. By lemma 1, there exists a vector $u \in Row(S)$ such that $u_t = 1$ and $C = \mathcal{R}_{\mathcal{I}}(u)$. Since $u \in Row(S) \iff u \in Null(N)$, u is a conical combination of the elementary flux modes of N. Note that the results of Müller and Regensburger [28] imply that since the space to which u belongs is linear (i.e., it does not need to satisfy any non-negativity constraints), this conical combination can be chosen to be **conformal**, meaning that there are no cancellations involved in any component. Let such a conformal conical

combination be given by

$$\forall 1 \leq i \leq k \quad u = \alpha_1 f_1 + \cdots + \alpha_k f_k \quad \text{where} \quad \alpha_i > 0 \tag{3.4}$$

Since all the coefficients are strictly positive in (3.4), we deduce that

$$\mathcal{R}_{\mathcal{I}}(u) = \mathcal{R}_{\mathcal{I}}(f_1) \cup \cdots \cup \mathcal{R}_{\mathcal{I}}(f_k).$$

Indeed, each $j \in \mathcal{R}_{\mathcal{I}}(u) \cap \mathcal{I}$ must have a negative component in at least one of the f_i, as otherwise the j-th component of the right-hand side of (3.4) will be non-negative, which gives the $\subseteq$ direction, and the fact that the combination is conformal gives the $\supseteq$ direction, as otherwise there would be a cancellation.

In particular, we deduce that $\mathcal{R}_{\mathcal{I}}(f_i) \subseteq \mathcal{R}_{\mathcal{I}}(u)$ for each $1 \leq i \leq k$. In this case, the minimality of C implies that either $\mathcal{R}_{\mathcal{I}}(f_i) = \mathcal{R}_{\mathcal{I}}(u)$ or f_i has a 0 in position t, for each $1 \leq i \leq k$. But since u has a 1 in position t, there must be at least one f_i in the first category, so that $\mathcal{R}_{\mathcal{I}}(f_i) = \mathcal{R}_{\mathcal{I}}(u) = C$ and therefore, $C \in \mathcal{C}$. Once again, by the minimality of C we conclude that $C \in \mathcal{C}'$ since it cannot be a superset of the $\mathcal{I}$-coordinated support of another vector in $\mathcal{C}$, concluding the proof. □

3.1.1 Limitations

Our method is limited to blocking one direction of a given reaction. However, in practice, blocking one direction of a given reaction is the typical objective [23]. To block multiple reactions it is possible to compute the MCSs of every target reaction, take unions of all possible combinations, then remove the supersets. However, this may not always be efficient.

A more critical issue is the possibility of generating a large number of non-minimal cut sets before the post-processing. The following Lemma shows that this type of blow-up can occur in theory:

Lemma 3 (Large number of supersets in the final step, proposed and proved by Leonid Chindelevitch, the thesis author's senior supervisor)**.** For every integer $k \geq 2$ there exists a network containing $k+2$ metabolites, $3k+3$ reactions and $2^{k-1}+1$ elementary vectors for the target reaction $t = 1$ that map to the exact same minimal cut set. This network is in canonical form, and is elementally balanced as per [46].

Proof. We construct the network as follows. There are two special metabolites, denoted M_I (initial) and M_F (final), and k intermediate metabolites, denoted M_i for $1 \leq i \leq k$. For each metabolite, we have an export reaction and an import reaction, with the export reactions for each intermediate metabolite coupled with an import of the final metabolite. Lastly, each intermediate metabolite except the first one can be transformed into the first one, M_1, which itself can also be transformed into the initial metabolite M_I. All reactions in the network are irreversible and all the stoichiometric coefficients are ± 1.

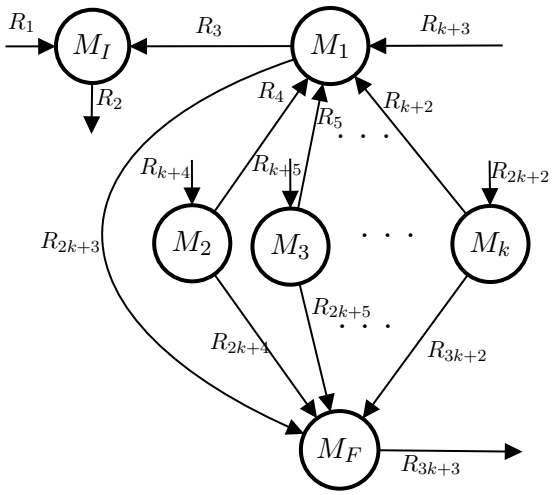


Figure 3.4: This sample network will have $2^{k-1}+1$ EFMs with $t=1$ which will map to a single cut set.

Figure 3.4 shows the constructed network.

We order these reactions as follows (for simplicity of argument):

$$\begin{aligned}
R_1 &: \emptyset \to M_I \\
R_2 &: M_I \to \emptyset \\
R_3 &: M_1 \to M_I \\
R_{3+i} &: M_{i+1} \to M_1 && 1 \le i \le k-1 \\
R_{k+2+i} &: \emptyset \to M_i && 1 \le i \le k \\
R_{2k+2+i} &: M_i \to M_F && 1 \le i \le k \\
R_{3k+3} &: M_F \to \emptyset.
\end{aligned}$$

The stoichiometric matrix then looks as follows (shown for $k=2$):

$$S = \begin{pmatrix} + & - & + & & & & & & \\ & & - & + & + & & - & & \\ & & & - & & + & & - & \\ & & & & & & + & + & - \end{pmatrix}$$

Here, a $+$ represents a 1 and a $-$ represents a -1. We now proceed to show each part of the desired statement:

- The network is elementally balanced because every reaction that is not pure import or pure export is an exchange of one metabolite for another in a 1-1 ratio, so we can consider each metabolite as containing exactly 1 atom.

