UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Propuestas y Análisis de Técnicas para la Generación de Subconjuntos Significativos de Soluciones de Redes Metabólicas

D. José Francisco Hidalgo Céspedes

2020

UNIVERSIDAD DE MURCIA

Facultad de Informática

PROPUESTAS Y ANÁLISIS DE TÉCNICAS PARA LA GENERACIÓN DE SUBCONJUNTOS SIGNIFICATIVOS DE SOLUCIONES DE REDES METABÓLICAS

Tesis Doctoral

Presentada por:
D. José Francisco Hidalgo Céspedes

Dirigida por:
Dr. D. Francisco de Asís Guil Asensio
Dr. D. José Manuel García Carrasco

Murcia, Diciembre de 2020

A Encarni y Valentina

Agradecimientos

Alguna vez he contado a quien me haya querido escuchar que yo me planteaba el trabajo de una tesis en Bioinformática como una donación intelectual. Como una forma de pagar tributo por tantas cosas que la vida me ha regalado. También como el resarcimiento de la espina que me dejó no haber optado por la carrera de Medicina en beneficio *in extremis* de la Ingeniería en Informática. Cerrando el documento y echando la vista atrás, no tengo claro si he saldado alguna de esas deudas.

Esta dedicatoria, si me otorga al menos la posibilidad de saldar otra deuda más grande. Si esta tesis tiene algún valor, quiero que el nombre de las personas que me han acompañado hasta aquí y que me inspiran a conquistar nuevas metas quede escrito junto al mío y para siempre.

A mi hija Valentina, con tu nacimiento encendiste la luz que me ha guiado a querer ser mejor persona y trabajar por un futuro mejor para todos nosotros. Mi deseo es que este trabajo haya contribuido a ello.

A mi esposa Encarni, por tu amor, generosidad con todos y por hacerme un regalo que nunca te podré pagar. Tu optimismo y ánimos constantes me han traído hasta aquí.

A mis padres Rodrigo y Maribel, por todos los esfuerzos que hicieron por criar y educar a sus cuatro hijos en salud y valores, y por darnos la oportunidad de ser lo que somos.

A mis hermanos Maria del Carmen, Francisco Javier y Juana Mari, que han sido y serán mis compañeros inseparables en el camino de la vida.

A mi abuelo Paco, al que le debo el cariño por la lectura y el conocimiento. Habría disfrutado mucho de este logro conmigo.

A José Manuel García Carrasco, por su enorme criterio investigador y científico y predisposición al trabajo. Por su empatía y paciencia. Por haberme prestado su fe en mí persona cuando yo perdí la mía propia durante el transcurso de este trabajo.

A Paco Guil, su precisión y profundidad analítica solo se ve superada por su humildad, paciencia y amor por la docencia. El salto cualitativo que ha aportado al resultado es incalculable.

Gracias.

Índice

Índice de figuras

Índice de Algoritmos

Resumen

El estudio de redes metabólicas aplicado a la Biotecnología y a la investigación de enfermedades es un tema de investigación de plena actualidad en nuestros días. La disponibilidad de modelos biológicos cada vez más precisos hace posible el uso de herramientas matemáticas e informáticas para dicho estudio que permiten analizar los posibles estados metabólicos de estas redes.

Es bien conocido que el conjunto de posibles estados de una red metabólica es un conjunto infinito que puede ser estudiado a través de un subconjunto finito de estados llamados modos elementales o EFMs de la red. Los modos elementales permiten representar todos los demás estados posibles en función de ellos. Por su importancia, se han propuesto una amplia variedad de métodos de extracción de EFMs basadas en estrategias y herramientas matemáticas diferentes. Cada método presenta sus propias ventajas e inconvenientes en términos de eficiencia, escalabilidad, etc. Atendiendo a la eficiencia, la herramienta matemática que ha demostrado el mejor comportamiento hasta la fecha es la optimización con programación lineal. Aún así, no es perfecta y los métdos basados en programación lineal presentan algunas limitaciones. Cabe destacar que, a pesar de todos los esfuerzos e independientemente del método usado, el problema de encontrar el conjunto de todos los EFMs de una red sigue abierto cuando se trata de redes grandes, por lo que es necesario centrarse en encontrar subconjuntos suyos, intentando asegurar la representatividad biológica del subconjunto obtenido.

Esta tesis contribuye al estudio de las redes metabólicas mediante la propuesta de diferentes estrategias, el análisis de los resultados y el estudio comparativo con otros algoritmos existentes. Ello culmina en la propuesta del algoritmo EFM-Ta de extracción de EFMs, que rompe la barrera de ratio ideal de eficiencia LP, batiendo en este aspecto a técnicas precedentes. Además del algoritmo EFM-Ta que mejora drásticamente la eficiencia de métodos anteriores, cabe destacar la inclusión de una técnica de estudio estadístico que supone un primer paso hacia el estudio de la tipología de los EFMs obtenidos mediante distintos métodos de extracción. Consideramos que este análisis es un paso adelante para un estudio que permitirá arrojar luz sobre la representatividad de distintos subconjuntos de EFMs.

Abstract

The study of metabolic networks applied to Biotechnology and to the investigation of diseases is a topic of current investigation in our days. In this sense, the availability of increasingly precise biological models makes possible the use of mathematical and computer tools for this study, which allows, in particular, to analyze the possible metabolic states of these networks.

It is well known that the set of possible states of a metabolic network is an infinite set, but that it can be studied through a finite subset of states (the so-called elementary modes or EFMs of the network) that allow us to represent all the other possible states depending on them. Because of this, different extraction methods have been proposed for EFMs based on different mathematical strategies and tools. Each proposed method has its own advantages and disadvantages in terms of efficiency, scalability, etc. If we look at the term efficiency, the tool that has shown the best performance to date is optimization with linear programming, although it also has its limitations. It should be noted that, despite all efforts and regardless of the method used, the problem of finding the set of all EFMs is still open when it comes to large networks, so it is necessary to focus on finding subsets of EFMs, trying to ensure the biological representativeness of the subset obtained.

This thesis contributes to the study of these topics based on the analysis of the different algorithms that can be proposed. It culminates in the proposal of the EFM-Ta algorithm for the extraction of EFMs, which breaks the barrier of the ideal ratio of efficiency to LP, beating in this respect previous techniques. In addition to this algorithm that dramatically improves the efficiency of previous methods, it is worth noting the inclusion of a statistical study technique on the typology of the EFMs obtained through different extraction methods. We consider that this analysis is a step forward for a study that will allow to shed light on the representativeness of different subsets of EFMs.

Capítulo 1

Introducción

1.1 Introducción

La Biología de Sistemas es hoy en día un campo de investigación activo que persigue una comprensión más profunda del metabolismo celular [77] mediante el uso de técnicas procedentes de la Biología, la Informática y las Matemáticas. El estudio detallado del metabolismo celular tiene fuertes implicaciones en el análisis y tratamiento de algunas enfermedades [7, 8, 32] o en Biotecnología [25, 44, 49, 81]. Un elemento clave para este área de conocimiento es la disponibilidad de modelos genómicos y de reconstrucciones metabólicas de organismos vivos [18,19,47,61,74], que permiten formas de experimentación *in silico* y extracción de conocimiento novedosas. La elaboración de modelos metabólicos de calidad es un proceso muy tasado [9,19,27,51,61,80] que parte de la información estequiométrica disponible sobre los procesos químicos que tienen lugar en una célula, y se circunscribe dentro de los límites de representación objetivo establecidos en laboratorio. El modelo metabólico a escala genómica se conoce como red metabólica (*Genome-Scale Metabolic Network*) o GSMN. Desde un punto de vista formal, un modelo GSMN es un modelo de red donde los metabolitos o nutrientes son transformados en otros metabolitos gracias a reacciones enzimáticas que ocurren en el interior de la célula. Una forma habitual de representar este modelo metabólico es por medio de una matriz cuyas filas representan los metabolitos y sus columnas a las reacciones, y que contiene los coeficientes estequiométricos usados en el modelo. Como toda red, presenta propiedades estructurales, entre las que se encuentran las que emergen en el análisis de los flujos metabólicos [71, 76].

Una vez construida la red como modelo biológico, una de las técnicas que puede ser usada para su estudio es la imposición de condiciones que restringen el espacio de soluciones para el problema a resolver. Esta técnica es conocida como modelado por restricciones (*Constraint-based modelling*) o *CBM* [10, 15, 53]. Estas restricciones pueden ser estequiométricas (basadas en los coeficientes estequiométricos calculados), termodinámicas (usando información sobre el tipo de reacciones químicas del modelo), o regulatorias (basadas en la regulación genética de la red) [2, 4, 45]. El tipo de restricciones que se imponen a una red dada en cada momento dependerá del problema

que se pretende abordar.

En este trabajo, el problema que analizaremos es el estudio de los posibles estados viables de la red usando técnicas *Flux balance analysis* (FBA) [3, 52, 76]. Usaremos restricciones estequiométricas y termodinámicas para estudiar las redes metabólicas como la extracción de partes relevantes de ellas o subredes (a las que llamamos modos o *pathways*) que cumplan restricciones equivalentes a las de la red original. Cada *pathway* se corresponde a un estado de la red metabólica en la que las reacciones y metabolitos no incluidos están inactivos. Excepto en casos triviales, el conjunto de *pathways* de una red metabólica es infinito, por lo que se han propuesto diferentes subconjuntos o familias (finitas) de estados a partir de los cuales se pueda calcular el conjunto completo de estados. De estos subconjuntos propuestos, nos centraremos en la familia de los *Elementary Flux Mode* o modos elementales [72]. Es bien conocido que el conjunto de EFMs de una GSMN es un conjunto finito de posibles estados o flujos que generan todo los demás posibles estados de la red calculados como combinación lineal convexa de ellos [16, 23, 38, 62, 66, 71–73].

Desafortunadamente, la cardinalidad de este conjunto es típicamente muy grande y solo ha podido ser calculado en pocos casos [33]. En los últimos años se han propuesto diferentes métodos para calcular el conjunto completo de EFMs de una red o, al menos, subconjuntos suficientemente grandes de ellos [16, 24, 36, 54, 55, 60, 63]. Estos métodos pueden ser divididos en dos grandes familias [57], de acuerdo a si se apoyan en propiedades asociadas a su forma de grafo (*path-finding methods*) [5, 35, 59, 75, 91], o emplean la información estequiométrica de la red. Dentro de esta segunda familia, podemos distinguir entre métodos basados en el algoritmo de doble descripcción (*Double Description Method* o DDM [22, 23]), programación lineal entera (método *Mixed Integer Linear Programming* o MILP [16, 62, 65]), o programación lineal (método de *Linear Programming* o LP [46, 78, 84]). Cada tipo de método tiene sus ventajas y desventajas. Una de las principales ventajas de los métodos DDM y los basados en MILP es que permiten obtener todos los EFMs de una red disponiendo de una condición de parada, es decir, detectan si todos los EFMs de la red han sido calculados. Por contra, los métodos basados en LP son más rápidos y eficientes en cómputo [11], pero una de sus debilidades es que no pueden asegurar haber encontrado todos los EFMs de la red. Nuestro trabajo se centra en este último tipo de método, intentando potenciar sus ventajas (eficiencia), así como analizar y paliar sus problemas (analizar el tipo de soluciones obtenidas y su representatividad).

