A fast method for extracting essential and synthetic lethality genes in GEM models

Francisco Guil [1] and José M. García [2]

[1]University of Murcia
[2]Affiliation not available

October 31, 2023

Abstract

The exploration and categorization of essential and synthetic lethality genes hold significant importance in seeking effective and targeted therapies for diverse ailments. This endeavor hinges upon genetic minimal cut sets (gMCSs), which also find utility in metabolic engineering. There have been various methods suggested for calculating gMCSs. Still, with the emergence of numerous new models and their growing intricacy, it has become vital to introduce new algorithms in this field. This paper presents a new algorithmic approach for computing gMCSs, which utilizes linear programming techniques to improve temporal efficiency. The key concept of the method is to use a k-representative subset to replace the target set with a smaller one.

A fast method for extracting essential and synthetic lethality genes in GEM models

Francisco Guil[#,1], José M. García[#,2]
Grupo de Arquitectura y Computación Paralela, Universidad de Murcia, Spain.
[1]fguil@um.es
[2]jmgarcia@um.es

***Abstract*—The exploration and categorization of essential and synthetic lethality genes hold significant importance in seeking effective and targeted therapies for diverse ailments. This endeavor hinges upon genetic minimal cut sets (gMCSs), which also find utility in metabolic engineering. There have been various methods suggested for calculating gMCSs. Still, with the emergence of numerous new models and their growing intricacy, it has become vital to introduce new algorithms in this field. This paper presents a new algorithmic approach for computing gMCSs, which utilizes linear programming techniques to improve temporal efficiency. The key concept of the method is to use a k-representative subset to replace the target set with a smaller one. Availability and implementation: Software and additional material is freely available at https://github.com/biogacop/fastMethod**

I. Introduction

The remarkable progress in DNA sequencing has paved the way for genome-scale models, constituting a significant breakthrough in genetics that opens up new avenues for research.

Constraint-based modeling has been developed as a generalized approach to generate and study these models. One of the key concepts in this approach is that of minimal cut sets (MCS) [Klamt, 2006, Hädicke and Klamt, 2011]. An MCS is a (minimal) set of reactions that, when inhibited simultaneously, prevent a specific task from being performed. It has been used to support the targeted design of microbial strains for bio-based production [Harder et al., 2016, Banerjee et al., 2020, von Kamp and Klamt, 2017, Alter and Ebert, 2019], to prevent the proliferation of certain bacteria (Guil et al. [2022]) or for the discovery of potential targets for cancer ([Tobalina et al., 2016, Apaolaza et al., 2017]). Implementing deletion strategies at a genetic level is often challenging due to conflicts when considering gene-protein-rules (GPRs) on a network. As a result, the concept of MCS has expanded to include a minimal genetic cut set (gMCS). A gMCS is a minimal set of genes that, when inhibited, prevent certain states or modes of the network+ [Machado et al., 2016, Apaolaza et al., 2017, Schneider et al., 2020]. This concept of gMCS is also referred to as essential or synthetic lethal genes in the biomedical field when there are one or more genes, respectively.

In recent years, several methods have been developed to compute gMCSs (refer to section II-B). However, most of these methods utilize MILP approaches for analysis, which can be expensive in terms of time and resources and may lack numerical stability. However, due to the growing intricacy and quantity of models accessible [Chen et al., 2022, Robinson et al., 2020], introducing fresh algorithms has become a crucial objective in this area. New methods must handle large networks, implicit targets and execute fast. This is particularly vital in medical contexts, where using multiple linked models to analyze differences between normal and pathological models is crucial, as mentioned in Foguet et al. [2022], Gustafsson et al. [2023].

Our paper introduces a new technique for calculating genetic minimal cut sets. We achieve this by limiting the search space and computing hitting sets for a specific subset of the target set. Our approach has proven highly effective for identifying gMCSs with fewer genes (typically 4 or 5). We have achieved efficiency rates that surpass those of previous methods by 10x to 30x.

The key concept is that of a k-representative subset of a target set **T** for a particular integer $k \geq 1$. A subset **T'** of **T** is said to be k-representative if it has precisely the same gMCSs of cardinality $\leq k$ than **T** so we can compute those gMCSs in **T'** instead of using the whole set **T**. We have created an algorithm that builds k-representative subsets iteratively. The process starts by computing a 1-representative subset and then successively extending it to k-representative subsets as k increases. The extensions are obtained by modifying the Berge algorithm Berge [1984] and filtering the resulting hitting sets using linear optimization problems (LP).

Substituting mixed-integer linear programming (MILP) problems with linear programming (LP) ones accelerates the method compared to previous proposed techniques.

As case studies, we use this new method to calculate synthetic lethalities of length ≤ 4 in two different networks: *iML1515*, a reconstruction model for *Escherichia coli* [Monk et al., 2017], and *Human1*, a unified human GEM lineage [Robinson et al., 2020]. We also calculate genetic interventions that ensure couple growth of biomass and ethanol in anaerobic conditions for the model *iJO1366*, another reconstruction model for *Escherichia coli* [Orth et al., 2011].

The *iML1515* and *iJO1366* models are available from BIGGs [Schellenberger et al., 2010], while the *Human1* model can be obtained from Metabolic Atlas [Li et al., 2023][1].

