Metabolic Pathway Analysis: from small to genome-scale networks

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium **(Dr. rer. nat.)**

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät
der Friedrich-Schiller-Universität Jena

von Dipl.-Ing. Luís F. D. P. de Figueiredo
geboren am 12. März 1982 **in** Lissabon, Portugal

Examiners / Gutachter:

1. Prof. Dr. Stefan Schuster (Friedrich-Schiller-Universität Jena, Jena)
2. PD Dr. Peter Dittrich (Friedrich-Schiller-Universität Jena, Jena)
3. Prof. Dr. Marie-France Sagot (INRIA Grenoble Rhône-Alpes, Université Claude Bernard, Lyon)

Date of the public defence / Tag der öffentlichen Verteidigung: 22.02.2011

The following work was developed in the context of the PhD Program in Computational Biology organized by the Instituto Gulbenkian de Ciência, Oeiras, Portugal. Financial support was given by the following Portuguese institutions: Fundação Calouste Gulbenkian, Fundação para a Ciência e Tecnologia, Academia Siemens and Fundação PT.

The author of this work received a PhD fellowship (SFRH/BD/32961/2006) from the Fundação para a Ciência e Tecnologia, co-financed by the POPH/FSE.

The research project associated with this work was carried at the Department of Bioinformatics from the School of Biology and Pharmacy of the Friedrich-Schiller-Universität Jena under the supervision of Prof. Dr. Stefan Schuster. Moreover, the BIOINF group from Prof. Dr. Stefan Schuster is a partner of the Jena Center for Bioinformatics.

Die vorliegende Arbeit wurde im Rahmen des Promotions-Programmes über *Computational Biology* angefertigt. Dieses Promotions-Programm wurde vom Instituto Gulbenkian de Ciência in Oeiras, Portugal organisiert. Die finanzielle Förderung erfolgte durch folgende portugiesisch Institutionen: Fundação Calouste Gulbenkian, Fundação para a Ciência e Tecnologia, Academia Siemens und Fundação PT.

PhD Program in Computational Biology

Der Autor dieser Arbeit erhielt ein Stipendium (SFRH/BD/32961/2006) der Fundação para a Ciência e Tecnologia, mitfinanziert von POPH/FSE.

Das Forschungs-Projekt in dieser Dissertation wurde am Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena unter der Leitung von Prof. Dr. Stefan Schuster angefertigt. Die BIOINF-Forschungsgruppe von Prof. Dr. Schuster ist ein Teil des Netzwerkes des Jena Center for Bioinformatics.

Abstract

Biology faced a revolution in the last century, in particular molecular biology. Starting in the beginning of the 20^{th} century from little knowledge on enzyme catalyzed reactions, structure of macromolecules and information coding, the scientific community is now able to generate thousands of terabytes of data quantifying biological processes. The need for mathematical modelling of these processes has grown alongside with the achievements in the experimental field leading to the appearance and development of new fields like systems biology. Systems biology aims at generating new knowledge through modelling and integration of experimental data in order to develop a holistic understanding of organisms.

In the first part of my PhD thesis, I compare two different levels of abstraction used for computing metabolic pathways, constraint-based and graph theoretical methods. I show that the current representations of metabolism as a simple graph correspond to wrong mathematical descriptions of metabolic pathways. On the other hand, the use of stoichiometric information and convex analysis as modelling framework like in elementary flux mode analysis, allows to correctly predict metabolic pathways. However, this approach does not scale up well with the size of the input network and therefore, graph-theory based methods have been developed to cope with the demands of systems biology for modelling genome-scale metabolic networks.

In the second part of the thesis, I present two of the first methods, based on elementary flux mode analysis, that can compute metabolic pathways in such large metabolic networks: the K-shortest EFMs method and the EFMEvolver method.

These methods contribute to an enrichment of the mathematical tools available to model cell biology and more precisely, metabolism. The optimization frameworks used are important to focus particular metabolic pathways present in the solution space, thereby allowing to deal with the combinatorial nature of elementary flux modes.

The application of these new methods to biotechnological problems is also explored in this part. I study the metabolic pathways involved in the production of L-lysine in two microorganisms, *Escherichia coli* and *Corynebacterium glutamicum*. Lysine is an essential amino acid for humans and of high commercial relevance. The K-shortest EFMs method predicts biological relevant pathways converting glucose into lysine with the shortest number of reactions steps. On the other hand, the EFMEvolver method can cope with a larger set of elementary flux modes, including also the shortest ones. All these pathways can be subdivided into four parts carrying important functional roles, such as glucose catabolism or lysine biosynthesis.

In the last part of my thesis, I give an overview of recent achievements in metabolic network reconstruction and constraint-based modelling as well as open issues. Perspectives on further extensions of constraint-based modelling to multicellular organisms are given. Moreover, I discuss possible strategies for integrating experimental data with elementary flux mode analysis. Further improvements in elementary flux mode computation on that direction are put forward.

Zusammenfassung

Im Laufe des letzten Jahrhunderts ereigneten sich in vielen Gebieten der Biologie revolutionäre Umbrüche, vor allem aber in der molekularen Zellbiologie. Zu Beginn des 20. Jahrhunderts hatte man wenig Kenntnisse über enzymatische Reaktionen, praktisch keine Vorstellung über die Strukturen von Makromolekülen und der Kodierung von genetischer Information. Heutzutage haben die Forscher jedoch die Möglichkeit tausende Terabytes an Daten über biologische Prozesse zu erfassen. Die Datenmenge, die man mit Hilfe der experimentellen Biologie gewann, machte eine gezielte mathematische Modellierung der biologischen Prozesse unabdingbar. In direkter Symbiose entwickelten sich beide Bereiche weiter und gingen im Forschungsgebiet der Systembiologie auf. Das große Ziel der Systembiologie ist es, ganzheitliches Wissen über Organismen zu erlangen. Die Forscher versuchen dies zu erreichen, indem sie in die Modelle experimentelle Daten integrieren.

Im ersten Teil meiner Dissertation vergleiche ich zwei unterschiedliche mathematische Abstraktionsstufen, die oft genutzt werden, um Stoffwechselwege zu berechnen; die constraint-basierte und die graphen-theoretische Methoden. Ich zeige, dass die derzeit genutzte Form Stoffwechselwege als vereinfachte Graphen zu beschreiben, zu einer mathematisch falschen Beschreibung der Stoffwechselwege führt. Auf der anderen Seite erlauben die stöchiometrischen Informationen und die Konvexanalyse als Modellierungswerkzeuge, wie bei der Elementarmoden-Analyse, eine korrekte Vorhersage der Stoffwechselwege. Dennoch kann dieser Ansatz bei zu großen Ausgangsnetzwerken nicht angewendet werden; darum wurden graphen-basierte Methoden entwickelt, die den Ansprüchen der Systembiologie an die

Modellierung von metabolischen Ganzzellmodellen genügen.

Im zweiten Teil meiner Dissertation stelle ich zwei der ersten Methoden vor, die auf der Elementarmoden-Analyse basieren und die Stoffwechselwege in solch großen Stoffwechselnetzwerken berechnen können: die Methode der *K-shortest EFMs* und die Methode der *EFMEvolver*. Diese Methoden leisteten einen Beitrag zu den heute genutzten mathematischen Methoden in der molekularen Zellbiologie, genauer gesagt, im Bereich des Stoffwechsels. Die verwendeten Optimierungswege sind dafür sehr wichtig, da sie es erlauben, einige im Lösungsraum liegende Teile des Stoffwechselweges genauer zu untersuchen und darüberhinaus auf die kombinatorische Natur der Elementarmoden einzugehen.

Des Weiteren wird in diesem Teil meiner Arbeit die Anwendung dieser neuen Methoden auf biotechnologische Probleme vorgestellt. Ich habe die Lysin-Stoffwechselwege der zwei Mikroorganismen *Escherichia coli* und *Corynebacterium glutamicum* untersucht. Lysin ist eine für den Menschen lebensnotwendige Aminosäure und mit größ kommerziell Wichtigkeit. Die Methode der *K-shortest EFMs* zeigt biologisch relevante Wege auf, um Glukose mit der kürzesten Anzahl an Reaktionsschritten in Lysin umzuwandeln. Außerdem kann die Methode der *EFMEvolver* eine größere Anzahl an Elementarmoden bewältigen, eingeschlossen die kürzesten. Alle diese Stoffwechselwege können in bis zu vier Teile untergliedert werden, von denen jeder eine wichtige funktionelle Rolle ausführt, beispielsweise den Abbau von Glukose oder die Synthese von Lysin.

Abschließend gebe ich einen Überblick über neueste Entwicklungen in der Rekonstruktion von metabolischen Netzwerken sowie constraint-basierten Modellierungen und erläutere aktuelle Fragestellungen. Des Weiteren wird ein Ausblick über Entwicklungsperspektiven der constraint-basierten Modellierungen hin zu vielzelligen Organismen gegeben. Außerdem erörtere ich, wie die in dieser Arbeit vorgestellten Methoden genutzt werden können, um experimentelle Daten in die Elementarmoden-Analyse zu integrieren. Und ich präsentiere weitere Entwicklungen in der Berechnung von Elementarmoden.

Sumário

A Biologia assistiu a uma revolução durante o século passado, em particular na área de Biologia Molecular. No início do século 20, a comunidade científica possuía um conhecimento limitado sobre atividades enzimáticas, estrutura de macromoléculas e o código genético. Hoje em dia, a comunidade científica tem a capacidade de gerar milhares de terabytes de dados quantificando todo o tipo de processos biológicos. A necessidade de modelar matematicamente estes processos biológicos tem acompanhado os desenvolvimentos no campo experimental dando origem a novas áreas de pesquisa como a Biologia de Sistemas[*]. Em Biologia de Sistemas o grande desafio consiste em gerar conhecimento através da modelação e integração de dados experimentais, com vista ao desenvolvimento de um conhecimento global sobre os organismos.

Na primeira parte do meu trabalho de doutoramento eu comparo dois níveis diferentes de abstração que podem ser usados para o estudo das vias metabólicas, nomeadamente análise baseada em restrições[†] e a teoria de grafos. Eu mostro que as representações matemáticas de redes metabólicas em forma de grafos simples são inapropriadas para o cálculo de vias metabólicas. Por outro lado, o uso de informação estequiométrica e a análise de espaços vetoriais convexos, como por exemplo na análise de modos elementares de fluxo[‡], permite prever corretamente vias metabólicas. Contudo, esta abordagem não está preparada para o estudo de redes metabólicas de grande dimensão, razão pela qual os métodos baseados em

[*] do inglês *Systems Biology*

[†] do inglês *constraint-based analysis*

[‡] do inglês *elementary flux modes*

teoria de grafos têm vindo a ser desenvolvidos de modo a preencher os requisitos necessários na área de Biologia de Sistemas para o estudo de redes metabólicas à escala genómica.

Na segunda parte da minha tese, eu apresento dois dos primeiros métodos, baseados na análise de modos elementares de fluxo, que conseguem calcular vias metabólicas em grandes redes metabólicas: o método *K-shortest EFMs* e o método *EFMEvolver*. Estes métodos são um contributo importante para o enriquecimento das ferramentas disponíveis para o estudo da biologia celular, mais precisamente do metabolismo. O uso de conceitos de otimização é extremamente importante para poder focar determinadas soluções do espaço vetorial e desta forma, lidar com a natureza combinatória dos modos elementares de fluxo.

Nesta parte da tese, eu exploro também a aplicabilidade destes métodos novos no estudo de problemas da área de Biotecnologia, nomeadamente, no estudo das vias metabólicas envolvidas na produção de L-lisina em dois microrganismos, *Escherichia coli* e *Corynebacterium glutamicum*. Lisina é um aminoácido essencial nos humanos e de enorme importância económica. O método *K-shortest EFMs* prevê as vias metabólicas mais curtas que levam a cabo a conversão de glucose em lisina e de relevância biológica. Por outro lado, o método *EFMEvolver* consegue calcular um maior número de modos elementares de fluxo, incluindo os mais curtos. Todas estas vias metabólicas podem ser divididas em quatro partes com diferentes funções, tais como o catabolismo da glucose ou a síntese de lisina.

Na última parte da minha tese, eu dou uma visão geral sobre os desenvolvimentos recentes na reconstrução de redes metabólicas, na análise baseada em restrições, assim como, alguns tópicos ainda em aberto. Perspetivas sobre futuras extensões da análise baseada em restrições ao estudo de organismos multicelulares são apresentadas. No final da tese, eu discuto algumas das possíveis estratégias para a integração de dados experimentais na análise de modos elementares de fluxo e dou algumas ideias para futuros melhoramentos no cálculo de modos elementares de fluxo com vista a esta integração.

Aknowledgments/ Danksagung/ Agradecimentos

This thesis would not be possible without the contribution of many people that took part in my life.

First, I would like to thank Stefan Schuster for accepting me as PhD student in his group and for guiding my studies in these four exciting years. I am thankful to him for his effort in introducing me to a vast scientific community, for sharing his thoughts and for giving me the freedom to persue my ideas. More importantly, I am grateful for having met a men with a genuine sincerity and friendship.

To Francisco Planes, who introduced me to the wonders of Mixed Integer Programming. For sharing his knowledge with me, for his challenges and for all the entusiasm that he put in every pice of work that we have done together: Muchas gracias y un fuerte abrazo!

I would like to thank Marie-France Sagot and Jorge Carneiro for giving me the opportunity to be part of the 1st edition of the PhD Program in Computational Biology at the Instituto Gulbenkian de Ciência (IGC) and for their support, presence and friendship. I am grateful to the wonderful colleagues that participated in this program, in particular the ones from the 1st edition, for the intense and enriching experience as for the good and bad moments. I thank all the people working at IGC for the stimulating environment and the fantastic conditions for

doing research.

I am thankful to all my colleagues at the BIOINF group from Stefan Schuster. My appreciation words include also all my former colleagues because they made my intregration in Jena very pleasant. I owe a special thanks to Beate Knoke and Jörn Behre for helping me with all sorts of problems from the academic to private life. I would also like to thank Christoph Kaleta for enriching scientific discussions, for nice team work and for introducing me to Java. Ich bin ihm auch dankbar, daß er mir gezeigt hat wie gut die Mischung aus Bier und Cola schmeckt (Diesel)! Ein besonderer Danke geht an alle Mitarbeiter des Lehrstuhls für Bioinformatik sowie Dr. Ina Weiß und Kathrin Schowtka.

I am thankful to the Friedrich-Schiller-Universität Jena for the wonderful research conditions, in particular, for the vast list of software and literature that are available to the students. Many thanks to PD Dr. Peter Dittrich and to Prof. Sebastian Böcker for the inspiring discussions. I would like to thank all the members of the Jena Center for Bioinformatics for the interesting seminars and the chances to present my work there.

I express my gratitude to all the co-authors of manuscripts where I have been involved in, for sharing their experience with me. A particular thanks to Prof. David A. Fell who gave me lectures while I was in the first year of the PhD Program in Computational Biology and with whom I discussed my first ideas about the PhD project. I would also like to thank Prof. John E. Beasley for the sharing his expertise in mathematical modelling.

Agradeço a todos os meus amigos portugueses que conheci em Jena e com quais pude partilhar a saudade e a emoção de ser Português: Artur, Ana, Luís (Vulcão), Helder, Filipa, Daniel e João. Estou também grato á comunidade Luso-Afro-Brasileira por toda a alegria que partilharam comigo. Um abraço especial para: Luís (Vulcão), Silvia e Marcelo Honda, Renata e Michael Gabler, Deia e Fred, Jan, Glaúcio, Turian da Silva, Stefano, Gabriel, Marcelão, Sidnei e Fabio Tuche.

Ich bedanke mich bei allen Freunden, die ich in Jena in den letzten vier Jahren kennengelernt habe. Ein besonderer Grüß geht an die Gruppe Capoeira Ibeca Jena, den Deutsch-Stammtisch und an alle Leute, die jedes Wochenende ob Regen oder Schnee mit mir zusammen Fußbal gespielt haben. Auch einen besonderen Gruß verdienen Valentina, Christoph, Marie und Tristan, Conrad, Declan, Fabio, Alex, Dave, Sebastian, Uli, Thilo, Grit und Sue. Ich werde euch alle sehr vermissen.

Estou eternamente grato aos meus pais pela educação e amor que me deram e pelo esforço que fizeram para que eu pudesse prosseguir os meus estudos. Ao meu irmão por ter aturado todas as minhas asneiras, pela amizade e camaradagem. Quero também agradecer à minha família em geral, por me ter apoiado ao longo de todos estes anos. Como não podia deixar de ser, agradeço a todos os meus amigos que me têm acompanhado ao longo da vida e cuja lista é demasiado extensa para citar aqui.

Last but not least, weil sie sehr wichtig für mich ist seit ich sie das erste Mal am 28. Oktober 2007 gesehen habe, y que desde entonces no salió de mi corazón, quero agradecer à minha namorada Maria pelo amor, carinho, compreensão e dedicação. Diese Arbeit ist auch Deine Arbeit.

Contents

Chapter 1

Introduction

Make everything as simple as possible,
but not simpler.

-Albert Einstein

The concept of a metabolic pathway has been changing since the discovery of the first pathways. In the second half of the 18th century, Louis Pasteur showed that the conversion from sugars into ethanol and lactic acid is carried anaerobically by microorganisms. Moreover, he showed that in the presence of oxygen the growth yield on sugar was up to 20 times greater aerobically than anaerobically (Barnett, 2005). The sequence of enzymatic reactions carrying the anaerobic conversion of sugars into ethanol or lactic acid was only complete in the 1930s with the establishment of glycolysis as a pathway catabolyzing carbohydrates (also named the Embden-Meyerhof-Parnas pathway).

At that time, researchers had no access to isolated enzymes. Instead, they used yeast juice or muscle extracts. The first crystallized enzyme, an urease, was obtain by Sumner (1926), even though these findings were not initially accepted by the scientific community (cf. Sumner, 1937). By 1938, already 10 enzymes were successfully crystallized including one of the enzymes of glycolysis, the alcohol dehydrogenase (Northrop and Herriott, 1938). Moreover, cellular processes such as respiration and fermentation were initially quantified by measuring the rates of gas exchange using the Warburg manometer (Barnett, 2005; Krebs, 1979). The manometer was later replaced by spectrophotometers, these allowing to quantify the amount of reduced pyridine nucleotides (Krebs, 1979).

The first formulation of the main aerobic pathway for glucose catabolism, later named the tricarboxylic acid cycle (TCA cycle; or also known as Krebs cycle), was published in 1937 by Krebs and Johnson (reprinted in Krebs and Johnson, 1980). This initial scheme, describing the oxidation of carbohydrates in pigeon breast muscle, would be subsequently elaborated but the essential aspects remained. The conclusive evidences of this cycle and its functioning were obtained some decades later with the introduction of isotopic labeling techniques (cf. Barnett, 2005).

In the second half of the 20^{th} century, the discovery of the structure and information coding of DNA enabled the development of the recombinant technology. This technology made microorganisms more amenable to manipulations laying down the foundations of a new field known as metabolic engineering (Bailey, 1991). Moreover, as the details on the central carbon metabolism were becoming increasingly clear and the aspects of metabolic regulation were starting to be characterized, biochemists wanted to understand the dynamics of metabolic fluxes. More precisely, their goal was to predict the rate limiting step in a metabolic pathway and therefore, avoiding complex and time consuming genetic manipulations.

The initial steps towards the quantification of metabolism dynamics were carried out in the 1960s by Higgins, who proposed a quantitative expression for the influence of an enzyme on the flux, the control strength (cf. Heinrich *et al.*, 1977; Fell, 1992). One decade later, two groups independently developed a theory to explain flux control, later known as Metabolic Control Analysis (MCA) (Kacser and Burns, 1973; Heinrich and Rapoport, 1974a,b). The summation theorem and the connectivity theorem are two of the most important outcomes of this theory. From the first theorem, one enzyme could not be said to be rate-limiting because there is a proportional relationship between the activity of an enzyme and the pathway flux. The second theorem explains how the enzyme kinetic properties impact the flux control. For further details on MCA see Heinrich and Schuster (1996) or Fell (2003).

The development of MCA can be associated with a primordial stage of the systems biology field, more precisely the systems biology of metabolic networks (Fell, 2005; Westerhoff and Hofmeyr, 2005). Indeed, the exact moment when the field of systems biology was created cannot be precisely defined (Westerhoff and Alberghina, 2005). The main agreement is that systems biology aims at discovering the principles underlying the emergence of functional properties of living organisms by investigating the interactions between the components of cellular networks and by the integration of computational methods with experimental efforts (Westerhoff and Alberghina, 2005; Klipp *et al.*, 2009). With MCA, metabolism was starting to be seen as a network of metabolites interconnected through enzymatic reactions, from which general laws could be extracted and therefore, MCA is also part of systems biology.

One major drawback of kinetic modelling is the lack of kinetic data for the analysis of large metabolic networks. This limitation was not compatible with the need to understand the impact of genetic manipulations carried in metabolic engineering. Thus, a qualitative but systematic study of metabolic networks was required. Seressiotis and Bailey (1986) presented the first algorithm and database for computing metabolic pathways. The algorithm of Seressiotis and Bailey (1986), based on the concepts of artificial intelligence, searched for sequences of enzymes converting a source metabolite to a target metabolite that fulfill the stoichiometric constraints of the reaction network. An initial database containing the description of 70 enzymes and approximately 100 substances (Seressiotis and Bailey, 1986) was later extended to 90 enzymes and 120 substances (Seressiotis and Bailey, 1988). Around the same time, the BRENDA (BRaunschweig ENzyme DAtabase) database was founded with the objective of collecting enzymatic and metabolic information from the literature (Schomburg *et al.*, 2002). BRENDA is nowadays one of the most important databases storing kinetic data from thousands of enzymes under several experimental conditions.

Few years later, Mavrovouniotis *et al.* (1990) presented an improved algorithm for computing metabolic pathways. The main advantage of this algorithm was that intermediate pathways that did not necessarily deal with the source and target compounds, were kept and extended during the search process. These pathways would be combined, in a later stage, with other solutions and hence, assuring that all the possible pathway combinations were explored. The network analyzed by Mavrovouniotis *et al.* (1990) was much larger than the ones used by Seressiotis and Bailey (1988), having more than the double number of reactions, 220 reactions, and around 4 times more metabolites, 400 metabolites.

The methods of Seressiotis and Bailey (1988) and Mavrovouniotis *et al.* (1990) were pioneering in what regards the computation of metabolic pathways. These methods were capable of computing genetically independent pathways fulfilling all stoichiometric constraints of the network. In other words, the support of the reaction set, corresponding to each pathway, is not a proper superset of the support of any other pathway, meaning that each pathway could not be decomposed into a smaller subset without violating the stoichiometric constraints. Nevertheless, the mathematical background of these methods was still very weak.

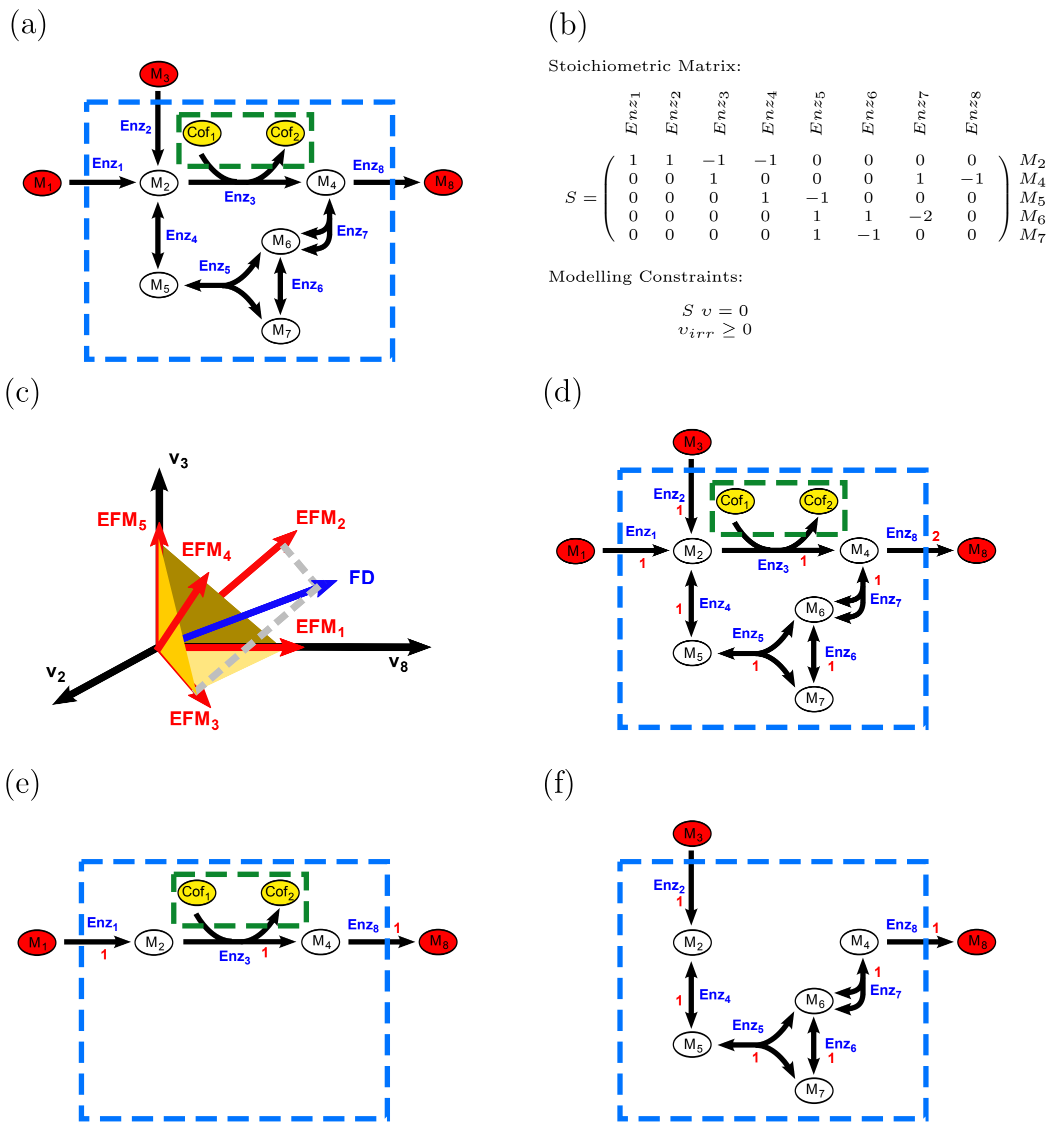
In the beginning of the 1990s, the new field of metabolic pathway analysis was created with the objective of obtaining a mathematical definition of metabolic pathways present in large networks of enzymatic reactions. The study of the structural properties of reaction networks, namely the network invariants (Lautenbach, 1973; Reder, 1988) and extreme currents (Clarke, 1988; Schuster and Schuster, 1993), served as a basis for the development of new concepts like elementary flux modes, extreme pathways and minimal T-invariants, see below. These theoretical frameworks enriched the modelling of metabolism in systems biology.

Given a metabolic system delimited by fixed boundaries, the set of enzymatic reactions that occur in it can be mathematically represented as a matrix, the stoichiometric matrix (Figure 1.1 (a) and (b)). The negative coefficients of the stoichiometric matrix correspond to the amounts of substrate consumed whereas the and positive ones to the amounts of product formed, in each enzymatic reaction. Since the time constants associated with growth are much larger than those associated with individual reaction kinetics, it is assumed that the metabolic system works at steady state. Although, in some cases, this assumption is not justified, in many other cases, it is. For example, when microorganisms are growing in a chemostat.

Due to thermodynamic constraints some reactions are irreversible and consequently, their fluxes can only have a positive value. Additional simplifications to the systems model can be performed in order to reduce the complexity of the problem, such as setting currency metabolites like cofactors to external (i.e., removing them from the stoichiometric matrix). In convex analysis, the solution space of this linear algebraic problem can be represented as a polyhedral cone, also know as a flux cone (Figure 1.1 (c)). If all reactions of the metabolic system are irreversible the flux cone is pointed, whereas in the presence of reversible reactions the cone is non-pointed only if there exists a reversible elementary flux mode (Wagner and Urbanczik, 2005; Larhlimi and Bockmayr, 2009).

According to the definition, an elementary flux mode is a minimal set of enzymatic reactions for which there is a flux distribution that fulfills both constraints, steady state and irreversibility (Schuster and Hilgetag, 1994; Schuster *et al.*, 2002a). The edges (or extreme rays) of the flux cone are elementary flux modes because they cannot be further decomposed (Pfeiffer *et al.*, 1999), for example, the EFMs 1, 3 and 5 in Figure 1.1 (c). There are, however, additional elementary flux modes that lay in the interior of the cone as a result of the convex combination of other elementary flux modes (Wagner and Urbanczik, 2005; Larhlimi and Bockmayr, 2009), for example, EFMs 2 and 4 in Figure 1.1 (c). The biological interpretation of an elementary flux mode is very close to the biochemical concept of a metabolic pathway, which helps to understand the success of this approach in the field of systems biology. The concept of elementary flux modes has been applied in the identification of new metabolic pathways (Poolman *et al.*, 2003). For example, a previously hypothesized pathway involving the glyoxylate shunt, some anapletoric reactions and the TCA cycle (Liao *et al.*, 1996), now known as the phosphoenolpyruvate-glyoxylate cycle (PEP-glyoxylate cycle), was predicted by elementary flux modes analysis (Schuster *et al.*, 1999) and later experimentally identified in hungry *Escherichia coli* (Fischer and Sauer, 2003). Moreover, it has been applied to the study of enzyme deficiencies and regulation of metabolic networks (Stelling *et al.*, 2002; Cakir *et al.*, 2004; Schuster and Kenanov, 2005) and to access structural properties of the metabolic networks such as the robustness and fragility (Stelling *et al.*, 2002; Wilhelm *et al.*, 2004; Behre *et al.*, 2008). Yet, elementary flux mode analysis has also an important role in metabolic engineering, in particular, it has been used in the optimization of biotechnological relevant strains (Trinh *et al.*, 2008; Trinh and Srienc, 2009; Teusink *et al.*, 2009). In the first part of this thesis, I show how a biological question such as the conversion of even-chain fatty acids into sugars, can be modeled using elementary flux mode analysis. This approach together with biochemical knowledge acquired during the last century, provides a fast and comprehensive way of explaining why certain conversions cannot occur in metabolism, solving in a few months a question that took more than half a century to be answered.