El punto crítico de las técnicas basadas en programación lineal está en generar diferentes problemas de optimización en base a programas lineales usando conjuntos de restricciones adicionales y variaciones de la función objetivo, de modo que cada uno de estos problemas genere un EFM de la red. Dado que variar el conjunto de restricciones y la función usada no garantiza la obtención de soluciones diferentes, mediremos la eficiencia de estas técnicas como la proporción de soluciones diferentes obtenidas frente a ejecuciones realizadas. Para ser más precisos, nos interesa minimizar el número de ejecuciones del método `simplex` necesarias para obtener cada EFM no repetido. Dentro de los algoritmos previamente propuestos, la mejor eficiencia obtenida es atribuida al método *treeEFM* [55] con un ratio de 1,3 (esto es, un EFM nuevo encontrado por cada

1,3 resoluciones de un programa lineal). Este ratio está muy próximo al ratio ideal *1*, lo que podría hacer pensar que el margen de mejora para métodos basados en LP es ya escaso.

El objetivo principal de esta tesis ha sido aportar mejoras substanciales a los métodos de extracción de EFMs basados en LP, en tres ejes complementarios: cantidad de soluciones, eficiencia y representatividad. Las sucesivas contribuciones y publicaciones representan el esfuerzo invertido en este objetivo.

1.2 Antecedentes

1.2.1 Sistema de ecuaciones estequiométricas

Cada reacción metabólica que tiene lugar en una célula se puede representar con su correspondiente ecuación estequiométrica. Concretamente, el cambio en la cantidad en la que un metabolito interno m_i está presente en la red en una fracción de tiempo se representa en la Ecuación 1.1 como función de las reacciones que pueden producir o eliminar dicho metabolito.

$$\frac{dm_i}{dt} = \sum_{j \in R} s_{ij} r_j \tag{1.1}$$

donde R es el conjuto de reacciones metabólicas de la red.

La consideración sistémica de las redes metabólicas permite establecer que la concentración de metabolitos permanece estable en un intervalo de tiempo suficientemente pequeño. Esta condición conocida como *estacionaria* o *steady-state* permite que la ecuación anterior pueda simplificarse, suponiendo que $\frac{dm_i}{dt} = 0$, dando lugar a la ecuación lineal homogénea 1.2.

$$\sum_{j \in R} s_{ij} r_j = 0 \tag{1.2}$$

Las ecuaciones anteriores pueden disponerse en una matriz S, conocida como matriz estequiométrica de la red [52], lo que permite plantear todas las condiciones de equilibrio en forma de sistema homogéneo de ecuaciones lineales (Ecuación 1.3).

$$S \cdot \mathbf{v} = \mathbf{0} \tag{1.3}$$

Una reacción química se dice irreversible si solamente puede darse en un determinado sentido. La existencia de reacciones irreversibles da lugar a un segundo tipo de condiciones necesarias, conocidas como ecuaciones de factibilidad termodinámica (*thermodinamic feasibility*) que obliga a las reacciones metabólicas irreversibles a producirse en el sentido químico correcto. Es conocido que cualquier reacción reversible puede ponerse como suma de dos reacciones irreversibles (representando los dos posibles sentidos de la misma). Habiendo realizado este procedimiento (y sustituyendo si es necesario algunas reacciones por sus opuestas), podemos suponer que todas las

reacciones de la red son irreversibles y que se dan en sentido positivo. Esto da lugar a la Ecuación 1.4.

$$v_r \geq 0, \quad \forall r \in R \tag{1.4}$$

Identificaremos un estado de la red con un vector v de cardinalidad $|R|$, siendo R el conjunto ordenado de reacciones metabólicas, que representa las respectivas tasas de ocurrencia de cada reacción presente en la GSMN. Los estados posibles de la red se corresponderán con aquellos vectores v que cumplan con las restricciones impuestas por las ecuaciones 1.3 y 1.4. Estos vectores o *flux vector* son llamados *modos*. Observemos que siempre existe al menos una solución de este sistema de ecuaciones que es la solución trivial, es decir, aquella en que ninguna reacción está activa.

Disponer del conjunto de soluciones de nuestras ecuaciones nos permite estudiar qué partes específicas de la red pueden estar activas conjuntamente. Para precisar esta idea introducimos el concepto de soporte de un modo:

Si v es un modo, su soporte $supp(v)$ se define como el conjunto de reacciones r que aparecen en v con valor positivo.

Por tanto, dado el modo v, éste representa un posible estado de la red en la que las reacciones que aparecen en su soporte están todas simultaneamente activas mientras que aquellas que no están en el soporte están todas inactivas. Si pensamos que una red es un grafo, un modo puede pensarse como un subgrafo del mismo que puede funcionar como una red metabólica cuya matriz estequiométrica está formada por las entradas de S correspondientes a reacciones y metabolitos presentes en el subgrafo.

1.2.2 Modos de Flujo Elemental (*Elementary Flux Modes*)

Como ya hemos comentado, para una red no trivial el conjunto de modos es infinito. Esto hace que sea importante restringirnos a subconjuntos finitos suyos con la condición de que a partir de ellos sea posible reconstruir todos los modos de la red. Dentro de los subconjuntos propuestos [40, 41, 67, 86], en este trabajo vamos a centrarnos en los llamados modos elementales. Para un estudio detallado de otros conjuntos propuestos puede consultarse [40, 41, 67, 86].

Un modo v se dice un modo elemental o EFM si su soporte es mínimo, esto es, no hay otro modo no nulo v' tal que $supp(v') \subsetneq supp(v)$ [72].

Es sencillo demostrar que un modo con vector de flujo v es un EFM si y sólo si no puede ser representado como combinación lineal positiva de otros modos de la red [40].

Observemos que si v es un modo y λ un número real positivo, entonces $\lambda \cdot v$ es otro modo de la red. Diremos que estos dos modos son equivalentes y, en la práctica, identificarems v con $\lambda \cdot v$. Una vez hecha esta identificación, partiendo de la minimalidad e indescomponibilidad de los EFMs, se obtiene que un EFM viene caracterizado por su soporte (esto es, dos EFMs son distintos si y sólo si sus soportes son distintos). Este hecho permite afirmar que el conjunto de EFMs de una red es finito, ya que cada EFM

se corresponde a un subconjunto del conjunto de reacciones R y el conjunto de dichos subconjuntos es finito.

La relevancia biológica del conjunto de EFMs proviene de su finitud y de que forman un conjunto generador de todos los posibles vectores o modos posibles, en el sentido de que los posibles modos de la red son las combinaciones lineales con coeficientes no negativos de los EFMs de la misma. Esto es cierto incluso para aquellos no observados todavía en laboratorio [64, 71].

Esto ha hecho de los EFMs un tipo de subredes metabólicas cuyo análisis ha sido reconocido ampliamente [17, 64] y ha dado lugar a la publicación de diferentes técnicas diseñadas para enumerar los EFMs de una GSMN [16, 24, 54, 63]. Asimismo, se han desarrollado familias de algoritmos especializados dirigidos a asegurar algunas características en los modos obtenidos, como la presencia de conjuntos específicos de reacciones en los mismos.

Cualquier algoritmo dirigido al cálculo de EFMs debe contar con un medio para asegurar que la solución encontrada es realmente un EFM. En este sentido, observemos que las ecuaciones 1.3 y 1.4 permiten solamente comprobar si un vector v es un modo de la red. Para comprobar la condición de minimalidad que caracteriza un EFM puede usarse la siguiente caracterización en función de su soporte y de la matriz estequiométrica

Para cualquier modo v, sea S_v la submatriz de S construida tomando solo las columnas correspondientes a las reacciones contenidas en $supp(v)$. Si todas las reacciones de la red son irreversibles, el modo v es un EFM si y solo si $rank(S_v) = k - 1$ para $k = |supp(v)|$ [38, 79].

El principal obstáculo del uso de EFMs está en el alto coste computacional de enumerarlos todos debido a la alta cardinalidad del conjunto de EFMs para redes grandes [40]. Esto hace que dicho cálculo sea imposible excepto en redes pequeñas. Por contra, si solamente calculamos subconjuntos de este conjunto total de EFMs no es fácil saber hasta qué punto el subconjunto de EFMs obtenido es lo suficientemente representativo, y en consecuencia, saber cuánta relevancia biológica hemos extraído.

1.2.3 Programación lineal

Como se ha mencionado arriba, la extracción de EFMs se aprovecha de técnicas de optimización como la programación lineal (LP). Existen muchas librerías, algunas de ellas gratuitas, que implementan soluciones eficientes de programación lineal [21, 34]. Estas librerías, cuyo motor se conoce habitualmente como *solver*, incluyen en particular implementaciones del método `simplex` (que es en el que nos basamos en este trabajo).

La programación lineal parte de una serie de restricciones de las que cada una de ellas puede ser una igualdad o desigualdad lineal. Su propósito es encontrar el mínimo (o máximo) de una determinada función (también lineal) dentro del conjunto de vectores que satisfacen las restricciones.

En nuestro caso, las restricciones iniciales del problema de optimización a resolver

se plantean sobre la base de la matriz estequiométrica y las restricciones estacionaria (Ecuación 1.3) y termodinámicas (Ecuación 1.4). Estas restricciones definen un poliedro cónico de soluciones que se corresponden con los posibles modos de la red [15,48,82,83, 87]. No hay, en principio, una función objetivo estándar a optimizar ni una dirección clara (minimización o maximización) para el proceso de optimización. Una técnica usada a menudo es plantear el problema de minimización de la suma de las reacciones que pueden estar presentes en la solución, pero otras alternativas son igualmente posibles. Intuitivamente, la función objetivo alcanzaría su óptimo reduciendo el número de reacciones con valor positivo con lo que la solución obtenida sería un modo (al cumplir las restricciones) y tendría un número de reacciones minimal (por tanto un EFM). La Ecuación 1.5 muestra el programa lineal básico a partir de una red metabólica con matriz S.

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{n} v[i] \\ \text{subject to} \quad & S \cdot \mathbf{v} = \mathbf{0} \\ & v[i] \geq 0 \qquad \forall r_i \in Irr \end{aligned} \tag{1.5}$$

Esta aproximación intuitiva solamente proporciona soluciones triviales ya que siempre aparece como solución aquella en la que todas las variables son cero (solución trivial) y esta solución proporciona el único mínimo de la función usada. Si queremos obtener soluciones no nulas, partiendo de esta formulación estándar básica de problema LP, debemos introducir restricciones adicionales que no sean satisfechas por esta solución trivial. En función del problema a resolver, las restricciones adicionales pueden asegurar la presencia de algunas reacciones metabólicas ($J = \{r_{i_1}, \cdots, r_{i_k}\}$) con valor no nulo en los flujos obtenidos (Ecuación 1.6). Este tipo de restricciones adicionales son conocidas como *restricciones positivas* [1, 90]. En sentido opuesto, podemos imponer restricciones que asegurarían que ciertas reacciones son nulas en el flujo (Ecuación 1.7) o *restricciones negativas*. Observemos que las restricciones negativas sí son satisfechas por la solución trivial mientras que las positivas no lo son. Esto obliga a que nuestros problemas de programación lineal propuestos deban incluir siempre al menos una restricción adicional positiva.

$$\sum_{i_j \in J_1} v[i_j] = 1 \text{ (para restricciones positivas)} \tag{1.6}$$

$$\sum_{i_j \in J_2} v[i_j] = 0 \text{ (para restricciones negativas)} \tag{1.7}$$

Una condición necesaria pero no suficiente para obtener diferentes modos es formular programas lineales distintos. Por tanto, para extraer tantos modos como sean necesarios (N), debemos plantear diferentes programas lineales y resolverlos en sucesivas iteraciones en las que aplicamos el método `simplex`, tal como ilustra el Algoritmo 1.