II. Material and Methods

A. Metabolic networks

A metabolic network is represented by a tuple comprising three elements: M, R, and S. M and R are sets that represent the metabolites and reactions, respectively. Meanwhile, S is a stoichiometry matrix that belongs to $M_{m\times n}(\mathbb{R})$. This matrix serves as a link between the metabolites and reactions in the network. The values of m and n represent the number of internal metabolites and reactions, respectively. Each state of the network is represented by a flux vector $v \in \mathbf{R}^n$, where v_i represents the activity level of reaction r_i.

The variation in concentrations of the metabolites is summarized in Equation (1)

$$\frac{dx}{dt} = S \cdot v \tag{1}$$

The equation representing the steady-state constraint is (2), where internal metabolite concentrations remain constant over time.

$$S \cdot v = 0 \tag{2}$$

Only internal metabolites must be included as rows in this formulation's stoichiometric matrix S.

Each reaction in R has upper (u_i) and lower (l_i) bounds on its reaction rates v_i. Therefore, equation 3 imposes a set of constraints on any flux vector.

$$l_i \leq v_i \leq u_i;\ \forall r_i \in R \tag{3}$$

A vector is called feasible or a network mode if it satisfies equations (2) and (3). The collection of all modes of the network is referred to as its feasible cone and is denoted by

$$C = \{v \in \mathbf{R}^n \mid S \cdot v = 0,\ l_i \leq v_i \leq u_i,\ \forall r_i \in R\}$$

Given a mode $v \in C$, its support is the set of reactions that appear with nonzero flux in v:

$$supp(v) = \{r_i \in R \mid v_i \neq 0\}$$

B. Genetic minimal cut sets

A cut set for a target set of modes T is a set of reactions that hit all the target modes. In other words, C intersects with every element e in T. A Minimal Cut Set (MCS) C is one where no proper subset of C is also a cut set for **T** [Hädicke and Klamt, 2011].

Metabolic networks often contain gene information in the form of gene protein rules (GPRs). GPRs for a reaction, denoted as r, are Boolean expressions indicating which gene combinations must be active to allow flux through r.

These GPRs extend the concept of minimal cut sets to that of genetic minimal cut sets, gMCS, [Machado et al., 2016, Apaolaza et al., 2017].

A genetic cut set (gCS) for a target set **T** is a set of genes G that, when knocked out, renders none of the elements $t \in \mathbf{T}$ a valid mode for the modified network. A gCS for **T** is minimal, a gMCS, if it does not contain any proper subset that is also a gCS for **T**.

C. Computing gMCSs

There have been multiple methods suggested for calculating MCSs and gMCSs. A straightforward method is to test all gene combinations of increasing length to detect and filter the minimal gCS by inclusion. This approach is limited to small cardinalities, typically no more than two genes, but can be enhanced by narrowing the search space. This field of study originated with the SL-Finder algorithm [Suthers et al., 2009] and has since been applied to other algorithms like fastSL [Pratapa et al., 2015] and rapidSL [Dehghan Manshadi et al., 2022].

A second approach relies on computing (a base of) the target set and computing MCSs as hitting sets of their supports, often using a variation of Berge's Algorithm [Berge, 1984, Jungreuthmayer et al., 2013a,b]. These methods often rely on a large base set, making it impossible to use them in large networks [Schneider et al., 2020].

New methods were developed to utilize the relationship between cut sets of the network and elementary flux modes of a specific dual network. This approach

[1] https://github.com/SysBioChalmers/Human-GEM

was initially suggested in Ballerstein et al. [2012]. This method can also be applied to gMCs by adding genes to the network, as shown in Machado et al. [2016] and Schneider et al. [2020]. Alternatively, a gene matrix can be introduced, as described in Apaolaza et al. [2017]. However, this later technique is currently restricted to analyzing synthetic lethalities. All of these methods use the K-shortest EFM algorithm (De Figueiredo et al. [2009]), which relies on solving mixed integer linear problems (MILP), resulting in high time and resource costs.

1) Computing gMCSs for reactions and modes: We begin our approach by examining the process of identifying gCSs and gMCSs for individual reactions.

Usually, the GPR for a given reaction $r \in R$ comes in two forms: a sum of products (Disjunctive Normal Form or DNF) or a product of sums (Conjunctive Normal Form or CNF).

- Notice that if the GPR is in DNF form, any gCS can be found by identifying the hitting set for the sets of genes corresponding to the summands.
- If the GPR for r is in DNF form, then its gCSs are the supersets of the sets of genes corresponding to each factor.

For our purposes, the second type of expression is more desirable. Moreover, converting GPRs from DNF form to CNF form is easy with the following algorithm.

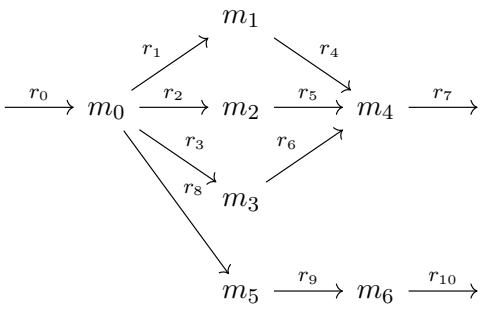
- Begin with a reaction r that has a GPR in DNF form.
- Express each term as a collection of genes.
- The GPR for r can be expressed as the product of sums of minimal hitting sets. These sets can be computed using Berge's Algorithm.

After being transformed into CNF form, the gMCSs for r are the factors of its GPR that do not contain any other factor as a subset.

It's important to note that if we want to focus on gCSs with length $\leq k$ for some natural number $1 \leq k \in \mathbb{N}$, we can limit ourselves to factors containing at most k terms.