The initial algorithms for computing elementary flux modes were based on the Gaussian elimination method with appropriate extensions to comply with irreversibility and non-decomposability (Schuster and Hilgetag, 1994; Pfeiffer *et al.*, 1999). Later, the running time for computing elementary flux modes was decreased with the introduction of the null space algorithm (Wagner, 2004; Urbanczik and Wagner, 2005) followed by a decrease in the memory requirement achieved with the representation of elementary flux modes as bit patterns (Gagneur and Klamt, 2004; Klamt *et al.*, 2005). Currently, the most efficient algorithm to compute elementary flux modes makes use of a new recursive enumeration strategy and of bit pattern trees to speed up the search of subsets (Terzer and Stelling, 2006,



$$S = \begin{array}{c} \\ \left(\begin{array}{cccccccc} Enz1 & Enz2 & Enz3 & Enz4 & Enz5 & Enz6 & Enz7 & Enz8 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -2 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{array} \right) \begin{array}{c} \\ M_2 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \end{array} \end{array}$$

$$S\ v = 0$$
$$v_{irr} \geq 0$$

Figure 1.1: Mathematical representation of a metabolic system through the convex basis. (a) Metabolic systems delimited by boundary (dashed blue lines). Additional assumption considering currency cofactors as external (dashed green lines). (b) Mathematical representation of reaction equations through the stoichiometric matrix and constraints defining the flux cone. Fluxes associated to irreversible reactions (Enz$_1$, Enz$_2$, Enz$_3$ and Enz$_4$) are constraint to the positive orthant. (c) Visualization of mathematical representation of the metabolic system: the solution space in the form of a flux cone; the elementary flux modes - red arrows; a flux distribution - blue arrow. (d) A flux distribution existing in the solution space. This flux distribution is not an elementary flux mode because it can be decomposed in two simpler solutions, elementary flux mode (e) and (f) which cannot be simplified without violating the steady-state constraint.

2008). A list of available software tools for elementary flux mode analysis is presented in Table 1.1. These tools also serve as computational library for other tools in systems biology. Unfortunately, there is not thorough analysis on the complexity of these algorithms, in spite the recent efforts to define the complexity class associated with the enumeration of elementary flux modes, see below.

Tool	Link	Reference
CellNetAnalyser	www.mpi-magdeburg.mpg.de/ projects/cna/cna.html	Klamt *et al.* (2007)
COPASI	www.copasi.org	Hoops *et al.* (2006)
efmtool	www.csb.ethz.ch	Terzer and Stelling (2008)
METATOOL	http://pinguin.biologie.uni-jena.de/bioinformatik/networks/	von Kamp and Schuster (2006)
ScrumPy	http://mudshark.brookes.ac.uk/index .php/Software/ScrumPy	Poolman (2006)
SNA	http://www.bioinformatics.org/ project/?group_id=546	Urbanczik (2006)
YANAsquare	http://yana.bioapps.biozentrum.uni-wuerzburg.de	Schwarz *et al.* (2007)

Table 1.1: Available tools for Elementary Flux Mode Analysis.

All these 'classical' approaches, implemented in the tools listed in Table 1.1, perform the full enumeration of elementary flux modes present in a given metabolic system. This fact becomes problematic when larger metabolic systems are considered because the number of elementary flux modes increases exponentially with the network size (Klamt and Stelling, 2002). Indeed, the combinatorial complexity of metabolic pathways was well explored in the algorithm of Mavrovouniotis *et al.* (1990), to assure the complete enumeration of all metabolic pathways. However, with the increase of the network size this property of metabolic pathways represents a major bottleneck in metabolic pathway analysis.

The concept of extreme pathways is similar to that of elementary flux modes. In the extreme pathways approach the metabolic network is reconfigured. More precisely, reversible reactions inside the metabolic system are decoupled into two irreversible reactions (Schilling *et al.*, 2000; Klamt and Stelling, 2003) whereas reversible reactions that cross the boundary of the system, called exchange reactions, remain as reversible reactions, when applicable. Due to this reconfiguration, extreme pathways correspond to the extreme rays of the reconfigured polyhedral flux cone (Wagner and Urbanczik, 2005). The main advantage of this approach over elementary flux modes is that the number of extreme pathways is smaller than the number of elementary flux modes if the exchange reactions of the system are reversible (Klamt and Stelling, 2003; Papin *et al.*, 2004). However, when combining extreme pathways together some reversible exchange reactions can cancel out, limiting the full evaluation of the network properties (Klamt and Stelling, 2003; Papin *et al.*, 2004). For a detailed comparison between elementary flux modes and extreme pathways see Klamt and Stelling (2003) or Wagner and Urbanczik (2005).

Recently, Acuña *et al.* (2009) have shown that counting the number of elementary flux modes given an input metabolic system is #P-complete (read 'number P-complete'), by reduction to the problem of counting perfect matchings in bipartite graphs. This class is associated with enumerating problems and is designed to reflect the additional difficulty of the enumeration (Valiant, 1979b,a; Garey and Johnson, 2000). Indeed, there have been attempts to predict the number of elementary flux modes (Klamt and Stelling, 2002), or of extreme pathways (Yeung *et al.*, 2007) given an input network, but these values remain rough estimations. Moreover, finding the shortest elementary flux mode is NP-hard[§] (Acuña *et al.*, 2009). Unfortunately, very few is known about the complexity of enumerating all elementary flux modes, besides that the main hurdle is when irreversible reactions are added to the metabolic system (Acuña *et al.*, 2009; Larhlimi and Bockmayr, 2009).

Another way of representing a metabolic system is that of a graph. The simplest graph one can use to describe a metabolic network is the unipartite graph in which nodes correspond to metabolites and reactions to edges connecting two nodes each, or vice versa. These are often called compound graphs in case metabolites are

[§] the initials NP stand for *nondeterministic polynomial* nodes or reaction graphs on the other case (cf. Lacroix *et al.*, 2008). Alternative representations are hypergraphs, where edges (reactions) connect two or more nodes (metabolites) and bipartite graphs, where reactions and metabolites are two different types of nodes (Klamt *et al.*, 2009). In the latter, the edges may be seen as the interaction between metabolites and the enzymes. A particular bipartite graph representation of a metabolic network is done when applying the theory of Petri nets (Petri, 1962) to the study of metabolic networks (Reddy *et al.*, 1993; Hofestädt, 1994). In Petri net analysis, the nodes corresponding to the metabolites are called places and the reaction nodes are the transitions (see Figure 1.2). In the Petri net approach all reversible reactions are decoupled into two irreversible reactions, including exchange reactions (as opposite to extreme pathway approach where only internal reactions are decoupled) and consequently, two different transitions are associated to them. The quantities of each metabolite existing in a given moment correspond to tokens and the distribution of tokens over the places represents the marking of the network. The marking of the network characterizes a certain state of the metabolic system (Figure 1.2). In order to account for the stoichiometric constraints, the arcs (i.e., edges of the graph) connecting a place to a transition have given weights which equal the stoichiometric coefficients of the metabolites in the reaction equations. Thus, the transition can only be fired when the correct proportion of tokens in the pre-places is available and fulfills the amount of tokens required by the arcs. When a transition is fired, the tokens from the pre-places can flow through the transition towards the post-places.

According to the Petri net analysis, a metabolic pathway is defined as a minimal T-invariants (also known minimal-support invariant) (Reddy *et al.*, 1993). In other words, the support of a T-invariant corresponds to the firing count vector of transitions which have to fire, in order to obtain the initial marking of the Petri net again. This support is said to be minimal if it does not contain any other trivial support vector. The computation of the minimal T-invariant requires the use of the incidence matrix (Lautenbach, 1973), which in constraint-based analysis

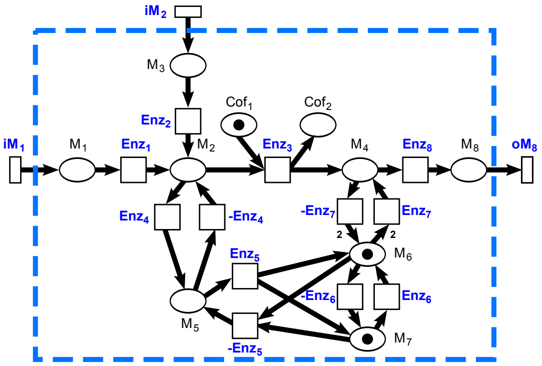


Figure 1.2: Petri net representation of the metabolic system depicted in Figure 1.1 (a) with a given marking. Squares correspond to the transitions, ellipses to places and the dots are the tokens. Artificial transitions without pre-places are used in order to model the environment (iM$_1$,iM$_2$ and oM$_8$).

corresponds to the stoichiometric matrix. Indeed, it can be shown that when all reactions are irreversible the minimal-support invariants correspond to elementary flux modes (Koch *et al.*, 2005).

A new revolution in biology occurred in the middle of the 1990s, giving a big impulse to the holistic thinking of systems biology. For the first time a genome of a free living organism, *Haemophilus influenzae* Rd, was completely sequenced (Fleischmann *et al.*, 1995) followed by several others, in particular the genomes of the model organisms *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996) and *Escherichia coli* (Blattner *et al.*, 1997). Moreover, the data generated by the sequencing technology boosted the development of metabolic databases. The aim in building databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Goto *et al.*, 1997; Ogata *et al.*, 1999) and the EcoCyc (Karp and Mavrovouniotis, 1994; Karp *et al.*, 1996) is to merge this new genetic data with the metabolic data generated during last century and by doing so, obtain further insight in the function of the encoded genes. An additional role of these databases is to collect published experimental data and to structure the metabolic knowledge in a comprehensive manner, where metabolic pathways are at the higher functional level (Kanehisa *et al.*, 2006; Kanehisa and Goto, 2000; Karp, 2001).

With this increase in the amount of genomic and metabolic data available through databases like KEGG, BRENDA and EcoCyc, the reconstruction of larger metabolic networks modelling the entire metabolism of an organism became possible. These networks, so called genome-scale networks, contain the entire metabolism encoded in the genome of a given organism and therefore, can be used to assess the metabolic capabilities of that organism (Edwards and Palsson, 1999). The first genome-scale networks of model organisms *H. influenzae* (Edwards and Palsson, 1999), *E. coli* (Edwards and Palsson, 2000) *S. cerevisiae* (Förster *et al.*, 2003) were published by the turn of the 20th century, following a similar progress to that of the published sequence genomes (Figure 1.3). An important feature of these models is the definition of the biomass equation. This equation is specific for each microorganism and aims at simulating the requirement in precursors, co-factors and energy for the growth of the microorganism (Feist *et al.*, 2007, 2009). The reconstruction process has been reviewed elsewhere (Feist *et al.*, 2009; Ruppin *et al.*, 2010) and there exists also a protocol for that (Thiele and Palsson, 2010).

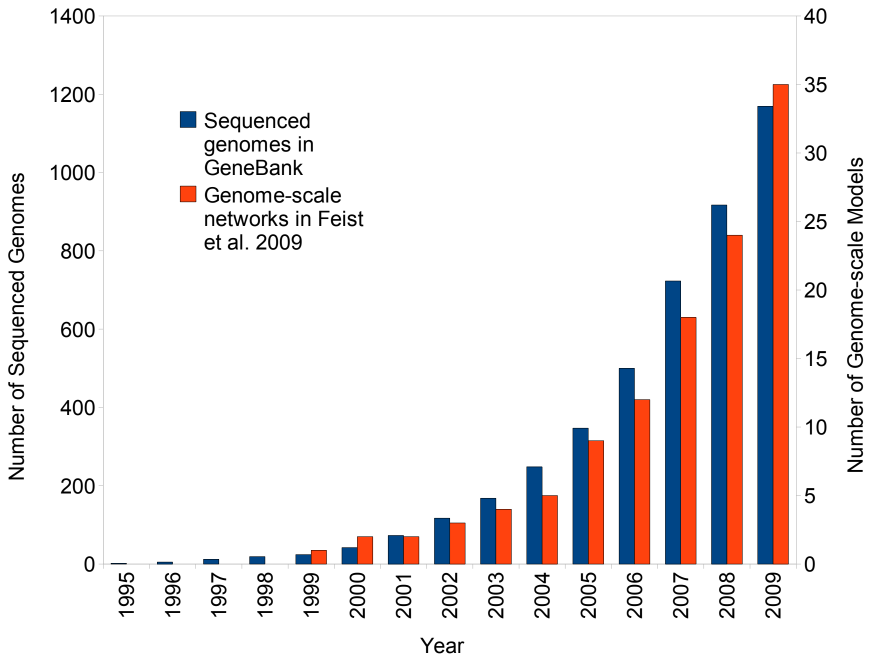


Figure 1.3: Progress of the published genome sequences and reconstruction of genome-scale networks. Data sources: www.genomesonline.org and Feist *et al.* (2009)

Organism	Genes	React.	Metabol.	Ref.
Escherichia coli	1260	2077	1668	Feist *et al.* (2007)
Saccharomyces cerevisiae	750	1149	1061	Duarte *et al.* (2004)
Arabidopsis thaliana	--	1406	1253	Poolman *et al.* (2009)
Homo sapiens	1496	3311	2766	Duarte *et al.* (2007)

Table 1.2: Genome-scale networks of some model organisms and their properties in terms of genes, reactions and metabolites.

The size and complexity of these networks (see Table 1.2) limit their analysis using metabolic pathway analysis methods like elementary flux modes and extreme pathways, as mentioned above. The largest model, in the sense of cell-scale modelling, where these methods have been successfully applied is the human red blood cell (erythrocyte) metabolic network (Wiback and Palsson, 2002; Cakir *et al.*, 2004; Schuster and Kenanov, 2005). However, erythrocytes have a reduced metabolism when compared with that of *E. coli* or yeast. The full enumeration of elementary flux modes in *E. coli* is often performed in a medium-scale network, focusing on the central carbon metabolism (Klamt and Stelling, 2002; Trinh *et al.*, 2008; Terzer and Stelling, 2008). There have been attempts to perform elementary flux mode or extreme pathway analysis in large-scale metabolic networks such as the ones from the human pathogens *Mycoplasma pneumoniae*, *Helicobacter pylori* or *H. influenzae*. In order to achieve that, the complexity of the network was reduced using some of the following strategies: use a small set of external metabolites to allow the computation of pathways consuming and producing only a specific set of metabolites (Price *et al.*, 2002; Terzer and Stelling, 2008); subdivide the metabolic network into smaller subnetworks, analyzing each subsystem separately (Schilling and Palsson, 2000; Schuster *et al.*, 2002b); use simplified representations of pathway charts from KEGG database to compute functional modes (Schwarz *et al.*, 2007). However, such approaches are prone to bias, leading to incorrect predictions (Kaleta *et al.*, 2009b). The second part of the work presented in this thesis, deals exactly with the computation of elementary flux modes in genome-scale metabolic networks. In particular, the methods presented here allow to compute and sample from the solution space, elementary flux modes that are involved in a particular biological process.

The main method supporting the use and further development of genome-scale models is Flux Balance Analysis (FBA) (Edwards and Palsson, 1999, 2000). The first applications of FBA were performed in small models of bacterial and animal metabolism (Papoutsakis and Meyer, 1985; Fell and Small, 1986; Watson, 1986; Majewski and Domach, 1990). The fact that FBA relies on linear programming to solve a given optimization criteria allows the scalability of this method to large-scale networks. In FBA, the same steady state and irreversibility constraints from methods based in convex analysis are used (see Figure 1.1 (b)). Additional constraints limit the reaction fluxes to a physiological admissible value and constraint the uptake/excretion rates of some metabolites (Savinell and Palsson, 1992a). The inclusion of these constraints changes the shape of the solution space to a bounded cone, or polytope (Price *et al.*, 2004). Moreover, the inclusion of uptake/excretion rates allows the integration of experimentally measured fluxes, like the uptake of carbon source and the secretion of by-products which usually can be measured (Savinell and Palsson, 1992b).

Another important aspect of FBA is the objective function because it can be used to explore the metabolic capabilities of the network by focusing on a given physiological state of the cell, such as the optimal biomass production, or maximization of adenosine triphosphate (ATP) production. The definition of the objective function is one of the bottlenecks in FBA because microorganisms can have complex objectives requiring the optimization of more than one function or, in some cases, the physiological state of the cell is better characterized by suboptimal solutions (Schuster *et al.*, 2008; Schuetz *et al.*, 2007; Teusink *et al.*, 2009). These issues are reviewed in more detail in Chapter 5. On the other hand, the solution obtained with FBA corresponds to an optimal flux distribution and not exactly to a metabolic pathway. Indeed, the optimal flux distribution can be a combination of several elementary flux modes or extreme pathways (Wiback *et al.*, 2003). Moreover, there are, in general, several alternative optimal flux distributions for a given objective function (Lee *et al.*, 2000; Mahadevan and Schilling, 2003). The methods presented in this thesis can be extended in order to cope with the alternative optimal flux distributions or with the suboptimal solutions. They can be used to predict which elementary flux modes are likely to take part in them or which are responsible for the suboptimal states.

Nevertheless, FBA has a broad range of applications in metabolic engineering, like strain optimization and growth medium design, in medicine, such as drug target identification or development of tissue specific metabolic networks, as well as in systems biology, like genome annotation refinement and analysis of high-throughput data (cf. Feist and Palsson, 2008; Raman and Chandra, 2009; Gianchandani *et al.*, 2010). Furthermore, FBA is a central framework to many other methods of constraint-based analysis enabling the study of genome-scale metabolic networks in terms of their topologies, the use of these networks for strain optimization or for integration of regulatory data (cf. Price *et al.*, 2004; Feist and Palsson, 2008; Gianchandani *et al.*, 2010). Indeed, these methods together with genome-scale networks are a paradigm in systems biology (Westerhoff and Palsson, 2004).

Recently, more graph theoretical approaches have been developed to derive properties from metabolic networks through the study of its topology. Jeong *et al.* (2000) have shown that metabolic networks have similar topological features to the real-world networks, such as friendship networks or electrical power grids, being quite distinct from random networks. The topology of these networks was named scale-free and the connectivity of metabolites in these networks follows a power-law distribution, meaning that any two nodes in the system can be connected by relatively short paths along existing links. This structural property would help to explain the robustness and error-tolerance of metabolic networks and support the hypothesis that these networks evolved towards the minimization of transition time between metabolic states (Wagner and Fell, 2001). Moreover, the metabolic networks show a hierarchical structure with an embedded modularity (Ravasz *et al.*, 2002). The modularity of metabolic networks could have emerged through copying and reusing existing modules and motifs, generating the scale-free topology where highly connected metabolites, also known as network *hubs* (Table 1.3), play an important role. This hierarchical view of metabolism corroborates with the hypothesis that intermediate metabolism recapitulates the evolution of biochemistry suggested by Morowitz (1992) in which some of these metabolic *hubs* were present in early life forms.

E. coli		*S. cerevisiae*		*A. thaliana*		*H. sapiens*	
Metab.	C [%]	Metab.	C [%]	Metab.	C [%]	Metab.	C [%]
$H^+_{[c]}$	43.65	$H^+_{[c]}$	37.65	O_2	13.73	$H^+_{[c]}$	18.76
$H_2O_{[c]}$	25.36	$H_2O_{[c]}$	19.22	NADPH	13.02	$H_2O_{[c]}$	12.2
$ATP_{[c]}$	16.31	$ATP_{[c]}$	12.35	$NADP^+$	13.02	$ATP_{[c]}$	6.86
$H^+_{[p]}$	13.72	$H^+_{[m]}$	9.13	ATP	10.53	$H^+_{[g]}$	6.83
$Pi_{[c]}$	13.04	$ADP_{[c]}$	8.78	CO_2	10.31	$H^+_{[m]}$	6.28
$ADP_{[c]}$	12.56	$Pi_{[c]}$	8.26	PPi	8.61	$H_2O_{[l]}$	5.59
$H_2O_{[p]}$	8.13	$H^+_{[e]}$	6.7	ADP	8.32	$Pi_{[c]}$	5.44
$PPi_{[c]}$	6.26	$PPi_{[c]}$	6.17	Pi	8.25	$ADP_{[c]}$	5.01
$NAD^+_{[c]}$	5.58	$NADP^+_{[c]}$	6	NAD^+	7.33	$Na^+_{[c]}$	4.17
$NADH_{[c]}$	5.25	$NADPH_{[c]}$	5.83	NADH	7.11	$Na^+_{[e]}$	4.08

Table 1.3: Top 10 of the highly connected metabolites in four genome-scale networks presented in Table 1.2. The first columns corresponds to metabolite abbreviations and column C to the metabolite relative connectivity in percentage (i.e., the percentage of reactions in the genome-scale network where a given metabolite takes part as substrate or product). The metabolite name abbreviation is followed by the compartment information, for compartmentalized models. Compartment nomenclature: [c] - cytoplasm, [e] - extra cellular compartment, [p] - periplasm, [m] - mitochondrion, [g] - golgi apparatus, [l] - lysosome.

The central question of my PhD work is, *How to predict metabolic pathways in large-scale metabolic networks?* We have seen throughout this Introduction that the definition of a metabolic pathway has been changing during the last century. Looking at this definition in biochemistry textbooks we find:

> *Metabolic pathways are series of consecutive enzymatic reactions that produce specific products.*
>
> Voet and Voet, *Biochemistry*, 3^{rd} edition, 2004

> *Metabolism is essentially a linked series of chemical reactions that begins with a particular molecule and converts it into some other molecule or molecules in a carefully defined* [pathway].
>
> Berg, Tymoczko and Stryer, *Biochemistry*, 6^{th} edition, 2007

> *The thousands of enzyme-catalyzed chemical reactions in cells are functionally organized into many different sequences of consecutive reactions called pathways, in which the product of one reaction becomes the reactant in the next.*
>
> Nelson and Cox, *Lehninger: Principles of Biochemistry*, 3^{rd} edition, 2003

We have also seen that in systems biology there are three different mathematical frameworks for defining metabolic pathways (Figure 1.4), namely: graph theoretical methods, constraint-based analysis and kinetic modelling. Note that this division is not straightforward, at least concerning the division between graph-theoretical methods and constraint-based analysis. More precisely, in the Petri net approach the metabolic network is treated as bipartite graph but, the methodology to compute the minimal T-invariants is similar to the one used in constraint-based modeling, requiring also the stoichiometric information in form of an incidence matrix. The ideal method to predict metabolic pathways would be by means of kinetic modelling. However, the lack of data has slowed down the development of large-scale kinetic models. Thus, in this work I will focus mainly on constraint-based analysis, more precisely, elementary flux mode analysis, and a comparison between that and graph theoretical approaches will be carried.

In Chapter 2, I will show how to formulate a metabolic model for studying a specific biochemical problem. Moreover, I will compare the elementary flux mode

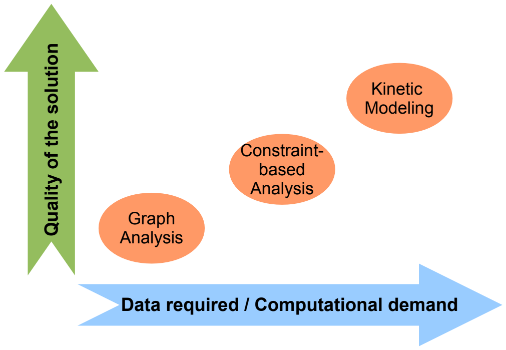

Quality of the solution
Kinetic Modeling
Constraint-based Analysis
Graph Analysis
Data required / Computational demand

Figure 1.4: Schematic positioning of three different approaches for metabolic pathway prediction according to data requirement/computational demands and the quality of the results obtained.

approach with two recently developed methods for metabolic pathway prediction based on graph theory, PathFinding (Croes *et al.*, 2005, 2006) and Pathway Hunter Tool (Rahman *et al.*, 2005). Two benchmark problems illustrating well known biochemical problems were used in this comparison. One of the problems corresponds to an old question in biochemistry dealing with the conversion of fatty acids into sugars, which was answered by Weinman *et al.* (1957). Even though these benchmark problems put in evidence some of the issues associated with graph theoretical methods, also pointed out by Arita (2004), the controversy around their use in the prediction of metabolic pathways remained (Faust *et al.*, 2009a). Thus, a third benchmark problem with applications in medicine and that can be studied at the cell scale by elementary flux modes analysis is presented in the second part of Chapter 2. This benchmark problem shows the role of adenine supply to human erythrocytes and will be analyzed more in detail throughout the Discussion.

In Chapter 3, the K-shortest EFM method is presented. This method allows for the first time the computation of elementary flux modes directly from a genome-scale network. Taking into account the problems mentioned above relative to the full enumeration of elementary flux modes in large-scale networks and to the issues associated with complexity-reduction strategies, this new method marks a milestone in the development of algorithms for the computation of elementary flux modes. The integer linear programming formulation plays an important role in the K-shortest EFM method, allowing to focus only on the solutions of interest avoiding the computation of all elementary flux modes. We first test the K-shortest EFM in a small metabolic network in order to evaluate its capabilities in enumerating all elementary flux modes, or simply a subset of elementary flux modes of interest given an input metabolic system. Then, we explore the potential application of this method in the fields of biotechnology and metabolic engineering, in particular, in the analysis of the pathways leading to the synthesis of the amino acid L-lysine. Lysine is an essential amino acid for humans and is acquired in the diet. Moreover, the industrial production of lysine is around 750.000 tons per year (Wittmann and Becker, 2007).We analyze the lysine biosynthesis pathways in two genome-scale metabolic networks, the curated network of *E. coli* (Feist *et al.*, 2007) and the initial draft of the genome-scale network of *Corynebacterium glutamicum* (Kjeldsen and Nielsen, 2008). The computed elementary flux modes address very well the pathways already described in literature. Moreover, the set of reactions in these elementary flux modes can be subdivided into functional subsets carrying out the catabolism of glucose, the biosynthesis of lysine, the assimilation of ammonium and the balancing of cofactors.

The K-shortest EFM method leads to a paradigm shift in the algorithms computing elementary flux modes. After this major breakthrough, my attention focused on improving the running time of the enumeration process and the sampling of elementary flux modes. This effort resulted in the development of a new and more efficient method, the EFMEvolver, that is presented in Chapter 4. This method combines the genetic algorithm framework with linear programming. Genetic algorithms are often used in optimization problems, in particular for multi-objective optimizations, and their main feature is the efficiency in exploring the solution space (Eiben and Smith, 2003). Furthermore, they have been successfully applied in many areas of biology, for example in protein design (Voigt *et al.*, 2002), strain optimization (Patil *et al.*, 2005), RNA structure prediction (Kashtan *et al.*, 2007) and also for pathway analysis (Boghigian *et al.*, 2010). The linear program formulation allows to efficiently compute a single elementary flux mode given an input network. Consequently, we extend the already broad spectrum of applications of linear optimization in systems biology.

The efficiency of the EFMEvolver is initially tested in a medium-sized metabolic network developed by Behre *et al.* (2008) where all the elementary flux modes can be enumerated. Again, the genome-scale models of *E. coli* and *C. glutamicum*, used in Chapter 3, are important to access the computational power of this new method. The degree of completion of both genome-scale networks is evident from this analysis. While for the network of *C. glutamicum* the typical saturation curve of genetic algorithms is apparent in all simulations, for *E. coli* we are far from computing all the elementary flux modes.

In Chapter 5, a review on the reconstruction of genome-scale metabolic models and on the use of constraint-based analysis to study metabolic network properties is carried. The issues associated with constraint-based analysis are discussed more in detail. Alternative modelling frameworks based in game theory are put forward highlighting the positive contribution of this approach in the study of metabolic robustness of microorganisms. New prospects in constraint-base analysis are delineated, in particular, the application of this modelling tool to study multi-cellular organisms.

Chapter 2

Benchmarking metabolic pathway analysis tools

Recently, graph theoretical methods have been developed to study the topology of large-scale metabolic networks, avoiding the issue of enumerating all possible pathways using convex analysis. In de Figueiredo *et al.* (2008), two recently developed tools for metabolic pathway prediction, PathFinding (Croes *et al.*, 2005, 2006) and Pathway Hunter Tool (Rahman *et al.*, 2005), are compared with METATOOL (von Kamp and Schuster, 2006). Relevant problems in biology are used as benchmarks to access the quality of the solutions produced by these methods. In this work, I developed the metabolic models representing the biological problems and formulated the same problems in terms of a search query given as input to the graph based tools. I performed all the simulation and analysis of the results. Moreover, I was involved in the production of the manuscript.

Due to a mistake by the production office of Bioinformatics, the Figure 3 in de Figueiredo *et al.* (2008) was incorrectly published. Bioinformatics published the complete article again as erratum in the Bioinformatics' first issue of 2009 (de Figueiredo *et al.*, 2009b). Here, I decided to present the original article published in 2008 and an erratum of Figure 3.

In (de Figueiredo *et al.*, 2009c), we express our main concern regarding the use of PathFinding for the prediction of metabolic pathways, given the controversy around the comparison performed in de Figueiredo *et al.* (2008) (Faust *et al.*, 2009a). Additionally, we show that there are other biological problems that can be used as benchmark and we clarify the limitations of elementary-flux mode analysis and of PathFinding. We also give some suggestions for improving the comparison of new pathway prediction methods (de Figueiredo *et al.*, 2009c).

The Supplementary material of de Figueiredo *et al.* (2008) can be found on pages 109 ff.

LF de Figueiredo, S Schuster, C Kaleta and DA Fell. *Bioinformatics*, 24(22):2615-2621, Nov 2008.

BIOINFORMATICS **ORIGINAL PAPER** *Vol. 24 no. 22 2008, pages 2615–2621*
doi:10.1093/bioinformatics/btn500

Systems biology

Can sugars be produced from fatty acids? A test case for pathway analysis tools

Luis F. de Figueiredo[1,*], Stefan Schuster[1,*], Christoph Kaleta[2] and David A. Fell[3]

[1]Department of Bioinformatics, [2]Bio Systems Analysis Group, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany and [3]School of Life Sciences, Oxford Brookes University, Headington, Oxford, OX3 0BP, UK

Received on July 24, 2008; revised and accepted on September 18, 2008

Advance Access publication September 19, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: In recent years, several methods have been proposed for determining metabolic pathways in an automated way based on network topology. The aim of this work is to analyse these methods by tackling a concrete example relevant in biochemistry. It concerns the question whether even-chain fatty acids, being the most important constituents of lipids, can be converted into sugars at steady state. It was proved five decades ago that this conversion using the Krebs cycle is impossible unless the enzymes of the glyoxylate shunt (or alternative bypasses) are present in the system. Using this example, we can compare the various methods in pathway analysis.