Algorithm 1: Programa para extraer modos de una red metabólica con LP

Data : Matrix S, seed's length L, extractions N
Result: set Z of solutions

Function *runExperiment(S,L,N)*
  $nR \leftarrow S.columns()$;
  for $i \in (1..N)$ **do**
    $s \leftarrow GenerateSeed(nR, L)$;
    $lp \leftarrow poseLinearProgram(S, s)$;
    $sol \leftarrow simplex(lp)$;
    $Z \leftarrow Z + [sol]$;
  end
  return Z;

Cada iteración usa un conjunto diferente de restricciones adicionales que se conoce como *seed* o semilla por ser el germen de la respectiva solución obtenida. Una diferencia sustancial entre diferentes métodos de extracción de modos en esta familia de técnicas radica en la forma en que se produce la modificación del programa lineal en cada iteración, consistente en la estrategia de selección de semillas, así como en la introducción de variaciones en la función objetivo.

Haber planteado un problema de miniminización tampoco asegura que el modo obtenido sea siempre un EFM. Un resultado bien conocido [54] asegura que si todas las reacciones son irreversibles y se plantea un programa lineal viable añadiendo una única restricción positiva y cualquier número de restricciones negativas, la solución obtenida mediante el algoritmo `simplex` es siempre un EFM. En el mismo artículo se incluyen algunas técnicas basadas en programación entera que permiten ampliar este resultado.

1.3 Dificultades prácticas de la extracción de EFMs

Como ya se ha comentado, la principal dificultad en el cálculo de EFMs es que su número crece exponencialmente con el tamaño de la red. A esto hay que añadir, para cada técnica específica, las dificultades añadidas que provienen del instrumento matemático central o de la propia estrategia. En el caso que nos ocupa, los métodos basados en programación lineal presentan problemas comunes que restan eficiencia, siendo los dos principales la aparición de la misma solución repetidas veces y la posible inviabilidad del programa lineal planteado. A continuación repasaremos brevemente algunos aspectos relacionados con dichos problemas.

1.3.1 Reacciones bloqueadas

Cuando planteamos un programa lineal hemos de introducir restricciones adicionales y ello nos puede conducir a plantear problemas inviables (*infeasible*). Ello se debe a que nuestro problema contiene restricciones adicionales incompatibles entre sí. Cada

iteración sin solución es una pérdida en el ratio de eficiencia que debemos evitar, y por ello cualquier técnica que busca EFMs mediante LP debe dedicar una parte de su estrategia a evitar la inviabilidad.

La inviabilidad puede estar latente desde el inicio en el propio modelo [80, 92]. La construcción de un modelo estequiométrico de la célula de un organismo vivo conlleva implícitamente el establecimiento de los límites de lo que entra dentro del modelo y se considera interno, y lo que es externo al modelo [19, 27, 61, 80]. En pro del significado biológico y como instrumento académico, el modelo de una red metabólica puede incluir elementos externos. Estos elementos externos no pueden participar del equilibrio estacionario dentro de los límites del modelo. Cuando ese elemento externo es un metabolito estaremos hablando de metabolitos externos y debemos de tener cuidado de no incluirlos dentro de las condiciones de equilibrio que dan lugar a la matriz estequiométrica.

En el caso de reacciones, nos encontraremos con reacciones que no forman parte de ningún modo (en particular, no forman parte de ningún EFM) de la red. Este tipo de reacciones se llaman reacciones bloqueadas. Se puede detectar si una reacción es bloqueada usando programación lineal.

Sea $s \in R$ una reacción, s se dice que es una *reacción bloqueada* si al añadir una única restricción adicional positiva (Ecuación 1.8) al programa lineal básico (Ecuación 1.5) se produce un programa lineal inviable.

$$v_s = 1 \tag{1.8}$$

La Ecuación 1.9 muestra el programa que habría que ejecutar por cada reacción individual s para ver si está bloqueada.

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{n} v_i \\ \text{subject to} \quad & S \cdot \mathbf{v} = \mathbf{0} \\ & v_i \geq 0 \qquad \forall r_i \in R \\ & v_s = 1 \end{aligned} \tag{1.9}$$

Sea $LPfeasible$ una función que usando programación lineal según la Ecuación 1.9 devuelve si un programa es viable o no. Entonces podemos usar el Algoritmo 2 para calcular el conjunto B de reacciones bloqueadas.

Algorithm 2: Cálculo de reacciones bloqueadas

Data : Matrix S, set of reactions R
Result: Set B of infeasible reactions

$I \leftarrow []$;
for $s \in R$ **do**
  if $notLPfeasible(S, s)$ **then**
    $B \leftarrow B + [s]$;
  end
end

R' es el nuevo conjunto de reacciones no bloqueadas en el modelo. Se podría mejorar el desempeño del Algoritmo 2 omitiendo el test sobre cualquier reacción aparecida en la solución de una iteración anterior.

$$R' = R - B \tag{1.10}$$

Usando R' en lugar de R estamos simplificando la matriz S a la porción del sistema de ecuaciones que puede producir programas lineales viables. En términos de modelo, podemos considerar que hemos curado el modelo para su posterior uso.

Una vez se han eliminado las reacciones bloqueadas del modelo, se debe examinar la matriz estequiométrica para eliminar condiciones de equilibrio ligadas a metabolitos externos. Para ello bastará con eliminar aquellas filas de la matriz cuyas entradas no nulas sean todas del mismo signo.

En general, en ausencia de reacciones bloqueadas, la inviabilidad es producida por la imposición de restricciones adicionales contradictorias. Esta contradicción puede aparecer incluso si solamente se incluyen restricciones adicionales positivas como puede observarse en el siguiente ejemplo:

Consideremos la red metabólica de la Figura 1.1, con tres metabolitos $\{m_1, m_2, m_3\}$ (de los cuales el primero y el tercero son externos) y reacciones $\{r_1, r_2\}$ con matriz estequiométrica $S = (1 - 2)$

Figura 1.1: Red metabólica simple con 3 metabolitos

La condición de equilibrio para el metabolito m_2 obliga a que en cualquier modo de la red se tenga $v_1 = 2v_2$. Por tanto si imponemos simultáneamente las restricciones positivas $v_1 = 1$ y $v_2 = 1$ obtendremos un programa lineal:

$$\begin{array}{ll} \text{Minimize} & v_1 + v_2 \\ \text{subject to} & S \cdot \mathbf{v} = \mathbf{0} \\ & v_1 \geq 0 \quad v_2 \geq 0 \\ & v_1 = 1 \\ & v_2 = 1 \end{array} \tag{1.11}$$

que es inviable.

En los casos en que se use solamente una restricción positiva para asegurar que las soluciones obtenidas son EFMs, la aparición de problemas inviables se deberá a la imposición de restricciones positivas y negativas incompatibles entre sí. El caso más habitual está ligado al conocido concepto de *flux coupling* [13, 42, 69]. Recordemos que dadas dos reacciones r_i y r_j, se dice que r_i implica a r_j si en cualquier modo en que la reacción r_i esté activa, r_j también debe estarlo. En este caso las restricciones adicionales $r_i = 1$ y $r_j = 0$ son incompatibles entre sí. Como consecuencia, aquellas técnicas de calculo de EFMs que no tengan en cuenta los posibles *couplings* entre las reacciones tendrán una proporción de programas lineales inviables mayor.

Por todo lo comentado, es recomendable que cualquier método incluya búsqueda de reacciones bloqueadas, detección de metabolitos externos y análisis de las relaciones de *flux coupling* entre reacciones.

1.3.2 Reacciones reversibles

Las reacciones metabólicas que aparecen en una red pueden ser reversibles e irreversibles. Esta distinción es importante ya que la restricción termodinámica fuerza a que las reacciones irreversibles presentes en el soporte de la solución tengan valores positivos. Sin embargo, esta limitación no se puede imponer a las reacciones reversibles que pueden tener valores positivos o negativos atendiendo a que pueden ocurrir en los dos sentidos de la fórmula estequiométrica.

Si todas las reacciones son irreversibles, minimizar funciones lineales cuyos coeficientes sean positivos es siempre un problema acotado. Por contra, la presencia de reacciones estequiométricas reversibles convierte usualmente el problema de minimización en no acotado, por lo que no nos garantizamos encontrar soluciones al intentar resolverlo. Como se ha comentado, una técnica habitual es desdoblar las reacciones reversibles en pares de reacciones irreversibles representando los dos posibles sentidos de la misma [56, 70, 79]. Este proceso de desdoblar las reacciones reversibles hace que aumente el tamaño de la matriz resultante y la complejidad del sistema de ecuaciones a resolver, pero a cambio obtenemos problemas de minimización acotados. Un enfoque diferente al expuesto se puede encontrar en [66].

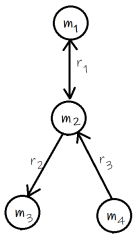
En nuestro caso, este desdoblamiento es un paso previo [14] en la técnica de extracción usada. Hay que tener en cuenta que, si se desdoblan las reacciones reversibles, el proceso de detección de reacciones bloqueadas debe ser posterior a este desdoblamiento puesto que alguna de las reacciones desdobladas resultantes puede estar bloqueada (falsas reversibles).

Sin pérdida de generalidad, podemos considerar que el conjunto de reacciones de la red no contiene reacciones bloqueadas, las reacciones reversibles están desdobladas, y la matriz estequiométrica S solamente incluye información sobre metabolitos internos.

Observemos que en un EFM no pueden aparecer juntas las dos reacciones resultantes del desdoblamiento de una misma reacción reversible. El funcionamiento interno del método `simplex` hace que esto no ocurra en el problema inicial, pero puede ser inducido por las restricciones adicionales que impongamos. Esto debe tenerse en cuenta al incluir restricciones adicionales para evitar la aparición de falsos EFMs que cumplen las restricciones estequiométricas y termodinámicas, pero cuya aparente minimalidad del soporte está ligada a las restricciones impuestas.

Este comportamiento puede ser observado con un ejemplo. Consideremos la red metabólica de la Figura 1.2. La reacción r_1 es reversible y todos los coeficientes estequiométricos valen 1.

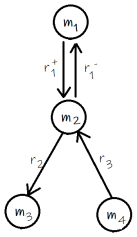
Figura 1.2: Red metabólica con una reacción reversible.



Esta red tiene tres EFMs con soportes $\{r_1, r_2\}$ (actuando r_1 en este caso en la dirección desde m_1 hacia m_2), $\{r_1, r_3\}$ (r_1 en la dirección opuesta) y $\{r_2, r_3\}$.

Si la reacción r_1 es reemplazada con dos reacciones irreversibles r_1^+ y r_1^-, se obtiene el grafo siguiente de la Figura 1.3.

Figura 1.3: Red metabólica con la reacción reversible desdoblada



Para el orden de las variables (r_1^+, r_1^-, r_2, r_3), (considerando solamente metabolitos internos) se deriva la matriz estequiométrica

$$S = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

Entonces, el problema LP que se obtiene al añadir la restricción $r_1^+ = 1$ conduce al vector solución $(1, 1, 0, 0)$ que no corresponde con un EFM.

Como hemos comentado, el problema radica en que no deben aparecer en un mismo vector las dos reacciones que proceden del desdoblamiento de una reacción reversible. Esto no se puede evitar usando una única restricción positiva adicional al LP básico, por lo que debería de incluirse también la restricción negativa

$$r_1^- = 0$$

.

Nuestro siguiente resultado muestra otra forma de evitar este comportamiento no deseado.

Teorema 1. *Sea M una red metabólica. Sin pérdida de generalidad, supongamos que las reacciones han sido reordenadas de modo que $\{r_1, \cdots, r_k\}$ son irreversibles y $r_{k+1}, \cdots, r_n$ son reversibles. Para cada reacción reversible r_i, se incorporan un par de reacciones irreversibles $\{r_i^+, r_i^-\}$ que representan los dos posibles sentidos del flujo de la reacción r_i. Escogemos una lista de números no negativos $\{a_1, \cdots, a_k\}$ tal que al menos uno de ellos no es cero. Considérese el subconjunto $V \subset \mathbb{R}^n$ definido por las siguientes restricciones*

$$\begin{aligned} S \cdot v &= 0 \\ v &\geq 0 \\ \sum_{i=1}^{k} a_i \cdot v_i &= 1 \end{aligned} \tag{1.12}$$

Si $V \neq \emptyset$ entonces cualquier punto extremo de V corresponde a un EFM de la red.