Example 1:

1) Consider the metabolic model containing 5 metabolites, 8 reactions, and 9 genes given in the following figure



The GPRs for these reactions can be found in Table I.

Reaction	GPR	Reaction	GPR
r_0	g_0	r_5	g_4
r_1	g_1	r_6	g_4
r_2	g_1	r_7	$g_0 \wedge (g_5 \vee g_6 \vee g_7)$
r_3	g_2	r_8	g_8
r_4	g_3	r_9	$(g_1 \wedge g_2) \vee (g_3 \wedge g_4)$

Table I
THE GENE PROTEIN RULE FOR ALL REACTIONS IN THE MODEL IS PRESENTED IN CNF FORM.

The GPR for reaction r_7 is in conjunctive normal form (CNF). Its greatest minimal cut sets (gMCSs) are the factors $\{g_0\}$ and $\{g_5, g_6, g_7\}$.
However, the GPR for r_9 is expressed in disjunctive normal form (DNF). According to the algorithm described above, it can equivalently be stated as $(g_1 \vee g_3) \wedge (g_1 \vee g_4) \wedge (g_2 \vee g_3) \wedge (g_2 \vee g_4)$, so it has four gMCSs corresponding to its factors.

2) The model *Human1* version 1.16 contains 13085 reactions, of which 8091 have an associated GPR rule. Out of these, 8087 rules are in DNF form, while only 4 are in CNF form.
The GPR of reaction MAR07161 is in CNF form and includes twenty-eight factors with a length of 1, one factor with a length of 2, one with a length of 3, and one with a length of 4. As a result, there are thirty-one gMCSs for this reaction with lengths 1, 2, 3, and 4, each corresponding to the different factors.
The GPR for reaction MAR04611 is in DNF form and consists of two summands, each of which has four genes. Three genes are present in both summands. After being converted to CNF, it contains four factors that correspond to its gMCSs. These gMCSs consist of three gMCSs of length 3 and one gMCS of length 2.

We will use $gMCS(r)$ to denote the set of gMCSs associated with reaction r, and $gMCS^k(r)$ to stand for the set of gMCSs for r with length $\leq k$.

In order to expand the examination of gCSs to modes, it should be noted that a group of genes, denoted as G, is classified as a gCS for a given mode e only if there exists a reaction r within the support of e where G is also considered a gCS for r.

We use $gMCS(e)$ to denote the set of gMCSs for a mode e, and $gMCS^k(e)$ for the set of gMCSs for e with length $\leq k$.

We introduce the concept of a reduced set of genes to characterize gMCSs for a given mode.

Definition 2.1: A set of sets of genes, S, is considered reduced if each subset C in S has no proper subset C' in S.

Given a set of gene sets S, its reduction is defined as $S' = \{C \in S \mid \text{C is minimal}\} \subset S$.

The gMCSs for a specific mode e are characterized by Proposition 2.1.

Proposition 2.1: Let $e \in C$ be a mode of the network. Then

- $gMCS(e)$ is the reduction of
$$\bigcup_{r \in supp(e)} gMCS(r)$$
.
- For any $1 \leq k \in \mathbb{N}$, $gMCS^k(e)$ is the reduction of
$$\bigcup_{r \in supp(e)} gMCS^k(r)$$
.

2) Calculating gMCSs for sets of modes.: We can now compute gCSs and gMCSs for a target set of modes.

For gCSs, this extension is simple.

Proposition 2.2: The set of gCSs for a target set of modes **T** is the intersection of the gCSs of each mode in **T**.

$$gCS(T) = \bigcap_{e \in T} gCS(e)$$

However, this statement no longer applies to gMCSs. To clarify this question, let's rephrase our definition of a gCS for the target set **T**. A set of genes C is a gCS for **T** if it is a gCS for any element e in **T**. In other words, for each mode e in **T**, there exists at least one element G' in $gMCS(e)$ such that G is a superset of G'.

This description can be viewed as an extension of the idea of a hitting set.

Definition 2.2: We say that C is a hitting set for **T** if, for any $e \in$ **T**, there exists an element $G' \in gMCS(e)$ such that G' is a subset of G.

GCSs refer to the hitting sets for T. At the same time, gMCSs can be recognized as the minimal hitting sets for T. Identifying gMCSs with minimal hitting sets has the advantage of being easily adaptable to the Berge algorithm for their detection (refer to Algorithm 1).

Algorithm 1: Modified Berge algorithm

```
Data: A set of modes T
Result: The set of gMCSs for T
Initialization CS = {∅};
for e ∈T do
    for C ∈ CS do
        Check if C is a superset of some
         C' ∈ gMCS(e) ;
        if False then
            for C' ∈ gMCS(e) do
                CS ← C ∩ C';
            end
            Remove C from CS
        end
    end
end
```

Example 2:

Let's continue with Example 1. Consider the set of modes denoted by **T** as follows:

$$\{\{r_0, r_1, r_4, r_7\}, \{r_0, r_2, r_5, r_7\}, \{r_0, r_3, r_6, r_7\}\}$$

For all reactions except r_7, the genetic support consists only of its associated gene. For reaction r_7, the genetic support is

$$gMCS(r_7) = \{\{g_0\}, \{g_5, g_6, g_7\}\}$$

We have the following for any mode in "T":

- $gMCS(\{r_0, r_1, r_4, r_7\}) = \{\{g_0\}, \{g_1\}, \{g_3\}, \{g_5, g_6, g_7\}\}$
- $gMCS(\{r_0, r_2, r_5, r_7\}) = \{\{g_0\}, \{g_1\}, \{g_4\}, \{g_5, g_6, g_7\}\}$
- $gMCS(\{r_0, r_3, r_6, r_7\}) = \{\{g_0\}, \{g_1, g_2\}, \{g_3, g_4\}, \{g_5, g_6, g_7\}\}$

It is easy to use Algorithm 1 to check that **T** has exactly five gMCSs: $\{\{g_0\}, \{g_1, g_2\}, \{g_1, g_4\}, \{g_3, g_4\}, \{g_5, g_6, g_7\}\}$.