- The network is in canonical form because every reaction can be active and no pair of reactions is constrained to have proportional fluxes; this is evidenced by the following flux modes:

$$\begin{aligned}
&R_1 + R_2 && \\
&R_2 + R_3 + R_{k+3} && \\
&R_1 + 2R_2 + R_3 + R_{k+3} && \\
&R_2 + R_3 + R_{3+i} + R_{k+3+i} && 1 \leq i \leq k-1 \\
&R_{k+2+i} + R_{2k+2+i} + R_{3k+3} && 1 \leq i \leq k
\end{aligned}$$

  This set of fluxes includes every reaction at least twice, and in at least two of these the sets of other active reactions are disjoint. The only exceptions are R_1 and R_3, which need R_2 to be active in order to occur, but the first and third flux modes (respectively second and third flux modes) show that their fluxes are not proportional to that of R_2 or to each other; and and the reactions R_{3+i} and R_{2k+2+i} for $1 \leq i \leq k-1$, both of which need R_{k+3+i} to be active in order to occur, but not in a fixed ratio, as evidenced by taking linear combinations of the last two sets of flux vectors.

- The stoichiometric matrix has full row rank, i.e., no metabolite generates a redundant constraint, because every metabolite except M_F has a pure import reaction, while M_F has a pure export reaction.

- There is a unique MCS for target reaction R_1, namely, R_2. This is because R_2 is the only reaction consuming M_I (recall that all reactions are irreversible). The first row of S, $u := e_1 - e_2 + e_3$ produces this MCS via its negative support.

- Lastly, there are 2^{k-1} additional vectors in the row space of S that produce supersets of this MCS via their negative support. The first one is obtained by adding the second row of S to u, replacing it by

$$v_\emptyset := e_1 - e_2 + e_4 + e_5 + \cdots + e_{k+2} + e_{k+3} - e_{2k+3},$$

  and then picking any subset P of the set of $k-1$ entries $E := \{4, 5, \ldots, k+2\}$ to form a new vector v_P, as follows.

  Let $3 + j \in P$ be an element of the chosen subset, where $1 \leq j \leq k-1$. We will replace e_{3+j} with $e_{k+3+j} - e_{2k+3+j}$ via the addition of the $(j+2)$-nd row of S (corresponding to the intermediate metabolite M_{j+1}) to the starting vector. Indeed, this row contains three nonzero entries: a -1 from reaction R_{3+j} (which cancels out the 1 in position e_{3+j}), as well as another 1 from reaction R_{k+3+j} and another -1 from reaction R_{2k+3+j}. We do this addition

independently for each element of P to get v_P (if $P = \emptyset$ we get $v_\emptyset$). It is easy to check that v_P has support:

$$\{1, 2, 2k+3\} \cup \{3+j \mid 3+j \notin P\} \cup$$
$$\{3+k+j \mid 3+j \in P\} \cup \{3+2k+j \mid 3+j \in P\}.$$

No proper subset of this support can produce a non-trivial vector in the row space of S, as it is impossible by construction to add a linear combination (possibly with negative coefficients) of the rows of S to v_P without adding any new elements to its support, so each v_P is elementary. Furthermore, the negative support of v_P is:

$$\{2, 2k+3\} \cup \{3+2k+j \mid 3+j \in P\},$$

which is a strict superset of the negative support $\{2\}$ of u.

□

3.1.2 Advantages

An advantage of the MCS^2 approach is that we find the MCSs directly, without first computing the EFMs of the original network. Also, we do not need to reconfigure or alter the stochiometric matrix; every step is performed directly on the original stoichiometric matrix or its null space matrix. Network compression or reduction may be done as a preprocessing step before going through the main procedure, but these are only used to reduce the running time and space and are optional. These advantages are shared with the target-specific dual network approach.

However, there are additional advantages that MCS^2 has over this method. First, the null space matrix is typically smaller than the original matrix, especially if the original matrix is nearly full-rank, while the target-specific dual network has a matrix that is always larger. This difference in input size can lead to substantial resource savings during EFM computation. Second, and perhaps most importantly, the dual network is independent of the target reaction in our method, while it is not with the target-specific approach. This means that we can calculate these EFMs once and use them for any given target reaction to be blocked.

3.2 Generating small MCSs

An alternative strategy for computing EFMs is via mixed integer linear programming (MILP), particularly when only a few small MCSs are required instead of a full enumeration [32, 33]. Recall that EFMs are minimal-support vectors in the null space. Our method for finding small MCSs, which we call MCS^2-MILP, similarly looks for vectors with minimal coordinated support in the row space, which is to say, EFMs with minimal coordinated support in the dual network.

Lemma 4 (MCSs of a target set of reactions in a fully irreversible metabolic network)**.** Let S be the stoichiometric matrix of a fully irreversible metabolic network $\mathcal{M}$. Let T be a set of target reactions. Then C is a cut set for all the reactions in T if and only if there exist a vector $u \in Row(S)$ such that $T \subseteq \mathcal{R}_+(u)$ and $\mathcal{R}_-(u) = C$.

Proof. We need to show that every cut set for the target reactions arises from a vector in the row space with the described constraints, and every vector in the row space satisfying those constraints maps to a cut set.

Let C be a cut set for all reactions in T. Therefore, C is a cut set for each reaction in $T = \{t_1, t_2, ..., t_k\}$ individually. Based on Lemma 1, there exist vectors $u_1, u_2, ..., u_k \in Row(S)$ such that $t_i \in \mathcal{R}_+(u_i)$ and $\mathcal{R}_-(u_i) = C$ for $1 \leq i \leq k$. In other words, for all vectors u_i $(1 \leq i \leq k)$ the only negative elements are the ones with indices belonging to C, and all other elements are non-negative, with a strictly positive value in the one with index t_i in the vector u_i, for $1 \leq i \leq k$. If we define the vector $u := u_1 + u_2 + ... + u_k$, then $\mathcal{R}_-(u) = C$ and $T \subseteq \mathcal{R}_+(u)$, and u is clearly in $Row(S)$.