Obsérvese que en este resultado la restricción positiva solo contiene reacciones irreversibles. Este Teorema mejora el conocido resultado en [54] y evita situaciones como la que aparecía en el ejemplo anterior. Una prueba de este resultado ha sido publicada como material complementario de [88].

Los posibles EFMs falsos producidos por el desdoblamiento de reacciones reversibles tienen siempre longitud 2, pero esto no quiere decir que todos los EFMs de longitud 2 vengan de estos desdoblamientos. Por ejemplo, el modelo de *E.coli core* [51] disponible en BiGG [50] incluye el 2-ciclo formado por FRD7 y SUCDi. Este 2-ciclo es un EFM aunque no sea fisiológicamente relevante (E.coli Core Model for Beginners -PART 1- [26, 31, 85]). Estos 2-ciclos tienden a aparecer repetidamente en los procesos de extracción de EFMs [58], por lo que es conveniente identificarlos previamente y, una vez calculados, tratar las reacciones implicadas como si provinieran del desdoblamiento de una reacción reversible.

1.3.3 Soluciones repetidas y el ratio de eficiencia de un método de extracción de EFMs

La eficiencia en tiempo de un método de extracción de EFMs puede depender en gran medida de la configuración de hardware usada y del software empleado (tanto del solver como del lenguaje de programación empleado). Por ello, para los métodos basados en programación lineal es mejor estudiar el ratio de eficiencia en función del número de ejecuciones del solver `simplex` necesarias para obtener cada solución diferente [55]. En este sentido, obtener una solución diferente en cada ejecución del método `simplex` puede ser considerado el ideal (es decir, el ratio ideal es 1) y la eficiencia del método será mejor cuanto menor sea este ratio.

Normalmente la incidencia de soluciones repetidas tiende a ser alta, sobre todo cuando ya se ha obtenido un número sustancial de EFMs en el experimento. Hay varios factores que pueden favorecer el aumento de estas repeticiones. A veces el motivo por el que no conseguimos evitar las repeticiones se debe principalmente a que, a pesar de que el programa lineal propuesto sea diferente, los gradientes dados por las funciones conducen a un mismo óptimo. También existe una fuerte componente estructural de la red como la que se manifiesta en los *flux couplings* [13,69] que convierten en equivalentes a conjuntos de restricciones adicionales aparentemente diferentes entre sí.

Estos factores, y seguramente otros que aún no son conocidos, hacen que los ratios de eficiencia tiendan a ser peores de lo deseable y su mejora es un problema interesante aunque difícil de abordar. A lo largo de esta tesis mostraremos diferentes técnicas que conducen a algoritmos con ratios que representan una clara mejora respecto a otros métodos propuestos anteriormente.

1.4 Modificación de programas lineales

Teniendo en cuenta todo lo anterior, los métodos de extracción de EFMs basados en LP parten del programa lineal inicial e intentan modificarlo para obtener en cada pasada un EFM diferente. Estas variaciones pueden hacerse modificando el sistema de ecuaciones añadiendo nuevas ecuaciones que representan restricciones adicionales, o influyendo en el problema mediante la modificación de la función objetivo.

1.4.1 Restricciones adicionales

Sin pérdida de generalidad, el programa lineal básico se conforma sin reacciones bloqueadas. Restringir adicionalmente un programa lineal consiste en añadir nuevas ecuaciones. Una parte nuclear de las técnicas de extracción de modos es la de proponer restricciones sobre el programa lineal básico conducentes a producir soluciones no repetidas. Hay que tener en cuenta que la cantidad de ecuaciones adicionales tiene un gran impacto en el tipo de soluciones y en la eficiencia obtenida.

Como hemos comentado, el uso de una única restricción adicional positiva permite plantear problemas diferentes en cada iteración que conducen a la obtención de

EFMs. Genéricamente las restricciones adicionales en los problemas de optimización planteados tienen la forma expresada en las Ecuaciones 1.6 y 1.7. Una ventaja adicional de usar única restricción positiva en una red que no contiene reacciones bloqueadas, es que con ella siempre se construye un programa viable.

Por contra, la imposición de restricciones negativas junto con esta restricción positiva puede dar lugar a construir programas inviables. Si, a pesar de esto, deseamos usar restricciones negativas, podemos suponer que es necesaria una única restricción de este tipo. Esto se debe a que poner dos restricciones negativas del tipo

$$\sum_{i \in I_1} v_i = 0$$

$$\sum_{i \in I_2} v_i = 0$$

es equivalente a imponer la única restricción:

$$\sum_{i \in I_1 \cup I_2} v_i = 0$$

Dado que nunca impondremos dos restricciones positivas para estar seguros de obtener un EFM, y que siempre debemos incluir al menos una, nos encontramos con dos posibles situaciones:

- Usar una restricción positiva y ninguna negativa. La ventaja es que obtendremos siempre problemas LP viables. En cambio, debemos tener cuidado con las reacciones que incluimos en esta restricción (recordemos que si usamos reacciones reversibles podemos encontrar falsos EFMs de longitud 2).

- Combinar simultáneamente una restricción positiva y una negativa. Esta modalidad nos plantea el problema de la posible compatibilidad entre ellas. Este problema no es de fácil solución, ya que conocer la compatibilidad entre las dos restricciones es equivalente a conocer los *cut sets* de la red [37, 39]. De hecho, si tenemos dos restricciones

  $$\sum_{i \in I_1} v_i = 1$$

  $$\sum_{i \in I_2} v_i = 0$$

  Estas dos restricciones son compatibles si y sólo si I_2 no es un *cut set* para I_1. Conocer todos los *cut sets* de la red es un problema de complejidad equivalente a conocer sus EFMs [6]. Sin embargo, esta vía permite incluir reacciones procedentes de desdoblar reversibles en la restricción positiva (siempre que incluyamos su inversa en la restricción negativa).

Finalmente indicar que un uso importante de las restricciones positivas es condicionar las soluciones a que contengan una reacción o metabolito objetivo (*target*). Por tanto, si queremos considerar reacciones *target* que sean reversibles, debemos usar una combinación entre una restricción positiva y otra negativa.

1.4.2 Importancia de las funciones objetivo

Otra técnica para obtener diferentes problemas de optimización es alterar el programa lineal cambiando la función objetivo. Podemos pensar que esta función era originalmente el sumatorio de todas las variables con coeficiente unidad. Por tanto, intuitivamente esa función alcanza su mínimo cuando el número de sumandos no nulos es mínimo y se espera que eso se traduzca en que las soluciones sean EFM. Cuando se opta por tener una única restricción positiva, la solución obtenida sabemos que es un EFM, por lo que podemos modificar la función objetivo propuesta para obtener diferentes problemas de optimización.

Hay diversas formas de modificar la función objetivo a minimizar:

- Sustituirla por una función del tipo

$$f(v) = \sum_{i \in R} \lambda_i v_i$$

  para cualquier conjunto de números positivos λ_i.

  Este tipo de funciones tienen propiedades semejantes a la original, pero su gradiente no es el mismo. Al usar diferentes funciones esperamos que los mínimos correspondientes sean diferentes y obtengamos nuevos EFMs.

- No usar todas las reacciones en la función, sino solamente un subconjunto. Puede interpretarse como una restricción negativa *soft*, en el sentido de intentar obtener EFMs que tengan inactivas todas o alguna de las reacciones que aparecen en dicha función. La idea tras esta modificación es que el mínimo de la función se obtendría cuando quitamos todas esas reacciones pero, al no saber si esto posible, se plantea el problema para tender a quitar el mayor número posible de ellas.

- Usar una función general del tipo

$$f(v) = \sum_{j \in J} \lambda_j v_j$$

  para cualquier conjunto de números $\lambda_j \neq 0, J \subset R$.

  Este tipo de funciones debe usarse con cuidado ya que, dependiendo de como se hayan incluido los límites de las variables en el solver, puede dar lugar a problemas no acotados o a falsos EFMs. Para evitar este comportamiento se deben incluir en la restricción positiva todas aquellas variables que aparezcan en la función con coeficientes positivos.

Un caso extremo del tercer tipo de modificación expuesta anteriormente se da cuando todos los coeficientes son negativos. Este caso es equivalente a plantear un problema de maximización.

Además de buscar EFMs diferentes en cada iteración, los motivos para condicionar la función objetivo pueden ser buscar penalizar o promocionar la aparición de un subconjunto de reacciones en los EFMs obtenidos [89].

1.5 Representatividad de un conjunto de EFMs

Ya hemos comentado que, para la mayor parte de las GSMN disponibles, ninguno de los métodos desarrollados permiten calcular el conjunto completo de EFMs de la red. Una ventaja de los métodos basados en optimización (tanto LP como MILP) es la probabilidad de construir subconjuntos grandes de EFMs a partir de los cuales intentar inferir propiedades de la red. Sin embargo, para que estas inferencias tengan validez el conjunto obtenido debe ser suficientemente representativo respecto a la propiedad que se quiere estudiar.

La producción de *biomasa* es una reacción artificial que representa el crecimiento de la célula. Los valores estequiométricos de esta reacción destacan por no ser valores enteros y es, en parte, porque representan con precisión la proporción de los metabolitos presentes en la célula [12,20]. No es deseable una representatión insuficiente de la misma cuando analizamos aspectos vitales del comportamiento metabólico. Supongamos que estamos estudiando la producción de biomasa en el modelo *E.coli core* [51] disponible en BiGG [50]. Esta red tiene 100.274 EFMs [67]. El histograma de la Figura 1.4 muestra la distribución de sus longitudes (por claridad, se ha excluido el único 2-ciclo de esta red formado por SUCDi y FRD7).

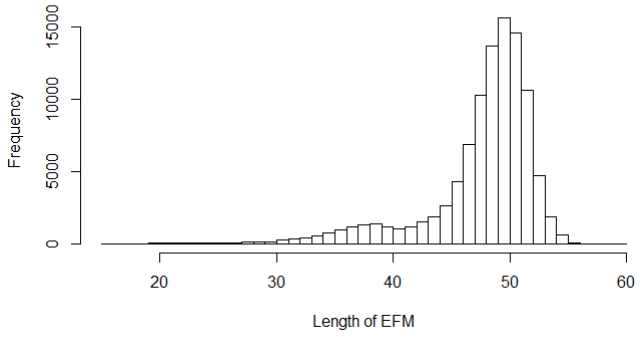


Figura 1.4: Histograma de longitudes de los EFMs en E.coli core

Este histograma es bimodal debido a que en esta red los EFMs en los que no participa la reacción de biomasa tienen una longitud menor. Esto puede observarse en la Figura 1.5 en la que se muestran superpuestos al histograma anterior los correspondientes a las longitudes de EFMs con y sin reacción de biomasa activa.

Si usamos un método de extracción de EFMs que prime encontrar EFMs de longitud pequeña, podríamos inferir equivocadamente que la mayor parte de los EFMs no contienen esta reacción. Este es efectivamente el caso, el gráfico 1.6 muestra la proporción de EFMs que contienen biomasa a lo largo de la extracción de todos los EFMs de la red usando un algoritmo de extracción basado en el propuesto en [88].