D. Using gMCSs to tackle various genetic strategies

The target set, **T**, usually contains undesired steady-state fluxes that require elimination. A list of supports does not explicitly define this set, but rather by inequalities set forth by a matrix $T \in \mathbf{R}^{s \times n}$ and a vector $t \in \mathbf{R}^s$. The set can be defined as **T**= $\{v \in C \mid T \cdot v \leq t\}$ [Schneider et al., 2020].

Let's start by discussing how to block the biomass reaction. Our focus is on all modes where there is a nonzero flow through the biomass reaction, denoted as $r_{biomass}$. This region is not defined by a restriction of the form $T \cdot v \geq 0$. we need to first calculate the highest flow

through $r_{biomass}$, which results in a value of $r^{max}_{biomass}$. We will block modes with biomass values exceeding $p \cdot r^{max}_{biomass}$, where p is a proportion that we will set at $p = 0.01$.

$$\mathbf{T}=\{v \in C \mid r_{biomass} \geq p \cdot r^{max}_{biomass}\}$$

Imagine that we aim to promote the coupled growth of biomass and a byproduct that is linked to exchange reaction r. Our goal is to remove any pathways that enable the flow of flux through $r_{biomass}$ without also carrying flux through r. We will determine the highest possible flux that can pass through $r^{max}_{biomass}$ and use a target set that accounts for various definitions of flux coupling [Schneider et al., 2020].

$$\mathbf{T}= \{v \in C \mid v_{biomass} \geq p \cdot r^{max}_{biomass},\, v_{ethanol} = 0\}$$

In this case, the variable p represents the same proportional constant as in the previous scenario.

It's important to note that any genetic intervention targeting this set falls into one of two categories:

- Interventions blocking the biomass reaction.
- Interventions that ensure ethanol production whenever $r_{biomass} > p \cdot r^{max}_{biomass}$.

We only need interventions in the second class, but it's easier to calculate all and filter by the maximal biomass flux in any genetic intervention.

E. Computing gMCSs for large target sets

Consider the target set given by $\mathbf{T}= \{v \in C \mid T \cdot v \leq t\}$. The primary challenge in this process is to calculate all the modes (or EFMs) present in **T**, which can be a time-consuming task for most networks (Schneider et al. [2020]). When dealing with computable target sets, they are often too large to use the methods previously discussed effectively.

1) k-representative subsets: To solve this issue, we recommend computing a smaller subset of modes called **T'** from the larger set **T**. By doing this, we can use the gMCSs of **T'** as an in-between step to identify the gMCSs for **T**.

To begin, we need to analyze the relationship between the gMCSs of a target set **T** and a subset **T'**$\subset$**T**.

Proposition 2.3: The following properties apply to any **T'** subset of **T**:

- Any gCS for **T** is also a gCS for **T'**
- Any gCS for **T** must contain a gMCS for **T'**
- If a gMCs for **T'** is a gCS for **T** then it is also a gMCS for **T**

According to Proposition 2.3, we can calculate the gMCSs for **T** by first computing the genetic minimal cut sets for **T'** and then eliminating those that are not minimal among the genetic cut sets for **T**. To guarantee that the resulting gMCs are minimal, we should compute them in increasing length. Thus, for every computed gCS, we can verify that it does not include any other gMCS with a smaller length.

However, there is a potential obstacle we need to avoid. Not all gMCSs for **T'** of a given length are also gMCSs for **T**. This means that if we are computing gCSs for T' to use as potential gMCSs for **T** of a given length k, we cannot eliminate those that contain any gMCS for **T'** of length $\leq k-1$. We need to identify the gMCSs for T and exclude any gCSs for T' that have them. If C' is a gMCS for **T'** whose length is less than k, then any superset C of C' will also be **T'**. This set may not be filtered out, so we will need to consider it as a potential candidate for being a gMCS for **T**. As a result, we will receive a large number of candidates to be gMCSs for **T**.

Example 3: Following with Example 1, consider **T'**=$\{\{r_0, r_1, r_4, r_7\}\} \subset$ **T**. Suppose we are interested in computing those gMCSs of length ≤ 2 for **T**.

We compute the gMCSs of length 1 for **T'** as $\{\{g_0\}, \{g_1\}, \{g_4\}\}$. Only $\{\{g_0\}\}$ is a gMCS for **T**.

Let's calculate the gCSs of length 2 for T', which are:

$$\{\{\{g_0, g_1\}, \{g_0, g_2\}, \{g_0, g_3\}, \{g_0, g_4\}, \{g_0, g_5\}, \{g_0, g_6\}$$

$$, \{g_0, g_7\}, \{g_1, g_2\}, \{g_1, g_3\}, \{g_1, g_4\}, \{g_1, g_5\}, \{g_1, g_6\}$$

$$, \{g_1, g_7\}, \{g_4, g_2\}, \{g_4, g_3\}, \{g_4, g_5\}, \{g_4, g_6\}, \{g_4, g_7\}\}$$

To simplify the gCS for **T**, we can eliminate any gCS that only contains the unique gMCS of length 1, which is g_0. However, we cannot eliminate gCS that contain either g_1 or g_2 as they are not gMCSs for **T**. For instance, $\{g_1, g_2\}$ is a legitimate gMCS for **T** even though it includes g_1.