Now, let u be a vector in $Row(S)$ such that $T \subseteq \mathcal{R}_+(u)$ and $\mathcal{R}_-(u) = C$. Then $t_i \in \mathcal{R}_+(u)$ for all $1 \leq i \leq k$. Based on Lemma 1, $C = \mathcal{R}_-(u)$ is a cut set for the reaction t_i, for each $1 \leq i \leq k$. Therefore, C is a cut set for all the reactions in T, completing the proof. □

Based on this Lemma we are able to find minimal cut sets for every set of target reactions without the restriction of only blocking one direction of a reaction. Since reversible reactions can be split into two reactions after reconfiguration, we can block a reversible reaction in one direction or in both directions.

Let S' be the $m \times n'$ reconfigured matrix of the $m \times n$ stoichiometric matrix S with irreversible reactions $\mathcal{I}$. Since all the values in the stoichiometric matrix are proportions of consumed and produced metabolites, we can scale each row of S to have only integer entries without changing its structural properties.

We now describe how to encode the problem of finding the smallest MCS for a target set T as a mixed-integer linear program (MILP). Let $v \in \mathbb{Z}^{n'}$ be a vector in the row space of the reconfigured matrix S' corresponding to the smallest MCS for target reaction set $T = \{t_1, t_2, ..., t_k\}$. Then there exists a vector $y \in \mathbb{Z}^m$ s.t $y^T S = v$. If we define $r^+, r^- \in \{0,1\}^{n'}$ as the indicator vectors of the positive and negative supports of v, respectively, we may force v_i to be non-positive if r_i^+ is 0, and force it to be non-negative if r_i^- is 0, by adding the following constraints using r_i^- and r_i^+ as indicator variables[17]:

$$r_i^+ = 0 \implies v_i \leq 0 \qquad\qquad r_i^- = 0 \implies v_i \geq 0 \tag{3.5}$$

There must also be positive values in the target positions:

$$r_i^+ = 1 \qquad\qquad \forall\, i \in T \tag{3.6}$$

These constraints also ensure that $v = 0$ is not in our feasible space. To make v a vector in the row space of S' we need to add the y_j variables, namely, the entries of a vector y with size m. The constraint $y^T S' = v$ then ensures that v is an element of the row space of S'.

The objective function for finding the smallest minimal cut set is:

$$\text{minimize} \sum_{i=1}^{n'} r_i^-, \tag{3.7}$$

since the cut set is precisely the negative support of v, i.e r^-.

Suppose that we have found the smallest MCS $\mathcal{C} \subsetneq \{1, 2, ..., n'\}$. To find the next smallest MCS, we need to exclude $\mathcal{C}$ and all its supersets from our feasible space. This is achieved by the following constraint:

$$\sum_{i \in \mathcal{C}} r_i^- \leq |\mathcal{C}| - 1 \tag{3.8}$$

We can keep excluding newly found MCSs and thus enumerate them in order of increasing size. As we stated above, in most scenarios we only wish to block an irreversible reaction or one direction of a reversible reaction. In those cases, we can avoid re-configuring the network to have a smaller stoichiometric matrix. Let t be the only target reaction. Instead of the constraints (3.6), we only need one constraint $r_t^+ = 1$ if we want to block it in forward direction, and we need the constraint $r_t^- = 1$ if we need to block it in the reverse direction. The objective function (3.7) and constraints (3.8) can be updated as follows to reflect the coordinated support instead of the negative support:

$$\text{minimize}\Big(\sum_{\substack{i=1 \\ i \neq t}}^{n} r_i^- + \sum_{\substack{i \in \mathcal{I} \\ i \neq t}} r_i^+\Big) \qquad\qquad \sum_{i \in \mathcal{C}} r_i^- + \sum_{i \in \mathcal{C} \cap \mathcal{I}} r_i^+ \leq |\mathcal{C}| - 1$$

Unlike [43], our problem formulation does not require any additional constraints, because they only reduce a part of the feasible space of our problem without affecting the optimum objective value. This concerns constraints such as $r_i^+ + r_i^- \leq 1$ and $r_i^+ \implies v_i > \epsilon$.

Chapter 4

Implementation Details

Except where noted, the implementations we discuss are in MATLAB. Each method that we consider requires an extreme ray computation, with the underlying cone varying. We used FluxModeCalculator's EFM generator [41] for this purpose. Note that the optimized Berge algorithm implemented by CellNetAnalyzer [22] uses the older EFM finder of CellNetAnalyzer by default. However, we observed that it is a slower implementation of an identical calculation, so we rewrote this part to use FluxModeCalculator in order to make a fair comparison. The MCS^2 method and the target-specific dual network method both need to remove redundant supersets from the obtained extreme rays, since the desired minimality is not with respect to the full support of the vector, but only a specific part of it (the $\mathcal{I}$-coordinated support for the former, and support in the v-coordinates for the latter). We use an implementation in Java whose time complexity is $O(N^2)$ for a collection of N sets. All stoichiometric matrices are compressed by the Mongoose [5] before processing, which converts them to a canonical form.

Since the null space is needed for the row space method, we calculate the null space basis matrix using Mongoose [5]. Since finding the MCSs in every method takes several seconds to several minutes, and the computation time of the null space basis matrix is less than a second in every case, we ignore this component of it. The reduced matrix given by Mongoose (for Berge and MFK), the target-specific dual matrix (for the target-specific dual method), and the null space basis matrix (for MCS^2) get further compressed by FluxModeCalculator before processing. For the Berge algorithm we used CellNetAnalyzer [22]. We also used an existing implementation of the improved modified Fredman-Khachiyan (MFK) algorithm [35]. However, we implemented the target-specific dual method from scratch using MATLAB and the source code of FluxModeCalculator. All the enhancements mentioned in the supplementary material of the original paper [1] were implemented as well.

We used CPLEX [16] to solve the MILPs. The implementation was done via the Java API and has been implemented for single target reactions without network reconfiguration, and for multiple target reactions with network reconfiguration. Since the stoichiometric matrices needed to contain only integers, we used the `integralize` function of MONGOOSE [5] to multiply each row by the

smallest possible integer that makes all the values integer (which is the least common multiple of the denominators of its entries). We also tested the results of our MILP in small networks against other implementations to make sure that the results are consistent. The implementation of all the methods and the MILP version of our method are publicly available at `https://github.com/RezaMash/MCS` under the GNU 3.0 license. Some of them require the use of non-public modules available for academic use, such as CellNetAnalyzer [22] and CPLEX [16].