Si hubiéramos obtenido los primeros 20,000 EFMs (esto es, aproximadamente un 20 % de todos los EFMs, lo que es mucho más de lo que puede obtenerse en redes

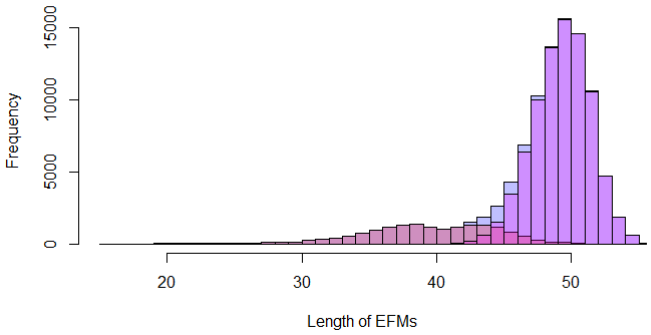


Figura 1.5: Histogramas de longitudes de los EFMs en E.coli core con superposición de histogramas con la reacción de biomasa y sin ella.

grandes) hubiéramos llegado a la conclusión de que la biomasa se encuentra activa en menos de la mitad de los casos (un 43.66 % frente al 83.37 % real). Este error hubiera sido mayor si nuestro conjunto de EFMs obtenidos hubiese sido más pequeño.

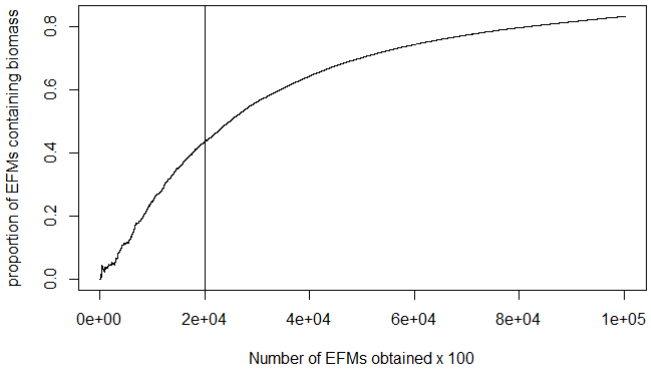


Figura 1.6: Progresión de extracción EFMs y sus longitudes en el algoritmo EFM-Ta.

Dado que en la mayor parte de los casos no se puede calcular la lista completa de EFMs, estos comportamientos anómalos son difíciles de detectar. Pero, al menos, debería prestarse atención a dos características importantes:

- Comparar la tipología de los EFMs obtenidos con diferentes medios de extracción.
- Para un mismo método, comprobar la estabilidad del tipo de EFMs obtenidos durante la realización de experimentos conducentes a la obtención de subconjuntos de EFMs de cardinalidad alta.

Cualquiera de los dos aspectos anteriores puede llevarse a cabo mediante técnicas estadísticas de comparación de muestras. Las posibles características a analizar dependerán del problema a tratar y pueden incluir, como en el ejemplo anterior, las longitudes de los EFMs obtenidos o las proporciones de algunas o todas las reacciones implicadas. Dentro de las muchas herramientas disponibles se puede usar tests de ajuste (como el test χ^2) o, dado que seguramente las características usadas no serán independientes, tests no paramétricos como el test de Wilcoxon.

1.6 Objetivos de esta tesis

En los últimos años se han propuesto diferentes métodos tendentes a la enumeración del conjunto de EFMs de una GSMN. Sin embargo, el crecimiento exponencial del número de estos EFMs conforme crece el tamaño de la red hace que haya una continua necesidad de mejorar dichos métodos haciéndolos cada vez más eficientes. En la mayoría de los casos es imposible obtener el conjunto completo de los EFMs de una red, por lo que los métodos disponibles solamente permiten obtener subconjuntos del total. Un aspecto importante asociado a este problema es la necesidad de estudiar la representatividad del subconjunto de EFMs obtenido respecto a alguna propiedad que se desee analizar.

Por todo ello, los objetivos principales de esta tesis son:

- Analizar las posibles formas de modificar el problema LP inicial asociado a la red metabólica para así clarificar las posibles estrategias que conduzcan al cálculo de EFMs basados en LP.

- Hacer un análisis previo de los tipos de reacciones que pueden aparecer en una red como paso previo al análisis de las modificaciones que son admisibles en el programa LP inicial y, en base a ello, mejorar la eficiencia de los métodos propuestos.

- Analizar el papel de la función objetivo en la creación y mejora de métodos de extracción de EFMs basados en LP.

- Analizar en profundidad el método `simplex` usado en estos métodos de forma que se puedan plantear métodos más eficientes de extracción de EFMs, incluso rompiendo la barrera del ratio ideal de eficiencia 1 planteado en el Capítulo 1.3.3.

- Iniciar el estudio, usando técnicas estadísticas, de la representatividad de un conjunto de EFMs, al menos dentro de los límites comentados en la sección 1.5 de esta introducción. Como hemos observado anteriormente, este estudio es fundamental ya que, de no realizarse, cualquier inferencia sobre la red en base a subconjuntos de EFMs calculados con cualquier método de extracción puede ser errónea debido a los sesgos producidos por el método empleado.

1.7 Estructura de esta tesis

La estructura de esta tesis por compendio de artículos intenta abordar y resolver los objetivos que acabamos de proponer. La tesis está dividida en tres grandes capítulos:

- Capítulo 1- Introducción y análisis de las principales técnicas usadas durante la realización de este trabajo. Material necesario para motivar y contextualizar las investigaciones realizadas, y exponer los retos a los que nos hemos enfrentado.
- Capítulo 2- Publicaciones que componen esta tesis doctoral. En este capítulo se ofrecen los 3 artículos publicados y que dan lugar a la tesis por compendio, más un artículo adicional (el primero) que completa perfectamente el contenido mostrado en los otros tres artículos.
  - Capítulo 2.1.- Artículo publicado por *J. F. Hidalgo, F. Guil y J.M. García*, ***Improving the performance of pathway extraction methods by infeasibilities removal***. Este artículo analiza el tipo de restricciones que se pueden imponer a un problema LP para la obtención de EFMs y estudia el impacto en la eficiencia de realizar un estudio previo de los tipos de reacciones implicadas.
  - Capítulo 2.2.- Artículo publicado por *J. F. Hidalgo, Jose A. Egea, F. Guil y J.M. García*, ***Improving the EFMs quality by augmenting their representativeness in LP methods***. Se recoge un primer estudio sobre la influencia del método escogido para extraer conjuntos de EFMs en la representatividad del conjunto obtenido.
  - Capítulo 2.3.- Artículo publicado por *F. Guil, J. F. Hidalgo y J.M. García*, ***Boosting the extraction of Elementary Flux Modes in Genome-ScaleMetabolic Networks using the Linear Programming approach***. En este artículo se presenta un nuevo método de extracción de EFMs basado en el análisis del tableau final del solver `simplex`. Dadas las dificultades de trabajar directamente con este tableau (debidas a su tamaño), se propone un algoritmo basado en una variante de este método que puede implementarse a partir de la solución final y de la matriz sparse del problema. Se demuestra que este algoritmo mejora la eficiencia de los métodos propuestos anteriormente.
  - Capítulo 2.4.- Artículo publicado por *J. F. Hidalgo, F. Guil y J.M. García*, ***Flux Coupling and the Objective Functions' Length in EFMs***. Artículo dedicado a estudiar las posibles modificaciones en la función objetivo usadas para obtener EFMs. Se centra en la segunda estrategia comentada, es decir, en disminuir el número de reacciones usadas en la función, y analiza el impacto sobre la eficiencia del método al elegir de forma adecuada la longitud del conjunto de dichas reacciones.
- Capítulo 3.- La tesis se cierra con un resumen de las principales aportaciones realizadas. Asimismo, se presentan también todas las contribuciones que se han

realizado relacionadas con esta tesis y se comenta su contenido. Finalmente se presenta una lista de tópicos cuyo estudio esperamos que proporcione en el futuro un mejor entendimiento de algunos de los temas y problemas tratados en esta tesis.

Capítulo 2

Publicaciones que componen esta Tesis Doctoral

2.1 *Improving the performance of pathway extraction methods by infeasibilities removal*

Título	*Improving the performance of pathway extraction methods by infeasibilities removal*
Autores	J. F. Hidalgo, F. Guil y J.M. García
Tipo	*Conferencia*
Conferencia	*International Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018)*
Páginas	121–136
Año	2018
Mes	Abril
URL	`http://iwbbio.ugr.es/Proceedings_ExtendedAbstract.pdf`
Estado	Publicado

Detalles de la conferencia
Nombre: *International Conference on Bioinformatics and Biomedical Engineering*
ISBN: 978-84-17293-36-9
Editorial: Godel S.L.
Índicios de calidad: Ratio de aceptación: 40 % (122/309). Posición: CORE B.
Página Web: `http://iwbbio.ugr.es/pdf/Proceedings_ExtendedAbstract.pdf`

Relación de Autores	
Nombre	**Jose F. Hidalgo** Universidad de Murcia
Nombre	**Dr. Francisco Guil** Universidad de Murcia
Nombre	**Dr. José M. García** Universidad de Murcia

Contribución del Doctorando
José F. Hidalgo Céspedes, declara haber participado en el diseño del método, en la redacción del artículo y haber realizado la programación y realizado los test que aparecen en el artículo *Improving the performance of pathway extraction methods by infeasibilities removal*.

Abstract

Motivation: Genome-scale metabolic network (GSMN) analysis requires efficient calculation of modes and specifically elementary flux modes (EFMs). State of the art methods for retrieving EFMs and pathways for biological studies use a linear programming formulation which is usually solved using the *simplex* method. One of the main drawbacks is the existence of large amounts of infeasible solutions, being an important fraction of them hard-coded in the metabolic reconstruction itself.

Results: Here we propose a method to avoid many of the possible infeasible solutions that can appear extracting modes. Furthermore, the computational cost that can be saved by applying this method can easily be estimated. Results over different case studies found in the literature are provided withing the paper.

2.2 *Improving the EFMs quality by augmenting their representativeness in LP methods*

Título	*Improving the EFMs quality by augmenting their representativeness in LP methods*
Autores	J. F. Hidalgo, Jose A. Egea, F. Guil y J.M. García
Tipo	*Publicación*
Publicación	*BMC Systems Biology*
Páginas	123–131
Año	2018
Mes	Noviembre
DOI	`https://doi.org/10.1186/s12918-018-0619-1`
URL	`https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-018-0619-1`
Estado	Publicado

Detalles de la revista

Nombre: *BMC Systems Biology*
ISSN: 1752-0509
Editorial: BIOMED Central Ltd.
Índicios de calidad: Factor de impacto de 2.0418 (ISI-JCR 2018). Cuartil: Q2: Mathematical & Computational Biology (22/59).
Web: `https://bmcsystbiol.biomedcentral.com/`

Relación de Autores	
Nombre	**Jose F. Hidalgo** Universidad de Murcia
Nombre	**Dr. Jose A. Egea** Universidad de Politécnica de Cartagena
Nombre	**Dr. Francisco Guil** Universidad de Murcia
Nombre	**Dr. José M. García** Universidad de Murcia

Contribución del Doctorando
José F. Hidalgo Céspedes, declara haber: participado en el diseño del método, en la redacción del artículo y haber realizado la programación y realizado los test que aparecen en el artículo *Improving the EFMs quality by augmenting their representativeness in LP methods.*

Abstract

Background: Although cellular metabolism has been widely studied, its fully comprehension is still a challenge. A main tool for this study is the analysis of meaningful pieces of knowledge called modes and, in particular, specially interesting classes of modes such as pathways and Elementary Flux Modes (EFMs). Their study often has to deal with issues such as the appearance of infeasibilities or the difficulty of finding representative enough sets of modes that are free of duplicated information. Mode extraction methods usually incorporate strategies devoted to mitigate these phenomena but they still get a high ratio of repetitions in the set of solutions.

Results: This paper presents a proposal to improve the representativeness of the full set of metabolic reactions in the set of computed modes by penalizing the eventual high frequency of occurrence of some reactions during the extraction. This strategy can be applied to any linear programming based extraction existent method.