Results: Elementary modes analysis (EMA) of a set of enzymes corresponding to the Krebs cycle, glycolysis and gluconeogenesis supports the scientific evidence showing that there is no pathway capable of converting acetyl-CoA to glucose at steady state. This conversion is possible after the addition of isocitrate lyase and malate synthase (forming the glyoxylate shunt) to the system. Dealing with the same example, we compare EMA with two tools based on graph theory available online, PathFinding and Pathway Hunter Tool. These automated network generating tools do not succeed in predicting the conversions known from experiment. They sometimes generate unbalanced paths and reveal problems identifying side metabolites that are not responsible for the carbon net flux. This shows that, for metabolic pathway analysis, it is important to consider the topology (including bimolecular reactions) and stoichiometry of metabolic systems, as is done in EMA.

Contact: ldpf@minet.uni-jena.de; schuster@minet.uni-jena.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

While the conversion of carbohydrates into fatty acids is experimentally well established, the existence of the converse transformation has long been discussed in biochemistry. This question was posed around the turn of the 19th century by Chaveau. While Pflüger stated that fat was the main source of sugar in diabetes, Lusk wrote that this was a figment of imagination (cf. Weinman *et al.*, 1957). This controversy intensified in 1922 with the discovery of insulin and the extended work on diabetes.

In the 1950s, experiments using a new method involving isotopically labelled compounds started to reveal the mechanism by which carbons of fatty acids are incorporated in carbohydrates. Experiments showed that labelled carbons arrived at glucose when the system was supplied with ^{14}C-labelled fatty acids. The Krebs cycle (tricarboxylic acid cycle) seemed to play a key role in this process (Weinman *et al.*, 1957). Nevertheless, these experiments were not conclusive because the Krebs cycle, as other metabolic pathways, does not operate alone and the net synthesis was yet to be proved.

In 1957, the question around the net synthesis of carbohydrates from fatty acids being the most important constituents of lipids started being answered by Weinman *et al.*, who formulated an algebraical treatment of the problem and proved that fatty acids cannot give rise to a net gain of carbohydrate running along the Krebs cycle. The main conclusions from their work was that fatty acids can enter in the metabolite pool of the Krebs cycle but the net synthesis of glucose is due to an influx of other intermediates in the Krebs cycle, such as amino acids or lactic acid (Weinman *et al.*, 1957).

Since fatty acids in living organisms usually contain an even number of carbon atoms with the most common numbers being 16 and 18 (cf. Stryer, 1995), Weinman only analysed that case. Here, we will do the same, by considering acetyl-CoA (AcCoA) as the initial substrate. AcCoA results from the degradation of even-chain fatty acids and ketogenic amino acids (cf. Stryer, 1995). In the case where odd-chain fatty acids occur, such as in some plants and marine organisms (cf. Voet and Voet, 2004), a minor fraction of the products of β-oxidation of these acids is propionyl-CoA, which can, via succinyl-CoA, be converted to pyruvate and, thus, to glucose. Moreover, for both chain lengths, glucose can be produced from glycerol, which is part of phospholipids and triglycerides.

Also in 1957, Kornberg and Madsen (1957) published a paper describing the discovery of the ‘glyoxylate bypass’, an alternative route from isocitrate to malate. The key enzymes in this pathway are isocitrate lyase (formerly called isocitritase), which cleaves isocitrate, and malate synthase (formerly called malate synthetase), which catalyzes the condensation of AcCoA and glyoxylate to malate. This new route enables the conversion of acetate—and

*To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

therefore fatty acids—to carbohydrates with a stoichiometry of 1 mol of oxaloacetate (OAA) per 2 mol of AcCoA.

This discovery reopened the question about the possibility of transforming fatty acids into sugars, though in another perspective. It was connected to the new question of whether the glyoxylate cycle is present in humans. The first experiments showed the presence of the glyoxylate cycle in microbes and plants (Kornberg and Beevers, 1957; Kornberg and Madsen, 1957). Madsen was the first to report that the glyoxylate cycle is not present in animal tissue even under conditions in which one might expect it to occur like hibernating mammals and chick embryos because these must use their fat reservoirs (Madsen, 1958). The only clade of animals where the glyoxylate shunt was detected is that of the nematodes, where a bifunctional malate synthase/isocitrate lyase enzyme occurs (Liu *et al.*, 1995). The question around the presence of the glyoxylate cycle in animal tissues remains open since some authors claim the presence of isocitrate lyase and malate synthase (Davis and Goodman, 1992; Ganguli and Chakraverty, 1961; Goodman *et al.*, 1980; Jones, 1980; Morgunov *et al.*, 2005; Popov *et al.*, 2005) although the coding sequence of these enzymes in humans remains unknown and there is no homology with known sequences. Kondrashov *et al.* (2006) found the sequence of malate synthase, but not isocitrate lyase, in some animals besides nematodes.

Today, there is increased knowledge of biochemical networks, and genome-scale metabolic models have been established. But are we able to really handle such networks? Researchers pay special attention to topological properties of the metabolic model in order to redefine what metabolic pathways are. Recently, several methods have been proposed for determining metabolic pathways in an automated way based on network topology (Beasley and Planes, 2007; Croes *et al.*, 2005, 2006; Rahman *et al.*, 2005; Schuster *et al.*, 1999, 2000). It is of interest to see whether these methods can help answering the question posed in the title of this article and, in particular for didactic purposes in biochemistry, to revisit the study by Weinman *et al.* (1957).

The term ‘elementary flux mode’ refers to a minimal group of enzymes that can operate at steady state with all the irreversible reactions used in the right direction (Schuster *et al.*, 1999, 2000). If only the enzymes belonging to one elementary mode (EM) are operative and, thereafter, one of the enzymes is inhibited, then the remaining enzymes can no longer be operational because the system cannot any longer maintain a steady state. Several software tools were established for computing EMs, for example, METATOOL 5.0 (von Kamp and Schuster, 2006). Elementary modes analysis (EMA) has been applied to various systems (Cakir *et al.*, 2004; Carlson and Srienc, 2004; Poolman *et al.*, 2003; Schwartz *et al.*, 2007; Stelling *et al.*, 2002; Wilhelm *et al.*, 2004). Also the Krebs cycle, glyoxylate shunt and adjacent reactions have been analysed by that method earlier, though not with the objective of the present article (Schuster *et al.*, 1999). A concept related to that of EMs is that of extreme pathways (Schilling *et al.*, 2000). A comparison of the two concepts was made by Klamt and Stelling (2003).

Any stationary flux distribution in the living cell is a linear combination of EMs (Schuster *et al.*, 1999). Therefore, if there is no EM consuming a given substrate or synthesizing a desired product, then we can conclude that there is no stationary flux distribution that would be able to consume that substrate or leading to that product.

Graph theory is another approach to studying metabolic networks based on the concept that these networks can be described as a simple graph (where nodes and edges represent metabolites and reactions, respectively) or as a bipartite graph (where two or more nodes, metabolites, connect to a common node of a second type, representing a reaction/enzyme). While EMA and graph-theoretical analyses of metabolic networks use the same input information, they usually produce different, complementary outputs (which should be consistent, though). A general comparison between the two methods can be found in Planes and Beasley (2008).

Based on graph-theoretical approaches, several computer programs have been presented. Pathway Hunter Tool (PHT; Rahman *et al.*, 2005) and PathFinding (Croes *et al.*, 2005, 2006) are freely available web tools, which can be used to reconstruct and analyse the shortest path connecting two metabolites. PHT uses a fingerprint algorithm to calculate the similarity between two molecules and in this way automatically assigns side metabolites (like ATP, ADP, water). Then a breadth-first-search algorithm calculates the shortest path between the seed and sink metabolites.

The approach underlying PathFinding (Croes *et al.*, 2005, 2006) is based on the connectivity of metabolites which is used to calculate the weight of paths between two metabolites or two reactions. Metabolites with a high connectivity will reduce the score of the path. The reason is that cofactors such as ATP are usually highly connected and should not be considered as intermediates on metabolic paths.

Here, we compare different tools for pathway finding, Metatool, PHT and PathFinding, by applying them to carbon metabolism in view of the question whether sugars can be produced from even-chain fatty acids. In Section 2, the reaction scheme to be analysed will be outlined. In Section 3, the results of the various tools will be presented and compared. The EMA of the lipid–sugar system considerably extends a preliminary analysis presented recently (Schuster and Fell, 2007). A final conclusion will be given in Section 4.

2 METHODS

The system under study is composed of reactions present in the Krebs cycle, which is the pathway for the oxidation of AcCoA and, thus, even-chain fatty acids, and the reactions in glycolysis and gluconeogenesis, responsible for the catabolism and anabolism, respectively, of glucose. The initial model draft was reconstructed on the basis of the human model present in the KEGG database (Aoki-Kinoshita, 2006). The model was refined and completed with some anaplerotic reactions using biochemistry textbooks (Michal, 1999; Nelson and Cox, 2000; Voet and Voet, 2004). The hypothesis of a carbon net flux using amino acids was also tested and external reactions were added to the first model enabling the influx of glutamate, aspartate and alanine. Using the first model as a template, a second model was generated by adding reactions catalyzed by isocitrate lyase and malate synthase to test the hypothesis that the glyoxylate cycle enables a net flux of carbons from AcCoA to α-D-glucose-6-phosphate (G6P), see Figure 1.

For the methods of EMs, the reader is referred to Schuster *et al.* (1999, 2000) and Gagneur and Klamt (2004). For computing EMs, we used the program METATOOL 5.0 (von Kamp and Schuster, 2006), which implements an algorithm proposed by Urbanczik and Wagner (2005). The reaction list of the complete model containing the glyoxylate cycle is represented in the Supplementary Material.

In order to study automated pathway generating tools, we queried PathFinding (Croes *et al.*, 2005, 2006) and PHT (Rahman *et al.*, 2005)

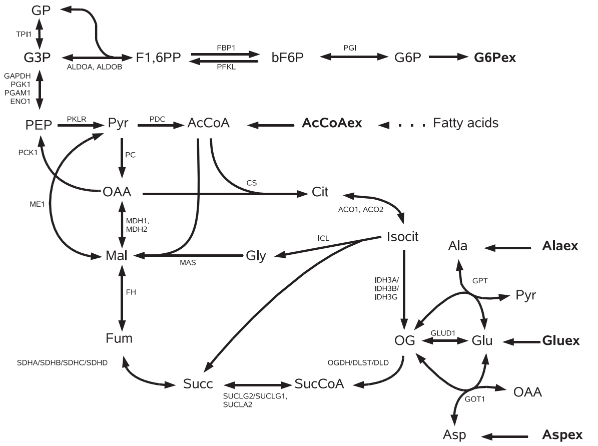


Fig. 1. Complete model of human glycolysis/gluconeogenesis and Krebs cycle, containing reactions of the glyoxylate cycle. For abbreviations of metabolites, see Supplementary Material. Enzyme abbreviations are according to the database ExPASy. When separated by commas or slashes, they correspond to isoenzymes or multi-enzyme complexes, respectively. The external metabolites, that is, the source and sink metabolites are represented in bold-face or are omitted (e.g. the cofactors such as ATP, ADP and NAD). Reactions between PEP and G3P are lumped into one reaction.

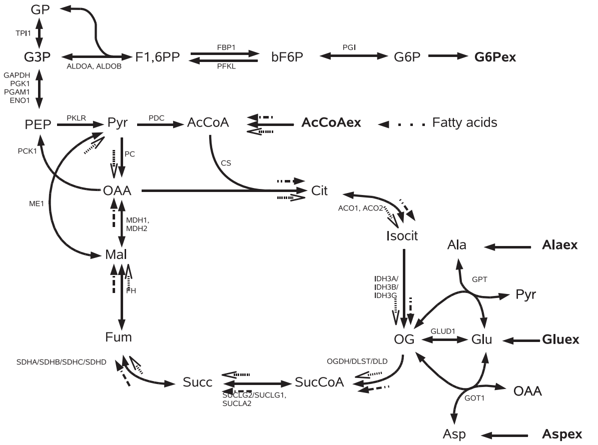


Fig. 2. Two EMs of the model without glyoxylate cycle and no entry of external amino acids, where AcCoA is consumed (empty and full dashed arrows). The mode shown in full dashed arrows is the usual Krebs cycle.

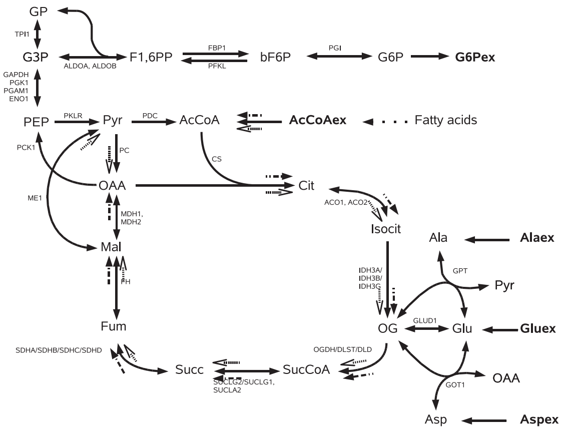


Fig. 3. Two EMs of the model with glyoxylate cycle and no external reaction of amino acids, where AcCoA is consumed (empty and full dashed arrows) and G6P produced. The mode shown in full dashed arrows is the usual glyoxylate cycle.

for a possible connection between AcCoA (KEGG entry C00024) and G6P (KEGG entry C00668).

3 RESULTS AND DISCUSSION

3.1 EMs analysis

The first model containing no glyoxylate cycle, and with no influx of amino acids, resulted in six EMs. None of these produces G6P. Two of these consume AcCoA, go along the Krebs cycle, produce GTP, NADH and CO_2 (Fig. 2). The absence of EMs producing G6P and, thus, of an enzyme set able to synthesize G6P from AcCoA at steady state supports the hypothesis that it is impossible to synthesize glucose from fatty acids using the Krebs cycle and the gluconeogenic reactions only. This can be understood by inspecting Figure 2. To consume 1 mol of AcCoA, 1 mol of OAA is needed. Going around the Krebs cycle, this produces 1 mol of OAA. To produce G6P via PEP, one more mole of OAA would be needed. This cannot be formed at steady state, though. Another explanation is that two carbons enter the Krebs cycle by AcCoA and two leave it in the form of CO_2 (not shown in the Figures). Therefore, no carbon net flux can go to glucose. Nevertheless, if AcCoA is radioactively labeled, some of the labeled carbons flow to G6P because there is a connected route linking AcCoA with G6P and because some carbon atoms are actually transferred along the entire route. For example, if carbon 1 in acetate is labelled, then tracer is detected at carbons 3 and 4 in glucose (Weinman *et al.*, 1957).

Then, we allowed for a carbon influx into the Krebs cycle from an additional source, for example, amino acids because this had also been analysed in Weinman *et al.* (1957). To simulate this, we added external reactions that enable the influx of the glucogenic amino acids glutamate, aspartate or alanine into the system (extended first model). This increased the number of EMs to 18. Among these, five modes connect one of the amino acids each to G6P using at least OAA or 2-oxoglutarate as intermediaries (Supplementary Material). Thus, glucogenic amino acids can really generate a carbon flux towards G6P synthesis. The number of modes using an influx of AcCoA remained the same and none of those modes could synthesize G6P.

The second model contains the glyoxylate cycle, yet no influx from amino acids. This model gives rise to 11 EMs, two of which convert AcCoA to G6P, using isocitrate lyase and malate synthase in the glyoxylate shunt. Moreover, these two modes use part of the Krebs cycle (Fig. 3). The two modes differ in the use of the malic enzyme (ME1) and pyruvate carboxylase (PC) versus malate dehydrogenase (MDH). These results reinforce the hypothesis that the synthesis of glucose from fatty acids through the Krebs cycle is possible in the presence of enzymes from the glyoxylate cycle.

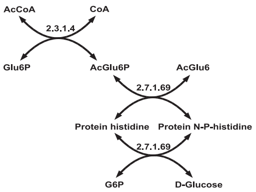


Fig. 4. Best scored path obtained from PathFinding with query 1. It involves part of the phosphotransferase system (PTS) present in bacteria only and is, thus, not relevant for humans. Moreover, no atom is transferred from AcCoA to G6P on that route. Weight = 178.0.

3.2 Graph-theoretical analysis

PathFinding at first glance has one major disadvantage when compared with PHT because it does not have an option to choose between different organisms. Therefore, we filtered the results by choosing, from the output, those enzymes that are present in humans. That information can easily be obtained from KEGG.

We queried PathFinding (April, 2008) to indicate 50 paths leading from AcCoA to G6P, and PathFinding is indeed able to detect that many. Figure 4 shows the path with the best score. However, from the molecular point of view, this path is not valid because it consumes D-glucose to produce G6P and, in the second and third reactions, only orthophosphate is transferred. In fact, not a single atom from AcCoA is transferred to G6P on that route. All of the paths generated for the first query are not present in humans, as results from a check with KEGG data.

Now we tried to find paths present in humans by splitting the path into two, choosing an intermediary metabolite that would connect both paths. The first metabolite chosen was (s)-malate (KEGG entry C00149) which takes part in the Krebs cycle and in the malate–aspartate shuttle as a precursor of OAA. Other metabolites chosen were phosphoenolpyruvate (PEP, KEGG entry C00074) and pyruvate (KEGG entry C00022) which are central metabolites in glycolysis and gluconeogenesis (Table 1). The only query that did not output any result was query 6 (data not shown). The number of the paths (within the output list) present in human is represented in Table 1. The only paths connecting AcCoA to G6P were obtained combining query 4 with query 5, using PEP as intermediary.

Regarding the weight range of the paths, the lower the weight is, the more significant should be the path. For the paths shown in Table 1, the weight range seems to be in an acceptable range because the weights of the two paths resulting from query 7, which correspond to gluconeogenesis, are 210 and 211.

In the results of PathFinding, the connection between different reactions is established by cofactors, such as ITP, IDP, dATP and dADP. However, these compounds are not responsible for the carbon net flux. Figure 5 represents one of the possible connections between AcCoA to G6P when PEP is predefined as an obligatory intermediate. All the other possible paths are combinations between paths of queries 4 and 5. Additionally, it can be noted in Figure 5 that neither of the depicted paths is balanced at steady state.

PHT is easier to handle due to the organism selection option which enables one to choose only paths present in humans. Two other features of this algorithm are ‘Atom Mapper’ (molecular local similarity) and ‘Atom Tracer’ (molecular global similarity), which can be used to improve the results quality though they may not work properly when metabolites do not have a defined structure, like macromolecules. In our analysis, activating both features simultaneously did not produce any paths. For this reason, we tested different combinations of these molecular similarity options (Supplementary Material). In Figure 6, the results obtained with PHT (April, 2008) by switching the ‘Atom Mapper’ *on* and leaving ‘Atom Tracer’ switched *off* and by switching both options *off* are shown.

The results from PHT are better regarding side metabolites because the chemical structure information is used to identify them (Fig. 6). The results of this algorithm were analysed by EMs. However, no such mode could be found, that is, there is no enzyme set capable of converting AcCoA to G6P. The path in Figure 6a resembles gluconeogenesis but is not balanced at steady state. This can clearly be seen in the figure because glycerone phosphate (GP)

Table 1. Queries of PathFinding and retrieved paths present in humans

			Paths present in humans	
Query	Start	Stop	Number	Weight Range
1	AcCoA	G6P	—	(178–203)
2	AcCoA	Mal	—	(85–204)
3	Mal	G6P	43	101
4	AcCoA	PEP	2; 16; 17	199–206
5	PEP	G6P	6; 16; 17; 35	78–87
6	AcCoA	Pyr	—	—
7	Pyr	G6P	44; 47	210–211

Weight ranges of paths not present in humans are given in parentheses. Tool options: Maximum weight = 2500; Maximum metabolic steps = 50; Mode = Weighted; Number of pathways = 50.

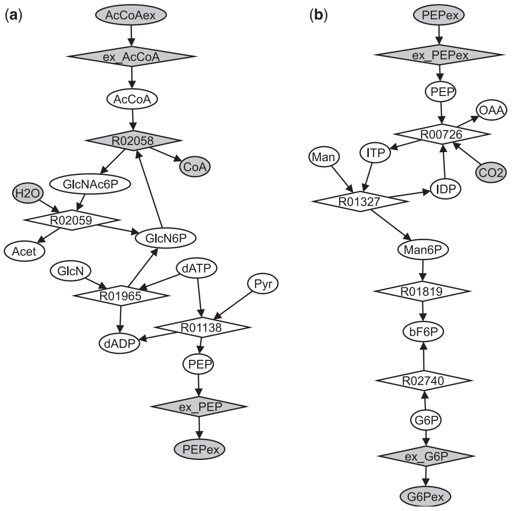


Fig. 5. (**a**) Path 2 from the results of PathFinding query 4; (**b**) Path 16 from the results of PathFinding query 5. Grey ellipses, external metabolites; white ellipses, internal metabolites; grey diamonds, irreversible reactions; white diamonds, reversible reactions. External reactions ex_AcCoA, ex_PEP, ex_G6P added *a posteriori* for EMA.

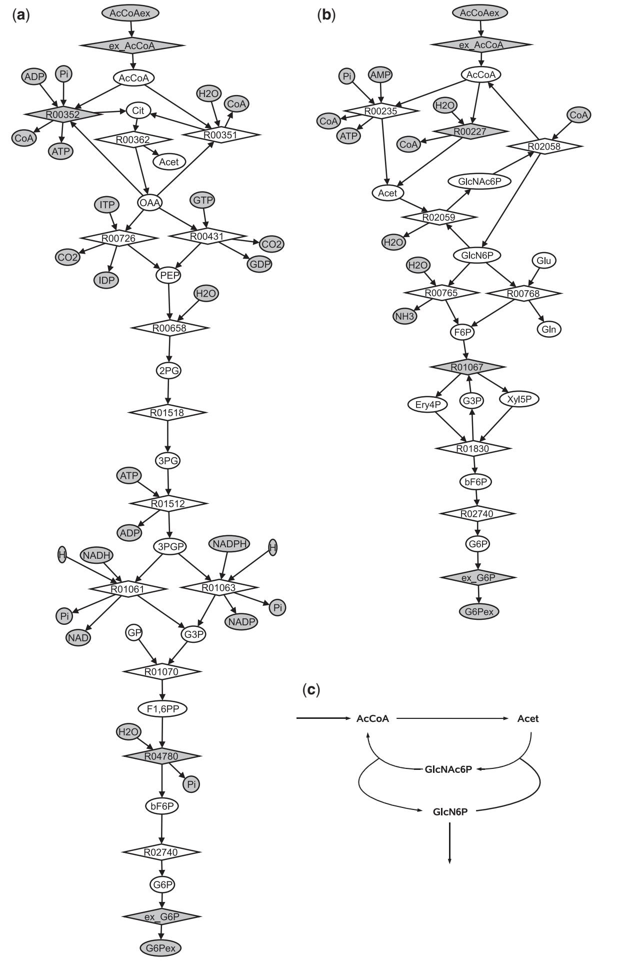


Fig. 6. (**a**) First result obtained by PHT for the conversion of AcCoA to G6P, with the features 'Atom Mapper' and 'Atom Tracer' switched *off*. (**b**) Result for the same query with the features 'Atom Mapper' and 'Atom Tracer' switched *on* and *off*, respectively. Grey ellipses, external metabolites; white ellipses, internal metabolites; grey diamonds, irreversible reactions, white diamonds, reversible reactions. External reactions ex_AcCoA and ex_G6P added *a posteriori* for EMA. (**c**) Scheme of part of the path shown in (b). There is no steady-state conversion of AcCoA to GlcN6P. For simplicity, the external metabolites were removed and reactions R00227 and R00235 were combined into one.

would be consumed in that path but not replenished. This imbalance can be resolved by including triose-phosphate isomerase, which interconverts G3P and GP. However, EMA shows that even in that case, transforming AcCoA to G6P is impossible at steady state because OAA is not balanced.

The path in Figure 6b can be shortened if we take into account the different levels of specificity at which substances are indicated in the KEGG database. In reaction R01067, generic D-fructose 6-phosphate (F6P) is indicated, while in reactions R01830 and R02740, β-D-fructose 6-phosphate (bF6P) is given. Even if reaction R01067 uses both the α and β forms of F6P, the detour via reactions R01067 and R01830 (both of which refer to transketolase, EC 2.2.1.1) is unnecessary, since bF6P spontaneously anomerises to a mixture of α and β F6P (cf. Stryer, 1995). That means, F6P could be converted directly to G6P by phosphogluco-isomerase (R02740). Moreover, from the structure of the path in Figure 6b, it is possible to identify the cycle schematically represented in Figure 6c . Equal amounts of D-glucosamine 6-phosphate (GlcN6P) are produced and consumed in the cycle, so that no drain to synthesize F6P is possible. Therefore, it cannot function as a pathway at steady state because GlcN6P cannot be balanced.

Looking carefully at the metabolite chemical structure in the cycle shown in Figures 6b and c, it can be seen that the atoms from the acetyl group transferred from AcCoA are not present in GlcN6P, which is connected to the rest of the path linking to G6P.

To demonstrate the generality of our results, we have checked another example, which concerns the question whether a pathway connecting G6P with pyruvate in bacteria lacking phosphofructokinase and G6P dehydrogenase exists. Pollack *et al.* (1997) proposed that such a pathway would exist in *Mycoplasma hominis*. Since *M.hominis* is not completely sequenced, its metabolism is not available from KEGG or similar databases. However, the completely sequenced *Bordetella pertussis* is comparable because in its genome, genes for phosphofructokinase and for the enzymes of the oxidative pentose pathway were not found (Armstrong and Gross, 2007). For (whatever) bacteria lacking the above-mentioned enzymes, an EMA had been performed in Schuster *et al.* (1999). It shows that G6P cannot then be converted to G3P at steady state by the glycolysis/pentose phosphate pathway system and, thus, neither to pyruvate, although there is a connected route between them via the non-oxidative pentose phosphate pathway. Interestingly, both PHT and PathFinding output such a route (results given in the Supplementary Material).

4 CONCLUSIONS

It has long been considered that given an input of AcCoA from the breakdown of fatty acids or ketogenic amino acids, it is impossible for animals (except nematodes) to achieve net synthesis of glucose from this precursor by the Krebs cycle and gluconeogenesis. Although ^{14}C-labelled isotopes can pass along this apparent pathway, animals cannot make glucose from two-carbon precursors in substantial amounts at a sustained steady state.

By applying the method of EMs, we have here substantiated this fact and that, when the set of enzymes involved in the glyoxylate shunt are added, the system can synthesize glucose out of AcCoA. Green plants, many bacteria (cf. Stryer, 1995) and fungi (cf. Deacon, 2006) harbour that shunt and are indeed capable of converting AcCoA into glucose at steady state.

We have elaborated on an earlier sketch of a pathway analysis of the lipid-to-sugar transformation (Schuster and Fell, 2007). Among other extensions, we have here studied the possibility of amino acid consumption, have compared several path finding methods and have given a historical review of the subject. It should be noted that we have restricted our analysis to the Krebs cycle (optionally allowing the influx of amino acids), glyoxylate shunt and gluconeogenesis. It cannot be excluded that a conversion of fatty acids into sugars is found when larger (perhaps genome-scale) metabolic networks in animals are studied. Indeed, already Weinman *et al.* (1957)

mentioned the possibility of a conversion via acetone or acetoacetyl-CoA (see below), and this has been supported by subsequent studies (Hetenyi and Ferrarotto, 1985; Reichard *et al.*, 1979).

Moreover, there appear to be various pathways alternative to the glyoxylate shunt or even the Krebs cycle in some bacteria (cf. Ensign, 2006). For example, in *Rhodobacter sphaeroides*, 2 mol of AcCoA can be condensed to acetoacetyl-CoA and converted further to malate and succinate in a series of condensation, rearrangement and carboxylation reactions (cf. Ensign, 2006). In some Archaeans, such as *Ignicoccus hospitalis*, AcCoA can be carboxylated by pyruvate synthase to give pyruvate (Jahn *et al.*, 2007).

The stoichiometry plays an important role in the question under study putting in evidence a molecular constraint. The acetyl group, which is a two-carbon group, enters the Krebs cycle as AcCoA and in two successive reactions, catalyzed by isocitrate dehydrogenase and α-ketoglutarate dehydrogenase, two carbons are converted into carbon dioxide and leave the cycle, although these are not the same atoms (Weinman *et al.*, 1957). Thus, the net carbon balance of an entire turn of the Krebs cycle is zero and the only way to synthesize glucose is to circumvent these decarboxylations or add a carbon source other than AcCoA. Another explanation of the role of the glyoxylate shunt is that it balances synthesis and use of OAA (see Section 3.1). In the absence of the glyoxylate shunt, a net flux of carbons from other carbon sources, like glucogenic amino acids, to G6P via the Krebs cycle is possible, in agreement with the work by Weinman *et al.* (1957).

The results presented above also indicate that automated pathway analysis is difficult. This is due to errors in metabolic databases (Poolman *et al.*, 2006), to ontological problems such as pointed out in Section 3.2 for the α and β forms of F6P, and to combinatorial explosion in large networks (Klamt and Stelling, 2002). Therefore, we advocate that, at the present stage, metabolic networks constructed by extraction from databases should be checked carefully.

The information about network properties obtained by EMA is complementary to that derived from graph theory-based methods because of the high frequency of reactions with more than one substrate or product (e.g. bimolecular reactions) in metabolic networks. Due to the presence of such reactions, connectedness of a network does not necessarily imply a steady-state flow. Metabolic networks are more complicated than graphs in the sense of graph theory. Mathematically, they are hypergraphs.

Several authors have used graph-theoretical concepts to define metabolic pathways (Croes *et al.*, 2005, 2006; Jeong *et al.*, 2000; Ma and Zeng, 2003; Ma *et al.*, 2004; Rahman *et al.*, 2005; Seo *et al.*, 2001). In large-scale networks, these methods are indeed easier to apply than stoichiometric methods. However, paths traced on graphs may not be competent metabolic pathways. This is illustrated by the example of conversion of fatty acids into sugars. To make a distinction between (a) connected routes in the sense of graph theory and (b) pathways that are able to carry a net flux at steady state, a distinction in terminology appears to be necessary and helpful. The terms path and pathway could be used for (a) and (b), respectively (cf. Beasley and Planes, 2007; Planes and Beasley, 2008). Routes detected by graph theory are of interest, for example, for the flow of radioactive tracer.