Chapter 5

Main Results

In this section we summarize the performance of MCS^2 and MCS^2-MILP in comparison to the other methods.

Table 5.1: Result of running the methods on the *hepatic polyamine and sulfur amino acid network* [31]. $m = 53, n = 73$; target reaction 1.

All times are in seconds.	Optimized Berge	Improved MFK	Target-specific Dual network	MCS^2 Dual network
extreme ray computation	270.2	270.2	1191.9	79.8
secondary process time	>18000	>18000	591.3	157.4
total time	>18000	>18000	1783.2	237.2

Table 5.2: Result of running the methods on the *kinetic model of yeast* network [37] with $m = 295, n = 285$; all the reactions were used as targets.

All times are in seconds.	Optimized Berge	Improved MFK	Target-specific Dual network	MCS^2 Dual network
extreme ray computation	86.0	86.0	>18000	53.0
secondary process time	>18000*	>18000	-	13.6
total time	>18000*	>18000	>18000	66.6

*: Berge computed the MCSs for the first five reactions before running out of time. The MFK and target-specific dual methods were not able to finish the computation of the MCSs even for the first reaction.

We ran the implementations on the first five models in our database in the GitHub repository. We compared the set of MCSs in these five small examples to confirm that all the implementations

Table 5.3: Result of running the methods on *Fernandez2006 ModelB* [7] with $m = 75, n = 152$; target reaction 1.

All times are in seconds.	Optimized Berge	Improved MFK	Target-specific Dual network	MCS^2 Dual network
extreme ray computation	99.5	99.5	>18000	>18000
secondary process	2.1	1445.1	-	-
total time	101.6	1544.6	>18000	>18000

This is an example where MCS^2 and the target-specific dual methods could not finish in time, while the Berge and MFK methods reported all 194,689 MCSs for the compressed network's reaction 1 fairly quickly.

Table 5.4: Result of running MCS^2-MILP and MCSEnumerator on the *E coli iAF1260* network with 2382 reactions (981 reactions after compression).

Method used	Average Number of MCSs	Average time for shortest MCS	Number of targets MILP failed on
MCS^2-MILP	12.74 MCSs	4.45 seconds	17
MCSEnumerator	12.07 MCSs	5.22 seconds	13

produced same results. For the other results presented in the tables we again checked the number of MCSs reported by implementations if they finished, and the numbers matched in all cases. We then ran the methods on several models from the BioModels database [25]. There were a few models on which our method either was not able to finish in the given time (5 hours) or took much longer to report the MCSs, while the optimized Berge was able to finish in time and beat our method (see Table 5.3 for an example). This is due to the large number of supersets generated in that example by MCS^2. However, MCS^2 always performed better than the target-specific dual network approach, despite all its suggested enhancements being implemented. In addition, as can be seen for the *hepatic polyamine and sulfur aminoacid combined* model [31], the Berge and MFK methods could not finish in five hours, but MCS^2 generated results in 4 minutes, and the dual method in 30 minutes.

The first task of every method is an extreme ray computation, which for Berge and MFK is the well-known EFM computation. Berge and MFK then proceed to generate the MCSs through dualization, while the secondary process of the target-specific dual network and MCS^2 approaches is removing the redundant cut sets. In the first two provided examples, the target is the forward direction of the first reaction. Table 5.2 shows the computation time for calculating the MCSs for all possible target reactions. In the *kinetic model of yeast metabolic network*, described in [37], our method's advantage is clear - it was able to finish computing the MCSs for all the reactions in under 14 seconds. Note that the dimensions stated in the tables are the ones before compression is

Table 5.5: Number of non-minimal cut sets production for Target-specific dual method and the MCS^2 method and the ratio between them

Model Name	nMCS in Ballerstein	nMCS in MCS^2	ratio
hepatic polyamine and sulfur amino acid	12,321,792	3,537,920	3.482778582
Ung2008_EGFR_Endocytosis	65,408	63,360	1.032323232
kinetic model of yeast	>50,000,000	263424	>190
Kotte2010_Ecoli_Metabolic_Adaption	>40,000,000	222,976	>180
Fernandez2006_ModelA	48832	720	67.82222222

Table 5.6: First 10 smallest Minimal Cut Sets for Li2012 Model [26] when reaction 10 is the target reaction.

Minimal Cut Set number	Members of the Minimal Cut Set
1	6, 28, 29, 30, 139, 182, 213, 377
2	5, 6, 7, 8, 132, 164, 196, 372
3	5, 6, 7, 8, 69, 70, 71, 72, 164, 196, 372
4	6, 28, 29, 30, 70, 89, 96, 98, 182, 213, 377
5	5, 6, 7, 8, 132, 164, 290, 291, 292, 293, 372
6	6, 28, 29, 30, 139, 182, 291, 310, 317, 319, 377
7	5, 6, 7, 8, 132, 196, 227, 228, 229, 230, 356, 372
8	6, 28, 29, 30, 139, 182, 213, 358, 405, 424, 431, 433
9	5, 6, 7, 8, 132, 164, 196, 404, 405, 406, 407, 525
10	5, 6, 7, 8, 132, 164, 196, 404, 405, 406, 407, 523

applied. The conclusion is there are models for which it was not feasible to enumerate the full set of MCSs for a given target reaction before our work, but it is feasible now with MCS^2.

Looking at table 5.5 you can see the summary of non-minimal cut sets produced by the Target-specific dual method and MCS^2 method. The greater sign means the job of producing cut sets (the first part of the algorithm) did not finish in job and this is the number of non-minimal cut sets produced in that time.

We ran the MILP versions on larger networks alongside the MILP version of the target-specific dual approach, as described in [43]. This version is also a part of CellNetAnalyzer and is believed to be the state-of-the-art for extracting some of the smallest MCSs in increasing order of size. We were able to compute 100 MCSs for reaction 10 (the first reaction with at least 100 MCSs) in the Li2012 Model [26], which has 578 reactions after compression. The time required by both approaches on the first 40 MCSs is shown in Figure 5.2. You can also see the growth of the MCSs size in 5.3. To see the first 10 MCSs, take look at table 5.6. You can see many of them shares some essential reaction and they have many reactions in common.