Conclusions: Our strategy enhances the quality of a set of extracted EFMs favouring the presence of every reaction in it and improving the efficiency by mitigating the occurrence of repeated solutions. The new proposed strategy can complement other EFMs extraction methods based on linear programming. The obtained solutions are more likely to be diverse using less computing effort and improving the efficiency of the extraction.

2.3 *Boosting the extraction of Elementary Flux Modes in Genome-Scale Metabolic Networks using the Linear Programming approach*

Título	*Boosting the extraction of Elementary Flux Modes in Genome-Scale Metabolic Networks using the Linear Programming approach*
Autores	F. Guil, J. F. Hidalgo y J.M. García
Tipo	*Publicación*
Publicación	*Bioinformatics*
Año	2020
Mes	Abril
DOI	`10.1093/bioinformatics/btaa280`
URL	`https://doi.org/10.1093/bioinformatics/btaa280`
Estado	Publicado

Detalles de la revista

Nombre: *Bioinformatics*
ISBN: 978-84-17293-36-9
Editorial: BIOMED Central Ltd.
Índicios de calidad: Factor de impacto de 5.610 (ISI-JCR 2019). Cuartil: Q1: Mathematical & Computational Biology (3/59).
Web: `https://academic.oup.com/bioinformatics`

Relación de Autores	
Nombre	**Dr. Francisco Guil** Universidad de Murcia
Nombre	**Jose F. Hidalgo** Universidad de Murcia
Nombre	**Dr. José M. García** Universidad de Murcia

Contribución del Doctorando
José F. Hidalgo Céspedes, declara haber: participado en el diseño del método, en la redacción del artículo y haber colaborado con la programación y los test que aparecen en el artículo *Boosting the extraction of Elementary Flux Modes in Genome-Scale Metabolic Networks using the Linear Programming approach.*

Abstract

Motivation: Elementary flux modes (EFMs) are a key tool for analyzing genome-scale metabolic networks (GSMNs), and several methods have been proposed to compute them. Among them, those based on solving Linear Programming (LP) problems are known to be very efficient if the main interest lies in computing large enough sets of EFMs.

Results: Here, we propose a new method called EFM-Ta that boosts the efficiency rate by analyzing the information provided by the LP solver. We base our method on a further study of the final Tableau of the simplex method. By performing additional elementary steps and avoiding trivial solutions consisting of 2-cycles, we obtain many more EFMs for each LP problem posed, improving the efficiency rate of previously proposed methods by more than one order of magnitude.

2.4 *Flux Coupling and the Objective Functions' Length in EFMs*

Título	*Flux Coupling and the Objective Functions' Length in EFMs*
Autores	F. Guil, J. F. Hidalgo y J.M. García
Tipo	*Publicación*
Publicación	*Metabolites*
Páginas	489
Año	2020
Mes	November
DOI	`https://doi.org/10.3390/metabo10120489`
URL	`https://www.mdpi.com/2218-1989/10/12/489/pdf`
Estado	Publicado

Detalles de la revista
Nombre: *Metabolites*
ISSN: 2218-1989
Editorial: Multidisciplinary Digital Publishing Institute
Índicios de calidad: JCR Category Rank: 95/297 (Q2) in 'Biochemistry & Molecular Biology'.
Web: `https://www.mdpi.com/journal/metabolites`

Relación de Autores	
Nombre	**Jose F. Hidalgo** Universidad de Murcia
Nombre	**Dr. Francisco Guil** Universidad de Murcia
Nombre	**Dr. José M. García** Universidad de Murcia

Contribución del Doctorando
José F. Hidalgo Céspedes, declara haber: participado en el diseño del método, en la redacción del artículo y haber colaborado con la programación y los test que aparecen en el artículo *Flux Coupling and the Objective Functions' Length in EFMs*.

Abstract

Motivation: Structural analysis of constraint-based metabolic network models attempts to find the network's properties by searching for subsets of suitable modes or Elementary Flux Modes (EFMs). One useful approach is based on Linear Program (LP) techniques, which introduce an objective function to convert the stoichiometric and thermodynamic constraints into a linear program (LP), using additional constraints to generate different non-trivial modes. This work introduces FLFS-FC (Fixed Length Function Sampling with Flux Coupling), a new approach to increase the efficiency of generation of large sets of different EFMs for the network. FLFS-FC is based on the importance of the length of the objective functions used in the associated LP problem and the imposition of additional negative constraints.

Results: Our proposal overrides some of the known drawbacks associated with the EFM extraction, such as the appearance of unfeasible problems or multiple repeated solutions arising from different LP problems.

Capítulo 3

Conclusiones y vías futuras

Este capítulo ofrece un resumen de la principales aportaciones de los artículos incluidos en la tesis así como da idea de las posibles líneas de trabajo que completarán y ampliarán estos resultados.

3.1 Conclusiones científicas

El primer capítulo nos ha permitido dar un somero resumen de las principales ideas que nos llevaron a abordar los problemas centrales de este trabajo. Al final de la misma, y a modo de sumario, se expusieron cuáles eran estos.

En el segundo Capítulo encontramos los artículos que contienen las principales aportaciones de esta tesis a los temas planteados. El orden que emplearemos aquí no es exactamente el mismo en que aparecen en la tesis (cronológico), pero creemos que se adapta mejor a tener una visión de conjunto de la misma.

- En el Capítulo 2.1 abordamos el estudio de las posibles restricciones adicionales que pueden usarse para obtener nuevos EFMs. Se estudia en mayor profundidad el problema de las reacciones bloqueadas, su posible origen, así como los métodos para detectar y eliminar dichas reacciones del problema. Este Capítulo termina con el análisis del impacto en la eficiencia de los métodos de extracción de EFMs que tiene dicha eliminación.

- En el Capítulo 2.4 nos enfrentamos a un estudio semejante al del Capítulo 2.1, pero centrándonos en este caso en la importancia de la modificación de la función objetivo. Se estudia la mejora de eficiencia cuando se modifica un programa lineal usando las restricciones adicionales o la función objetivo. Para este último enfoque, se analiza en detalle una de las posibles modificaciones (restringir el número de reacciones a incluir en ella con coeficientes no nulos) y cómo afecta a la eficiencia de nuestros métodos el hecho de elegir más o menos reacciones para formar parte de dicha función. Este artículo incluye también el análisis de cuál es la mejor longitud de función posible para algunas situaciones concretas. Así

mismo se realiza un estudio del impacto que tiene en la eficiencia la inclusión de restricciones negativas.

- Capítulo 2.2 presenta un primer estudio de la influencia del método de extracción elegido en el tipo de EFMs obtenidos. Se proponen dos métodos sencillos (extracción usando funciones escogidas de forma totalmente aleatoria y funciones con coeficientes elegidos en función de EFMs anteriores) y se demuestra que los conjuntos de reacciones obtenidos son diferentes en una medida que es estadísticamente significativa.

- Finalmente el capítulo 2.3 presenta un nuevo algoritmo de extracción de EFMs (llamado EFM-Ta) que mejora de forma muy significativa la eficiencia de métodos anteriores. Este nuevo método se basa en la inspección del tableau final del método simplex. En los casos en los que trabajar con ese tableau final sea imposible o muy costoso computacionalmente, se muestra un algoritmo que puede usarse de forma alternativa. Se comprueba la eficiencia de este método en diversas redes de tamaño grande y se analiza cómo adaptarlo al problema más específico de encontrar EFMs que contienen una determinada reacción objetivo.

En resumen, creemos que las aportaciones expuestas en la sección anterior deben ser vistas como pasos interesantes en el estudio de los temas que nos proponíamos abordar.

3.2 Principales aportaciones

Al hilo de los objetivos planteados en la Introducción para esta tesis, pasamos a enumerar las principales aportaciones de este trabajo:

- *Analizar las posibles formas de modificar el problema LP inicial asociado a la red metabólica para así clarificar las posibles estrategias que conduzcan al cálculo de EFMs basados en LP.*

  Se han analizado las dos estrategias que conducen a realizar modificaciones en el problema lineal original: introduciendo restricciones adicionales o cambiando la función objetivo. Como se comenta en el Capítulo 2.4, el cambio en la función objetivo tiene un mayor impacto en la eficiencia del método, pero el ideal es un método que incluya ambas modificaciones. A lo largo de las diversas aportaciones se ha analizado también las limitaciones que deben imponerse a estas modificaciones.

- *Hacer un análisis previo de los tipos de reacciones que pueden aparecer en una red como paso previo al análisis de las modificaciones que son admisibles en el programa LP inicial y, en base a ello, mejorar la eficiencia de los métodos propuestos.*

  En este sentido se ha analizado la influencia de eliminar las reacciones bloqueadas y metabolitos externos antes de proceder a la extracción de EFMs. En segundo

lugar se ha estudiado la interrelación entre las reacciones provenientes de desdoblar reacciones reversibles y las estrategias basadas en imponer una restricción positiva adicional. Se ha mostrado cómo evitar 2-ciclos en el proceso de extracción (tanto los que corresponden a EFMs como los falsos EFMs producidos por el proceso de dedoblamiento). Se ha analizado cómo este estudio previo mejora claramente la eficiencia.

- *Analizar el papel de la función objetivo en la creación y mejora de métodos de extracción de EFMs basados en LP.*

  Ya hemos comentado que la modificación adecuada de la función a modificar es la que produce un mayor impacto en la eficiencia del proceso de extracción. Como comentamos, una posible estrategia es modificarla a base de incluir en ella solamente subconjuntos de reacciones. Esta modificación puede verse como una restricción soft en el sentido de intentar eliminar todas las reacciones que forman parte de este conjunto. Pero quedaba la cuestión de si la cardinalidad de estos conjuntos jugaba algún papel en cuanto a la eficiencia de nuestro proceso. En el Capítulo 2.4 se procede a este análisis, llegando a la conclusión de que esta cardinalidad es un fáctor determinante en dicha eficiencia y se apunta a cómo obtener, para cada red particular, la cardinalidad óptima.

- *Analizar en profundidad el método simplex usando en estos métodos de forma que se puedan plantear métodos más eficientes de extracción de EFMs, incluso rompiendo la barrera del ratio ideal de eficiencia 1 planteado en el Capítulo 1.3.3.*

  Como comentamos, el ratio de eficiencia de un método de extracción de EFMs mediante LP se define como el número de programas LP necesarios para obtener un nuevo EFM. En este sentido, podría parecer que el ratio ideal es 1. Sin embargo, un análisis detallado del método simplex permite ir más allá de este ratio ideal, ya que se puede aprovechar cada solución obtenida, junto con el tableau final que el método proporciona, para obtener otras soluciones a partir de ésta. El Capítulo 2.3 presenta dos métodos de conseguir esto mediante un paso adicional de pivoteo sobre el tableau final o usando la solución y la matriz estequiométrica en forma sparse. Este método, llamado EFM-Ta, presenta un ratio de eficiencia mucho mejor que otros métodos propuestos anteriormente.

- *Iniciar el estudio, usando técnicas estadísticas, de la representatividad de un conjunto de EFMs, al menos dentro de los límites comentados en 1.5 en esta introducción.*

  Este estudio es fundamental ya que, de no realizarse, cualquier inferencia sobre la red en base a subconjuntos de EFMs calculados con cualquier método de extracción puede ser errónea debido a los sesgos producidos por el método empleado. Un primer paso de dicho estudio es ser capaces de determinar si dos métodos de extracción producen subconjuntos que tengan la misma tipología. Para abordarlo hemos usado técnicas de comparación de muestras (concretamente el test de Wilcoxon). El resultado obtenido es que, como era de esperar, el método

de extracción influye de modo estadísticamente significativo en el tipo de EFMs obtenidos.