It's important to note that there is a significant issue. If we denote the number of genes in the network as s, for each $C \in gMCS(T')\backslash gCS(T)$ of length k,' we can obtain up to $\binom{s-1}{k-k'}$ candidates may belong in $gMCS^k(T)$.

To solve this issue, we propose the idea of a k-representative subset **T'**$\subset$ **T**.

Definition 2.3: A set of modes **T'** is k-representative if any gMCS for **T'** of length less than or equal to k is also a gMCS for **T**.

Example 4: The subset T' defined in Example 3 was not 1-representative because there were gMCSs of length 1 for T' that are not gMCSs for T.

On the other hand, **T''** = $\{\{r_0, r_3, r_6, r_7\}\}$ is 1-representative, because it has only one gMCS of length 1, $\{g_0\}$, which is also a gMCS for **T**.

According to Definition 2.3, if a subset **T'** $\subset$ **T** is k-representative of the target region **T**, then a gMCS of length $k+1$ for **T'** is also a gMCS for **T** if it is a gCS for **T**.

*2) Computing gMCSs for **T** by using k-representative subsets:* We have all the necessary components to begin computing gMCSs for target set T.

We begin by finding a k-saturated subset T' of T for a particular integer k, where k is a natural number. To accomplish this, choose any nonzero mode, e.

Please note that if $gMCS(e) = \emptyset$, then according to 2.3, there are no gMCS for **T** as well. In all other scenarios, determine the smallest length of elements in $gMCS(e)$ as k and identify all gMCSs with length k. While $\{e\}$ may not be k-representative, it can be easily extended to a k-representative set **T'**. Start by creating $T' = \{e\}$. Knock out all genes in any $C_i \in gMCS^k(e)$ and choose a linear function f on the reactions. Then, pose the LP problem:

$$\begin{aligned} \text{Maximize} \quad & f \\ \text{subject to} \quad & S \cdot v = 0 \\ & v_i \geq 0 \qquad \forall r_i \in Irr \\ & T \cdot v \geq t \end{aligned} \tag{4}$$

If the problem cannot be solved, then C_i becomes a gMCS for **T**. If we do find a mode $e^i \in$ **T** where C_i is not a cut set, then we add e^i to **T'**.

We have computed all the gMCSs of length k for T and a k-saturated subset T'.

Proposition 2.4: The subset **T'** that has been obtained is a subset of **T** that is k-saturated.

Proof: It is worth noting that, according to Proposition 2.3, any set C^i belonging to $gMCS^k(T')$ can also be found in $gMCS^k(\{e\})$. If C^i was not a gMCS for T, we know that there is a mode e^i such that C^i is not a gCS for e^i. This contradicts the assumption that C^i belongs to $gMCS(T')$. □

The steps to construct k, all $gMCS^k(T)$, and the k-saturated subset $T' \subset T$ are outlined in Algorithm 2.

Suppose we have a subset **T'**$\subset$ **T** that is k-representative for $i < k \in \mathbb{N}$ and we have already computed all the gMCs for **T** of length $\leq k$. We can expand the set T' to T" such that T" is a (k+1)-representative subset of T while computing the gMCs for T of length k+1.

- Begin by setting T'' equal to T'.
- Compute all gCSs of length $k+1$ for T' using Algorithm 1, and then filter them by removing any gCSs containing an element of $gMCS^k(T) = gMCS^k(T')$. The remaining gCSs are candidates for $gMCS^{k+1}(T)$.
- Proceed as in Algorithm 2 to check if candidate C_i is in $gMCS^{k+1}(T)$.
- In any other case, find a mode $e^i \in T$ such that C_i is not a gCS for e^i. Add it to T''

The summarized process can be found in Algorithm 3.

Algorithm 2: Algorithm for computing all k gMCSs for **T** while also constructing a k-representative set

Data: A set of modes **T**
Result: A k-representative set of modes **T'** and $gMCS^k(T)$
Initialization: **T'**={e} for some $e \in$ **T**;
$k = min(length(C) \mid C \in gMCS(e))$;
for $C \in \bigcap_{e \in T'} gMCS^k(e)$ **do**
  Check if C is a cut set for **T**;
  if *False* **then**
    Find $e' \in$ **T** such that C is not a cut set for e';
    T'$\leftarrow$ **T'**$\cup\{e'\}$;
  else
    C is a gMCs for **T** of length k
  end
end

Algorithm 3: algorithm to extend a k-representative set of modes to a (k+1)-representative set

Data: A k-representative set of modes **T'** and $gMCS^k(T)$
Result: A (k+1)-representative set of modes **T"** and $gMCS^{k+1}(T)$
Initialization: **T"=T'**;
Compute $gMCS^{k+1}(T'')$
for $C^i \in gMCS^{k+1}(T'')$ **do**
  Check if C^i is a cut set for **T**;
  if *False* **then**
    Find $e^i \in$ **T** such that C^i is not a cut set for e^i;
    T"$\leftarrow$ **T"**$\cup\{e^i\}$;
  else
    If C^i does not contain any gMCS of length $\leq k$, then C^i is a gMCS for **T** of length $k+1$.
  end
end

Combining both methods yields an algorithm that finds all gMCSs for **T** of length $\leq k$ given integer $k \geq 1$. Additionally, the algorithm computes a k-representative subset **T'** of **T**.