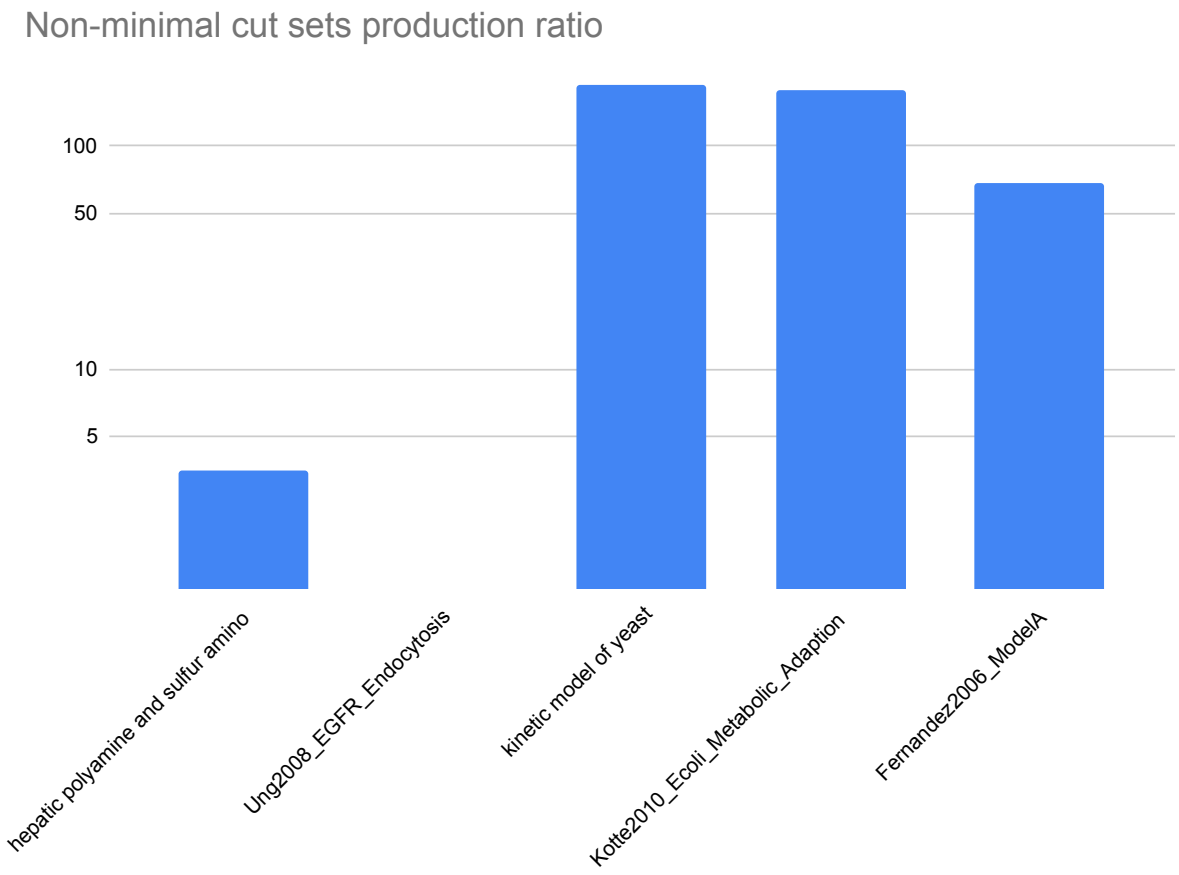


Figure 5.1: Ratio of number of non-minimal cut sets produced by the Target-Specific Dual method to number of non-minimal cut sets produced by the MCS^2 method

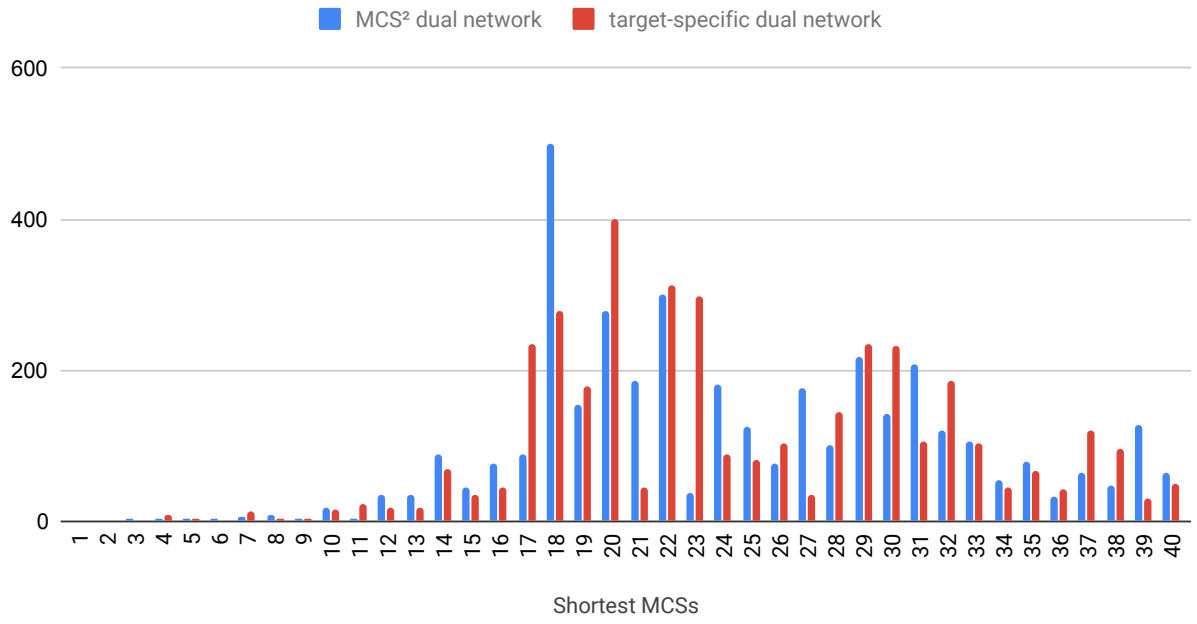


Figure 5.2: Time (in seconds) for computing each of the 40 smallest MCSs for reaction 10 (the first reaction which has at least 100 MCSs) of the Li2012 Calcium-mediated synaptic plasticity model [26]