We here critically examined two tools for finding paths, PathFinding and PHT. They did succeed in finding paths connecting AcCoA to glucose. However, none of them is a biochemically relevant pathway. Though the paths generated by these algorithms are connected they cannot, at steady state, synthesize G6P out of AcCoA and some do not even realize an overall transfer of carbon atoms. This example illustrates that if only the connectedness of the graph is considered and the stoichiometric constraints are neglected, then it is likely that non-functional pathways will be postulated. Another example is monosaccharide metabolism in *M. hominis* and *B. pertussis*, for which graph-theoretical methods again predict invalid pathways from G6P to pyruvate.

Another drawback of the graph-theoretical approaches mentioned above (methods using bipartite graphs excepted) is that cycles cannot be easily obtained because they search for linear paths that connect metabolite A to metabolite B not taking into account metabolites that are not synthesized by the path. Nevertheless, as the paths found in the results of PathFinding and PHT (see Section 3.2) show, certain types of cycles can be obtained. One type can occur where there is more than one reaction synthesizing the same product using the same substrate (like the reaction converting 3PGP into G3P in Fig. 6a). Another type can be obtained when one of the substrates in a path (such as GlcNAc6P in Fig. 6b) occurs as a product of a reaction further down in the path. As observed in the above results of the programs PHT and PathFinding, it is not possible to obtain a non-trivial cycle or cyclic pathways like the Krebs cycle using these algorithms, probably also due to the fact that they search for the shortest pathway or the pathway with the lowest weight. It is a well-known biochemical fact that complex metabolisms involve cyclic pathways, such as the Krebs cycle or the urea cycle. Therefore, algorithms for detecting them are useful.

One option for using graph-theoretical methods also for detecting pathways is to use the theory of Petri nets, which are bipartite graphs. Metabolites and reactions are then represented by two different types of nodes (cf. Koch *et al.*, 2005; Zevedei-Oancea and Schuster, 2003). Another option (used here) is to choose an algebraic treatment such as in EMA, which properly takes into account stoichiometry. The problem of combinatorial explosion in large networks could be solved by using linear programming approaches, by which only specific pathways are computed (Beasley and Planes, 2007; Feist and Palsson, 2008; Fell and Small, 1986). However, a fully automated solution cannot easily be achieved by such approaches either because the proper definition of side metabolites is context dependent.

ACKNOWLEDGEMENTS

The authors would like to thank three anonymous referees for very helpful comments. The groups in which L.F.F., S.S. and C.K. work belong to the Jena Centre for Bioinformatics (JCB). The scientific atmosphere in the JCB has significantly stimulated this work.

Funding: Portuguese entities: Fundação Calouste Gulbenkian, FCT and Siemens SA Portugal (PhD grant SFRH/BD/32961/2006 to L.F.F.). Biotechnology and Biosciences Research Council (BB/E00203X/1 to D.A.F.).

Conflict of Interest: none declared.

REFERENCES

Aoki-Kinoshita,K.F. (2006) Overview of KEGG applications to omics-related research. *J. Pesticide Sci.*, **31**, 296–299.

Armstrong,S.K. and Gross,R. (2007) Primary metabolism and physiology of *Bordetella* species. In Locht,C. (ed.) *Bordetella: Molecular Microbiology*, ch. 8. 1 edn. Horizon Bioscience, Norwich, pp. 165–190.

Beasley,J.E. and Planes,F.J. (2007) Recovering metabolic pathways via optimization. *Bioinformatics*, **23**, 92–98.

Cakir,T. *et al*. (2004) Metabolic pathway analysis of enzyme-deficient human red blood cells. *BioSystems*, **78**, 49–67.

Carlson,R. and Srienc,F. (2004) Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: identification of reactions. *Biotechnol. Bioeng.*, **85**, 1–19.

Croes,D. *et al*. (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res.*, **33**(Web Server issue), W326–W330.

Croes,D. *et al*. (2006) Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.*, **356**, 222–236.

Davis,W.L. and Goodman,D.B. (1992) Evidence for the glyoxylate cycle in human liver. *Anat. Rec.*, **234**, 461–468.

Deacon,J. (ed.) (2006) *Fungal Biology*. 4th edn. Blackwell Publishing, Oxford.

Ensign, S.A. (2006) Revisiting the glyoxylate cycle: alternate pathways for microbial acetate assimilation. *Mol. Microbiol.*, **61**, 274–276.

Feist,A.M. and Palsson,B.O. (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.*, **26**, 659–667.

Fell,D.A. and Small,J.R. (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.*, **238**, 781–786.

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

Ganguli,N. and Chakraverty,K. (1961) Evidence for malic synthetase in animal tissues. *J. Am. Chem. Soc.*, **83**, 2581–2583.

Goodman,D.B. *et al*. (1980) Glyoxylate cycle in toad urinary bladder: possible stimulation by aldosterone. *Proc. Natl Acad. Sci. USA*, **77**, 1521–1525.

Hetenyi,G. and Ferrarotto,C. (1985) Gluconeogenesis from acetone in starved rats. *Biochem. J.*, **231**, 151–155.

Jahn,U. *et al*. (2007) Insights into the autotrophic CO_2 fixation pathway of the archaeon *Ignicoccus hospitalis*: comprehensive analysis of the central carbon metabolism. *J. Bacteriol.*, **189**, 4108–4119.

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Jones,C.T. (1980) Is there a gloxylate cycle in the liver of the fetal guinea pig? *Biochem. Biophys. Res. Commun.*, **95**, 849–856.

Klamt,S. and Stelling,J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.*, **29**, 233–236.

Klamt,S. and Stelling,J. (2003) Two approaches for metabolic pathway analysis? *Trends Biotechnol.*, **21**, 64–69.

Koch,I. *et al*. (2005) Application of Ketri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, **21**, 1219–1226.

Kondrashov,F.A. *et al*. (2006) Evolution of glyoxylate cycle enzymes in metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol. Direct*, **1**, 31.

Kornberg,H.L. and Beevers,H. (1957) The glyoxylate cycle as a stage in the conversion of fat to carbohydrate in castor beans. *Biochim. Biophys. Acta*, **26**, 531–537.

Kornberg,H.L. and Madsen,N.B. (1957) Synthesis of C4-dicarboxylic acids from acetate by a glyoxylate bypass of the tricarboxylic acid cycle. *Biochim. Biophys. Acta*, **24**, 651–653.

Liu,F. *et al*. (1995) Bifunctional glyoxylate cycle protein of *Caenorhabditis elegans*: a developmentally regulated protein of intestine and muscle. *Dev. Biol.*, **169**, 399–414.

Ma,H. and Zeng,A.-P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**, 270–277.

Ma,H.-W. *et al*. (2004) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, **20**, 1870–1876.

Madsen,N.B. (1958) Test for isocitritase and malate synthetase in animal tissues. *Biochim. Biophys. Acta*, **27**, 199–201.

Michal,G. (1999) *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. 1st edn. Spektrum Akademischer Verlag, Heidelberg, Berlin.

Morgunov,I.G. *et al*. (2005) Evidence of the glyoxylate cycle in the liver of newborn rats. *Med. Sci. Monit.*, **11**, BR57–BR60.

Nelson,D.L. and Cox,M.M. (2000) *Lehninger Principles of Biochemistry*. 3rd edn. W. H. Freeman, New York.

Planes,F.J. and Beasley,J.E. (2008) A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Brief. Bioinformatics*, **9**, 422–436.

Pollack,J.D. *et al*. (1997) The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.*, **23**, 269–354.

Poolman,M.G. *et al*. (2003) Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *Eur. J. Biochem.*, **270**, 430–439.

Poolman,M.G. *et al*. (2006) Challenges to be faced in the reconstruction of metabolic networks from public databases. *IEE Proc. Syst. Biol.*, **153**, 379–384.

Popov,V.N. *et al*. (2005) Comparative analysis of the glyoxylate cycle clue enzyme isocitrate lyases from organisms of different systemic groups. *J. Evol. Biochem. Physiol.*, **41**, 507–513.

Rahman,S.A. *et al*. (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, **21**, 1189–1193.

Reichard,G.A. *et al*. (1979) Plasma acetone metabolism in the fasting human. *J. Clin. Invest.*, **63**, 619–626.

Schilling,C.H. *et al*. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229–248.

Schuster,S. and Fell,D. (2007) Modelling and simulating metabolic networks. In Lengauer,T. (ed.) *Bioinformatics: From Genomes to Therapies*. Vol. 2. Wiley-VCH, Weinheim, pp. 755–806.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Schuster,S. *et al*. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Schwartz,J.-M. *et al*. (2007) Observing metabolic functions at the genome scale. *Genome Biol.*, **8**, R123.

Seo,H. *et al*. (2001) Graph-theoretical identification of pathways for biochemical reactions. *Biotechnol. Lett.*, **23**, 1551–1557.

Stelling,J. *et al*. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.

Stryer,L. (1995) *Biochemistry*. 4th edn. W.H. Freeman and Company, New York.

Urbanczik,R. and Wagner,C. (2005) An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, **21**, 1203–1210.

Voet,D. and Voet,J. (2004) *Biochemistry*. 3rd edn. Wiley, USA.

von Kamp,A. and Schuster,S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22**, 1930–1931.

Weinman,E.O. *et al*. (1957) Conversion of fatty acids to carbohydrate: application of isotopes to this problem and role of the Krebs cycle as a synthetic pathway. *Physiol. Rev.*, **37**, 252–272.

Wilhelm,T. *et al*. (2004) Analysis of structural robustness of metabolic networks. *Syst. Biol. (Stevenage)*, **1**, 114–120.

Zevedei-Oancea,I. and Schuster,S. (2003) Topological analysis of metabolic networks based on Petri Net theory. *In Silico Biol.*, **3**, 323–345.

Erratum for Figure 3

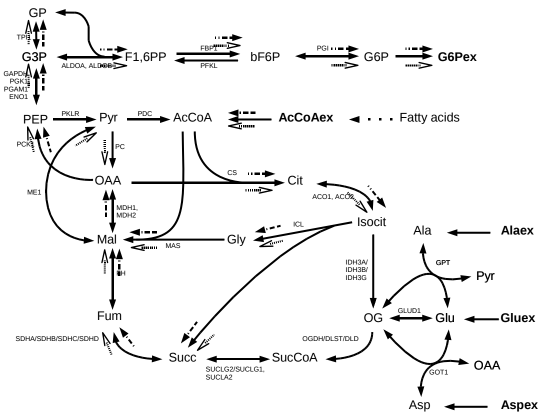


Fig. 3. Two EMs of the model with glyoxylate cycle and no external reaction of amino acids, where AcCoA is consumed (empty and full dashed arrows) and G6P produced. The mode shown in full dashed arrows is the usual glyoxylate cycle.

LF de Figueiredo, S Schuster, C Kaleta and DA Fell. *Bioinformatics*, 25(24):3330-3331, Dec 2009.

BIOINFORMATICS LETTER TO THE EDITOR
Vol. 25 no. 24 2009, pages 3330–3331
doi:10.1093/bioinformatics/btp591

Systems biology

Response to comment on 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'

Luis F. de Figueiredo[1,*], Stefan Schuster[1,*], Christoph Kaleta[1] and David A. Fell[2]

[1]Department of Bioinformatics, Friedrich-Schiller University Jena, Ernst-Abbe-Platz 2,07743 Jena, Germany and [2]School of Life Sciences, Oxford Brookes University, Headington, Oxford, OX3 0BP, UK

Received on August 18, 2009; revised on October 2, 2009; accepted on October 8, 2009

Advance Access publication October 13, 2009

Associate Editor: Alfonso Valencia

Contact: Luis.deFigueiredo@uni-jena.de; Stefan.Schu@uni-jena.de

The main points of criticism by Faust *et al.* (2009) concerning the work in de Figueiredo *et al.* (2009a) are the following: the cases presented are biased; different tools should be supplied with the same input networks; the number of study cases should be representative; and tools should be evaluated by neutral assessors.

We think that the models in de Figueiredo *et al.* (2009a) illustrate concrete biological problems that are very well documented. In particular, the conversion of fatty acids into sugars was not so trivial to be answered and is historically relevant (cf. Weinman *et al.*, 1957). Today, many biochemistry textbooks dedicate at least one paragraph explaining why there is no net conversion of acetyl-CoA to glucose via the tricarboxylic acid cycle in vertebrates (cf. Nelson and Cox, 2000; Stryer, 1995; Voet and Voet, 2004). Thus, new tools for metabolic pathway prediction have to be able to answer these problems correctly.

Analysing the same network used in (de Figueiredo *et al.*, 2009a) by Path Finding (Croes *et al.*, 2006), paths converting acetyl-CoA into glucose are computed, even though this is impossible for humans in the network studied. Thus, the critique of Faust *et al.* (2009) does not bring anything new to what was discussed in de Figueiredo *et al.* (2009a).

It is often said that the size of the input network limits the computation of elementary flux modes (EFMs; Faust *et al.*, 2009; Papin *et al.*, 2003, 2004). Indeed, the enumeration of all EFMs in genome-scale models with the existing methods is difficult (Klamt *et al.*, 2007; Schwarz *et al.*, 2005; Terzer and Stelling, 2008; von Kamp and Schuster, 2006). However, there are approaches to compute at least a subset of EFMs in such models (Acuña *et al.*, 2009; de Figueiredo *et al.*, 2009b; Kaleta *et al.*, 2009), for example, the shortest EFMs (de Figueiredo *et al.*, 2009b).

Regarding the number of test cases presented in (de Figueiredo *et al.*, 2009a), we are sure they are representative of the issue that is discussed in that article. More cases exist, for example, the conversion of hypoxanthine into ATP in human erythrocytes, for which EFM analysis can be performed at the cell level (Schuster and Kenanov, 2005).

Of course, a comparison between tools is preferably made by neutral assessors. However, it is usual in bioinformatics that authors who have established a new tool compare their method with others (Klamt *et al.*, 2007; Urbanczik and Wagner, 2005; Wagner and Urbanczik, 2005). In addition, one co-author (C.K.) has written an article where EFMs are compared with chemical organizations (Kaleta *et al.*, 2006).

In the letter by Faust *et al.* (2009), it is argued that an incorrect definition of internal metabolites in EFM analysis, can generate wrong pathway predictions. The study of any biochemical system requires the definition of the system's boundary (see, e.g. Schilling and Palsson, 1998). Thus, the definition of internal and external metabolites in EFM analysis is nothing more than the definition of the boundary conditions found in many other modelling methods.

Faust *et al.* (2009) say that the steady-state constraint is not always an appropriate assumption. Although, in some cases, this assumption is not justified, in many other cases, it is. Accordingly, it is used in many approaches such as Metabolic Control and Flux Balance Analyses. Faust *et al.* (2009) cite the work of Teusink *et al.* (2000) to support their statement. However, that work shows that the experimental system does reach a steady state. Additionally, we do not think that the study summarized in Table 1 of Faust *et al.* (2009) is exhaustive enough with respect to EFMs (cf. Trinh *et al.*, 2009, for a review).

Nevertheless, the communication from Faust *et al.* (2009) raises an important point concerning the comparison of tools for metabolic pathway prediction. It is suggested to follow a CASP-like protocol to evaluate the methods for metabolic pathway prediction and that this task should be performed by an independent committee.

The validation process of some of the new tools for metabolic pathway prediction has been performed using the pathway information present in metabolic pathway databases (Blum and Kohlbacher, 2008; Croes *et al.*, 2006). A CASP-like protocol, to be developed, has to take into account the fact that these databases contain errors (Likić, 2006; Poolman *et al.*, 2006). On the other hand, only very well-documented pathways are stored in these databases. Many pathways are missing due to the lack of information or simply due to the fact that this classification is performed manually. Furthermore, these databases do not contain all the functional modes of a pathway within a metabolic network, e.g. all the five functional modes of the pentose phosphate pathway in conjunction with glycolysis (Schuster *et al.*, 2000; Stryer, 1995). Moreover, EFM analysis has been successful in predicting relevant, hitherto unknown pathways, for example, the catabolic PEP-glyoxylate cycle in *Escherichia coli* (Schuster *et al.*, 1999), which was later found in

[*]To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

experiment (Fischer and Sauer, 2003) and that, to our knowledge, is not present in KEGG nor in MetaCyc (Caspi *et al.*, 2008; Kanehisa *et al.*, 2008). In conclusion, it is a challenge to represent the combinatorial multitude of biochemical pathways in metabolic databases (Sauer, 2006).

Funding: PhD grant SFRH/BD/32961/2006 (to L.F.F.); by the following Portuguese entities: Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (to L.F.F.); Biotechnology and Biosciences Research Council (BB/E00203X/1 to D.A.F.); German Ministry for Research and Education (BMBF) within the Forsys Partner initiative (to C.K.).

Conflict of Interest: none declared.

REFERENCES

Acuña,V. *et al*. (2009) Modes and cuts in metabolic networks: complexity and algorithms. *Biosystems*, **95**, 51–60.

Blum,T. and Kohlbacher,O. (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J. Comput. Biol.*, **15**, 565–576.

Caspi,R. *et al*. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–D631.

Croes,D. *et al*. (2006) Inferring meaningful pathways in weighted metabolic networks. *J. Mol. Biol.*, **356**, 222–236.

de Figueiredo,L.F. *et al*. (2009a) Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, **25**, 152–158.

de Figueiredo,L.F. *et al*. (2009b) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, [Epub ahead of print, doi:10.1093/bioinformatics/btp564, September 30].

Faust,K. *et al*. (2009) In response to "can sugars be produced from fatty acids? a test case for pathway analysis tools". *Bioinformatics*, [Epub ahead of print, doi: 10.1093/bioinformatics/btp557, September 23].

Fischer,E. and Sauer,U. (2003) A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*. *J. Biol. Chem.*, **278**, 46446–46451.

Kaleta,C. *et al*. (2006) Analyzing molecular reaction networks: from pathways to chemical organizations. *Mol. Biotechnol.*, **34**, 117–123.

Kaleta,C. *et al*. (2009) Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.*, **19**, 1872–1883.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Klamt,S. *et al*. (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst. Biol.*, **1**, 2.

Likić,V.A. (2006) Databases of metabolic pathways. *Biochem. Mol. Biol. Educ.*, **34**, 408–412.

Nelson,D.L. and Cox,M.M. (2000) *Lehninger Principles of Biochemistry*. 3rd edn. W. H. Freeman, New York.

Papin,J.A. *et al*. (2003) Metabolic pathways in the post-genome era. *Trends Biochem. Sci.*, **28**, 250–258.

Papin,J. *et al*. (2004) Comparison of network-based pathway analysis methods. *Trends Biotechnol.*, **22**, 400–405.

Poolman,M.G. *et al*. (2006) Challenges to be faced in the reconstruction of metabolic networks from public databases. *IEE Proc. Syst. Biol.*, **153**, 379–384.

Sauer,U. (2006) Metabolic networks in motion: ^{13}C-based flux analysis. *Mol. Syst. Biol.*, **2**, 62.

Schilling,C.H. and Palsson,B.Ø. (1998) The underlying pathway structure of biochemical reaction networks. *Proc. Natl Acad. Sci. USA*, **95**, 4193–4198.

Schuster,S. and Kenanov,D. (2005) Adenine and adenosine salvage pathways in erythrocytes and the role of *S*-adenosylhomocysteine hydrolase. A theoretical study using elementary flux modes. *FEBS J.*, **272**, 5278–5290.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Schuster,S. *et al*. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Schwarz,R. *et al*. (2005) YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, **6**, 135.

Stryer,L. (1995) *Biochemistry*. 4th edn. W.H. Freeman and Company, New York.

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

Teusink,B. *et al*. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.*, **267**, 5313–5329.

Trinh,C.T. *et al*. (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl. Microbiol. Biotechnol.*, **81**, 813–826.

Urbanczik,R. and Wagner,C. (2005) An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, **21**, 1203–1210.

Voet,D. and Voet,J. (2004) *Biochemistry*. 3rd edn. Wiley, Hoboken, USA.

von Kamp,A. and Schuster,S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22**, 1930–1931.

Wagner,C. and Urbanczik,R. (2005) The geometry of the flux cone of a metabolic network. *Biophys. J.*, **89**, 3837–3845.

Weinman,E.O. *et al*. (1957) Conversion of fatty acids to carbohydrate: application of isotopes to this problem and role of the Krebs cycle as a synthetic pathway. *Physiol. Rev.*, **37**, 252–272.

Chapter 3

First approach to genome-scale metabolic networks

In the previous chapter we have seen that stoichiometry is important in the prediction of metabolic pathways but on the other hand, there is no algorithm capable of computing elementary flux modes in genome-scale networks. In de Figueiredo *et al.* (2009a), we present an integer linear program that computes the K-shortest elementary flux modes in genome-scale metabolic networks. This mathematical model enables to focus on a subset of elementary flux modes of interest, producing or consuming a given metabolite, avoiding the full enumeration of all the elementary flux modes. Taking advantage of this feature, we analyze the 10-shortest elementary flux modes producing a biotechnological relevant amino acid, lysine, in two genome-scale metabolic networks. In this work, I implemented the mathematical model, performed all the simulations and supported the improvement of the model. I was also involved in the production of the manuscript, more precisely, the presentation and discussion of the results.

The Supplementary material of de Figueiredo *et al.* (2009a) can be found on pages 124 ff. Furthermore, the source code for a tool computing elementary flux modes using the K-shortest EFM method is in the Supplementary material, on page 144.

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 23 2009, pages 3158–3165
doi:10.1093/bioinformatics/btp564

Systems biology

Computing the shortest elementary flux modes in genome-scale metabolic networks

Luis F. de Figueiredo[1,2], Adam Podhorski[3], Angel Rubio[3], Christoph Kaleta[1], John E. Beasley[4], Stefan Schuster[1] and Francisco J. Planes[3,*]

[1]Friedrich-Schiller-University Jena, 07743 Jena, Germany, [2]PhD Program in Computational Biology, Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal, [3]CEIT and TECNUN, University of Navarra, 20016 San Sebastián, Spain and [4]Brunel University, Uxbridge, UB8 3PH, UK

Received on May 11, 2009; revised on September 10, 2009; accepted on September 25, 2009

Advance Access publication September 30, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Elementary flux modes (EFMs) represent a key concept to analyze metabolic networks from a pathway-oriented perspective. In spite of considerable work in this field, the computation of the full set of elementary flux modes in large-scale metabolic networks still constitutes a challenging issue due to its underlying combinatorial complexity.

Results: In this article, we illustrate that the full set of EFMs can be enumerated in increasing order of number of reactions via integer linear programming. In this light, we present a novel procedure to efficiently determine the K-shortest EFMs in large-scale metabolic networks. Our method was applied to find the K-shortest EFMs that produce lysine in the genome-scale metabolic networks of *Escherichia coli* and *Corynebacterium glutamicum*. A detailed analysis of the biological significance of the K-shortest EFMs was conducted, finding that glucose catabolism, ammonium assimilation, lysine anabolism and cofactor balancing were correctly predicted. The work presented here represents an important step forward in the analysis and computation of EFMs for large-scale metabolic networks, where traditional methods fail for networks of even moderate size.

Contact: fplanes@tecnun.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In recent years, different approaches have been proposed to investigate the structure of complex metabolic networks (Price *et al.*, 2004). In particular, elementary flux modes (EFMs) have attracted increasing interest. An EFM is defined as a minimal set of enzymes that operates at steady state with all irreversible reactions used in the appropriate direction (Schuster and Hilgetag, 1994; Schuster *et al.*, 2000). An analogous concept in Petri net theory is provided by the minimal T-invariants (Koch *et al.*, 2005). The relevance of EFMs for various applications has been recently reviewed (Trinh *et al.*, 2009). EFM analysis has proved useful in elucidating novel metabolic pathways in addition to textbook knowledge, e.g. a new catabolic pathway that degrades glucose via the glyoxylate shunt (Fischer and Sauer, 2003; Liao *et al.*, 1996; Schuster *et al.*, 1999). Several software packages for computing EFMs have been developed, e.g. METATOOL (von Kamp and Schuster, 2006), CellNetAnalyzer (Klamt *et al.*, 2007), YANAsquare (Schwarz *et al.*, 2007) and efmtool (Terzer and Stelling, 2008). However, EFM analysis suffers from an important drawback: the number of EFMs grows exponentially with network size (Klamt and Stelling, 2002). For instance, more than two million EFMs have been reported for the metabolic network describing the central metabolism in *Escherichia coli*, which contains 110 reactions (Gagneur and Klamt, 2004). Despite a number of attempts to cope with such complexity (Dandekar *et al.*, 2003; Klamt *et al.*, 2005; Schuster *et al.*, 2002; Terzer and Stelling, 2008; Teusink *et al.*, 2006), computing the full set of EFMs in large metabolic networks still constitutes a challenging issue.

Based on the work of Beasley and Planes (2007), we show here that the full set of EFMs can be enumerated via integer linear programming. Technically, our approach produces EFMs in increasing order of number of reactions by solving a sequence of discrete optimization problems. Thus, it is promising to start with the shortest, second shortest, etc., overall called K-shortest EFMs. The 'K-shortest' concept has been previously used in the context of graph theory and paths (see, for illustration, Planes and Beasley, 2009), but not in the context of EFMs. Acuña *et al.* (2009) have recently suggested that finding short EFMs should become interesting if size is considered a relevant criterion. Also, in Mavrouniotis *et al.* (1990), biochemical pathways (not EFMs) are obtained in increasing length order.

Detection of K-shortest EFMs is of interest for several biological applications. Experimentally, it is expensive and laborious to overexpress a large number of enzymes. On the other hand, since the highest increase in pathway flux is achieved if all enzymes (Kacser and Acerenza, 1993) or (at least) a considerable number of enzymes in a pathway (Fell and Thomas, 1995; Niederberger *et al.*, 1992) are overexpressed, shorter pathways are better suited as a target for genetic manipulation. Moreover, shorter pathways can carry higher fluxes (Meléndez-Hevia *et al.*, 1994; Pfeiffer and Bonhoeffer, 2004).

The use of integer linear optimization makes our procedure more flexible than previous approaches found in the literature

[*]To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

(Schilling *et al.*, 2000; Schuster *et al.*, 2000), which require the computation of the full set of EFMs before any further analysis can be performed. Instead, our method allows us to directly explore the K-shortest EFMs related to a particular problem of interest, e.g. the K-shortest EFMs that consume/produce a particular metabolite.

In order to illustrate the applicability of our approach, we here analyse the K-shortest EFMs producing lysine in two different genome-scale metabolic networks, *E.coli* K-12 MG1655 (Feist *et al.*, 2007) and *Corynebacterium glutamicum* ATCC 13032 (Kjeldsen and Nielsen, 2009). Lysine is one of the essential amino acids in humans and is also used as supplement in animal feeds. The industrial production of lysine has a long history in biotechnology (Tosaka *et al.*, 1983; Wendisch *et al.*, 2006). Studying the production of lysine has been essential for the rational design of optimized strains. Nowadays, *C.glutamicum* is the organism of choice for lysine overproduction due to the higher yields obtained with it. The capability for producing lysine has been previously examined from a pathway oriented perspective (de Graaf, 2000; Mavrovouniotis *et al.*, 1990; Schuster *et al.*, 2007). However, these studies were not conducted at the genome-scale. Therefore, the results presented here extend these studies to a larger scale.

2 METHODS

The mathematical model proposed below formulates the task of finding EFMs as a sequence of optimization problems. Our method starts from the basis that the flux mode involving the minimum number of reactions must be elementary. We here refer to it as the shortest EFM. Accordingly, we first define the constraints and the function (objective) to be optimized that allows us the calculation of the shortest EFM. Based on this optimization model, we then show how to calculate the K-shortest EFMs. Finally, extensions of the K-shortest to other problems of interest are presented.

We mean here by 1-shortest EFM, the EFM containing the minimum number of reactions; 2-shortest EFM, the EFM containing the second minimum number of reactions, etc. We may have multiple EFMs containing the same minimum number of reactions. If this occurs, they are counted separately with different K values. The enumeration order of equally long EFMs depends on the actual implementation of the mathematical model and the solving procedure.

As noted above, EFMs are defined as minimal sets of enzymes in steady state (Schuster *et al.*, 2000). The meaning of 'minimal' in the definition of EFMs refers to the non-decomposability condition, i.e. the addition of an enzyme would turn the EFM into non-elementary. In contrast, we here refer the 1-shortest EFM as to the EFM that contains the (global) minimum number of enzymes.

2.1 Shortest EFM

Assume we have a metabolic network that comprises R reactions and C compounds. Here we decompose reversible reactions into two opposing reaction steps. Thus, we can regard all fluxes as taking positive values. Let s_{cr} be the stoichiometric coefficient associated with compound c ($c=1,\dots,C$) in reaction r ($r=1,\dots,R$). As usual in the literature (Schilling *et al.*, 2000; Schuster and Hilgetag, 1994), substrates and products have negative and positive stoichiometric coefficients, respectively. The matrix containing all these coefficients is called the stoichiometric matrix.

A zero-one (binary integer) variable is assigned to each reaction, namely $z_r = 1$ if reaction r ($r=1,\dots,R$) is active in the EFM, 0 otherwise. In addition, each reaction has an associated non-negative (integer) flux t_r. As we are studying structural properties of metabolic networks, it is appropriate to use integer fluxes. If the coefficients of the stoichiometric matrix (s_{cr}) take integer values, as it is assumed here and in many other approaches such as Petri net theory (Koch *et al.*, 2005), then the relative fluxes carried by EFMs can also be described using integer values. In addition, our computational experience reveals that the K-shortest method is more expensive when fluxes are allowed to be non-integer.

For the optimization model we need constraints relating the reaction variables z_r and t_r:

$$t_r \le M z_r \qquad r=1,\dots,R \tag{1}$$

$$z_r \le t_r \qquad r=1,\dots,R \tag{2}$$

Equation (1) ensures that no flux traverses a reaction r if $z_r = 0$. Equation (2) guarantees that t_r is non-zero if $z_r = 1$. Note here that in the case a reaction r is active ($z_r = 1$), its associated (integer) flux value t_r can take any value from the interval $[1, M]$, M being a large constant value. This does not constitute an issue if M is a sufficiently large value.