Example 5: Let's utilize the model that was previously established in Example 2.

We obtain all gMCSs for T of length ≤ 3 using Algorithms 2 and 3. We start by taking $e = \{r_0, r_1, r_4, r_7\}$, $T' = \{e\}$.

Let's compute $gMCS(e)$ which is equal to $\{\{g_0\}, \{g_1\}, \{g_3\}, \{g_5, g_6, g_7\}\}$. Since $k = 1$, we can get $gMCS^1(e)$ which is equal to $\{\{g_0\}, \{g_1\}, \{g_3\}\}$.

We have three candidates for being gMCSs for **T** of length $k = 1$. After solving the corresponding LP problems, we obtain the following:

- $\{g_0\}$ is a gMCS for **T**.
- $\{g_1\}$ is not a gMCS for **T**. We obtain mode $e^1 = \{r_0, r_2, r_5, r_7\} \in \mathbf{T}$ with $\{g_1\}$ not being a gCS for e^1. We actualize **T'**=$\{e, e^1\}$.
- $\{g_3\}$ is not a gMCS for **T'**.

We obtained a unique gMC of length 1, $\{g_0\}$, and the 1-representative subset **T'**=$\{e^1, e^2\}$, where $e^1 = \{r_0, r_1, r_4, r_7\}$ and $e^2 = \{r_0, r_3, r_6, r_7\}$.

First, we need to calculate $gMCSs^2(T')$, and then we can expand **T'** to a 2-representative subset of **T**.

We get

$$supp^2_G(e) = \{\{g_0\}, \{g_1\}, \{g_3\}\}$$

$$supp^2_G(e^1) = \{\{g_0\}, \{g_2\}, \{g_4\}\}$$

Using Berge's algorithm, the mGCS for T' of length less than or equal to 2 are $\{\{g_0\}, \{g_1, g_2\}, \{g_1, g_4\}, \{g_3, g_2\}, \{g_3, g_4\}\}$. We check if any of them are a gCS for T.

- $\{\{g_1, g_2\}, \{g_3, g_4\}\}$ are gMCSs for **T**
- $\{g_2, g_3\}$ is not a gMCS. We can get a mode $e^2 = \{r_0, r_3, r_6, r_7\} \in T$ such that $\{g_3, g_2\}$ is not a gMCS for e^2. We add it to **T'**
- $\{g_1, g_4\}$ is not a gMCS for **T'**

We have completed the computation of $gMCS^2(T)$. Now, **T'** is a 2-representative subset of **T**.

Next, we calculate the gMCSs with a length of 3. Beginning with $T' = \{e, e^1, e^2\}$, we get:

$$supp^3_G(e) = \{\{g_0\}, \{g_1\}, \{g_3\}, \{g_5, g_6, g_7\}\}$$

$$supp^3_G(e^1) = \{\{g_0\}, \{g_2\}, \{g_4\}, \{g_5, g_6, g_7\}\}$$

$$supp^3_G(e^2) = \{\{g_0\}, \{g_1, g_2\}, \{g_3, g_4\}, \{g_5, g_6, g_7\}\}$$

By employing Berge's algorithm once more, the gMCS for **T'** with length ≤ 3 are

$$\{\{g_0\}, \{g_1, g_2\}, \{g_1, g_4\}, \{g_5, g_6, g_7\}\}$$

They are all gMCS for **T**. Therefore,

$$gMCS^3(T) = \{\{g_0\}, \{g_1, g_2\}, \{g_1, g_4\}, \{g_5, g_6, g_7\}\}$$

III. Results

A. Test bed

The evaluation platform was a double-socket Intel Xeon Gold 6226R with 384 GB RAM running on a CentOS 8.2 (4.18.0 kernel) provided by the Research Group of the High-Performance Computer Architecture (GACOP) of the University of Murcia (Spain). The network models were analyzed using the COBRApy package with Python 3.6 kernel in a Jupyter Notebook. We utilized Cplex version 12.10 to solve the LP problems that were associated ((https://www.ibm.com/academic/topic/data-science)).

B. Network models used as case studies

For our case study, we selected three network models to test our algorithm. We utilized two of them, namely *iML1515* and *Human1*, to compute their synthetic lethalities. On the other hand, we used the third one to determine genetic interventions that can guarantee growth coupling between biomass and ethanol.

The first reconstruction model used is *iML1515* for Escherichia coli, which has 2712 reactions, 1877 metabolites, and 1516 genes. This model can be found in Schellenberger et al. [2010] and was previously used in Schneider et al. [2020] to compare their algorithm's efficiency in computing gMCs to the one proposed in Apaolaza et al. [2017]. In their study, they set the maximum length of the computed gMCs to k=4. We also used this maximum length for all our cases.

We utilized the unified human GEM lineage, known as *Human1*, as our second model. Its latest release (version 1-16) can be obtained from https://github.com/SysBioChalmers/Human-GEM and consists of 13085 reactions, 8499 metabolites, and 2897 genes.

Finally, we have explored genetic modifications to promote the production of ethanol from biomass in oxygen-free environments using the *iJO1366* model. This model, like the one for Escherichia coli that can be downloaded from BIGG, encompasses 2583 reactions, 1805 metabolites, and 1367 genes.

C. Obtained metabolic interventions.

We begin our study using the *iML1515* reconstruction model for Escherichia coli. As demonstrated in Schneider et al. [2020], this model contains 889 gMCSs that are of a length less than or equal to 4.