Mixed-integer linear programming was used to find a subset of the MCSs [36]. The target-specific dual method has previously been used for this task, in a method called MCSEnumerator [43, 42]. As its authors state, not all the EFMs in the dual space result in valid MCSs, but by adding the appropriate constraints, one can remove the redundant results from the ILP's feasible space. To get a sense of how our approach, MCS^2-MILP, performs compared to MCSEnumerator, we implemented the MILP described in [43], currently part of CellNetAnalyzer. We ran our implementation of MCSEnumerator [43] and MCS^2-MILP on *E.coli iAF1260* for the sake of comparison, which showed a similar performance, as shown in Table 5.4. This table contains the result of running MILPs for each reaction as a target reaction once per iteration. In each iteration we restricted the MILPs to not spend more than one minute on finding MCSs. Table 5.7 shows the result of the same experiment repeated with models chosen from the BiGG database [19].

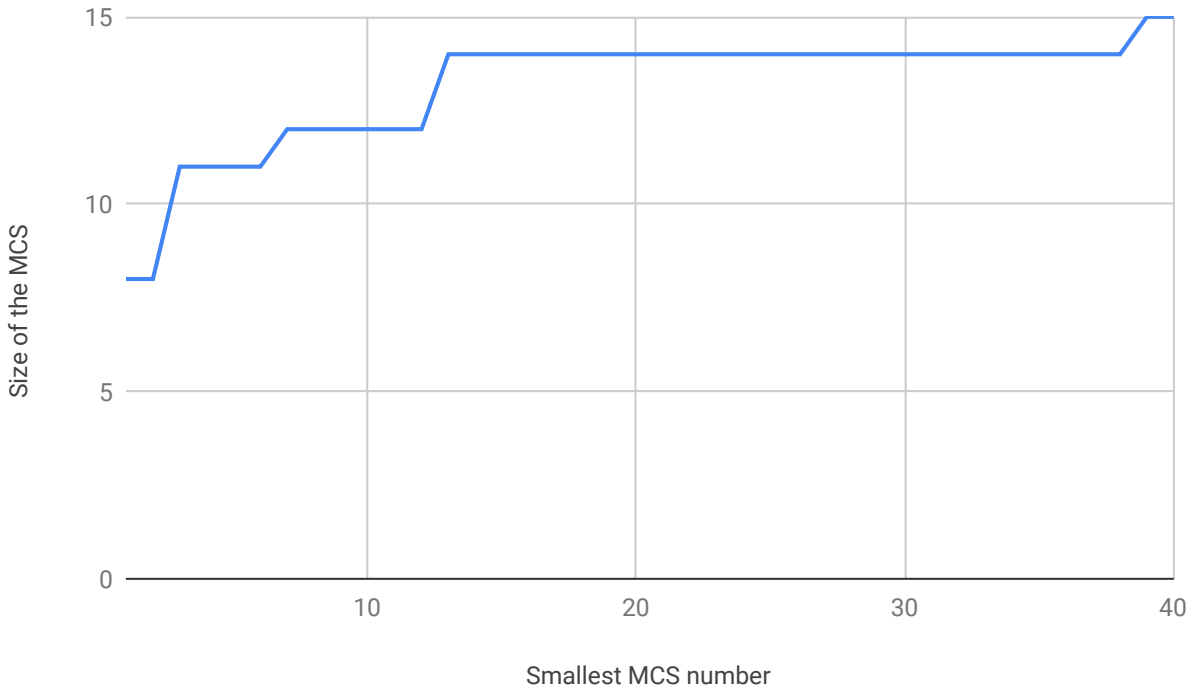


Figure 5.3: Size of the 40 smallest MCSs for reaction 10 (the first reaction which has at least 100 MCSs) of the Li2012 Calcium-mediated synaptic plasticity model [26]

Table 5.7: Result of running MCS^2-MILP and MCSEnumerator on the models from the BiGG database which initially have 2000 to 2600 reactions.

Model ID	Average time for shortest MCS for MCSEnumerator	Average time for shortest MCS for MCS^2-MILP	Reactions before (after) compression
iJO1366	4.66 seconds	3.98 seconds	2583 (1106)
iRC1080	7.12 seconds	7.19 seconds	2191 (1080)
STM_v1_0	1.82 seconds	1.83 seconds	2545 (1031)
iSbBS512_1146	14.10 seconds	19.03 seconds	2591 (1018)
iAF1260	5.22 seconds	4.45 seconds	2382 (981)
iSDY_1059	8.00 seconds	9.63 seconds	2539 (942)
iYL1228	1.88 seconds	2.11 seconds	2262 (805)

Chapter 6

Additional Results

The following claim, Cut Set Split Conjecture, is equivalent to the fact that for a given metabolic network we can ignore the reversibility of each reaction and then look for (Minimal) Cut Sets and then we can extract the a true (Minimal) Cut Set for the original Network from it.

Claim 1 (Splitting a cut-set to block opposite directions separately)**.** Let S be a stoichiometric matrix of a Metabolic network with reversible reaction Rev and irreversible reactions $Irrev$. Let $i \in Rev$. Then $C \subseteq Rev$ is a cut-set for i if and only if exist a split of each reversible reaction ζ into two positive and negative direction ζ^+, ζ^- such that C^+ (containing only positive direction of reactions in C) is a cut-set for reaction i^+ and C^- (containing only negative direction of the reactions in C) is a cut-set for reaction i^-.