In our model, reversible reactions are decomposed into two irreversible reactions, and therefore, we define the set B=$\{(\alpha,\beta)|$ reaction α and reaction β are the reverse of each other, $\alpha<\beta\}$.

$$z_\alpha + z_\beta \le 1 \qquad \forall\,(\alpha,\beta)\in B \tag{3}$$

Equation (3) ensures that a reaction and its reverse do not appear in an EFM.

The steady-state condition is critical for the definition of EFMs and it is formulated as

$$\sum_{r=1}^{R} s_{cr} t_r = 0 \qquad \forall c \in I \tag{4}$$

where I is the set of internal compounds. As opposed to internal compounds, external compounds are excluded from being balanced, because they are exchange metabolites between the outside and the system under study or they belong to metabolic pools whose concentration is assumed constant. They typically represent consumed substrates, excreted products and cofactors. We denote the set of external compounds by E.

In order to avoid the trivial solution ($z_r = t_r = 0$, r=1, ..., R), we require that at least one reaction is active:

$$\sum_{r=1}^{R} z_r \ge 1 \tag{5}$$

Equations (1–5) define the flux modes solution space for a particular metabolic network. In order to calculate the shortest EFM, we minimize the number of reactions:

$$\text{minimize} \sum_{r=1}^{R} z_r \tag{6}$$

As noted above, EFMs cannot be decomposed into smaller entities without violating the steady-state assumption, Equation (4). This is referred as to the non-decomposability (elementary) condition (Schuster and Hilgetag, 1994). In essence, this condition implies that no subset of reactions of an EFM can perform at steady state. We ensure that the non-decomposability condition is satisfied by minimizing the number of active reactions involved in the solution flux mode. Clearly, the flux mode involving the minimum number of reactions will be non-decomposable.

2.2 *K*-shortest EFMs

The mathematical optimization model given above [objective function (6) subject to Equations (1)–(5)], once solved, allows us to obtain the shortest EFM. In order to find the K-shortest EFM, we need to add further constraints to eliminate the $(K-1)$-shortest EFMs from the set of solutions. To illustrate this, suppose we are interested in finding the 2-shortest EFM. Let Z_r^1be the binary solution associated with the shortest EFM, where Z_r^1 equals to 1 if reaction r is active, 0 otherwise. We need to eliminate the shortest EFM from the set of solutions. To do this we add the following constraint to our previous formulation:

$$\sum_{r=1}^{R} Z_r^1 z_r \le \left(\sum_{r=1}^{R} Z_r^1\right) - 1 \tag{7}$$

The left-hand side of Equation (7) determines the number of reaction variables in the current solution that were active in the 1-shortest EFM solution. The right-hand side is the number of reactions that were active in the 1-shortest EFM less one. The inequality states that the number of active reactions repeating from the 1-shortest EFM should be less by at least one than the total number of active reactions in that EFM. This ensures that, once we solve our model, the new solution found does not contain the shortest EFM. This also guarantees that the shortest EFM can never occur as a part of any other flux mode. In essence, we remove the shortest EFM from the solution space. In the general case, the $K-1$ shortest EFM solution is eliminated before the K-th solution is computed and clearly the optimization problem for the K-th shortest EFM accumulates constraints from all $(1,\dots,K-1)$ previous solutions, i.e. in order to find the K-shortest EFM, we need to include EFM elimination constraints related to the first $(K-1)$ shortest EFMs:

$$\sum_{r=1}^{R} Z_r^k z_r \leq \left(\sum_{r=1}^{R} Z_r^k\right) - 1 \qquad k=1,\dots,K-1 \tag{8}$$

where Z_r^k is the binary solution for the k-shortest EFM.

Note here that the K-shortest EFMs described above are also elementary. For an indirect proof, suppose that the K-shortest EFM (once solved) is not elementary, i.e. it contains a subset of reactions satisfying Equations (1–5) and (8). Since we are constructing EFMs in increasing order of the number of reactions they contain, we must have encountered the EFM corresponding to this subset before. However, then we would have added a constraint, as described in Equation (8), preventing it from ever appearing as a subset in future EFMs. So it cannot in that case ever be found as part of the K-shortest EFM, which contradicts the original assumption. Thus, every EFM we find must be elementary.

2.3 Extensions to K-shortest EFMs

Our procedure can be applied to enumerate all EFMs, namely by constructing them one by one. This is not particularly efficient for small-scale metabolic networks when compared with existing methods. The main advantage of our mathematical optimization model is that, by adding new constraints, special subsets of EFMs (of particular biomedical or biotechnological interest) can be found without having to first compute all EFMs as is the case in existing methods (Klamt *et al.*, 2005; Schilling *et al.*, 2000; Schuster *et al.*, 2000; Terzer and Stelling, 2008). Below, we present some of these constraints that can be easily added to our formulation.

Genome-scale metabolic networks are typically compartmentalized models, in the simplest case containing the extracellular compartment and cytosol. We assume that metabolites in the extracellular compartment can be taken up or secreted as by-products, therefore these metabolites can be set to be external. We denote U the set of extracellular metabolites defining the growth medium. In the case an extracellular metabolite c is not included in the medium set, we need to avoid this compound to be consumed. Equation (9) describes how this constraint is incorporated into our model.

$$\sum_{r=1}^{R} s_{cr} t_r \geq 0 \qquad \forall c \in E, c \notin U \tag{9}$$

We may also need to find the K-shortest EFMs that produce a particular external compound, μ. To do so, we need to add the following constraint:

$$\sum_{r=1}^{R} s_{\mu r} t_r \geq 1 \tag{10}$$

This can be easily reformulated if we want an external compound μ to be used as substrate, as observed in Equation (11).

$$\sum_{r=1}^{R} s_{\mu r} t_r \leq -1 \tag{11}$$

Note here that Equation (5) can be dropped from the formulation if we include Equations (10) or (11), as both already require at least one compound to be produced or consumed, respectively, hence at least one reaction must be active. In addition, the non-decomposability condition is not guaranteed when more than one constraint based on Equations (10) or (11) is included in the formulation. For example, if we apply constraint (10) for metabolites μ_1 and μ_2, i.e. finding solutions to our model that produces μ_1 and μ_2, then we might obtain solutions containing two EFMs, namely one producing μ_1 and another producing μ_2. For this reason, in this article, we restrict our analysis to EFMs forced to produce/consume one metabolite. Equation (9) does not alter the non-decomposability condition.

2.4 Integer programming

Our mathematical optimization model given above for computing the K-shortest EFMs [objective function (6) subject to Equations (1–5) plus elimination constraints (8) and perhaps constraints (9–11)] is an integer linear program. Algorithmically such programs are solved by linear programming based tree search (Pardalos and Resende, 2002). Various free and commercial software tools are available to perform this task. We used ILOG CPLEX®.

3 RESULTS

We applied our method to three different metabolic networks. Firstly, we examined a well-known metabolic network that contains the tricarboxylic acid (TCA) cycle and some adjacent reactions (Schuster *et al.*, 1999). Since this metabolic network is of moderate size, the full set of EFMs can be obtained using classic methods (Schuster *et al.*, 1999). We used it as a benchmark to validate the capabilities of our method. Then, we applied our method to study the production of lysine in two different genome-scale metabolic networks, *E.coli* K-12 MG1655 (Feist *et al.*, 2007) and *C.glutamicum* ATCC 13032 (Kjeldsen and Nielsen, 2009). Details of the three metabolic networks can be found in the Supplementary Material.

3.1 TCA cycle network

For the TCA cycle network, our method correctly enumerated, in increasing order of number of reactions, all 16 EFMs previously determined in Schuster *et al.* (1999). Details on the 16 EFMs are shown in Table 1. The shortest EFM contains two reactions, which are catalyzed by enzymes Pck and Ppc. The 2-shortest EFM also has two reactions. The 16-shortest EFM involves 13 reactions. These results confirm the applicability of our method.

We compared the computation time of our method with METATOOL (version 5.1) for this particular small network. Our method turned out to be less efficient than METATOOL, though both methods take <1 s (data not shown). However, as will be shown below, our method is particularly suitable for large-scale metabolic networks, where classical methods for EFMs computation are not applicable.

In addition, we extended the analysis by calculating the subset of EFMs that produces succinyl-CoA (SucCoAxt). This is done by incorporating a constraint based on Equation (10) for SucCoAxt into the K-shortest EFMs formulation. Our method directly enumerated the six EFMs producing SucCoAxt without having to first compute the full set of EFMs, as typically done by METATOOL and classic methods (Table 1).

Table 1. Full set of EFMs in the TCA cycle metabolic network

K	*L*	Enzyme set	SCA
1	2	Pck; Ppc	–
2	2	Pps; Pyk	–
3	5	AlaCon; Eno; Gdh; IlvE_AvtA; Pyk	–
4	5	AspC; AspCon; Eno; Gdh; Ppc	–
5	5	AspA; AspC; Fum; Gdh; Mdh	–
6	7	Eno; Ppc; SucCoAcon; -Fum; -Mdh; -Sdh; -SucCD	1
7	8	AspA; AspC; Eno; Gdh; Ppc; SucCoAcon; -Sdh; -SucCD	2
8	9	AceEF; Acn; 2 Eno; GltA; Icd; Ppc; Pyk; SucAB; SucCoAcon	3
9	9	AceEF; Acn; 2 Eno; Gdh; GltA; GluCon; Icd; Ppc; Pyk	–
10	10	2 AceEF; Acn; 2 Eno; GltA; Icl; Mas; Mdh; 2 Pyk; SucCoAcon; -SucCD	4
11	11	AceEF; Acn; Eno; Fum; GltA; Icd; Mdh; Pyk; Sdh; SucAB; SucCD	–
12	11	2 AceEF; Acn; Eno; Fum; GltA; Icl; Mas; 2 Mdh; Pck; 2 Pyk; Sdh	–
13	12	2 AceEF; Acn; 3 Eno; GltA; Icl; Mas; Ppc; 2 Pyk; 2 SucCoAcon; -Fum; -Sdh; -2 SucCD	5
14	13	3 AceEF; 2 Acn; 3 Eno; Fum; 2 GltA; Icd; Icl; Mas; 2 Mdh; 3 Pyk; Sdh; SucAB; SucCoAcon	6
15	13	3 AceEF; 2 Acn; 3 Eno; Fum; Gdh; 2 GltA; GluCon; Icd; Icl; Mas; 2 Mdh; 3 Pyk; Sdh	–
16	13	2 AceEF; Acn; AspC; AspCon; 2 Eno; Fum; Gdh; GltA; Icl; Mas; 2 Mdh; 2 Pyk; Sdh	–

K: the order by which EFMs are computed; *L*: the number of reactions in each EFM; SCA—order by which EFMs producing SucCoAxt are computed. Reversible reactions active in the opposite direction have a minus sign before the flux value.

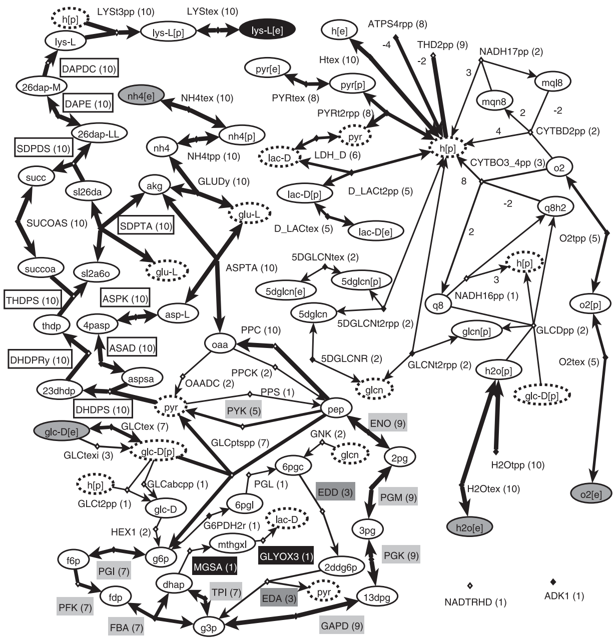


Fig. 1. Merged representation of the 10-shortest EFMs producing lysine in *E.coli* when cofactors are set as internal metabolites. Ellipses represent metabolites and arrows reactions. Stoichiometric coefficients higher than one are represented next to the edge linking the respective metabolite. Dashed ellipses are duplicated metabolite nodes, light grey ellipses are medium metabolites and the black ellipse is the target metabolite. Numbers in brackets after enzyme abbreviations correspond to the number of EFMs where these are present. Thickness of the arrows is proportional to this number. Boxed enzyme abbreviations represent the lysine biosynthetic pathway (Cohen and Saint-Girons 1987, Wittmann and Becker, 2007), enzyme abbreviations in light grey, in dark grey and black correspond to glycolysis, the Entner–Doudoroff pathway and the methylglyoxal bypass, respectively. The following metabolite nodes in the cytosolic compartment were removed from the representation for better visualization: atp, adp, amp, nad, nadh, nadp, nadph, h, coa, h2o, pi, co2. Note here that abbreviations are the same as in the original network (see Feist *et al*., 2007). Thus, reactions involving only these removed metabolites may seem disconnected from the sub-network when they are actually connected, e.g. NADTRHD.

3.2 Genome-scale metabolic networks

We calculated the *K*-shortest EFMs that produce lysine in the genome-scale metabolic networks of *E.coli* and *C.glutamicum* with $K = 10$. These metabolic networks differ in the number of reactions and metabolites, as well as in the level of accuracy. During the computation of 10-shortest EFMs some errors in the *C.glutamicum* network were identified. In particular, an error in reaction dapB was responsible for a null lysine net synthesis. More details as to errors can be found in the Supplementary Material.

The *E.coli* network is larger than the *C.glutamicum* network. For this reason, the *E.coli* metabolic network represents a greater challenge in the computation of 10-shortest EFMs. Our method successfully computed them, though the difference in the computation time is significant (see Supplementary Material). We used glucose and ammonium as carbon and nitrogen sources, respectively, for both metabolic networks. See Supplementary Material for exact definition of the medium set, *U*. A sufficiently large *M* value is needed to ensure that no EFM information is lost. We conducted experimentation for different *M* values (see Supplementary Material) and selected $M = 10\,000$, since no change in the *K*-shortest EFMs solution was found with respect to smaller *M* values. This selected value is similar to that proposed in previous studies (Kjeldsen and Nielsen, 2009; Vallino and Stephanopoulos, 1993).

We first applied our mathematical model to the metabolic network of *E.coli*. Figure 1 shows a merged representation of the 10-shortest EFMs producing lysine in *E.coli*. The shortest EFMs are mainly fermentation modes and therefore, they require higher fluxes on glucose catabolism (see Supplementary Material for more information about the fluxes and the reaction sets). The combinatorial effect seen in EFM analysis can be immediately observed. This is particularly apparent for transport reactions. For example, there are two different reactions for the uptake of glucose (glc-D) from the extracellular compartment to the periplasm, specifically GLCtex and GLCtexi. Thus, there will be at least two EFMs among the 10-shortest EFMs that differ only in the use of one of these two reactions while the rest of the enzyme set remains the same. Such combinatorial features can also be found in the other *K*-shortest EFMs.

A detailed analysis of Figure 1 reveals that there are three major pathways for glucose catabolism: glycolysis, the Entner–Doudoroff (ED) pathway and the methylglyoxal bypass. Glycolysis provides higher quantities of ATP but does not produce any NADPH and therefore the periplasmic NAD(P) transhydrogenase, THD2pp,

LF de Figueiredo, A Podhorski, A Rubio, C Kaleta, JE Beasley, S Schuster and FJ Planes. *Bioinformatics*, 25(23):3158-3165, Dec 2009.

is required to reduce NADP by oxidizing NADH. The ED pathway can use two different precursors of 6-phospho-D-gluconate (6pgc), namely, 6-phospho-D-glucono-1,5-lactone (6pgl) and D-gluconate (glcn). In case 6pgl is used as precursor the oxidative part of the pentose phosphate (PP) pathway produces NADPH and therefore, the THD2pp is not required in this mode, in contrast with the rest of EFMs. When the methylglyoxal bypass is used there is a very low ATP yield from glucose catabolism and therefore, the ATP synthase, ATPS4rpp, has a higher flux when compared with the other modes. It should be noted that this pathway, though possible, is very unlikely to be the main catabolic route of glucose due to the toxicity of methylglyoxal (Subedi *et al*., 2008).

In *E.coli*, ammonium assimilation can be carried out via the glutamine synthetase/glutamate synthase (GLNS/GLUSy) cycle or exclusively using glutamate dehydrogenase (GLUDy). The GLNS/GLUSy cycle constitutes the main ammonium assimilation route even for growth conditions with high extracellular ammonium content (Yuan *et al*., 2006). In the 10-shortest EFMs, the assimilation of ammonium is however conducted by GLUDy, which involves fewer steps and consumes less ATP. The other route would appear for EFMs containing 40 reactions.

In addition, it is well-known that *E.coli* has only one pathway for lysine biosynthesis using aspartate and pyruvate as precursors (Cohen and Saint-Girons, 1987; Wittmann and Becker, 2007). This is also observed in the left upper corner in Figure 1, where the thickness of the involved arrows is maximal, i.e. they appear in all 10-shortest EFMs.

On the right-hand side of Figure 1, there are many reactions around the periplasmatic proton node, h[p]. These reactions are mainly involved in the establishment of a proton gradient so that ATP and NADPH can be produced. We assumed that cofactors are buffered in the metabolic network and set them as external metabolites. We repeated our K-shortest procedure ($K = 10$) and found that the shortest EFM involves 27 reactions, as opposed to the case described above where the shortest EFMs involved 38 reactions.

In Figure 2, there are no EFMs producing by-products such as lactate or pyruvate. The main reason is that there is no need of fermentative modes or other modes producing cofactors in small reaction steps and with high fluxes, since cofactors are now external metabolites. The catabolism of glucose in Figure 2 is again accomplished by the same three pathways: glycolysis, the ED pathway and the methylglyoxal bypass. Combinations of these three pathways are also found in the 10-shortest EFMs, e.g. in the 7-shortest EFM, the ED pathway is combined with the triose phosphate part of glycolysis, while in the 10-shortest EFM the ED pathway is combined with the methylglyoxal bypass. There is a detour to the classical glycolysis described in textbooks, via dihydroxyacetone (dha). This detour has been recently hypothesized by van Winden *et al.* (2003). However, the use of dha as intermediate is questionable due to its toxicity and possible conversion to methylglyoxal (Molin *et al*., 2003; Subedi *et al*., 2008).

The results also show that, with glycolysis as single catabolic pathway, it is possible to produce one mole of lysine per mole of glucose consumed, consuming four moles of NADPH and one mole of ATP and producing two moles of NADH. Thus, from a molecule containing six carbon atoms, glucose, it is possible to produce another six-carbon molecule, lysine, requiring two NADPH for ammonium assimilation, plus two NADPH and one ATP for the intermediate metabolites inter-conversion. However, due to the carboxylation and decarboxylation reactions, this 1:1 conversion cannot be deduced directly from the number of carbons.

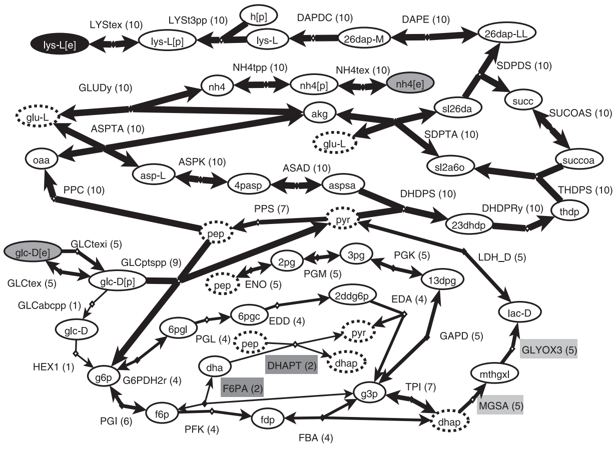


Fig. 2. Merged representation of the 10-shortest EFMs producing lysine in *E.coli* when cofactors are set as external metabolites. Enzyme abbreviations in light grey and dark grey represent the methylglyoxal bypass and a detour of the classical glycolysis over dha, respectively. The following metabolite nodes in the cytosolic compartment were removed from the representation for better visualization: atp, adp, amp, nad, nadh, nadp, nadph, h, coa, h2o, pi, co2. Note here that abbreviations are the same as in the original network (see Feist *et al*., 2007).

A similar analysis was conducted for *C.glutamicum*. We found that the shortest EFM contains 33 reactions when cofactors are set to internal. The shortest EFMs for *C.glutamicum* are not fermentative (Fig. 3) in contrast to *E.coli* (Fig. 1) and the main route for glucose catabolism is the PP pathway. A reasonable question that can be posed is why there is no fermentative mode in the shortest EFMs for *C.glutamicum*. This is due to the fact that the reaction catalyzed by lactate dehydrogenase, which reduces pyruvate to lactate, is not present in the metabolic network, nor any other pathway linking pyruvate to lactate. Note, however, that such reaction is present in the genome annotation of this organism and there is experimental data on lactate dehydrogenase mutants (Inui *et al*., 2004).

In Figure 3, it is also apparent that the main variability in the 10-shortest EFMs is in the balancing of cofactors and there are no alternative pathways for glucose catabolism in comparison to the 10-shortest EFMs of *E.coli* (Fig. 1). This fact can be attributed to the differences in the metabolic networks caused by evolution. While in *E.coli* the ED pathway and the methylglyoxal bypass are present, to date they have not been identified in *C.glutamicum* (Eggeling and Bott, 2005). Moreover, there are differences in some anaplerotic reactions. Nevertheless, there is also an evident difference in the accuracy of both networks, since the number of reactions in the metabolic network of *E.coli* is almost 5-fold higher while the size of the genome and the number of predicted proteins for both organisms is of the same order of magnitude (Blattner *et al.,* 1997; Kalinowski *et al.,* 2003).

As mentioned above, the PP pathway is the only glucose catabolic pathway present in the EFMs, which is due to the requirement of redox anabolic power. An alternative pathway would have been the TCA cycle or anaplerotic reactions between oxaloacetate, malate,

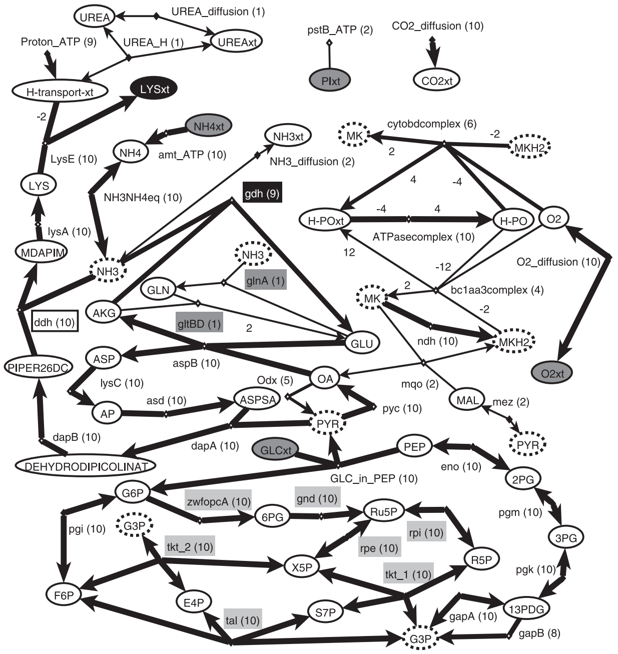


Fig. 3. Merged representation of the 10-shortest EFMs in *C.glutamicum* producing lysine and with cofactors as internal metabolites. Boxed enzyme abbreviation is characteristic for *C.glutamicum* (Eggeling, 1994; Wittmann and Becker, 2007), enzyme abbreviations in light grey, dark grey and black represent the PP pathway, the longest and the shortest pathways for ammonium assimilation, respectively. The following metabolite nodes, in the cytosolic compartment, were removed from the representation for better visualization: ATP, ADP, NAD, NADH, NADP, NADPH, H-transport, COA, PI, CO_2.

phosphoenolpyruvate and pyruvate. However, the presence of the complete TCA cycle requires more enzymes to reduce NADP using glucose. Experimentally, the PP pathway also has a more important role in NADPH synthesis than the TCA cycle. Indeed, metabolic flux analyses have shown that ∼70% of the NADPH is generated by the PP pathway and the remaining 30% by isocitrate dehydrogenase of the TCA cycle (Eggeling and Bott, 2005).

Possible NADPH regenerating cycles, involving anaplerotic reactions, which are often mentioned in the literature (cf. Wittmann and Becker, 2007), are not found with this function. Instead, they can only convert NADPH into NADH because in the genome-scale network the reactions mdh and mqo are set to irreversible forcing these cycles to be irreversible. The existence of two glyceraldehyde-3-phosphate dehydrogenases, gapA and gapB, also allows the conversion of NADPH into NADH, but not the reverse. If the reaction catalysed by lactate dehydrogenase is included in the metabolic network, the fermentative pathways are still not the shortest because there is no alternative to the PP pathway for NADPH synthesis, and therefore, the EFMs with this pathway are the shortest (data not shown).

Regarding the ammonium assimilation, it can be seen that a larger number of EFMs uses glutamate dehydrogenase (gdh) and only two EFMs use the glutamine synthase/glutamate synthase (glnA/gltBD) pathway. The appearance of a longer route is due to the fact that the 10-shortest EFMs in *C.glutamicum* have more widely distributed lengths than the 10-shortest EFMs in *E.coli*. Nevertheless, for *C.glutamicum*, the shorter pathway is more relevant at high ammonium concentrations (Eggeling and Bott, 2005).

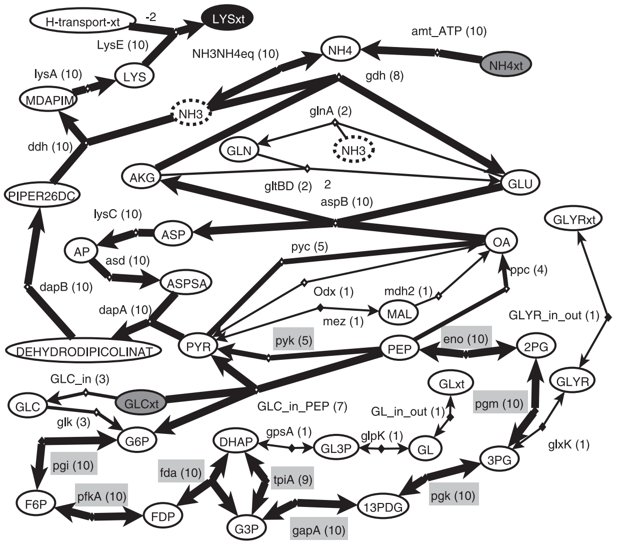


Fig. 4. Merged representation of the 10-shortest EFMs producing lysine in *C.glutamicum* and with cofactors as external metabolites. Enzymes with abbreviations in light grey represent glycolysis. The following metabolite nodes, in the cytosolic compartment, were removed from the representation for better visualization: ATP, ADP, NAD, NADH, NADP, NADPH, H-transport, COA, PI, CO_2.

If cofactors are set external, the PP pathway, the cycles converting NADPH to NADH and enzymes from the respiratory chain do not appear in the 10-shortest EFMs. Instead, glycolysis is the main route for glucose catabolism (Fig. 4). This pathway is indeed the shortest catabolic pathway in this network, as the ED pathway and the glyoxylate bypass are not present. The main variability in these EFMs is found in the synthesis of by-products such as glycerate and glycine and in the interconnection of the catabolic and anabolic part of the EFMs. The latter is evident by the detour made through malate (Fig. 4).

From Figures 3 and 4, it can be observed that the 10-shortest EFMs involve the shortest lysine biosynthetic pathway described in the literature (Wittmann and Becker, 2007). An alternative longer route does exist in *C.glutamicum*, which differs in three reactions and requires one additional reaction to balance succinate and succinyl-CoA, as shown in the 10-shortest EFMs of *E.coli* (Figs 1 and 2). This means that EFMs with higher length are needed so as to obtain the alternative pathway for lysine synthesis.

4 CONCLUSION

The computation of EFMs in genome-scale metabolic networks has been very difficult if not impossible so far. In order to explore the metabolic capabilities of a given organism via EFMs, often smaller sub-networks are delimited. However, the analysis of small sub-networks can be misleading (Kaleta *et al.* 2009; Terzer and Stelling, 2008) and therefore, the computation of EFMs in genome-scale networks is essential for a more comprehensive analysis of the metabolic capabilities of an organism. In such large networks, detecting short EFMs is of interest from the biological viewpoint. Experimentally, it is expensive to overexpress a large number of enzymes, so that shorter pathways are better suited for genetic manipulation. Moreover, shorter pathways usually carry higher fluxes.

In this article we showed that the full set of EFMs can be theoretically enumerated via discrete optimization. This is a promising development in EFM computation and it might serve as a basis for building new methods to explore the structure of large metabolic networks. We presented an effective method to compute the shortest EFMs even in genome-scale networks, as opposed to classic approaches, where EFM analysis cannot be accomplished. A clear advantage of our method in comparison to the classic approaches for EFMs computation is its inherent flexibility. Certainly, the use of optimization enables one to directly search for EFMs that produce/consume a certain metabolite or involve a particular reaction. For this reason the K-shortest EFMs is a suitable concept when exploration of a specific subset of EFMs is of interest.

It is beyond the scope of this article to analyse the run-time complexity of the algorithm. Interesting results in that direction have been presented by Acuña *et al.* (2009). Here we have shown by numerical examples that even for genome-scale networks, the K-shortest EFMs can be computed in reasonable time.

Our procedure was applied to find the 10-shortest EFMs that produce lysine in the genome-scale metabolic networks of *E.coli* and *C.glutamicum*. The computation of the 10-shortest EFMs in *C.glutamicum* was faster than in *E.coli*, mainly due to the difference in network complexity. The sets of reactions in the computed EFMs can be divided into four parts: catabolism of glucose; anabolism of lysine; ammonium assimilation and a subset responsible for cofactor balancing, when cofactors are set internal metabolites. This classification is in agreement with the presentation in many biochemical textbooks.