3.3 Publicaciones realizadas de esta tesis

A continuación, ofrecemos todas las publicaciones que se han realizado durante el desarrollo de este tesis doctoral:

1. José F. Hidalgo, Francisco Guil, and José M. García, (2015). *A new approach to obtain EFMs using graph methods based on the shortest path between end nodes*. Proc. de 3th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2015), Granada, Spain. Páginas: 641-649, 2015. Springer International Publishing. Ratio de aceptación: 50 %(134/268). Posición: B. https://doi.org/10.1007/978-3-319-16483-0_62 .

   Esta contribución a congreso constituye el comienzo del estudio de los elementos que constituyen el análisis de las redes metabólicas y la extracción de rutas elementales. Se estudia el problema desde el punto de vista de la red como digrafo dirigido. Se plantea un enfoque de la minimalidad de las rutas basado en el cálculo de caminos más cortos con el algoritmo de Dijkstra. No ha sido incluido en el compendio que conforma la tesis por ser un trabajo exploratorio de la materia, y por utilzar técnicas que no están alineadas con las que si hemos utilizado intensamente en las publicaciones más recientes y de mayor impacto editorial.

2. José F. Hidalgo, Francisco Guil, and José M. García, (2016). *Computing EFMs using balanced subgraphs and boolean logic*. En 4th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2016), Granada, Spain. Abril 2016.

   En esta ponencia describimos la característica de grafo balanceado desde el punto de vista del álgebra booleana. Un modo, elemental o no, cumple la condición necesaria de equivaler a un grafo balanceado. En este artículo se asienta la diferencia entre grafo balanceado y modo, y se justifica con la caracterización de la regla del rango.

   No ha sido incluido en el compendio que conforma la tesis por que finalmente los métodos basados en el álgebra booleana presentan una eficiencia considerablemente menores que los basados en LP, por lo que no ha supuesto una linea trabajo que contribuya al resultado de esta tesis.

3. José F. Hidalgo, Francisco Guil, and José M. García, (2016). *A new approach to obtain EFMs using graph methods based on the*

shortest path between end nodes. Genomics and Computational Biology, [S.l.], v. 2, n. 1, p. e30, sep. 2016. ISSN 2365-7154. Available at: https://genomicscomputbiol.org/ojs3/GCB/article/view/27. doi: https://doi.org/10.18547/gcb.2016.vol2.iss1.e30.

Este artículo es la secuela de la contribución al congreso IWBBIO 2016. Se trata de una revista open-access y de revisión entre pares. Contiene una versión extendida de la contribución de dicho congreso.

Con esta publicación se cierra la etapa de exploración de redes metabólicas con técnicas basadas en grafos.

4. Jose A. Egea, José F. Hidalgo, Francisco Guil y José M. García, (2016). *Cálculo de Modos de Flujo Elemental en redes metabólicas mediante búsqueda dispersa*. En XVII Conferencia de la Asociación Española para la Inteligencia Artificial, Salamanca (Spain). Septiembre 2016.

   Esta contribución presenta el algoritmo de búsqueda dispersa para la extracción de EFMs mediante la modificación la función objetivo de programas lineales que han producido EFMs. La modificación del gradiente de la función objetivo nos permite obtener nuevas soluciones sin incurrir en inviabilidad del nuevo problema planteado. Esta idea la volvemos a poner en juego con más claridad en artículos posteriores como un complemento a diferentes estrategias. El apartado 1.4.2 analiza algunas de sus conclusiones.

5. José F. Hidalgo, Jose A. Egea, Francisco Guil, and José M. García (2017) *Representativeness of a Set of Metabolic Pathways*. In: Rojas I., Ortuño F. (eds). 5th International Bioinformatics and Biomedical Engineering Congress (IWBBIO 2017), Granada, Spain. Ratio de aceptación: 40 % (122/309). Posición: B. https://doi.org/10.1007/978-3-319-56148-6_58. Lecture Notes in Computer Science, vol 10208. Springer, Cham.

   Esta contribución redunda en la modificación de la función objetivo para técnicas de extracción de EFMs con programación lineal. En este caso, se aplica al objetivo de acelerar la representatividad del conjunto de soluciones que vamos obteniendo. Damos una medida de cómo valorarlo con respecto a la propia evolución del algoritmo y con respecto al conjunto completo de soluciones en el hipotético caso de que estuviera disponible. Sobre la medida de la representatividad y el significado biológico y la falacia interpretativa en la que se puede incurrir sin ellos hemos presentado avances el capítulo 1.5.

6. José F. Hidalgo, Francisco Guil, and José M. García (2018) *Improving the performance of pathway extraction methods by infeasibilities removal*. 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO

2018), Granada, Spain. Abril 2018. Proceedings Extended abstracts, ISBN: 978-84-17293-36-9

Esta contribución a congreso muestra la importancia de la preparación de los modelos metabólicos como mecanismo para impulsar la eficiencia posterior en la búsqueda de EFMs. Es aplicable principalmente a técnicas basadas en programación lineal en las que se circunscribe esta tesis. Se alinea con el capítulo 1.3.1 sobre detección de reacciones bloqueadas y los efectos en términos de eficiencia de plantear programas lineales inviables.

7. José F. Hidalgo, Jose A. Egea, Francisco Guil, and José M. García, (2018). *Improving the EFMs quality by augmenting their representativeness in LP methods*. BMC Systems Biology, 12 (Suppl. 5), 101 (2018). https://doi.org/10.1186/s12918-018-0619-1

   Este artículo es la contribución ampliada de la presentada en IWBBIO 2017. Aparece el concepto de huella de un conjunto de soluciones como el atributo calculado para diferentes conjuntos de EFMs o soluciones y con el que poder establecer comparativas en términos de representatividad.

8. Francisco Guil, José F. Hidalgo, José M. García, (2020). *Boosting the extraction of Elementary Flux Modes in Genome-Scale Metabolic Networks using the Linear Programming approach*. Bioinformatics , 04 2020. doi: https://doi.org/10.1093/bioinformatics/btaa280

   En este artículo se presenta el método EFM-Ta (que significa *EFM using the Tableau*) que usa el algoritmo simplex para producir una solución óptima inicial que es un EFM y, tras ello, se realizan sencillos pasos para producir nuevos vértices, esto es, nuevos EFMs. Llamamos a esos nuevos vértices los *vértices adyacentes* de la primera solución. La mejor eficiencia obtenida anteriormente en la extracción de EFMs usando métodos basados en programación lineal se ve aumentada en 10x con nuestro método. El artículo detalla las diferentes técnicas y cómo pueden ser exportadas a otros métodos de extracción basados en LP.

9. Francisco Guil, José F. Hidalgo, and José M. García, (2020). *Flux Coupling and the Objective Functions' Length in EFMs*, Metabolites 10.12 (2020): 489. doi: https://doi.org/10.3390/metabo10120489

   Este artículo presenta una técnica heurística específica para modificar funciones objetivo de programas lineales que llamamos FLFS-FC. Los ingredientes principales son la elección de diferentes longitudes de función objetivo por su impacto observado en el conjunto resultantes de soluciones. Otro aspecto que introducimos es la incorporación de restricciones tanto positivas como negativas en nuestros problemas lineales. El método FLFS-FC mejora suficientemente los ratios de eficiencia obtenidos por otros grupos de investigación.

3.4 Vías futuras

Las aportaciones anteriores abren la posibilidad de ahondar en algunos de los problemas propuestas.

Un primer tema que pensamos que es de importancia central es continuar el estudio de la representatividad de los conjuntos de EFMs que se obtienen usando los diferentes métodos disponibles. Creemos esto porque, dado que la cantidad y calidad de modelos disponibles para diferentes GSMNs es cada vez mayor, es muy probable que el conjunto de EFMs no esté disponible para la mayoría de ellos a pesar de las mejoras en la eficiencia de los métodos de extracción. Como ya hemos comentado, es muy difícil extrapolar información de forma precisa al conjunto completo si solamente disponemos de subconjuntos de EFMs que no sabemos si son representativos. Creemos que se debe emprender un estudio sistemático para poder comparar los conjuntos obtenidos por distintos métodos. Esta comparación debe empezar por redes en las que ya es conocido el conjunto total de EFMs para después intentar extrapolar el conocimiento obtenido a redes más complejas.

También creemos que parte del estudio realizado en este trabajo podría ser extensible a otros enfoques, concretamente a aquellos basados en programación entera. Al igual que en métodos basados en LP, seguramente un análisis más detallado del algoritmo de optimización usado podría desembocar en algoritmos basados en MILP que tuvieran una eficiencia significativamente mejor que los desarrollados actualmente.

En línea con lo anterior, sería deseable poder extender la aplicación del algoritmo EFM-Ta de modo que se pudiera aplicar a otros contextos. Concretamente, sería deseable poder aplicarlo no solamente a la obtención de conjuntos de EFMs, sino también a la búsqueda de *Minimal Cut Sets* [37, 39], *Minimal Metabolic Behaviours* [43, 68] o *Minimum Set of Elementary Modes* [67].

Capítulo 4

Bibliografía

[1] V. Acuña, F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, and L. Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51–60, 2009.

[2] Mats Åkesson, Jochen Förster, and Jens Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metabolic engineering*, 6(4):285–293, 2004.

[3] Maciek R Antoniewicz. A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications. *Metabolic Engineering*, 2020.

[4] 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):1–9, 2017.

[5] Mona Arabzadeh, Morteza Saheb Zamani, Mehdi Sedighi, and Sayed-Amir Marashi. A graph-based approach to analyze flux-balanced pathways in metabolic networks. *Biosystems*, 165:40–51, 2018.

[6] 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.

[7] Susanna Bazzani. Article commentary: Promise and reality in the expanding field of network interaction analysis: Metabolic networks. *Bioinformatics and biology insights*, 8:BBI–S12466, 2014.

[8] Susanna Bazzani, Andreas Hoppe, and Hermann-Georg Holzhütter. Network-based assessment of the selectivity of metabolic drug targets in plasmodium falciparum with respect to human liver metabolism. *BMC systems biology*, 6(1):118, 2012.

[9] Scott A Becker and Bernhard O Palsson. Context-specific metabolic networks are consistent with experiments. *PLoS Comput Biol*, 4(5):e1000082, 2008.

[10] Aarash Bordbar, Jonathan M Monk, Zachary A King, and Bernhard O Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics*, 15(2):107–120, 2014.

[11] K-H Borgwardt. The average number of pivot steps required by the simplex-method is polynomial. *Zeitschrift für Operations Research*, 26(1):157–177, 1982.

[12] Anthony P Burgard and Costas D Maranas. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnology and bioengineering*, 82(6):670–677, 2003.

[13] Anthony P Burgard, Evgeni V Nikolaev, Christophe H Schilling, and Costas D Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome research*, 14(2):301–312, 2004.

[14] Jose Francisco Hidalgo Céspedes, Francisco De Asís Guil Asensio, and Jose Manuel García Carrasco. A new approach to obtain efms using graph methods based on the shortest path between end nodes. In *Bioinformatics and Biomedical Engineering*, pages 641–649. Springer, 2015.

[15] M.W. Covert and B.O. Palsson. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol. 221*, page 309–325, 2003.

[16] 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.

[17] Luis F De Figueiredo, Stefan Schuster, Christoph Kaleta, and David A Fell. Can sugars be produced from fatty acids? a test case for pathway analysis tools. *Bioinformatics*, 24(22):2615–2621, 2008.

[18] Xin Fang, Colton J Lloyd, and Bernhard O Palsson. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nature Reviews Microbiology*, 18(12):731–743, 2020.

[19] A M Feist, C S Henry, J L Reed, M Krummenacker, A R Joyce, P D Karp, L J Broadbelt, Vassily Hatzimanikatis, and B Ø Palsson. A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular Systems Biology*, 3(121), 2007.

[20] Adam M Feist and Bernhard O Palsson. The biomass objective function. *Current opinion in microbiology*, 13(3):344–349, 2010.