Proof. Unfortunately this claim is not necessarily correct when $Irrev \neq \varnothing$. See the figure 6.1. $C = \{6, 7, 8\}$ is a cut-set for $t = 3$. On the other hand, reaction r_3^+ is blocked if all of the r_6^+, r_7^+, and r_8^+ blocked. If any of r_6^+, r_7^+, and r_8^+ allowed, reaction r_3^+ can proceed. However, if only reactions r_6^-, r_7^-, and r_8^- are blocked, reaction r_3^- can proceed. Thus, the claim is false. □

Proof. The claim is correct when the $Irrev$ set is empty. The proof is based on lemma 2 and 3 in PhD thesis [4]. Let v be the vector in row-space of S described by lemma 2 of thesis [4]. If we kill

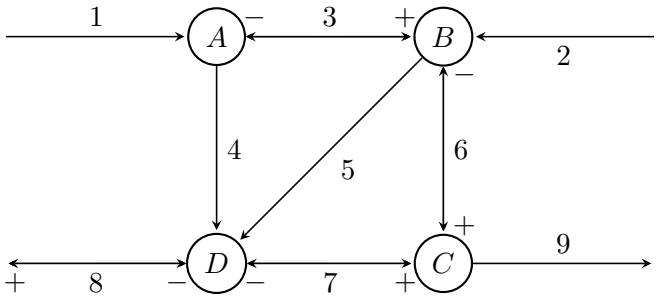


Figure 6.1: Reaction 6, 7, and 8 are a (minimal) cut-set for reaction 3. But the splitting claim won't hold here.

$\mathcal{R}_-(v)$ in forward direction and kill $\mathcal{R}_+(v) - \{i\}$ in backward direction, reaction i will be blocked in forward direction. Same is correct for forward direction of $\mathcal{R}_+(v) - \{i\}$, backward direction of $\mathcal{R}_-(v)$ and reaction i in backward direction. The rest of the proof is very similar to the proof of lemma 3 of thesis [4] □

We felt we should share this result, since it is possible others reach this claim while trying to convert a true theory for fully reversible/irreversible network to a general theory for networks with mixed reactions of reversible and irreversible ones. Especially because this is the smallest example we have found, we wanted to save the time for others.

Chapter 7

Conclusion and Future Work

One key advantage of our method is that it does not depend on the target reaction to construct the dual network. The computations for one target reaction can therefore be reused for a different target reaction. Furthermore, it tends to operate on a smaller network than the original.

One limitation to our method is that it is primarily designed for single target reactions (rather than a target containing a set of reactions), while both are just as easily handled by the competitor methods. Although MCS^2 does not find the MCSs for a set of reactions directly, it can easily find the MCSs for each reaction individually, then prune any supersets from the union of these MCSs.

An alternate strategy for computing MCSs is via mixed-integer linear programming (MILP), particularly when only a few short sets are required, rather than a complete enumeration [32, 33]. We showed that MCS^2 can be easily adapted to this task via the MCS^2-MILP method, which has shown performance not inferior to that of the state-of-the-art.

Another strategy is to alter the double description method to directly find rays with minimal coordinated support instead of minimal support, e.g., by ignoring some of dimensions of the reconfigured network. Here it is important to be careful about zero-cycle flux modes, which are flux modes that have fluxes in both direction of a split reversible reaction. These are not valid flux modes, but they do appear in the output of the double description method [10] and they may cause the omission of some rays which contain them in their support.

As we mentioned, there are many models for which our method outperforms all other existing methods, while for some models, the best performance is obtained by the Berge algorithm. The challenge is to find out what features of these models are different, and then to decide ahead of time what method to choose for a given model.

Our method is based on novel insights, and may be refined further. Possible additional sources of improvement include identifying and removing unwanted supersets during the execution of the double description method and optimizing the process of superset removal during post-processing. We believe that our method opens the door to further ideas exploring this different kind of duality between EFMs and MCSs, and deeper insights into the structure of metabolic network models.

Bibliography

[1] Kathrin Ballerstein, Axel von Kamp, Steffen Klamt, and Utz-Uwe Haus. Minimal cut sets in a metabolic network are elementary modes in a dual network. *Bioinformatics*, 28(3):381–387, 2011.

[2] Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45 of *North-Holland Mathematical Library*. Elsevier, 1984.

[3] Aarash Bordbar, Jonathan Monk, Zachary King, and Bernhard Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2):107, 2014.

[4] Leonid Chindelevitch. *Extracting Information from Biological Networks*. PhD thesis, MIT, 2010.

[5] Leonid Chindelevitch, Jason Trigg, Aviv Regev, and Bonnie Berger. An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models. *Nature Communications*, 5, 2014.

[6] Thomas Eiter, Kazuhisa Makino, and Georg Gottlob. Computational aspects of monotone dualization: A brief survey. *Discrete Applied Mathematics*, 156(11):2035–2049, 2008.

[7] Éric Fernandez, Renaud Schiappa, Jean-Antoine Girault, and Nicolas Le Novère. DARPP-32 is a robust integrator of dopamine and glutamate signals. *PLoS Computational Biology*, 2(12):e176, 2006.

[8] Michael Fredman and Leonid Khachiyan. On the complexity of dualization of monotone disjunctive normal forms. *Journal of Algorithms*, 21(3):618–628, 1996.

[9] Komei Fukuda and Alain Prodon. Double description method revisited. In *Franco-Japanese and Franco-Chinese Conference on Combinatorics and Computer Science*, pages 91–111. Springer, 1995.

[10] Julien Gagneur and Steffen Klamt. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5(1):175, 2004.

[11] Andrew Gainer-Dewar and Paola Vera-Licona. The minimal hitting set generation problem: algorithms and computation. *SIAM Journal on Discrete Mathematics*, 31(1):63–100, 2017.

[12] Andrew Gainer-Dewar and Paola Vera-Licona. The minimal hitting set generation problem: algorithms and computation. *SIAM Journal on Discrete Mathematics*, 31(1):63–100, 2017.

[13] Björn-Johannes Harder, Katja Bettenbrock, and Steffen Klamt. Model-based metabolic engineering enables high yield itaconic acid production by *Escherichia coli*. *Metabolic Engineering*, 38:29–37, 2016.