The catabolic subset converts glucose into aspartate and pyruvate, precursors of lysine, and plays an important role in cofactor supply, in particular of NADPH. In the genome-scale network of *E.coli*, a variety of pathway combinations exists for glucose catabolism because NADPH can be obtained via a NAD(P) transhydrogenase, whereas in the network of *C.glutamicum* the PP pathway is preponderant for NADPH supply. The cofactor balancing subset is more influenced by the catabolic subset than by the anabolic subset. The latter partially overlaps in the solutions of both organisms and does not change in the 10-shortest EFMs. Shorter routes are clearly favored by the K-shortest EFMs method and this fact is evident in the anabolic subset and ammonium assimilation subsets. When cofactors are removed from the balancing constraints, pathways with 100% yield are obtained, hence highlighting the impact of cofactors consumption/supply in lysine synthesis.

Finally, contrary to the widely held belief that the computation of EFMs in large-scale metabolic networks is impossible, the work presented here represents an important step forward.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the helpful comments made by three anonymous reviewers. The groups in which L.F.F., S.S. and C.K. work belong to the Jena Centre for Bioinformatics (JCB).

Funding: Portuguese entities: Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (PhD grant SFRH/BD/32961/2006 to L.F.F.).

Conflict of Interest: none declared.

REFERENCES

Acuña,V. *et al.* (2009) Modes and cuts in metabolic networks: complexity and algorithms. *Biosystems*, **95**, 51–60.

Beasley,J.E. and Planes,F.J. (2007) Recovering metabolic pathways via optimization. *Bioinformatics*, **23**, 92–98.

Blattner,F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Cohen,G.N. and Saint-Girons,I. (1987) Biosynthesis of threonine, lysine, and methionine. In Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella typhimurium—Cellular and Molecular Biology.* Vol. 1, 1st edn, American Society for Microbiology, Washington, pp. 429–444.

Dandekar,T. *et al.* (2003) A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems*, **70**, 255–270.

de Graaf,A.A. (2000) Metabolic flux analysis of *Corynebacterium glutamicum.* In Schügerl,K.B. and Bellgardt,K.H. (eds) *Bioreaction Engineering, Modelling and Control.* Springer, New York, pp. 506–555.

Eggeling,L. (1994) Biology of L-lysine overproduction by *Corynebacterium glutamicum*. *Amino Acids*, **6**, 261–272.

Eggeling,L. and Bott,M (2005) Handbook of *Corynebacterium glutamicum*. CRC Press, Boca Raton.

Feist,A.M. *et al.* (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Fell,D.A. and Thomas,S. (1995) Physiological control of metabolic flux: The requirement for multisite modulation. *Biochem. J.*, **311**(Pt 1), 35–39.

Fischer,E. and Sauer,U. (2003) A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*. *J. Biol. Chem.*, **278**, 46446–46451.

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

Inui,M. *et al.* (2004) Metabolic analysis of Corynebacterium glutamicum during lactate and succinate productions under oxygen deprivation conditions. *J. Mol. Microbiol. Biotechnol.*, **7**, 182–196.

Kacser,H. and Acerenza,L. (1993) A universal method for achieving increases in metabolite production. *Eur. J. Biochem.*, **216**, 361–367.

Kaleta, *et al.* (2009) Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res.*, **19**, 1872–1883.

Kalinowski,J. *et al.* (2003) The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J. Biotechnol.*, **104**, 5–25.

Kjeldsen,K.R. and Nielsen,J.(2009) *In silico* genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol. Bioeng.*, **102**, 583–597.

Klamt,S. and Stelling,J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.*, **29**, 233–236.

Klamt,S. *et al.* (2005) Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *Syst. Biol.*, **152**, 249–255.

Klamt,S. *et al.* (2007) Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol.*, **1**, 2.

Koch,I. *et al.* (2005) Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, **12**, 1219–1226.

Liao,J.C. *et al.* (1996) Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.*, **52**, 129–140.

Mavrovouniotis,M.L. *et al.* (1990) Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.*, **36**, 1119–1132.

Meléndez-Hevia,E. *et al.* (1994) Optimization of metabolism: the evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *J. Theor. Biol.*, **166**, 201–220.

Molin,M. *et al.* (2003) Dihydroxyacetone kinases in *Saccharomyces cerevisiae* are involved in detoxification of dihydroxyacetone. *J. Biol. Chem.*, **278**, 1415–1423.

Niederberger,P. *et al.* (1992) A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. *Biochem. J.*, **287**(Pt 2), 473–479.

Pardalos,P.M. and Resende,M.G.C. (2002) *Handbook of Applied Optimization*. Oxford University Press, New York, USA.

Pfeiffer,T. and Bonhoeffer,S. (2004) Evolution of cross-feeding in microbial populations. *Am. Nat.*, **163**, E126–E135.

Planes,F.J. and Beasley,J.E. (2009) Path finding approaches and metabolic pathways. *Disc. Appl. Math.*, **157**, 2244–2256.

Price,N.D. *et al*. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.*, **2**, 886–897.

Schilling,C.H. *et al*. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229–248.

Schuster,S and Hilgetag,C. (1994) On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, **2**, 165–182.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Schuster,S. *et al*. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, **18**, 326–332.

Schuster,S. *et al*. (2002) Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J. Math. Biol.*, **45**, 153–181.

Schuster,S. *et al*. (2007) Understanding the roadmap of metabolism by pathway analysis. In Weckwerth,W. (ed.), *Metabolomics – Methods and Protocols*, Vol. 358, Human Press, Totowa, New Jersey, pp. 199–226.

Schwarz,R. *et al*. (2007) Integrated network reconstruction, visualization and analysis using YANAsquare. *BMC Bioinformatics*, **8**, 313.

Subedi,K.P. *et al*. (2008) Role of GldA in dihydroxyacetone and methylglyoxal metabolism of *Escherichia coli* K12. *FEMS Microbiol. Lett.*, **279**, 180–187.

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

Teusink,B. *et al*. (2006) Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *J. Biol. Chem.*, **281**, 40041–40048.

Tosaka,O. *et al*. (1983) The production of L-lysine by fermentation. *Trends Biotechnol.*, **1**, 70–74.

Trinh,C.T. *et al*. (2009) Elementary mode analysis: A useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl. Microbiol. Biotechnol.*, **81**, 813–826.

Vallino,J. and Stephanopoulos,G. (1993) Metabolic flux distributions in *Corynebacterium glutamicum* during growth and lysine overproduction. *Biotechnol. Bioeng.*, **41**, 633–646.

van Winden,W.A. *et al*. (2003) Metabolic flux and metabolic network analysis of *Penicillium chrysogenum* using 2D [^{13}C, ^{1}H] COSY NMR measurements and cumulative bondomer simulation. *Biotechnol. Bioeng.*, **83**, 75–92.

von Kamp,A. and Schuster,S. (2006) Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, **22**, 1930–1931.

Wendisch,V.F. *et al*. (2006) Metabolic engineering of *Escherichia coli* and *Corynebacterium glutamicum* for biotechnological production of organic acids and amino acids. *Curr. Opin. Microbiol.*, **9**, 268–274.

Wittmann,C. and Becker,J. (2007) The L-lysine story: from metabolic pathways to industrial production. In Wendisch,V.F. (ed.) *Amino acid biosynthesis – Pathways, Regulation and Metabolic Engineering.* Springer, Heidelberg, pp. 39–70.

Yuan,J. *et al*. (2006) Kinetic flux profiling of nitrogen assimilation in *Escherichia coli*. *Nat. Chem. Biol.*, **2**, 529–530.

Chapter 4

Exploring elementary flux modes in genome-scale networks

While developing the K-shortest EFM method presented in the previous chapter, I examined the characteristics of the optimization framework and identified the key points of this methodology. I found that the iterative nature of the K-shortest EFM method was not required as long as we have a genetic algorithm to control the set of available reactions and an optimization problem to compute a single elementary flux mode given an input network. This idea leads to the development of the EFMEvolver presented in Kaleta *et al.* (2009a). Another improvement in this method is the optimization problem used for determining elementary flux modes, now expressed as a linear program that, given an input network, computes a single elementary flux mode. This simplification reduces the time required for computing an elementary flux mode. On the other hand, the genetic algorithm is used for exploring the space of elementary flux modes by constraining the set of available reactions. With this method we can compute larger sets of phenotypically distinct elementary flux modes. My contributions in this work, besides the “catalytic activity”for producing this new method, were the implementation of part of the genetic algorithm, the achievement of a number of the initial simulations and the preparation of the manuscript.

C Kaleta, LF de Figueiredo, J Behre and S Schuster. In *Lecture Notes in Informatics*, vol P-157, pp 179-189, Bonn 2009.

EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks

Christoph Kaleta*, Luís Filipe de Figueiredo*, Jörn Behre, and Stefan Schuster†

Department of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

Abstract: Elementary flux mode analysis (EFM analysis) is an important method in the study of biochemical pathways. However, the computation of EFMs is limited to small and medium size metabolic networks due to a combinatorial explosion in their number in larger networks. Additionally, the existing tools to compute EFMs require to enumerate all EFMs before selecting those of interest. The method presented here extends EFM analysis to genome-scale models. Instead of computing the entire set of EFMs an optimization problem is used to determine a single EFM. Coupled with a genetic algorithm (GA) this allows to explore the solution space and determine specific EFMs of interest. Applied to a network in which the set of EFMs is known our method was able to find all EFMs in two cases and in another case almost the entire set before aborted. Furthermore, we determined the parts of three metabolic networks that can be used to produce particular amino acids and found that these parts correspond to significant portions of the entire networks.
Availability: Source code and an executable are available upon request.

1 Introduction

In the post-genomic era, the analysis of metabolic networks is essential for molecular biology. These networks are complex and the subdivision of a network into pathways makes the analysis more comprehensive. However, the focus only on specific classically known pathways can conceal the view on the actual metabolic capabilities of an organism [KdFS09]. Thus, the construction of genome-scale metabolic networks that model the entire metabolism of organisms has come to importance [FP08].

A method that has been used to comprehensively studying pathways in metabolic networks is elementary flux mode analysis [SDF99]. Elementary flux modes (EFMs) are a systematic definition of the biological concept of a pathway. They correspond to minimal sets of reactions that can perform at steady state [SDF99]. EFM analysis has already been used to study biochemical relevant metabolic pathways [CS04, dFSKF09], to study metabolic network properties such as fragility and robustness [SKB$^+$02, BWvK$^+$08], and to optimize microorganisms with respect to the production of a certain metabolite [TUS08]. However, EFM analysis has been limited to small and medium scale networks because the number of EFMs grows exponentially with the size of the network [KS02]. For instance, Yeung *et*

*Both authors contributed equally

†Corresponding author (stefan.schu@uni-jena.de)

al. [YTP07] estimated that the number of extreme pathways [SLP00], a subset of EFMs, is at the order of 10^{29} for a genome-scale model of human.

Due to this problem, alternative approaches for the identification of pathways based on graph theory have been proposed [RAS^{+}05, CCWvH06, BK08]. These methods abstract from the metabolic network by converting it into a graph and consider only connected paths. While they operate efficiently in genome-scale metabolic networks, they bear the problem that a detected pathway does not automatically imply that a net-conversion of the source metabolite into a specific target metabolite is possible [PB08, dFSKF09].

Here we want to present a method that allows the enumeration of EFMs in genome-scale metabolic models. Starting from an initial pathway, the space of EFMs is explored using a genetic algorithm (GA). GAs have already been used in the analysis of metabolic networks to find combinations of gene knockouts that improve the production of a given metabolite [PRFN05]. We used benchmark models for EFM analysis to validate our new method and applied it to a study of amino-acid synthesis in genome-scale metabolic models.

2 Methods

The aim of our algorithm is, given a metabolic network and an input medium, to find all EFMs producing a certain metabolite. The employed strategy is based on the observation that gene knockouts can force an organism to use pathways alternative to those found under standard conditions. Thus, we are detecting EFMs by evolving a population in which each individual corresponds to a set of knockouts. However, instead of considering the knockouts of genes we here focus on the "knockout" of reactions. By searching for a specific EFM avoiding reactions that are knocked out and iterating over different sets of knockouts we are able to determine different EFMs.

2.1 Detecting a single EFM

A metabolic network comprising *m* metabolites and *n* reactions is defined by the *m*x*n* stoichiometric matrix $\mathbf{N}$. Each metabolite can be defined to be either internal or external. External metabolites differ to internal metabolites in that their concentration is assumed to be buffered by the system. Examples for such external metabolites are energy currency metabolites like ATP, NADH and FADH. Since their concentration is assumed to be constant they are not required to be balanced by an EFM.

To be an EFM, a flux $\mathbf{v} \in \mathbb{R}^n$ through a reaction network has to fulfill the following conditions: (1) steady-state condition, i.e., all internal metabolites are balanced; (2) irreversible reactions have positive fluxes; (3) non-decomposability of the enzyme set, i.e., the non-zero indices of one EFM cannot be a subset of the non-zero indices of another EFM. In our approach reversible reactions are decomposed into two irreversible reactions with opposite directions. Therefore, all fluxes have to be positive.

Given a set K of reactions to be knocked out and an index μ corresponding to a target reaction which produces a certain metabolite of interest, the optimization problem to compute an EFM can be formulated as a linear program by minimizing $\sum_{r=1}^{n} v_r$ subject to

$$\mathbf{Nv} = 0 \tag{1}$$

$$\mathbf{v} \geq 0 \tag{2}$$

$$v_\mu \geq 1 \tag{3}$$

$$\forall i \in K: \ v_i = 0 \tag{4}$$

Using eqs. 1 and 2 we only allow for a strictly positive flux $\mathbf{v}$ that obeys the steady-state condition. Eq. 3 forces the solution to have a positive flux through a given reaction which can be the outflow of the product of interest, i.e., if a solution exists, $\mathbf{v}$ produces the metabolite of interest. Eq. 4 guarantees that we only find a flux that does not use the reactions in K that are knocked out. By minimizing the overall flux and solving the linear program using the simplex algorithm [Sch98] we achieve that $\mathbf{v}$ corresponds to an EFM. This property of $\mathbf{v}$ will be shown in the following.

The solution space of the steady-state and the irreversibility condition (eqs. 1 and 2) in the space of possible fluxes $\mathbb{R}^n$ corresponds to a convex polyhedral cone $\mathcal{P}$ [GK04]. Since, we split reversible reactions, the extreme rays or spanning vectors of $\mathcal{P}$ correspond to the EFMs of the system. Furthermore, a knockout of a reaction only leads to the disappearance of some EFMs [SDF99]. Thus, for every K chosen, the cone is still spanned by EFMs and eq. 4 does not impact the property of the spanning vectors of $\mathcal{P}$ of being EFMs. Furthermore, eq. 3 cuts $\mathcal{P}$ with a hyperplane at $v_\mu = 1$ (Figure 1C). Since $\mathcal{P}$ is unbounded the edges of the solution space of eqs. 1 - 3 correspond to the intersection points between the EFMs defined by eqs. 1 as well as 2 and the hyperplane defined by $v_\mu = 1$. These points can each be written as the corresponding EFM multiplied with a scaling-factor. From linear programming it is known that the simplex algorithm used to solve such problems always returns a solution that can be found at the edges of the solution space [Sch98]. Thus, using the simplex algorithm and minimizing the objective function subject to eqs. 1 - 4 will always return an EFM.

In principle, the described linear program can find all EFMs by testing every possible set of knocked out reactions K. However, this is computationally inefficient and thus we will next outline an algorithm that allows to explore the space of EFMs more efficiently.

2.2 Genetic Algorithm

The aim of the GA is to test different sets of reactions to be knocked out in order to find all EFMs. Each such set of reactions corresponds to an individual. Each individual is represented by a binary genome $\mathbf{G}$ of length n, i.e., the number of reactions in the system. $G_i = 1$ indicates that reaction i can be used by that organism and $G_i = 0$ that this reaction is knocked out. From each genome an EFM can be derived by mapping $\mathbf{G}$ to the

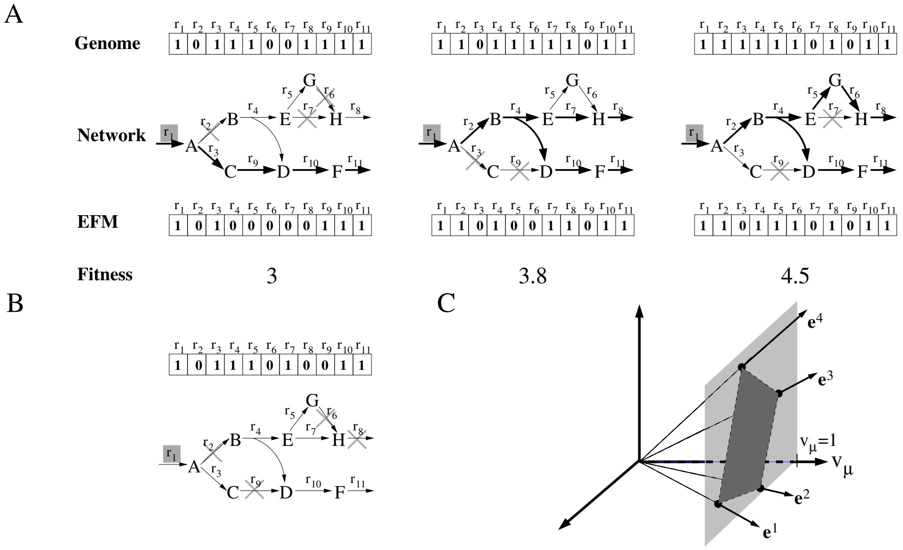


Figure 1: Scheme of the computation of EFMs. **A** Viable individuals. The target reaction $\mu = r_1$ is shaded in gray. In the upper row the genome of each individual is given. The second row indicates the reactions knocked out in the model and the third row the EFM obtained from the linear program. Even though the EFM of the third individual is also a valid EFM satisfying eqs. 1 - 4 for the second individual it is not minimal since the sum of fluxes is higher. The fourth row gives the fitness of each individual for a population containing the three depicted genomes. **B** Individual for which no EFM can be found. **C** Three-dimensional solution space of eqs. 1 - 3 for 3 reactions (not shown). The solution space is defined by the intersection of the solution space of eqs. 1 and 2, spanned by the EFMs $\mathbf{e}^1$ to $\mathbf{e}^4$, and the half-space defined by eq. 2. Optimal solutions of the linear program can always be found in the edges of the solution space (black circles).

set of knocked out reactions K and solving the linear program described in the previous section. Thus, we can obtain an EFM associated to an individual (Figure 1A and 1B). Solving the linear program described in the last section we can only find a single EFM. In consequence, by specifying different sets of reactions that should not be used by an EFM, that is, by knocking them out, we can sample EFMs.

Central for each GA is the definition of a fitness function that returns a numerical value indicating the quality of an individual. In contrast to other approaches the aim of the GA described here is not to find an individual that is optimal in some sense, but to detect all possible EFMs in a metabolic network. Thus, we attribute higher fitness to individuals whose associated EFMs use reactions which are not frequent in the EFMs of the population. Given a population $\mathbf{G}^1$, ..., $\mathbf{G}^s$ of individuals and the associated EFMs $\mathbf{e}^1$, ..., $\mathbf{e}^s$ the fitness $f(\mathbf{G}^k)$ of a particular individual $\mathbf{G}^k$ is defined by

$$f(\mathbf{G}^k) = \sum_{i=1}^{n} \frac{sign(e_i^k)}{\sum\limits_{j=1}^{s} sign(e_i^j)} \tag{5}$$

with $sign(x)$ returning '1' if x is non-zero, i.e., if a reaction is used, and '0' otherwise.

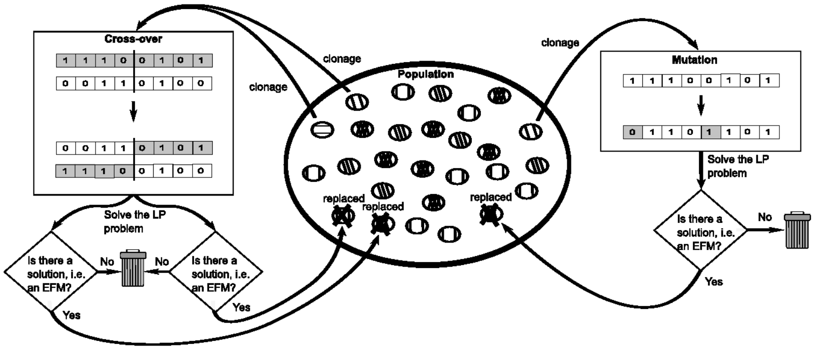


Figure 2: Setup of the GA. Individuals from the population are cloned and subsequently mutated or recombined. Afterward the viability of the individuals is tested by determining an EFM contained in them that uses the target reaction μ. If no such EFM is found, the individual is discarded. Otherwise it is reinserted by replacing a randomly chosen individual in the population.

For the GA we use the setup depicted in Figure 2. We assume a constant population size of s individuals and use two genetic operators: mutation and recombination. Before selecting individuals from the population it is decided whether a mutation or recombination should be performed. With probability $1 - p_{rec}$ one individual is mutated and with probability p_{rec} two individuals are recombined. To apply these operators, individuals are cloned from the population. By cloning we mean that an individual is selected and its genome copied creating a new individual. Thus, the original individual persists in the population. Given the individual fitness values $f(\mathbf{G}^1)$, ..., $f(\mathbf{G}^s)$ the probability of individual k to be cloned is proportional to its fitness:

$$P(\text{“Individual } k \text{ is cloned”}) = \frac{f(\mathbf{G}^k)}{\sum\limits_{i=1}^{s} f(\mathbf{G}^i)} \tag{6}$$

During a mutation event, after cloning a single individual, each position in its genome is mutated with probability p_{mut}. Subsequently, it is tested whether the new individual is “viable” by determining the EFM associated to it. If such an EFM is found, the individual is re-inserted into the population by replacing a randomly chosen individual. Furthermore, the EFM that has been found is compared to all previously found EFMs and is added to this set if it has not already been detected. If two individuals are recombined, they are first cloned and then the genomes are interchanged starting from a random position. Subsequently it is tested for both if they are viable, and, if this is the case, they are reinserted replacing two randomly chosen individuals of the population. Thus, EFMs are detected as a side product of checking the viability of new individuals.

An important advantage of GAs is that they can be easily parallelized by the use of separate threads that mutate, recombine, and test individuals. Thus, the multi-processor architecture of modern desktop PCs is fully exploited.

3 Results

We applied our method to compute EFMs producing lysine, threonine, and arginine in two metabolic networks of *Escherichia coli* and one metabolic network of *Corynebacterium glutamicum*. Especially for the industrial production of lysine *C. glutamicum* is of importance [WBE06]. The first network of *E. coli* has been presented in [BWvK$^+$08]. It comprises 220 reactions and models amino acid metabolism. This network has the advantage that we can compute EFMs using Metatool [vKS06]. The second network represents a genome-scale model of *E. coli* metabolism and comprises 3558 reactions [FHR$^+$07]. The model of *C. glutamicum* contains 641 reactions and has been presented in [KN09]. In order to avoid side-pathways used for the balancing of co-factors and to provide an input medium we set the metabolites ammonium, AMP, ATP, CO_2, coenzyme A, glucose, NAD^+, NADH, $NADP^+$, NADPH, oxygen, protons, and inorganic ions to external status. As parameters for the computation we used a population size of $s = 100$ individuals, a mutation rate of $p_{mut} = 0.01$ per reaction and a probability of $p_{rec} = 0.3$ for recombination events. Computations were performed on an Intel® Core$^{\text{TM}}$2 Quad Q9300 machine with 4096 MB RAM running Linux Kernel 2.6.25 and Java Hotspot VM version 1.6.0. *Clp* version 1.0.6 from the COIN-OR project [LH03] has been used to solve the linear programs. An overview on the results is given in Table 1 and Figure 3.

As a first benchmark we tested to what extend our method can recover EFMs in a system in which they are already known. The model of [BWvK$^+$08] contains 3436 EFMs producing lysine, 444 EFMs producing threonine and 27450 EFMs producing arginine. We found all EFMs producing threonine and lysine after 491 s and 4821 s, respectively. For arginine we recovered 95.6% of all EFMs after a running time of 7200 s. In comparison, Metatool 5.1 took only 61 s to find all 65840 EFMs. However, a direct run-time comparison even to the currently fastest algorithm for the enumeration of EFMs presented in [TS08] does not bear much meaning since these methods in general only return the entire set of EFMs. This is not practicable in genome-scale networks since the number of EFMs exceeds by far current limitations in memory and processing power [YTP07]. An interesting behavior of the GA can be observed from these experiments. First, the time-course shows a kind of saturation when having found most of the EFMs. Furthermore, we observe phases in which only few new EFMs are found and sudden jumps in which the number increases rapidly as in the case of threonine in the model of amino acid metabolism at $t = 320$ s. While this particular behavior is also observable in the case of lysine in the model of *C. glutamicum*, a saturation

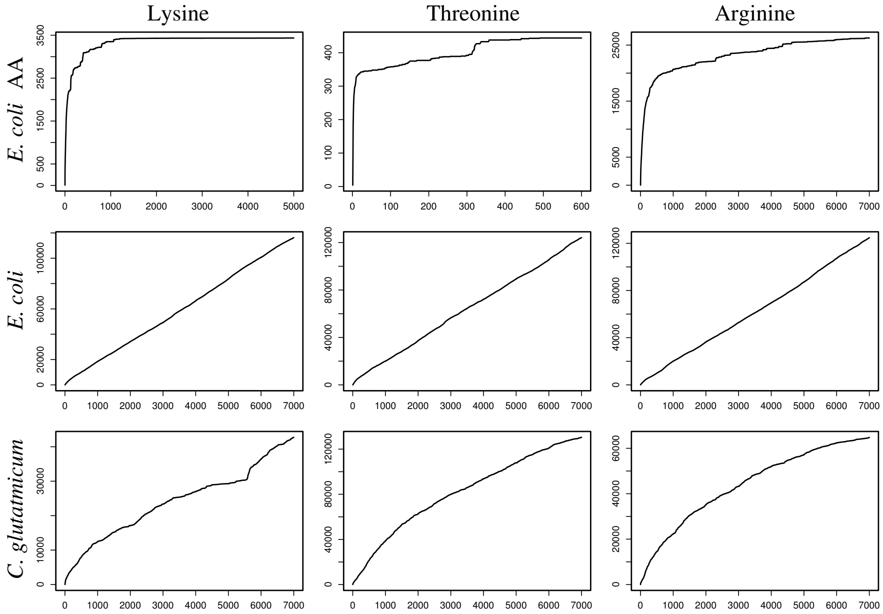


Figure 3: Time-course of the determination of EFMs for the three test-models: *E. coli* AA, [BWvK+08]; *E. coli*, [FHR+07]; *C. glutamicum*, [KN09]. The X-axis gives the running time in seconds and the Y-axis the number of EFMs found.

can be observed for the two other amino acids. In conjunction with the limited size of this model these results indicate that our method has already discovered a significant portion of all EFMs producing the three amino acids. In contrast, in the genome-scale system of *E. coli* we observe an almost linear increase in the number of EFMs without any saturation indicating that the number of EFMs existing in this model is much larger than the number already sampled.

Furthermore we tested the time required for the computation of 2000 EFMs in all models. We found the influence of network size on the running time much smaller than expected. Thus, it took on average 26.3 s to find 2000 EFMs in the model of *C. glutamicum* and 43 s in the genome-scale model of *E. coli* although both models differ more than five-fold in the number of reactions. This behaviour might be attributed to the simplex algorithm used to solve the linear programing problem described in Section 2.1. Since we are iteratively solving very similar problems and the simplex algorithm can start from a previous solution after changing some constraints, new solutions can be found very fast without need to consider the entire problem, but only a specific sub-part for which constraints were changed.

Another interesting aspect of the detected EFMs arises from the part of the network that can be used for the production of particular amino acids. For this analysis we combined

Model	# Rea.	AA	# EFMs	# Min.	CS	2000 EFMs
E. coli	220	Lysine	3436	16	94	95 s
AA metabolism		Threonine*	444	11	67	839 s
[BWvK$^+$08]		Arginine	26276	18	95	8 s
E. coli	3558	Lysine	118598	29	1826	49 s
Genome-scale		Threonine	126491	26	2084	38 s
[FHR$^+$07]		Arginine	127988	37	1895	42 s
C. glutamicum	641	Lysine	43115	23	240	28 s
Genome-scale		Threonine	131346	24	245	22 s
[KN09]		Arginine	65236	35	246	29 s

Table 1: Overview on computed EFMs. For each of the three test-models (number of reactions in the second column) the GA has been used to determine EFMs for the production of lysine, threonine and arginine (third column). The fourth column gives the number of EFMs detected after a running time of 7200 s. The fifth and sixth column indicate the minimal length of a detected EFM for the production of the given amino acid and the total number of different reactions used by all EFMs. The last column indicates the time required for the computation of 2000 EFMs averaged over 10 runs. In the case marked with *, the system only contained 444 EFMs.

all the computed EFMs for each test-case and determined the number of reactions used (Table 1). Furthermore, we determined the minimal number of reactions used by an EFM for the production of a given amino acid (Table 1). Combining all EFMs, the part of the metabolic network that can be used for the production of each amino acid varies in between 31% to 59% of the total network size. In consequence, there seems to be a great versatility in potential pathways. However, this versatility can be mostly attributed to the side-products of amino acid biosynthesis. For instance, in the production of lysine succinyl-CoA is converted to succinate. There are two ways of balancing succinyl-CoA and succinate. Either succinyl-CoA is additionally produced from the input medium and succinate is disposed through some other pathway, or succinate is reconverted into succinyl-CoA. Hence, we see a combinatorial explosion since the basic route producing lysine can be combined, on the one hand, with every pathway producing succinyl-CoA and consuming succinate. On the other hand this route can be combined with every possible pathway converting succinate into succinyl-CoA. This is also apparent from an analysis of the 64699 EFMs producing amino acids in the model of [BWvK$^+$08]. Here we found that 35% of the EFMs do not only produce a single, but several amino acids. These additional amino acids can serve as sinks for side-metabolites.

4 Discussion

In this work we have outlined a new approach based on a genetic algorithm (GA) that allows to determine EFMs using a specific reaction in genome-scale metabolic networks. Previous methods that are based on searching paths in a graph representation of a metabolic network only guarantee to find connected routes while EFMs correspond to routes of actual metabolic conversions [dFSKF09]. Computing EFMs in a network in which they also can be enumerated using deterministic algorithms we demonstrated that even large sets of EFMs can be recovered almost entirely. Comparing the time-course of the number of EFMs enumerated between a small and two large networks we concluded that we had already found a significant portion of all EFMs in a genome-scale model of *C. glutamicum* but only a small portion in a much larger model of *E. coli*. Analyzing the parts of the metabolic network which can be used by EFMs we found that they corresponded to 31% to 59% of the entire network even though individual pathways are usually much shorter. We attributed this result to the large variability of pathways that can be used to balance side-metabolites of amino acid biosynthetic pathways.