[21] J Forrest. Clp-coin-or linear program solver. In *DIMACS Workshop on COIN-OR, July*, pages 17–20, 2006.

[22] Komei Fukuda and Alain Prodon. Double description method revisited. In *Combinatorics and Computer Science*, volume 1120 of *Lecture Notes in Computer Science*, pages 91–111. Springer, 1995.

[23] J. Gagneur and S. Klamt. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5(175), 2004.

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

[25] Cameron Glasscock. Genetic tools and approaches for engineering metabolism and metabolic pathways. 2019.

[26] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdóttir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019.

[27] Christopher S Henry, Matthew D Jankowski, Linda J Broadbelt, and Vassily Hatzimanikatis. Genome-scale thermodynamic analysis of escherichia coli metabolism. *Biophysical journal*, 90(4):1453–1461, 2006.

[28] J. F. Hidalgo, J. A. Egea, F. Guil, and J. M. García. Representativeness of a set of metabolic pathways. In *Bioinformatics and Biomedical Engineering*, volume 10208, pages 659–667, Granada (Spain), April 2017. Springer International Publishing.

[29] José F. Hidalgo, José A. Egea, Francisco Guil, and José M. García. Improving the efms quality by augmenting their representativeness in lp methods. *BMC Systems Biology*, 12 (Suppl. 5):101:123–131, November 2018.

[30] Jose F. Hidalgo, Francisco Guil, and Jose M. Garcia. A new approach to obtain efms using graph methods based on the shortest path between end nodes. In Francisco Ortuño and Ignacio Rojas, editors, *Bioinformatics and Biomedical Engineering*, volume 9043 of *lnbi*, pages 641–649, Granada (Spain), April 2015. Springer International Publishing.

[31] H. Scott Hinton. E.coli core model for beginners (part 1). https://opencobra.github.io/cobratoolbox/latest/tutorials/.

[32] Peggy P Hsu and David M Sabatini. Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5):703–707, 2008.

[33] K. Hunt, J. Folsom, R. Taffs, and R. Carlson. Complete enumeration of elementary flux modes through scalable demand-based subnetwork definition. *Bioinformatics (Oxford, England)*, 30:1569–1578, 06 2014.

[34] IBM. IBM ILOG CPLEX Optimizer. http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/, Last 2010.

[35] Sudhakar Jonnalagadda and Rajagopalan Srinivasan. An efficient graph theory based method to identify every minimal reaction set in a metabolic network. *BMC systems biology*, 8(1):28, 2014.

[36] C. Kaleta, L. F. de Figueiredo, J. Behre, and S. Schuster. Efmevolver: Computing elementary flux modes in genome-scale metabolic networks. In *Lecture Notes in Informatics-Proceedings*, volume 157, pages 179–189, 2009.

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

[38] S. Klamt, J. Gagneur, and A. von Kamp. Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *IEE Proceedings - Systems Biology*, 152(4):249–255, 2005.

[39] S. Klamt and E. D. Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics, 20(2), 226–234*, 2004.

[40] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends Biotechnol*, 21:64–69, 2003.

[41] A. Larhlimi and A. Bockmayr. A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, 157:2257–2266, 05 2009.

[42] A. Larhlimi, L. David, J. Selbig, and Bockmayr A. F2C2: a fast tool for the computation of flux coupling in genome-scale metabolic networks. *BMC Bioinformatics 13:57*, 2012.

[43] Abdelhalim Larhlimi. *New concepts and tools in constraint-based analysis of metabolic networks*. PhD thesis, 2009.

[44] Sang Yup Lee, Hyun Uk Kim, Tong Un Chae, Jae Sung Cho, Je Woong Kim, Jae Ho Shin, Dong In Kim, Yoo-Sung Ko, Woo Dae Jang, and Yu-Sin Jang. A comprehensive metabolic map for production of bio-based chemicals. *Nature Catalysis*, 2(1):18–33, 2019.

[45] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS computational biology*, 10(4), 2014.

[46] Daniel Machado, Zita Soons, Kiran Raosaheb Patil, Eugénio C. Ferreira, and Isabel Rocha. Random sampling of elementary flux modes in large-scale metabolic networks. *Bioinformatics*, 28(18):i515–i521, 9 2012.

[47] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature biotechnology*, 35(1):81, 2017.

[48] Sayed-Amir Marashi, Laszlo David, and Alexander Bockmayr. Analysis of metabolic subnetworks by flux cone projection. *Algorithms for Molecular Biology*, 7(1):17, 2012.

[49] Jens Nielsen and Jay D Keasling. Engineering cellular metabolism. *Cell*, 164(6):1185–1197, 2016.

[50] University of California. Bigg models. http://http://bigg.ucsd.edu.

[51] Jeffrey D Orth, Ronan MT Fleming, and Bernhard Ø Palsson. Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide. *EcoSal Plus*, 4(1), 2010.

[52] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature biotechnology*, 28(3):245–248, 2010.

[53] B.O. Palsson. The challenges of in silico biology. *Nat. Biotechnol 18*, page 1147–1150, 2000.

[54] J Pey and FJ Planes. Direct calculation of elementary flux modes satisfying several biological constraints in genome-scale metabolic networks. *Bioinformatics (Oxford, England)*, 30(15):2197, 2014.

[55] J. Pey, J. A. Villar, L. Tobalina, A. Rezola, J. M. García, J. E. Beasley, and F. J. Planes. Treeefm: calculating elementary flux modes using linear optimization in a tree-based algorithm. *Bioinformatics*, 31(6):897–904, 2015.

[56] T. Pfeiffer, I. Sanchez-Valdenebro, J.C. Nuno, F. Montero, and S. Schuster. METATOOL: for studying metabolic networks. *Bioinformatics 15*, page 251–257, 1999.

[57] F.J. Planes and F.E. Beasley. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Briefings in bioinformatics*, 9(5):422–436, 2008.

[58] Francisco J Planes. *Metabolic pathway analysis via integer linear programming.* PhD thesis, Brunel University, School of Information Systems, Computing and Mathematics, 2008.

[59] Francisco J Planes and John E Beasley. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Briefings in Bioinformatics*, 9(5):422–436, 2008.

[60] Lake-Ee Quek and Lars K Nielsen. A depth-first search algorithm to compute elementary flux modes by linear programming. *BMC Systems Biology*, 8(1), 2014.

[61] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, Bernhard O Palsson, et al. An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9):R54, 2003.

[62] A. Rezola, L. F. De Figueiredo, M. Brock, J. Pey, A. Podhorski, C. Wittmann, S. Schuster, A. Bockmayr, and F. J. Planes. Exploring metabolic pathways in genome-scale networks via generating flux modes. *Bioinformatics*, 27:534–540, 2011.

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

[64] Alberto Rezola, Jon Pey, Luis Tobalina, Ángel Rubio, John E Beasley, and Francisco J Planes. Advances in network-based metabolic pathway analysis and gene expression data integration. *Briefings in bioinformatics*, 16(2):265–279, 2015.

[65] A. Röhl and A. Bockmayr. A mixed-integer linear programming approach to the reduction of genome-scale metabolic networks. *BMC bioinformatics*, 18(2), 2017.

[66] A. Röhl, T. Riou, and A. Bockmayr. Computing irreversible minimal cut sets in genome-scale metabolic networks via flux cone projection. *Bioinformatics*, 35(15):2618–2625, 2019.

[67] Annika Röhl and Alexander Bockmayr. Finding memo: Minimum sets of elementary flux modes. *Journal of Mathematical Biology*, 79(5):1749–1777, 2019.

[68] Annika Röhl, Tanguy Riou, and Alexander Bockmayr. Computing irreversible minimal cut sets in genome-scale metabolic networks via flux cone projection. *Bioinformatics*, 35(15):2618–2625, 2019.

[69] Johann M Rohwer, Stefan Schuster, and Hans V Westerhoff. How to recognize monofunctional units in a metabolic system. *Journal of Theoretical Biology*, 179(3):213–228, 1996.

[70] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[71] S. Schuster, T. Dandekar, and D. A Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology*, 17(2):53–60, 1999.

[72] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(02):165–182, 1994.

[73] Stefan Schuster, David A Fell, and Thomas Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326–332, 2000.

[74] Evgeni Selkov, Natalia Maltsev, Gary J Olsen, Ross Overbeek, and William B Whitman. A reconstruction of the metabolism of methanococcus jannaschii from sequence data. *Gene*, 197(1):GC11–GC26, 1997.

[75] H Seo, D-Y Lee, S Park, LT Fan, S Shafie, B Bertók, and F Friedler. Graph-theoretical identification of pathways for biochemical reactions. *Biotechnology letters*, 23(19):1551–1557, 2001.

[76] Jörg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002.

[77] George Stephanopoulos, Aristos A Aristidou, and Jens Nielsen. *Metabolic engineering: principles and methodologies*. Elsevier, 1998.

[78] S. Tabe-Bordbar and S. Marashi. Finding elementary flux modes in metabolic networks based on flux balance analysis and flux coupling analysis: application to the analysis of escherichia coli metabolism. *Biotechnol Lett*, 35:2039–2044, 2013.

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

[80] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, 2010.

[81] Cong T Trinh, Johnny Li, Harvey W Blanch, and Douglas S Clark. Redesigning escherichia coli metabolism for anaerobic production of isobutanol. *Applied and environmental microbiology*, 77(14):4894–4904, 2011.

[82] Robert Urbanczik. Sna–a toolbox for the stoichiometric analysis of metabolic networks. *BMC bioinformatics*, 7(1):129, 2006.

[83] Robert Urbanczik and Clemens Wagner. Functional stoichiometric analysis of metabolic networks. *Bioinformatics*, 21(22):4176–4180, 2005.

[84] Amit Varma and Bernhard O Palsson. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/technology*, 12(10):994–998, 1994.

[85] Nikos Vlassis, Maria Pires Pacheco, and Thomas Sauter. Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol*, 10(1):e1003424, 2014.

[86] Clemens Wagner and Robert Urbanczik. The geometry of the flux cone of a metabolic network. *Biophysical journal*, 89(6):3837–3845, 2005.

[87] Sharon J Wiback and Bernhard O Palsson. Extreme pathway analysis of human red blood cell metabolism. *Biophysical journal*, 83(2):808–818, 2002.

[88] Francisco Guil, José F Hidalgo, and José M García. Boosting the extraction of Elementary Flux Modes in Genome-Scale Metabolic Networks using the Linear Programming approach. *Bioinformatics*, 04 2020. btaa280.

[89] José F. Hidalgo, José A. Egea, Francisco Guil, and José M. García. Improving the efms quality by augmenting their representativeness in lp methods. *BMC Systems Biology*, 12 (Suppl. 5):101:123–131, November 2018.

[90] José F. Hidalgo, José A. Egea, Francisco Guil, and José M. García. Representativeness of a set of metabolic pathways. In Ignacio Rojas and Francisco Ortuño, editors, *Bioinformatics and Biomedical Engineering*, volume 10208, pages 659–667, Granada (Spain), April 2017. Springer International Publishing.

[91] José F. Hidalgo, Francisco Guil, and José M. García. A new approach to obtain efms using graph methods based on the shortest path between end nodes. In Francisco Ortuño and Ignacio Rojas, editors, *Bioinformatics and Biomedical Engineering*, volume 9043, pages 641–649, Granada (Spain), April 2015. Springer International Publishing.

[92] José F. Hidalgo, Francisco Guil, and José M. García. Improving the performance of pathway extraction methods by infeasibilities removal. In Daniel Castillo, Juan Manuel Gálvez, Maria José Sáez, Fernando Rojas, Luis Javier Herrera, and Ignacio Rojas, editors, *Proc. of the 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018)*, pages 121–136, Granada, Spain, April 2018. Editorial Godel S.L.