[14] Hassan Hartman, David Fell, Sergio Rossell, Peter Jensen, Martin Woodward, Lotte Thorndahl, Lotte Jelsbak, John Olsen, Anu Raghunathan, Simon Daefler, and Mark Poolman. Identification of potential drug targets in Salmonella enterica sv. Typhimurium using metabolic modelling and experimental validation. *Microbiology*, 160(6):1252–1266, 2014.

[15] Utz-Uwe Haus, Steffen Klamt, and Tamon Stephen. Computing knock-out strategies in metabolic networks. *Journal of Computational Biology*, 15(3):259–268, 2008.

[16] IBM. CPLEX optimizer.

[17] IBM. What are indicator variables? *ibm.com/indicator_constr*.

[18] Marcin Imielinski and Calin Belta. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Systems Biology*, 2(1):40, 2008.

[19] Zachary King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua Lerman, Ali Ebrahim, Bernhard Palsson, and Nathan Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44:D515–D522, 2015.

[20] Steffen Klamt. Generalized concept of minimal cut sets in biochemical networks. *Biosystems*, 83(2-3):233–247, 2006.

[21] Steffen Klamt and Ernst Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, 2004.

[22] Steffen Klamt, Julio Saez-Rodriguez, and Ernst Gilles. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, 1(1):2, 2007.

[23] Abdelhalim Larhlimi and Alexander Bockmayr. Minimal direction cuts in metabolic networks. In *AIP Conference Proceedings*, volume 940, pages 73–86. AIP, 2007.

[24] Nathan Lewis, Harish Nagarajan, and Bernhard Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291, 2012.

[25] Chen Li, Marco Donizelli, Nicolas Rodriguez, Harish Dharuri, Lukas Endler, Vijayalakshmi Chelliah, Lu Li, Enuo He, Arnaud Henry, Melanie Stefan, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4(1):92, 2010.

[26] Lu Li, Melanie Stefan, and Nicolas Le Novère. Calcium input frequency, duration and amplitude differentially modulate the relative activation of calcineurin and CaMKII. *PloS One*, 7(9):e43810, 2012.

[27] Radhakrishnan Mahadevan, Axel von Kamp, and Steffen Klamt. Genome-scale strain designs based on regulatory minimal cut sets. *Bioinformatics*, 31(17):2844–2851, 2015.

[28] Stefan Müller and Georg Regensburger. Elementary vectors and conformal sums in polyhedral geometry and their relevance for metabolic pathway analysis. *Frontiers in Genetics*, 7:90, 2016.

[29] Nathan Price, Jennifer Reed, and Bernhard Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886, 2004.

[30] Paul Pritchard. A simple sub-quadratic algorithm for computing the subset partial order. *Information processing letters*, 56(6):337–341, 1995.

[31] Armando Reyes-Palomares, Raúl Montañez, Francisca Sánchez-Jiménez, and Miguel Medina. A combined model of hepatic polyamine and sulfur amino acid metabolism to analyze S-adenosyl methionine availability. *Amino Acids*, 42(2-3):597–610, 2012.

[32] Alberto Rezola, Jon Pey, Luis de Figueiredo, Adam Podhorski, Stefan Schuster, Angel Rubio, and Francisco Planes. Selection of human tissue-specific elementary flux modes using gene expression data. *Bioinformatics*, 29(16):2009–2016, 2013.

[33] Alberto Rezola, Jon Pey, Luis Tobalina, Ángel Rubio, John Beasley, and Francisco Planes. Advances in network-based metabolic pathway analysis and gene expression data integration. *Briefings in Bioinformatics*, 16(2):265–279, 2014.

[34] Stefan Schuster, David Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(3):326, 2000.

[35] Nafiseh Sedaghat, Tamon Stephen, and Leonid Chindelevitch. Speeding up dualization in the Fredman-Khachiyan algorithm B. In *Leibniz International Proceedings in Informatics*, volume 103. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[36] Hyun-Seob Song, Noam Goldberg, Ashutosh Mahajan, and Doraiswami Ramkrishna. Sequential computation of elementary modes and minimal cut sets in genome-scale metabolic networks using alternate integer linear programming. *Bioinformatics*, 33(15):2345–2353, 2017.

[37] Natalie Stanford, Timo Lubitz, Kieran Smallbone, Edda Klipp, Pedro Mendes, and Wolfram Liebermeister. Systematic construction of kinetic models from genome-scale metabolic networks. *PloS One*, 8(11):e79195, 2013.

[38] Melanie I Stefan, Stuart J Edelstein, and Nicolas Le Novere. An allosteric model of calmodulin explains differential activation of pp2b and camkii. *Proceedings of the National Academy of Sciences*, 105(31):10768–10773, 2008.

[39] Marco Terzer and Jörg Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235, 2008.

[40] Cong T Trinh, Ross Carlson, Aaron Wlaschin, and Friedrich Srienc. Design, construction and performance of the most efficient biomass producing *E. coli* bacterium. *Metabolic Engineering*, 8(6):628–638, 2006.

[41] Jan Bert van Klinken and Ko Willems van Dijk. FluxModeCalculator: an efficient tool for large-scale flux mode computation. *Bioinformatics*, 32(8):1265–1266, 2015.

[42] Vítor Vieira, Paulo Maia, Isabel Rocha, and Miguel Rocha. Development of an integrated framework for minimal cut set enumeration in constraint-based models. In *10th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 193–201. Springer, 2016.

[43] Axel von Kamp and Steffen Klamt. Enumeration of smallest intervention strategies in genome-scale metabolic networks. *PLoS Computational Biology*, 10(1):e1003378, 2014.

[44] Axel von Kamp and Steffen Klamt. Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms. *Nature Communications*, 8:15956, 2017.

[45] Clemens Wagner. Nullspace approach to determine the elementary modes of chemical reaction systems. *The Journal of Physical Chemistry B*, 108(7):2425–2431, 2004.

[46] Hooman Zabeti, Tamon Stephen, Bonnie Berger, and Leonid Chindelevitch. A duality-based method for identifying elemental balance violations in metabolic network models. In *Leibniz International Proceedings in Informatics*, volume 113. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.