We computed all gMCs and recorded runtimes while executing the Algorithm ten times to ensure its correctness. Table II displays the number of gMCSs for each length and the time taken to compute all gMCs up to that length in seconds.

k	Number of gMCs	Min time	Max time	Meantime
1	196	10,41	10,65	10,54
2	78	21,26	21,86	21,59
3	119	36,02	37,26	36,68
4	496	316,58	424,94	364,19

Table II
TIME IN SECONDS FOR OBTAINING SYNTHETIC LETHALITIES OF LENGTH ≤ 4 IN *iML1515*

We computed all gMCs with length ≤ 4 for version 1.16 of the *Human1* model as a second example.

This model contains 133 gMCSs with a length of 4 or less. Table III displays the amount of gMCSs at each length. It also indicates the time in seconds needed to compute all gMCs with lengths up to this value.

k	Number of gMCs	Min time	Max time	Meantime
1	92	9.82	12.70	10.64
2	14	31.39	40.24	33.09
3	15	55.01	60.21	56.37
4	12	484,72	758,76	617,30

Table III
TIME IN SECONDS FOR OBTAINING SYNTHETIC LETHALITIES OF LENGTH ≤ 4 IN *Human1*

For our third example, we focused on exploring genetic interventions that could enable the coupling of ethanol and biomass growth under anaerobic conditions in the *iJO1366* model. To begin, we restricted the model's ability to have flux through the exchange reaction linked to O_2 in this particular model.

We found 195 genetic interventions of length 4 or less. We ran the Algorithm ten times and recorded its run times in seconds. The corresponding results are shown in Table IV.

k	Candidates	Interventions	Min time	Max time	Meantime
1	0	0	3,09	3,25	3.16
2	81	3	11.33	11.79	11.56
3	223	47	28.16	29.77	28.83
4	509	145	156.78	173.75	165.15

Table IV
TIME IN SECONDS FOR OBTAINING GENETIC INTERVENTIONS OF LENGTH ≤ 4 IN *iJO1366* FOR THE COUPLE GROWTH OF ETHANOL

You can access the software, models, and all computed gMCSs by visiting https://github.com/biogacop/fastMethod.

D. Discusion

The gMCS framework is a powerful tool for exploring genomic scale models.

When evaluating methods for computing gMCSs, it is important to consider their ability to handle large networks, implicit target sets and obtain reasonably small execution times. It would also be desirable if a single technique could address all related problems, including identifying synthetic lethal genes and suggesting genetic interventions for the growth coupling of byproducts.

Our algorithm has met the first two criteria and has been proven to work in various scenarios through different case studies. In regards to execution times, Schneider et al. [2020] reported that their approach required 65 minutes to compute the gMCSs, while the technique outlined in Apaolaza et al. [2017] took 163 minutes. It's important to note that our execution times are significantly better, with improvements of 10x and 27x, respectively.

There is no conclusive evidence regarding the necessary time for the other approaches when applied to all our case studies. Nevertheless, it is worth noting that although the runtimes for *Human1* may be longer than those for *iML1515*, the disparity is not significant considering the size differences between the two models. It's important to mention that the elapsed time is determined by both the model's size and the number of gMCSs being computed. Specifically, it relies on the size of the model and the number of gMCSs that have the resulting sets **T'** introduced by the algorithm.

IV. CONCLUSIONS

A Genetic Minimal Cut Set (gMCS) is a set of genes that, when inactivated, prevents any mode of a specific target set. The target set varies depending on the desired genetic intervention. This concept offers a comprehensive method of examining essential and synthetic lethal genes, as well as metabolic engineering techniques to guarantee growth coupling for desired byproducts.

Numerous algorithms have been created to calculate gMCSs, with the majority relying on developing a dual network and utilizing MILP approaches for analysis. Regrettably, these techniques may lack numerical stability and can be costly in both time and resources.

In this paper, we present a novel algorithm for computing gMCSs. The new approach employs linear programming techniques, resulting in more efficient runtimes. The main idea is to use a k-representative subset of the target set. This enables us to replace the target set with a smaller one, which is necessary when the set is too big or cannot be calculated. Our method uses LP problems,

eliminating the need for MILP techniques and resulting in shorter execution times.

For our Algorithm tests, we utilized three models: *iML1515* and version 1.16 of *Human1* to assess the effectiveness of our algorithm in computing essential and synthetic lethality genes in medium and large-scale networks. Lastly, we used the *iJO1366* model to examine potential genetic interventions for the growth coupling of ethanol and biomass under anaerobic restriction. We have demonstrated that our method meets the requirements and outperforms alternative techniques in the first case study.

The technique discussed in our paper has the potential to improve the use of gMCS in various fields, including medicine. This technique can assist in targeting a single essential gene or a combination of genes, synthetic lethality, which can help cure certain illnesses. Furthermore, this technique can also be used in biotechnology to enhance the creation of new byproducts.

Ethics approval and consent to participate

This article does not contain any studies with human participants or animals performed by any authors.

Conflict of interest

All authors declare that they have no conflict of interest.

Funding

Grant TED2021-129221B-I00 funded by MCIN/AEI/10.13039 /501100011033 and by the "European Union NextGenerationEU/PRTR". Additionally, this work has been partially funded by Campus de Excelencia Internacional de Ámbito Regional (CEIR) Campus Mare Nostrum (CMN).