There exist several alternative approaches that allow a similar analysis of pathways in genome-scale networks. They either decompose a large network into smaller subnetworks or consider the entire network. The former approaches bear the problem that they only consider a small network on the local scale and thus they can contain artificial pathways that do not appear on the scale of the entire system [KdFS09]. Among the latter approaches especially constrained based methods are of importance. Methods from this field that allow to perform a similar analysis are flux balance analysis (FBA, [VP94]), flux variability analysis (FVA, [MS03]), and stochastic sampling of the solution space of eqs. 1 - 3 with additional upper bounds on reaction fluxes [WFGP04]. However, FBA only returns a specific pathway optimizing a certain objective function [VP94] and flux variability analysis only determines the set of reactions that can take part in alternative optimal pathways, without allowing to identify these pathways [MS03]. Stochastic sampling in contrast is very similar to our approach, but returns solutions that lie within the solution space of eqs. 1 - 3. Thus, rather than EFMs fluxes that correspond to combinations of EFMs are returned.

Our method represents an important step towards the analysis of EFMs, and thus of pathways, in genome-scale metabolic networks. While we used a fitness function that selects for diversity one can think of other functions that can be used. Thus, it is of interest to analyze suboptimal EFMs for the production of some metabolite which are in a specific range of yield per mole of an input metabolite or fulfill additional criteria like the production of a certain side-metabolite. Furthermore, since EFMs correspond to the concept of minimal transition invariants (MTIs) in petri-nets [SPM$^+$00, KH08], our approach can also be useful to find MTIs in large petri-nets.

5 Acknowledgments

Financial support from the German Ministry for Research and Education (BMBF) to C.K. and J.B. (grant FKZ 0315285E and FKZ 0313078E), and from the Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (PhD grant SFRH/BD/32961/2006) to L.F.F. is gratefully acknowledged.

C Kaleta, LF de Figueiredo, J Behre and S Schuster. In *Lecture Notes in Informatics*, vol P-157, pp 179-189, Bonn 2009.

References

[BK08] T. Blum and O. Kohlbacher. Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J Comput Biol*, 15(6):565–576, 2008.

[BWvK^{+}08] J. Behre, T. Wilhelm, A. von Kamp, E. Ruppin, and S. Schuster. Structural robustness of metabolic networks with respect to multiple knockouts. *J Theor Biol*, 252(3):433–441, Jun 2008.

[CCWvH06] D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol*, 356(1):222–236, Feb 2006.

[CS04] R. Carlson and F. Srienc. Fundamental *Escherichia coli* biochemical pathways for biomass and energy production: Identification of reactions. *Biotechnol Bioeng*, 85(1):1–19, Jan 2004.

[dFSKF09] L. F. de Figueiredo, S. Schuster, C. Kaleta, and D. A. Fell. Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, 25(1):152–158, Jan 2009.

[FHR^{+}07] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

[FP08] A. M. Feist and B. Ø. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol*, 26(6):659–667, Jun 2008.

[GK04] J. Gagneur and S. Klamt. Computation of elementary modes: A unifying framework and the new binary approach. *BMC Bioinformatics*, 5:175, 2004.

[KdFS09] C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res Res*, 2009. Accepted.

[KH08] I. Koch and M. Heiner. *Biological Network Analysis*, chapter Petri Nets, pages 139 – 180. Wiley Book Series in Bioinformatics. Wiley & Sons, 2008.

[KN09] K. R. Kjeldsen and J. Nielsen. In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol Bioeng*, 102(2):583–597, Feb 2009.

[KS02] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233–236, 2002.

[LH03] R. Lougee-Heimer. The Common Optimization INterface for Operations Research: Promoting open-source software in the operations research community. *IBM J Res Dev*, 47(1):57–66, 2003.

[MS03] R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*, 5(4):264–276, Oct 2003.

[PB08] F. J. Planes and J. E. Beasley. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Brief Bioinform*, 9(5):422–436, Sep 2008.

[PRFN05] K. R. Patil, I. Rocha, J. Förster, and J. Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6:308, 2005.

[RAS^{+}05] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, 21(7):1189–1193, Apr 2005.

[Sch98] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, June 1998.

[SDF99] 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 Biotechnol*, 17(2):53–60, Feb 1999.

[SKB^{+}02] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, Nov 2002.

[SLP00] C. H. Schilling, D. Letscher, and B. Ø. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3):229–248, Apr 2000.

[SPM^{+}00] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar. Structural analysis of metabolic networks: Elementary flux modes, analogy to Petri nets, and application to *Mycoplasma Pneumoniae*. In *German Conference on Bioinformatics*, pages 115–120, 2000.

[TS08] Marco Terzer and Jörg Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235, Oct 2008.

[TUS08] C. T. Trinh, P. Unrean, and F. Srienc. Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Appl Environ Microbiol*, 74(12):3634–3643, Jun 2008.

[vKS06] A. von Kamp and S. Schuster. Metatool 5.0: Fast and flexible elementary modes analysis. *Bioinformatics*, 22(15):1930–1931, Aug 2006.

[VP94] A. Varma and B. Ø. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol*, 60(10):3724–3731, Oct 1994.

[WBE06] V. F. Wendisch, M. Bott, and B. J. Eikmanns. Metabolic engineering of *Escherichia coli* and *Corynebacterium glutamicum* for biotechnological production of organic acids and amino acids. *Curr Opin Microbiol*, 9(3):268–274, Jun 2006.

[WFGP04] S. J. Wiback, I. Famili, H. J. Greenberg, and B. . Palsson. Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *J Theor Biol*, 228(4):437–447, Jun 2004.

[YTP07] M. Yeung, I. Thiele, and B. Ø. Palsson. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics*, 8(1):363, 2007.

Chapter 5

Prospects of constraint-based analysis

The methods presented in the previous two chapters, namely the K-shortest EFMs and the EFMEvolver, are at the cutting edge of elementary flux mode analysis. An overview of the current achievements of metabolic network reconstruction, constraint-based analysis is given in this chapter. Moreover, Ruppin *et al.* (2010) show the increasing importance of game theory as a modelling framework in systems biology and how it can be used to study some of the issues associated with constraint-based analysis. Finally, future applications of these modelling approaches in the study of the metabolism of multi-cellular organisms are pointed out.

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks

Eytan Ruppin[1], Jason A Papin[2], Luis F de Figueiredo[3] and Stefan Schuster[3]

With the advent of modern omics technologies, it has become feasible to reconstruct (quasi-) whole-cell metabolic networks and characterize them in more and more detail. Computer simulations of the dynamic behavior of such networks are difficult due to a lack of kinetic data and to computational limitations. In contrast, network analysis based on appropriate constraints such as the steady-state condition (constraint-based analysis) is feasible and allows one to derive conclusions about the system's metabolic capabilities. Here, we review methods for the reconstruction of metabolic networks, modeling techniques such as flux balance analysis and elementary flux modes and current progress in their development and applications. Game-theoretical methods for studying metabolic networks are discussed as well.

Addresses
[1] Department of Computer Science, Tel Aviv University, Tel Aviv, Israel
[2] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908, USA
[3] Department of Bioinformatics, Friedrich Schiller University Ernst-Abbe-Pl. 2, 07743 Jena, Germany

Corresponding author: Schuster, Stefan (stefan.schu@uni-jena.de)

Current Opinion in Biotechnology 2010, **21**:502–510

This review comes from a themed issue on
Systems biology
Edited by Vitor Martins dos Santos and Jiri Damborsky

Available online 6th August 2010

0958-1669/$ – see front matter

© 2010 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.copbio.2010.07.002

Introduction

The study of metabolism has changed drastically during the last century. The concept of metabolic pathways was molded by the experimental methods available in the beginning of the 20th century resulting in a stepwise elucidation of metabolism (cf. [1•]). In the second half of the 20th century, the discovery of the structure and information coding of DNA laid the foundations for recombinant technology, making microorganisms more amenable to metabolic engineering [2].

The mathematical modeling and computer simulation of metabolic systems started with dynamic modeling [3•]. This is useful, for example, to simulate the occasional metabolic oscillations in bioreactors. Later, it was realized that the knowledge of kinetic parameters was insufficient in many cases and that a detailed dynamic simulation is often unnecessary [4,5]. Many specific questions such as the effect of the activation or overexpression of an enzyme can be tackled by specially tailored techniques such as Metabolic Control Analysis (MCA) (cf. [2,3•,4]). Moreover, the search for methods to analyze invariants of intracellular networks led to the development of the constraint-based modeling (CBM) approach, also called structural analysis or network analysis [4,5]. This requires even less input data than MCA. A subfield is called Metabolic Pathway Analysis [6,7,8••], in which the structure of pathways (routes) going through the system is detected and/or optimal flux distributions are calculated based on the stoichiometry of the network and the directionality of reactions, the knowledge of which is often available. Elementary flux modes (EFMs) [9••,10••] and extreme pathways [11••] were established as unbiased mathematical representations of metabolic pathways.

CBM comprises metabolic flux analysis (MFA) (cf. [2]) by which flux distributions can be predicted using flux measurements in addition to network stoichiometry. The limitations in measurements led to the inclusion of additional constraints coming from the Darwinian theory of optimization in evolution. Notably, optimality principles such as maximizing growth rate or given reaction fluxes at normalized input rate are widely used, thus enabling one to predict fluxes by linear programming [12••,13••]. This is the essence of Flux Balance Analysis (FBA), by which phenotypically relevant flux distributions in metabolic networks can be predicted (cf. [14,15]).

At the end of the 20th century, the development of new experimental technologies such as sequencing and chip technologies triggered the explosion of omics data. In metabolomics, several hundreds of metabolite concentrations can be measured simultaneously [16••,17]. In fluxomics, in contrast, it is difficult to measure more than a dozen fluxes simultaneously [18••]. After DNA microarrays, transcriptomics has recently reached a higher level by RNA-Seq technologies [19]. A secondary source is bibliomics (text mining). High amounts of data enabled one to reconstruct more complex metabolic networks reaching a genome scale for a rapidly increasing number of species (cf. [20••,21••]). In addition, recent efforts to integrate gene expression data [22,23••,24], sequencing data [25••], proteomics [17], and other omics data have enabled the generation of reconstructions unique to particular life-cycle stages, environments, and genetic backgrounds.

These large models together with CBM methods represent a key foundational advance in Systems Biology [5,14] and are essential for seeking comprehension of biological functioning throughout the integration of data with mathematical models. The modeling and computer simulation of metabolism in the genome and post-genome eras have been the subject of a number of reviews [6,7,20••,21••]. In this paper we will review recent advancements in the methods for metabolic network reconstruction, the tools available for their analysis and several applications. A special focus will be on game-theoretical methods (cf. [26]). Finally, perspectives on further developments in Systems Biology will be outlined.

Reconstruction of metabolic networks

The process of metabolic network reconstruction begins with the annotation data of the relevant genes (cf. [20••,27]). This annotation provides the 'parts list' for the network. The metabolic reactions that the associated gene products catalyze are delineated by incorporating data on the metabolites and stoichiometry from databases (e.g. ExPasy and KEGG) and the literature. The stoichiometric coefficients of the metabolites or compounds in the associated reactions are typically represented in a stoichiometric matrix, N (sometimes denoted S) with its rows corresponding to the metabolites and the columns representing the chemical transformations that the gene products (enzymes) catalyze. The usefulness of CBM in the reconstruction process is outlined in Box 1. Two key challenges are firstly, the integration of disparate high-throughput data and secondly, the inclusion of additional constraints to improve the predictive power. Recently, methods for the iterative refinement of the networks using high-throughput transcript data have been developed and used to significantly improve the reconstruction of *Chlamydomonas reinhardtii* and to identify key genes associated with biofuel production [25••].

Box 1 Most metabolic systems subsist at stationary states. Even if they oscillate (only very few do), the average reaction rates (on a sufficiently long time scale), ν, must fulfill the steady-state condition

$$N \cdot \nu = 0 \tag{1}$$

because, otherwise, the concentrations of metabolic intermediates would accumulate or be depleted in the long run. In addition, for some or all fluxes, inequality constraints can be written:

$$\nu_{i,\mathrm{inf}} \leq \nu_i \leq \nu_{i,\mathrm{sup}}. \tag{2}$$

For example, $\nu_{i,\mathrm{inf}}$ is zero for all irreversible reactions. For all reactions, $\nu_{i,\mathrm{sup}}$ can be given by the respective maximal velocity if it is known.

Central to FBA is an optimization principle

$$\text{maximize } \Sigma c_i \nu_i \tag{3}$$

subject to relations (1) and (2) [12••,13••,15]. The coefficients c_i denote the weights of the particular fluxes in the objective function, for example, the production of biomass. Typically, relation (2) includes one constraint that fixes or limits a relevant input flux, for example, glucose uptake.

The non-decomposability of elementary flux modes can be expressed by a constraint saying that the support of the flux vector is not a proper superset of the support of any other steady-state flux vector, ν':

$$S(\nu) = \{i : \nu_i \neq 0\} \text{ not proper superset of } S(\nu') \tag{4}$$

For a reconstructed network to be realistic, there must be a flux vector ν fulfilling Eqn (1) in Box 1 and covering practically all reactions. While there may be a few reactions that always subsist at thermodynamic equilibrium, the vast majority must be able to operate at non-equilibrium, that is, they must not be blocked due to missing reactions [28•] and have to fulfill mass conservation [29]. The coverage of a network by a flux distribution can be tested by a method called flux coupling analysis, which is based on linear programming [30••].

Genome-scale reconstructions have been assembled for organisms from all kingdoms: archaea, eukaryota, and bacteria (Figure 1), and include single-cell and multi-cell organisms (for references to specific reconstructions, see [21••]). Here, we mention only a few: *Escherichia coli* [31••], *Saccharomyces cerevisiae* [32••], *C. reinhardtii* [25••,33], *Arabidopsis thaliana* [34,35], mouse [36,37], and human [38••,39••]. The goal is to account for all the enzymes encoded in the entire genome. However, the term 'genome-*scale*' is to express the dimension in which this is done, which does not necessarily imply that this difficult task would be accomplished with 100% completeness. Questions such as whether substrate A can be transformed into product B can only be tackled exhaustively in a whole-cell model, which provides a further motivation for their reconstruction [40••].

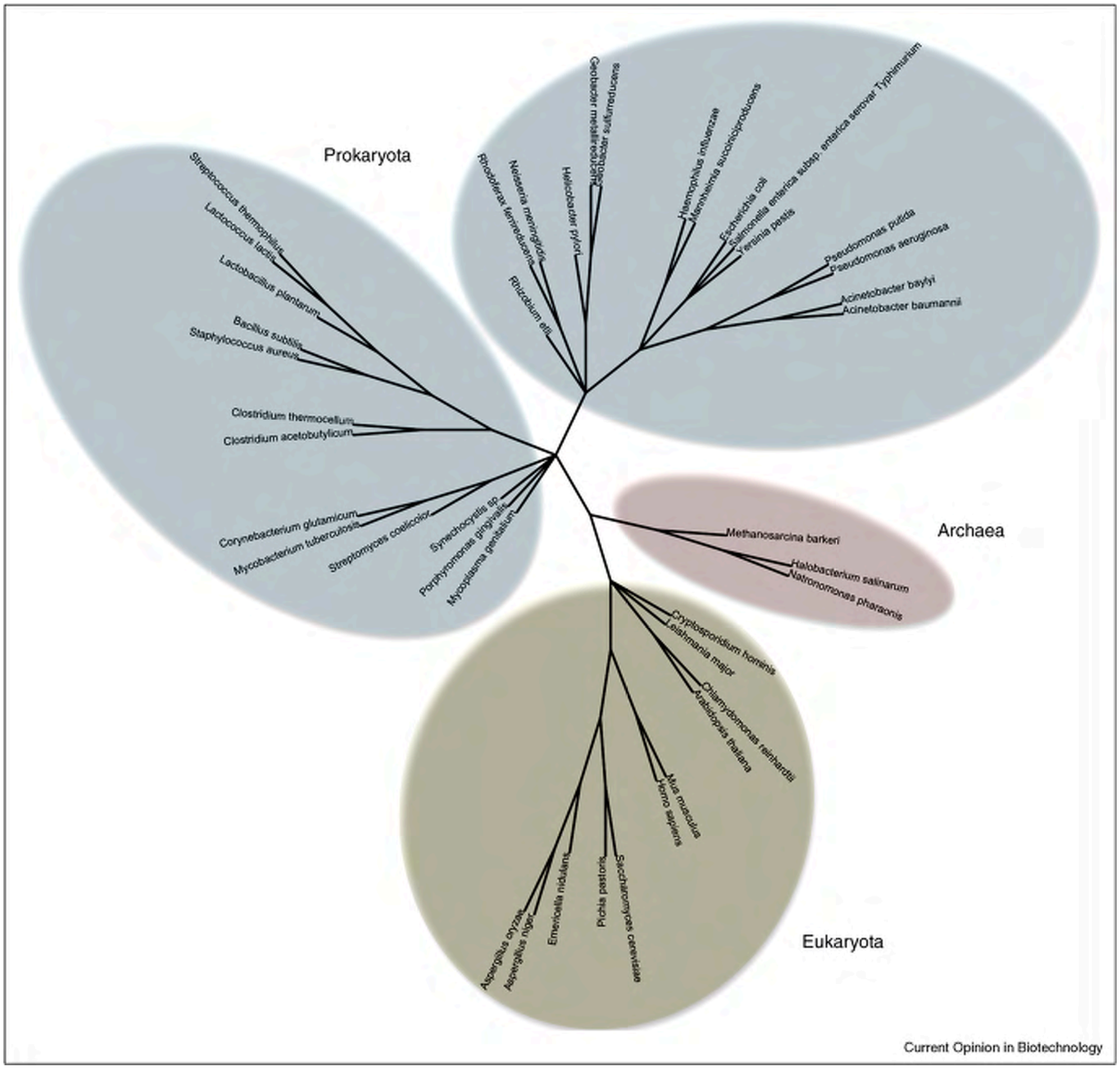
Some recent efforts are focused on automating the reconstruction process (cf. [20••,41]), aided by the development of computational platforms to manage the data associated with gene–enzyme-reaction associations and reaction stoichiometry. This has had some success, although there is typically a recognized need for manual curation efforts [28•,29,42]. A significant remaining challenge is the visualization of these networks, in spite of some recent efforts in this area [43].

Biotechnological applications of metabolic modeling

With the reconstructions that have been generated, the next important step is the development of analysis tools and frameworks to study functional properties of these networks [30••,40••,44•,45•]. Various tools for metabolic modeling have been established and refined. These include dynamic modeling, optimization, game-theoretical methods, FBA, Metabolic Pathway Analysis and others (cf. Introduction).

FBA is based on optimality principles (Box 1). It is a matter of debate whether FBA always gives sufficiently correct results [46,47], see below. Various objective functions have been compared by [48•]. However, more

Figure 1



Current Opinion in Biotechnology

Phylogenetic relationship between the organisms for which metabolic network reconstructions exist, generated with iTOL [84].

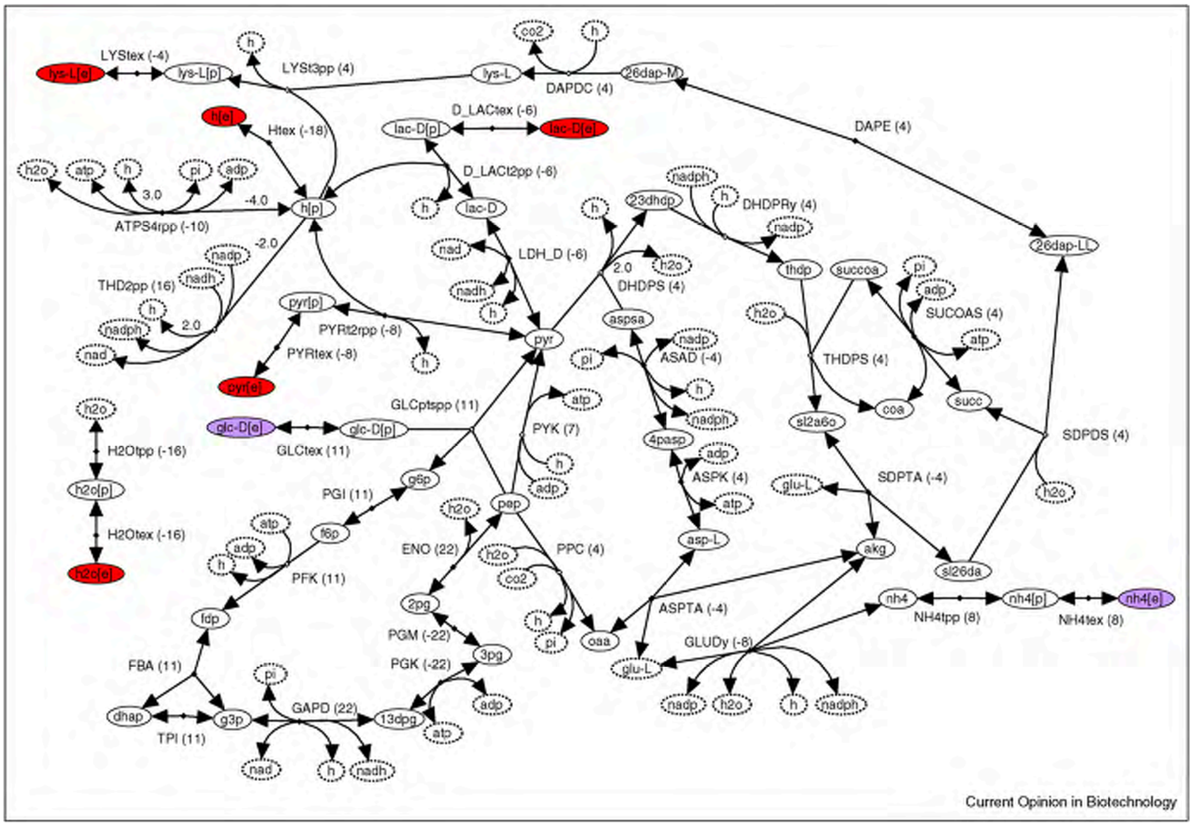
sophisticated methods usually require kinetic data, with the above-mentioned problem of incomplete data availability. Some current approaches attempt to guess flux values from thermodynamics [49], purely structural properties [45•] or RNA transcript data [23••].

A useful tool in Metabolic Pathway Analysis is based on the concept of EFMs [9••,10••]. An EFM is a minimal set of enzymes that can operate at steady-state such that all irreversible reactions involved proceed in the thermodynamically favored direction (Box 1). The related concept of minimal T-invariants had been established earlier in Petri net theory (cf. [50]).

Maximal yields can be computed by EFMs or by FBA. The latter methodology is particularly suitable in large-scale networks, in which EFM analysis meets the problem of combinatorial explosion. In small-sized and moderate-sized networks, in contrast, the set of EFMs provides a more comprehensive overview of the network's metabolic capabilities because it also comprises suboptimal pathways and pathways optimal with respect to other

E Ruppin, JA Papin, LF de Figueiredo and S Schuster. *Curr Opin Biotechnol*, 21(4):502-510, Aug 2010.

Figure 2



Current Opinion in Biotechnology

The shortest EFM producing lysine from glucose in *Escherichia coli* [53••]. The calculation was performed in the genome-scale network from [31••]. Violet nodes belong to the *in silico* growth medium, red nodes denote external metabolites. Duplicate nodes are dashed. Values in parentheses indicate reaction fluxes. For abbreviations, see [31••].

substrate-product pairs. Moreover, knockouts can easily by assessed by considering the remaining subset of EFMs. This sometimes leads to the counter-intuitive result that the average yield increases, as has been confirmed experimentally after the *in silico* analysis [8••].

By EFM analysis, previously unrecognized pathways can be detected (cf. [10••,55•], see also Box 2). Recently, we proposed two methods for pathway prediction in large-scale networks [40••,53••]. We computed the 10 shortest EFMs producing lysine in the genome-scale networks of *E. coli* and *Corynebacterium glutamicum* [53••] (Figure 2). Moreover, EFM analysis allows the quantification of robustness (see below).

There is a growing effort to use network models to identify drug targets and characterize mechanisms of disease. A recent study reconstructed and analyzed the metabolic networks of multiple strains of *Staphylococcus aureus* to identify novel drug targets [56]. A network-based pipeline

Box 2 The EFM method has manifold applications in biotechnology. First, it allows one to compute maximal molar yields (product-to-substrate ratios). For example, a previously undescribed pathway of efficient conversion of carbohydrate to oil in developing green plant seeds was detected [51••]. That pathway involves the pentose-phosphate pathway and the RUBISCO enzyme and provides 20% more acetyl-CoA for fatty acid synthesis than glycolysis. Trinh *et al.* [52•] designed, initially *in silico*, an *E. coli* strain with eight gene KO mutations. By EFM analysis, four pathways with non-growth-associated conversion of pentoses and hexoses into ethanol (important for biofuel production) at maximum theoretical yields and two pathways with tight coupling of growth with ethanol formation at high yields were obtained. Thereafter, they verified in experiment that the ethanol yields of the engineered strains closely matched the theoretical predictions. A third example is the EFM analysis of the synthesis of the commercially important amino acid, lysine (Figure 2) [7,53••,54•]. Depending on the bacterial species and on whether ATP was assumed to be sufficiently available, different maximum lysine-over-glucose yield values have been computed, for example 9:11 in *Corynebacterium glutamicum* when ATP must be regenerated by part of the glucose resource [54•]. There are many more studies in which EFMs were used, see [8••] for a recent review.

for identifying potential antimicrobials is being developed [57]. The human metabolic network reconstruction was analyzed to identify alternative enzyme targets for treating hyperlipidemia [39••]. It has also been recently used to predict biomarker changes characterizing a large set of different genetically inherited metabolic disorders [58].

Using metabolic models to study basic biological questions

Metabolic games

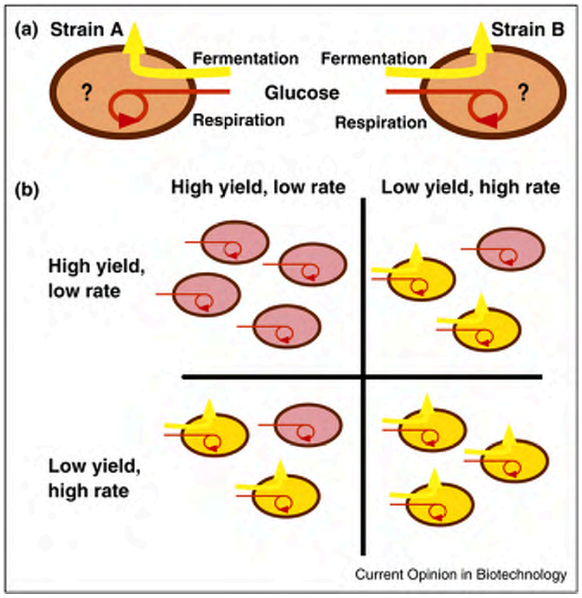
As has been seen above with FBA and EFMs, the concept of optimality has turned out to be extremely useful in understanding biological systems. Traditional optimization is, however, often insufficient for a deeper understanding of evolution. It usually neglects that the properties of the environment can change, and this in turn can change the optimal strategy. This is particularly important if the environment includes coevolving competitors that optimize their own metabolic capabilities. A mathematical framework to describe coevolution is provided by game theory (cf. [59]).

An illustrative example of the importance of competition in pathway evolution is the interplay between fermentation and respiration in ATP production [60••]. Several organisms and cell types such as *S. cerevisiae* and *E. coli* use respiro-fermentation at high glucose levels for degrading glucose to produce ATP, that is, respiration at maximum rate and fermentation in addition (Figure 3a), while many wild yeasts such as *Kluyveromyces marxianus* do fully respire glucose under aerobic conditions [61]. Respiration has a higher ATP-over-glucose yield but a relatively low rate in comparison to fermentation. When two species or strains compete for the same substrate, a typical game-theoretical situation arises (Figure 3b). The fitness of either organism depends not only on its own strategy (pathway usage in this case) but also on that of the other because both strategies affect the common nutrient pool.

To apply game theory requires little kinetic information. Not much more than the constraints (1) and (2) in Box 1 are necessary. In the example of fermentation and respiration, the upper limits given by relation (2) have to be chosen appropriately. In addition, the different yields of the pathways must be considered.

In order to use external glucose as economically as possible, it would be best if all organisms in a given habitat opted for respiration. The evolutionary reason for the profligate utilization of glucose by baker's yeast is that it thus out-competes organisms that operate more economically. In the terminology of game theory (cf. [59]), yeast cells are trapped in a Nash equilibrium (stable solution of a game) of a Prisoner's Dilemma (cf. [62•]). In the light of these results, FBA should be critically re-visited [46,47] although it was very successful in many cases [14,15,63••]. If maximum yield were the relevant criterion for the choice of pathway, respiration would always be chosen by *S. cerevisiae*. Fermentation was predicted only by FBA when additional constraints or specially tailored optimization principles were used [64]. Game-theoretical approaches can help predict flux distributions without additional corrections.

Figure 3



Game played by *S. cerevisiae* when growing at high glucose level under aerobic conditions. **(a)** The strains face the decision as to whether they use respiration (red arrow) or respiro-fermentation (red and yellow arrows). The higher rate of fermentation is indicated by a thicker arrow. **(b)** Schematic picture of the payoff matrix for this game. The best solution is to use respiration in order to have a higher yield. However, the respiro-fermenters would grow faster.

Several other metabolic systems have been analyzed by game theory, such as distinct regimes of glycolysis [65] and metabolic strategies in biofilms [66]. *S. cerevisiae* is involved in yet another interesting game. It concerns the extracellular enzyme, invertase, which generates glucose [67•]. A cheating strategy is to take up glucose while saving the metabolic costs of production and secretion of invertase. Gore *et al.* [68••] showed by experiments and a mathematical model that a stable coexistence between invertase-secreting and non-secreting yeast cells can be established (for an alternative mathematical model and the biotechnological relevance, see [69]). When the metabolic effort for exoenzyme production is low, all cells cooperate (harmony game); at intermediate costs, cooperators and cheaters coexist (hawk–dove game), while at high costs, all cells use the cheating strategy (Prisoner's Dilemma) [68••,69].