References

Tobias B Alter and Birgitta E Ebert. Determination of growth-coupling strategies and their underlying principles. *BMC bioinformatics*, 20(1):1–17, 2019.

Iñigo Apaolaza, Edurne San José-Eneriz, Luis Tobalina, Estíbaliz Miranda, Leire Garate, Xabier Agirre, Felipe Prósper, and Francisco J Planes. An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. *Nature communications*, 8(1):459, 2017.

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, 2012.

Deepanwita Banerjee, Thomas Eng, Andrew K Lau, Brenda Wang, Yusuke Sasaki, Robin A Herbert, Yan Chen, Yuzhong Liu, Jan-Philip Prahl, Vasanth R Singan, et al. Genome-scale metabolic rewiring to achieve predictable titers rates and yield of a non-native product at scale. *BioRxiv*, pages 2020–02, 2020.

Claude Berge. *Hypergraphs: combinatorics of finite sets*, volume 45. Elsevier, 1984.

Yu Chen, Feiran Li, and Jens Nielsen. Genome-scale modeling of yeast metabolism: retrospectives and perspectives. *FEMS Yeast Research*, 22(1):foac003, 2022.

L. F. De Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, and F. J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–3165, 2009.

Mehdi Dehghan Manshadi, Payam Setoodeh, and Habil Zare. Rapid-sl identifies synthetic lethal sets with an arbitrary cardinality. *Scientific reports*, 12(1):14022, 2022.

Carles Foguet, Yu Xu, Scott C Ritchie, Samuel A Lambert, Elodie Persyn, Artika P Nath, Emma E Davenport, David J Roberts, Dirk S Paul, Emanuele Di Angelantonio, et al. Genetically personalised organ-specific metabolic models in health and disease. *Nature Communications*, 13(1):7356, 2022.

Francisco Guil, Guillermo Sánchez-Cid, and José M García. Staphylococcus epidermidis rp62a's metabolic network: Validation and intervention strategies. *Metabolites*, 12(9):808, 2022.

Johan Gustafsson, Mihail Anton, Fariba Roshanzamir, Rebecka Jörnsten, Eduard J Kerkhoven, Jonathan L Robinson, and Jens Nielsen. Generation and analysis of context-specific genome-scale metabolic models derived from single-cell rna-seq data. *Proceedings of the National Academy of Sciences*, 120(6):e2217868120, 2023.

Oliver Hädicke and Steffen Klamt. Computing complex metabolic intervention strategies using constrained minimal cut sets. *Metabolic engineering*, 13(2):204–213, 2011.

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.

Christian Jungreuthmayer, Marie Beurton-Aimar, and Jürgen Zanghellini. Fast computation of minimal cut sets in metabolic networks with a berge algorithm that utilizes binary bit pattern trees. *IEEE/ACM transactions on computational biology and bioinformatics*, 10 (5):1–1, 2013a.

Christian Jungreuthmayer, Govind Nair, Steffen Klamt,

and Juergen Zanghellini. Comparison and improvement of algorithms for computing minimal cut sets. *BMC bioinformatics*, 14(1):1–12, 2013b.

S. Klamt. Generalized concept of minimal cut sets in biochemical networks. *Biosystems, 83(2–3), 233–247*, 2006.

Feiran Li, Yu Chen, Mihail Anton, and Jens Nielsen. Gotenzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Research*, 51(D1): D583–D586, 2023.

Daniel Machado, Markus J Herrgård, and Isabel Rocha. Stoichiometric representation of gene–protein–reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. *PLoS computational biology*, 12(10):e1005140, 2016.

Jonathan M Monk, Colton J Lloyd, Elizabeth Brunk, Nathan Mih, Anand Sastry, Zachary King, Rikiya Takeuchi, Wataru Nomura, Zhen Zhang, Hirotada Mori, et al. i ml1515, a knowledgebase that computes escherichia coli traits. *Nature biotechnology*, 35(10): 904–908, 2017.

Jeffrey D Orth, Tom M Conrad, Jessica Na, Joshua A Lerman, Hojung Nam, Adam M Feist, and Bernhard Ø Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism—2011. *Molecular systems biology*, 7(1):535, 2011.

Aditya Pratapa, Shankar Balachandran, and Karthik Raman. Fast-sl: an efficient algorithm to identify synthetic lethal sets in metabolic networks. *Bioinformatics*, 31(20):3299–3305, 2015.

Jonathan L Robinson, Pınar Kocabaş, Hao Wang, Pierre-Etienne Cholley, Daniel Cook, Avlant Nilsson, Mihail Anton, Raphael Ferreira, Iván Domenzain, Virinchi Billa, et al. An atlas of human metabolism. *Science signaling*, 13(624):eaaz1482, 2020.

J. Schellenberger, J. O. Park, T. M. Conrad, and B.Ø. Palsson. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC bioinformatics*, 11(1):213, 2010.

Philipp Schneider, Axel von Kamp, and Steffen Klamt. An extended and generalized framework for the calculation of metabolic intervention strategies based on minimal cut sets. *PLoS computational biology*, 16(7): e1008110, 2020.

Patrick F Suthers, Alireza Zomorrodi, and Costas D Maranas. Genome-scale gene/reaction essentiality and synthetic lethality analysis. *Molecular systems biology*, 5(1):301, 2009.

Luis Tobalina, Jon Pey, Alberto Rezola, and Francisco J Planes. Assessment of fba based gene essentiality analysis in cancer with a fast context-specific network reconstruction method. *PloS one*, 11(5):e0154583, 2016.

A von Kamp and S Klamt. Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms, 2017.