Several other biochemical examples including photosynthesis have been reviewed in [62•]. Moreover, the concept of Shapley value from game theory has been used in quantifying metabolic robustness [70], see below.

Robustness and its evolution

A general feature of living cells is their robustness to varying environmental conditions and genetic mutations. As metabolic network models provide an exciting opportunity to study genotype–phenotype relations on a genome scale, CBM models (and metabolic models in general) have been successfully used to study many facets of this fundamental relationship [71•,72••,73,74]. These studies have mainly asked two basic questions: Firstly, how did genetic robustness evolve? Does it have a direct adaptive value, or is it a consequence of environmental robustness, or perhaps just a side effect of other network properties? Secondly, what network mechanisms underlie the observed robustness — is it primarily due to gene duplications, to alternative metabolic pathways, or related to untested environmental conditions?

Since employing FBA in an exhaustive search of all gene knockout combinations cannot proceed beyond combinations of four knocked out (KO) genes, Deutscher et al. [75••] used a probabilistic approach. Thus, gene sets providing mutual functional backup until the depth of eight could be cataloged for *S. cerevisiae*. This has enabled them to characterize the '*k*-robustness' (the depth of backup interactions) of each gene, revealing that almost three quarters of yeast metabolic genes do participate in processes that are essential to growth in a standard laboratory environment, compared with only 13% previously found to be essential using single KOs. Optimization-based procedures for the exhaustive identification of multi-gene backup sets in genome-scale metabolic models have been recently developed [76], revealing new avenues available for redirecting metabolism, and uncovering complex patterns of gene interdependence. On the reverse side, genetic robustness may markedly hamper classical genetic studies using KOs to identify gene functionality, due to backups. Using the concept of Shapley value, Deutscher et al. [70] have shown that when assigning gene contributions for individual metabolic functions (such as the production of a given amino acid), the picture arising from single-perturbations is severely lacking and a multiple-perturbations approach turns out to be essential. Metabolic robustness under multiple KOs has also been studied in CBMs of several cell types by developing a robustness measure [73]. That measure is based on the ratio of the number of remaining EFMs after KO and the number of EFMs in the unperturbed situation.

Genetic interactions and network organization

By systematically generating double KOs of nonessential genes and assessing the resulting growth rate (fitness) of the organism, geneticists have traditionally identified both positive (alleviating) and negative (aggravating) genetic interactions, which has been a traditional tool for discovering functional relationships between genes. A comprehensive experimental screening for this in a whole organism is currently underway for yeast [77••]. Naturally, CBM models offer an opportunity to carry out such screens *in silico*. Segrè *et al.* [78] computed growth phenotypes of all single and double KOs of metabolic genes in *S. cerevisiae*, using FBA. The ensuing genetic interaction network could be clustered into modules composed of genes interacting with each other 'monochromatically' (i.e. with purely aggravating or purely alleviating interactions), emphasizing interactions between, rather than within, functional modules. Harrison *et al.* [79••] investigated the extent to which the functional impact of single and double KOs in yeast changes across different environments, employing FBA across 53 different conditions. The synthetic lethal (SL) predictions of the model were then validated by an *in vivo* double gene KO experiment and by literature search. The strong context dependency of the pattern of SL interactions observed suggests that the environment plays an important role in shaping genetic robustness.

From unicellular to multi-cellular organisms

The vast majority of the work on metabolic CBM performed up until now has focused on unicellular organisms. Naturally, in recent years, there have been attempts at extending these methods to study the metabolism of multicellular and multi-tissue organisms, a considerably greater challenge. CBM reconstructions of human metabolic networks were performed up to 2007 only for cell types and organelles with a very limited scope of metabolism [80,81•,82••]. A fundamental step forward has been presented in recent reconstructions of the global, generic human metabolic network based on an extensive evaluation of genomic and bibliomic data [38••,39••]. These networks include ~3000 reactions, ~2000 metabolites, and ~1500 genes mapped to the different reactions over 7 organelles. The generic model of [39••] helped identify a set of functionally related reactions involving glutathione metabolism that were causally related to hemolytic anemia, and another set of functionally related enzymes containing HMG-CoA reductase, a common target for the cholesterol lowering statins. This model, however, is not tissue or cell specific. More recent efforts have presented methods for inferring context-specific networks [23••,24], which can be utilized to infer large-scale descriptions of the human tissues' metabolism. Accordingly, Shlomi *et al.* [23••] have integrated tissue-specific gene and protein expression data to predict and validate versus publicly available data for the tissue-specific metabolic activity for 10 human tissues, identifying that post-transcriptional regulation plays a central role in shaping tissue-specific metabolic activity. Very recently, an extended approach of the latter has been used to build and study the first large-scale model of liver metabolism [83••].

Concluding remarks and future directions/ challenges

CBM methods are very useful for understanding the complex architecture of metabolism and for manifold biotechnological and medical applications. Even if kinetic parameters were to become available, an analysis of the network properties using tools of FBA and Metabolic Pathway Analysis often provides valuable insight before performing a dynamic simulation. As outlined above, the integration of omics data of different types into metabolic models has had much success. Nevertheless, its refinement and scaling-up certainly remains a challenge. Cell-specific and tissue-specific studies can now be performed for those multi-cellular organisms for which metabolic reconstructions are available, as is already being done for humans. Whole-cell modeling has also raised philosophical issues on what level of completeness can be reached in modeling.

More work in this field is also needed to study emergent properties, which is at the heart of Systems Biology, after the necessary assembly of the network constituents has been done. Another direction is Synthetic Biology. Specific goals are the design of minimal metabolisms (depending very much on the given set of nutrients) and minimal genomes. This could help design efficient microbes for biosyntheses. Game-theoretical methods, in particular, are helpful in assessing the impact of 'cheater' mutants in bioreactors, which may impair productivity.

Overall, one can safely maintain that the field of genome-scale metabolic modeling has undergone a tremendous development and growth in the last decade, in terms of the organisms spanned, the methodologies developed, and the themes covered. Certainly, if there is one specific field in systems biology where we have made significant strides towards the holy grail of generating a working cell *in silico*, this is the one.

Acknowledgements

Financial support from the following agencies is gratefully acknowledged: German-Israeli Foundation to ER and SS, Israeli Science Foundation (ISF) to ER, National Science Foundation CAREER award (0643548) to JP and German Ministry of Education and Research (BMBF) to SS. Financial support from Portuguese entities: Fundação Calouste Gulbenkian, Fundação para a Ciência e a Tecnologia (FCT) and Siemens SA Portugal (PhD grant SFRH/BD/32961/2006) to LF de F. The authors thank Christoph Kaleta and Steffen Klamt for helpful discussions.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Barnett JAG: **Glucose catabolism in yeast and muscle**. In *Selected Topics in the History of Biochemistry: Personal Recollections IX*, vol 44. Edited by Semenza G, Turner AJ. Elsevier; 2005:1-132.

This paper provides a historical perspective on glycolysis, being the first metabolic pathway formulated.

2. Stephanopoulos GN, Aristidou AA, Nielsen J (Eds): *Metabolic Engineering. Principles and Methodologies*. San Diego: Academic Press; 1998.

3. • Heinrich R, Rapoport SM, Rapoport TA: **Metabolic regulation and mathematical models**. *Prog Biophys Mol Biol* 1977, **32**:1-82.

An early review on mathematical modeling of metabolism.

4. Heinrich R, Schuster S: *The Regulation of Cellular Systems*. New York: Chapman and Hall; 1996.

5. Palsson BØ: *Systems Biology: Properties of Reconstructed Networks* Cambridge: Cambridge University Press; 2006.

6. Papin J, Stelling J, Price N, Klamt S, Schuster S, Palsson BØ: **Comparison of network-based pathway analysis methods**. *Trends Biotechnol* 2004, **22**:400-405.

7. Schuster S, von Kamp A, Pachkov M: **Understanding the roadmap of metabolism by pathway analysis**. *Methods Mol Biol* 2007, **358**:199-226.

8. •• Trinh CT, Wlaschin A, Srienc F: **Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism**. *Appl Microbiol Biotechnol* 2009, **81**:813-826.

Excellent review on elementary-mode analysis focusing on its biotechnological applications.

9. •• Schuster S, Hilgetag C: **On elementary flux modes in biochemical reaction systems at steady state**. *J Biol Syst* 1994, **2**:165-182.

Pioneering paper, in which elementary flux modes were introduced.

10. •• Schuster S, Dandekar T, Fell DA: **Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering**. *Trends Biotechnol* 1999, **17**:53-60.

Clear introduction into EFM analysis, with illustrative examples. Moreover, in this paper, a catabolic pathway comprising part of the TCA cycle, glyoxylate shunt and PEP carboxykinase had been predicted for *E. coli*. This prediction was confirmed experimentally later [55•].

11. •• Schilling CH, Letscher D, Palsson BØ: **Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective**. *J Theor Biol* 2000, **203**:229-248.

First paper on extreme pathways, to reduce the number of pathways necessary to describe the metabolic capabilities of a network.

12. •• Fell DA, Small JR: **Fat synthesis in adipose tissue. An examination of stoichiometric constraints**. *Biochem J* 1986, **238**:781-786.

Pioneering paper on FBA (although that term was not yet used then), including application to lipid metabolism.

13. •• Varma A, Palsson BØ: **Metabolic capabilities of *Escherichia coli*. I. Synthesis of biosynthetic precursors and cofactors**. *J Theor Biol* 1993, **165**:477-502.

An early paper in FBA, with mathematical foundations and illustrative examples.

14. Price ND, Reed JL, Palsson BØ: **Genome-scale models of microbial cells evaluating the consequences of constraints**. *Nat Rev Microbiol* 2004, **2**:886-897.

15. Lee JM, Gianchandani EP, Papin JA: **Flux balance analysis in the era of metabolomics**. *Brief Bioinform* 2006, **7**:140-150.

16. •• Çakır T, Patil KR, Önsan Zİ, Ülgen KÖ, Kırdar B, Nielsen J: **Integration of metabolome data with metabolic networks reveals reporter reactions**. *Mol Syst Biol* 2006, **2**:50.

Addresses the important issue of metabolic versus hierarchical control of metabolic fluxes, in an integrative computational and experimental manner.

17. Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T: **Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model**. *ISMB 2010 Bioinform* 2010, **26**:i255-i260.

18. •• Sauer U: **Metabolic networks in motion: ^{13}C-based flux analysis**. *Mol Syst Biol* 2006, **2**:62.

Excellent review on flux measurement and analysis.

19. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57-63.

20. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ:
•• **Reconstruction of biochemical networks in microorganisms**. *Nat Rev Microbiol* 2009, **7**:129-143.
Excellent review on metabolic reconstructions.

21. Oberhardt MA, Palsson BØ, Papin JA: **Applications of genome-**
•• **scale metabolic reconstructions**. *Mol Syst Biol* 2009, **5**:320.
Excellent review on metabolic reconstructions, listing the coverage of species.

22. Lewis NE, Cho BK, Knight EM, Palsson BØ: **Gene expression profiling and the use of genome-scale in silico models of *Escherichia coli* for analysis: providing context for content**. *J Bacteriol* 2009, **191**:3437-3444.

23. Shlomi T, Cabili MN, Herrgård MJ, Palsson BØ, Ruppin E:
•• **Network-based prediction of human tissue-specific metabolism**. *Nat Biotechnol* 2008, **26**:1003-1010.
The first paper presenting a distinct set of specific metabolic descriptions of a variety of human tissues, which are validated via comparison to pertaining tissue-specific data on the activity of disease genes, among others.

24. Becker SA, Palsson BØ: **Context-specific metabolic networks are consistent with experiments**. *PLoS Comput Biol* 2008, **4**:e1000082.

25. Manichaikul A, Ghamsari L, Hom EFY, Lin C, Murray RR,
•• Chang RL, Balaji S, Hao T, Shen Y, Chavali AK *et al.*: **Metabolic network analysis integrated with transcript verification for sequenced genomes**. *Nat Methods* 2009, **6**:589-592.
Influential paper delineating a pipeline for the iterative refinement of the networks with high-throughput transcript verification.

26. Pfeiffer T, Schuster S: **Game-theoretical approaches to studying the evolution of biochemical systems**. *Trends Biochem Sci* 2005, **30**:20-25.

27. Thiele I, Palsson BØ: **A protocol for generating a high-quality genome-scale metabolic reconstruction**. *Nat Protoc* 2010, **5**:93-121.

28. Poolman MG, Bonde BK, Gevorgyan A, Patel HH, Fell DA:
• **Challenges to be faced in the reconstruction of metabolic networks from public databases**. *IEE Proc Syst Biol* 2006, **153**:379-384.
Databases are central to the reconstruction of genome-scale networks. This article points out important errors present in databases which render the reconstruction process difficult.

29. Gevorgyan A, Poolman MG, Fell DA: **Detection of stoichiometric inconsistencies in biomolecular models**. *Bioinformatics* 2008, **24**:2245-2251.

30. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD: **Flux**
•• **coupling analysis of genome-scale metabolic network reconstructions**. *Genome Res* 2004, **14**:301-312.
Extending the concept of 'enzyme subsets', this paper introduces various refined concepts such as 'partially coupled reactions', taking into account irreversibility. This opened a new methodology — flux coupling analysis, and allowed the determination of blocked reactions.

31. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR,
•• Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**. *Mol Syst Biol* 2007, **3**:121.
Widely used genome-scale metabolic reconstruction of *E. coli*.

32. Duarte NC, Herrgård MJ, Palsson BØ: **Reconstruction and**
•• **validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model**. *Genome Res* 2004, **14**:1298-1309.
Widely used genome-scale metabolic reconstruction of *S. cerevisiae*.

33. May P, Christian J-O, Kempa S, Walther D: **ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii***. *BMC Genomics* 2009, **10**:209.

34. de Oliveira Dal'Molin CG, Quek L-E, Palfreyman RW, Brumbley SM, Nielsen LK: **AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis**. *Plant Physiol* 2010, **152**:579-589.

35. Poolman MG, Miguet L, Sweetlove LJ, Fell DA: **A genome-scale metabolic model of *Arabidopsis thaliana* and some of its properties**. *Plant Physiol* 2009, **151**:1570-1581.

36. Sheikh K, Förster J, Nielsen LK: **Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus***. *Biotechnol Prog* 2005, **21**:112-121.

37. Selvarasu S, Karimi IA, Ghim GH, Lee D-Y: **Genome-scale modeling and in silico analysis of mouse cell metabolic network**. *Mol Biosyst* 2010, **6**:152-161.

38. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I:
•• **The Edinburgh human metabolic network reconstruction and its functional analysis**. *Mol Syst Biol* 2007, **3**:135.
Widely used genome-scale reconstruction of human metabolism.

39. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD,
•• Srivas R, Palsson BØ: **Global reconstruction of the human metabolic network based on genomic and bibliomic data**. *Proc Natl Acad Sci U S A* 2007, **104**:1777-1782.
A seminal, pioneering study charting out the first genome-scale, generic model of human metabolism, and showing its potential utility for disease modeling and drug target identification.

40. Kaleta C, de Figueiredo LF, Schuster S: **Can the whole be less**
•• **than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns**. *Genome Res* 2009, **19**:1872-1883.
The new concept of elementary flux pattern can be used to find routes in subsystems that are compatible with flux distributions in genome-scale metabolic networks and, thus, to detect previously unknown pathways.

41. Reed JL, Famili I, Thiele I, Palsson BØ: **Towards multidimensional genome annotation**. *Nat Rev Genet* 2006, **7**:130-141.

42. Herrgård MJ, Fong SS, Palsson BØ: **Identification of genome-scale metabolic network models using experimentally measured flux profiles**. *PLoS Comput Biol* 2006, **2**:e72.

43. Schreiber F, Dwyer T, Marriott K, Wybrow M: **A generic algorithm for layout of biological networks**. *BMC Bioinformatics* 2009, **10**:375.

44. Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, Pinto JP,
• Nielsen J, Patil KR, Ferreira EC, Rocha M: **OptFlux: an open-source software platform for in silico metabolic engineering**. *BMC Syst Biol* 2010, **4**:45.
Describes a software that combines various methods for metabolic network analysis. It also gives an overview of other software packages for Systems Biology.

45. Hädicke O, Klamt S: **CASOP: a computational approach for**
• **strain optimization aiming at high productivity**. *J Biotechnol* 2010, **147**:88-101.
New method for designing genetic manipulation strategies, using elementary flux modes to weight the impact of blocking or increasing the flux through a given enzyme.

46. Teusink B, Wiersma A, Jacobs L, Notebaart RA, Smid EJ: **Understanding the adaptive growth strategy of *Lactobacillus plantarum* by in silico optimisation**. *PLoS Comput Biol* 2009, **5**:e1000410.

47. Schuster S, Pfeiffer T, Fell DA: **Is maximization of molar yield in metabolic networks favoured by evolution?** *J Theor Biol* 2008, **252**:497-504.

48. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of**
• **objective functions for predicting intracellular fluxes in *Escherichia coli***. *Mol Syst Biol* 2007, **3**:119.
Eleven objective functions in FBA are compared, showing that no single objective describes the flux states under all conditions.

49. Boghigian BA, Shi H, Lee K, Pfeifer BA: **Utilizing elementary mode analysis, pathway thermodynamics, and a genetic algorithm for metabolic flux determination and optimal metabolic network design**. *BMC Syst Biol* 2010, **4**:49.

50. Grafahrend-Belau E, Schreiber F, Heiner M, Sackmann A, Junker BH, Grunwald S, Speer A, Winder K, Koch I: **Modularization of biochemical networks based on classification of Petri net t-invariants**. *BMC Bioinform* 2008, **9**:90.

51. Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y: **Rubisco**
•• **without the Calvin cycle improves the carbon efficiency of developing green seeds**. *Nature* 2004, **432**:779-782.
Excellent application of elementary-mode analysis to plant metabolism, leading to discovery of previously undescribed, highly efficient pathway.

52. • Trinh CT, Unrean P, Srienc F: **Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses**. *Appl Environ Microbiol* 2008, **74**:3634-3643.

Convincing application of EFM analysis to a system of high biotechnological relevance, by predicting and verifying the effect of KOs. The design of such streamlined metabolic functionality is interesting in view of Synthetic Biology.

53. •• de Figueiredo LF, Podhorski A, Rubio A, Kaleta C, Beasley JE, Schuster S, Planes FJ: **Computing the shortest elementary flux modes in genome-scale metabolic networks**. *Bioinformatics* 2009, **25**:3158-3165.

The first method computing elementary flux modes in genome-scale networks, with application to lysine synthesis in *E. coli* and *C. glutamicum*.

54. • Wittmann C, Becker J: **The L-lysine story: from metabolic pathways to industrial production**. In *Amino Acid Biosynthesis — Pathways, Regulation and Metabolic Engineering.* Edited by Wendisch VF. Springer; 2007:39-70.

Nice overview of pathway analysis of lysine synthesis.

55. • Fischer E, Sauer U: **A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli***. *J Biol Chem* 2003, **278**:46446-46451.

Experimental detection of the PEP-glyoxylate cycle predicted theoretically earlier, for example, by [10••].

56. Lee DS, Burd H, Liu J, Almaas E, Wiest O, Barabási A-L, Oltvai ZN, Kapatral V: **Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets**. *J Bacteriol* 2009, **191**:4015-4024.

57. Shen Y, Liu J, Estiu G, Isin B, Ahn Y-Y, Lee D-S, Barabási A-L, Kapatral V, Wiest O, Oltvai ZN: **Blueprint for antimicrobial hit discovery targeting metabolic networks**. *Proc Natl Acad Sci U S A* 2010, **107**:1082-1087.

58. Shlomi T, Cabili MN, Ruppin E: **Predicting metabolic biomarkers of human inborn errors of metabolism**. *Mol Syst Biol* 2009, **5**:263.

59. Gintis H: *Game Theory Evolving*. Princeton: Princeton University Press; 2009.

60. •• Pfeiffer T, Schuster S, Bonhoeffer S: **Cooperation and competition in the evolution of ATP-producing pathways**. *Science* 2001, **292**:504-507.

A pioneering paper on game-theoretical analysis of metabolic networks.

61. Veiga A, Arrabaça JD, Maria C, Loureiro-Dias MC: **Cyanide-resistant respiration is frequent, but confined to yeasts incapable of aerobic fermentation**. *FEMS Microbiol Lett* 2000, **190**:93-97.

62. • Schuster S, Kreft J-U, Schroeter A, Pfeiffer T: **Use of game-theoretical methods in biochemistry and biophysics**. *J Biol Phys* 2008, **34**:1-17.

Overview of numerous applications of game theory.

63. •• Fong SS, Palsson BØ: **Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes**. *Nat Genet* 2004, 36:1056-1058.

This paper provided significant experimental evidence for the principle of optimality in metabolic networks as implemented in Flux Balance Analysis.

64. Çakır T, Efe C, Dikicioglu D, Hortacsu A, Kirdar B, Oliver SG: **Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains**. *Biotechnol Prog* 2007, **23**:320-326.

65. Aledo JC, Pérez-Claros JA, Esteban del Valle A: **Switching between cooperation and competition in the use of extracellular glucose**. *J Mol Evol* 2007, **65**:328-339.

66. Kreft JU: **Biofilms promote altruism**. *Microbiology* 2004, **150**:2751-2760.

67. • Greig D, Travisano M: **The Prisoner's Dilemma and polymorphism in yeast SUC genes**. *Proc R Soc B* 2004, **271**:S25-S26.

One of the first experimental studies on game-theoretical aspects of metabolism.

68. •• Gore J, Youk H, van Oudenaarden A: **Snowdrift game dynamics and facultative cheating in yeast**. *Nature* 2009, **459**:253-256.

Excellent combination of experimental and game-theoretical approaches to studying exoenzyme production.

69. Schuster S, Kreft J-U, Brenner N, Wessely F, Theissen G, Ruppin E, Schroeter A: **Cooperation and cheating in microbial exoenzyme production — theoretical analysis for biotechnological applications**. *Biotechnol J* 2010 doi: 10.1002/biot.200900303.

70. Deutscher D, Meilijson I, Schuster S, Ruppin E: **Can single knockouts accurately single out gene functions?** *BMC Systems Biol* 2008, **2**:50.

71. • Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J: **Robustness of cellular functions**. *Cell* 2004, **118**:675-685.

Very good review on robustness of intracellular networks.

72. •• Blank LM, Kuepfer L, Sauer U: **Large-scale ^{13}C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast**. *Genome Biol* 2005, **6**:R49.

A combined experimental and computational account aiming to decipher the underlying mechanisms of metabolic network robustness by comparing flux distributions in wild-type and knockout yeast strains.

73. Behre J, Wilhelm T, von Kamp A, Ruppin E, Schuster S: **Structural robustness of metabolic networks with respect to multiple knockouts**. *J Theor Biol* 2008, **252**:433-441.

74. Freilich S, Kreimer A, Borenstein E, Gophna U, Sharan R, Ruppin E: **Decoupling environment-dependent and independent genetic robustness across bacterial species**. *PLoS Comput Biol* 2010, **6**:e1000690.

75. •• Deutscher D, Meilijson I, Kupiec M, Ruppin E: **Multiple knockout analysis of genetic robustness in the yeast metabolic network**. *Nat Genet* 2006, **38**:993-998.

The first study to chart the backup architecture of genes using very large *in silico* knockout combinations and associate the latter with the genes' conservation and functional specificity.

76. Suthers PF, Zomorrodi A, Maranas CD: **Genome-scale gene/reaction essentiality and synthetic lethality analysis**. *Mol Syst Biol* 2009, **5**:301.

77. •• Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S *et al.*: **The genetic landscape of a cell**. *Science* 2010, **327**:425-431.

Comprehensive experimental study of double KOs and modularity in yeast.

78. Segrè D, DeLuna A, Church GM, Kishony R: **Modular epistasis in yeast metabolism**. *Nat Genet* 2005, **37**:77-83.

79. •• Harrison R, Papp B, Pál C, Oliver SG, Delneri D: **Plasticity of genetic interactions in metabolic networks of yeast**. *Proc Natl Acad Sci U S A* 2007, **104**:2307-2312.

A combined computational and experimental study showing the large extent of context-dependency of genetic interactions on the organism's environment.

80. Chan C, Berthiaume F, Lee K, Yarmush ML: **Metabolic flux analysis of cultured hepatocytes exposed to plasma**. *Biotechnol Bioeng* 2003, **81**:33-49.

81. • Vo TD, Greenberg HJ, Palsson BØ: **Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data**. *J Biol Chem* 2004, **279**:39532-39540.

Detailed reconstruction of the metabolic network in a cell organelle.

82. •• Thiele I, Price ND, Vo TD, Palsson BØ: **Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet**. *J Biol Chem* 2005, **280**:11683-11695.

One of the first demonstrations of the potential use of CBM models for modeling and studying human disease conditions, employing sampling methods to track the flux profile changes after different perturbations.

83. •• Jerby L, Shlomi T, Ruppin E: **Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism**. *Mol Syst Biol* 2010:6 doi: 10.1038/msb.2010.56.

Differing from [23••], this study presents a general approach for building tissue models, that can be perturbed to study the metabolic implications of disorders and other context-dependent alterations.

84. Letunic I, Bork P: **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation**. *Bioinformatics* 2007, **23**:127-128.

Chapter 6

Discussion

This thesis presents a collection of four papers dedicated to metabolic pathway analysis using the network stoichiometries to predict metabolic pathways. An additional review paper outlines the recent developments in metabolic network reconstruction, constraint-based modelling and game theory, and gives a perspective for further achievements. In the first part of the work, I evaluated some of the recently published methods for metabolic pathway prediction based on graph theory with elementary flux mode analysis. This analysis gave me a clear overview of the advantages and limitations of both modelling frameworks.

From the comparison performed in Chapter 2 it is evident that the tested tools based on graph theory can tackle the computational complexity of pathway prediction in large-scale metabolic networks and even at the database level but they are not able to correctly answer relevant biochemical questions. Some of these questions are part of the history of biochemistry. Thus, topological properties alone, without chemical constraints, are not enough to predict metabolic pathways. This was pointed out already by Arita (2004) and it is becoming more apparent in the literature with the application of these tools to biotechnological problems (Ranganathan and Maranas, 2010). Graph-theoretical methods give sometimes incorrect results. In addition to the examples shown in Chapter 2 and discussed below, there is a recently publihsed example: in the thiobutanoate pathway predicted by (Ranganathan and Maranas, 2010) there is no single carbon atom from pyruvate reaching the end product, 1-butanol. Nevertheless, these issues involving the use of graph-theoretical methods for pathway prediction remain a matter of debate (Faust *et al.*, 2009a; de Figueiredo *et al.*, 2009c).

In order to improve the quality of the predicted paths, one has to add more information to the graph representation, like the similarity between the chemical structure of metabolites connected by a given reaction. Indeed, this has been the strategy adopted in the improved version of PathFinding (Faust *et al.*, 2009b) and in the recently published MetaRoute (Blum and Kohlbacher, 2008a,b). In general the algorithms for computing chemical structure similarity are NP-hard, albeit the existence of heuristics to solve this problem in polynomial time (Raymond *et al.*, 2002; Hattori *et al.*, 2003; Akutsu, 2004). However, the mapping rules for a given reaction are not unique (Akutsu, 2004). This could be one explanation for the incorrect pathways predicted by Pathway Hunter Tool (Rahman *et al.*, 2005). Additionally, it is confirmed that there were some issues relative to the software used for the representation of chemical structures (Dr. Rahman, personal communication). Blum and Kohlbacher (2008b) overcome the issue relative to the non-uniqueness of the mapping rules by clustering reactions according to the Enzyme Commission (EC) number. The EC number is a nomenclature developed by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it corresponds to a hierarchical classification of enzymatic reactions according to the reaction mechanism (Webb, 1992). Reactions with the same first three EC digits have to have the same reaction mechanism and therefore, the mapping rules should be the same. In the long-term, the approach of Blum and Kohlbacher (2008b) will face a major drawback because the assignment of new EC numbers is hampered by the requirement of experimental validation. Thus, many of the recently discovered enzymes will not have an assigned EC number (cf. Kotera *et al.*, 2004; Egelhofer *et al.*, 2010). Moreover, the assignment of incomplete EC numbers generates incorrect annotations (Green and Karp, 2005;

Egelhofer *et al.*, 2010). More details on the EC number nomenclature limitations can be found in Babbitt (2003).

Faced with these limitations, the developers of the KEGG database found a different approach to characterize reaction mechanisms. The recently developed RPAIR database consists of chemical structure alignments between pairs of reactants, substrate - product, and of chemical structure transformation patterns occurring in all enzyme-catalyzed reactions present in KEGG (Oh *et al.*, 2007). The chemical structure alignments are computationally generated and subsequently, manually curated. Recently, Faust *et al.* (2009b) have shown that the predictions from PathFinding can be improved when a graph generated from these reaction pairs stored in RPAIR database is used as input network. In this graph, shortcuts through cofactors, such as ATP and NADH, are avoided and consequently, the prediction of metabolic paths in the core of the metabolic network is improved. Similarly to what was said for MetaRoute, also the coverage of RPAIR is limited which imposes constraints in the size of system under study (Faust *et al.*, 2009b).

Another important point concerning the tools based on graph theoretical methods should be discussed, more precisely the evaluation of the predicted paths. This evaluation has been made so far, using the pathway information from KEGG or EcoCyc databases (Faust *et al.*, 2009b; Blum and Kohlbacher, 2008b). As mentioned in the Introduction, the higher functional level in these databases corresponds to metabolic pathways. This information has been acquired during the last century and corresponds to well characterized pathways. Thus, new metabolic pathways are not depicted in these databases. Moreover, the pathways depicted in KEGG or EcoCyc do not always contain information about the functional modes of metabolic networks. For example, the pentose phosphate pathway described in Stryer (1995) has four functional modes and in other textbooks has even less modes (Michal, 1999; Voet and Voet, 2004). In KEGG, this pathway is depicted as a single map and in EcoCyc as three maps, being one of them an integrated view of the non-oxidative and the oxidative branches. The theoretical predictions carried out with convex analysis show that the pentose phosphate pathway, when integrated with glycolysis and fructose-1,6-bisphosphatase (EC 3.1.3.11) has five functional modes (Schuster *et al.*, 2000). It is, therefore, important to know whether the efficiency of tools for metabolic pathways prediction can be accessed using only the description made in these databases or, on the other hand, a collection of biological questions can be used to address the efficiency of such tools, as we performed in Chapter 2.

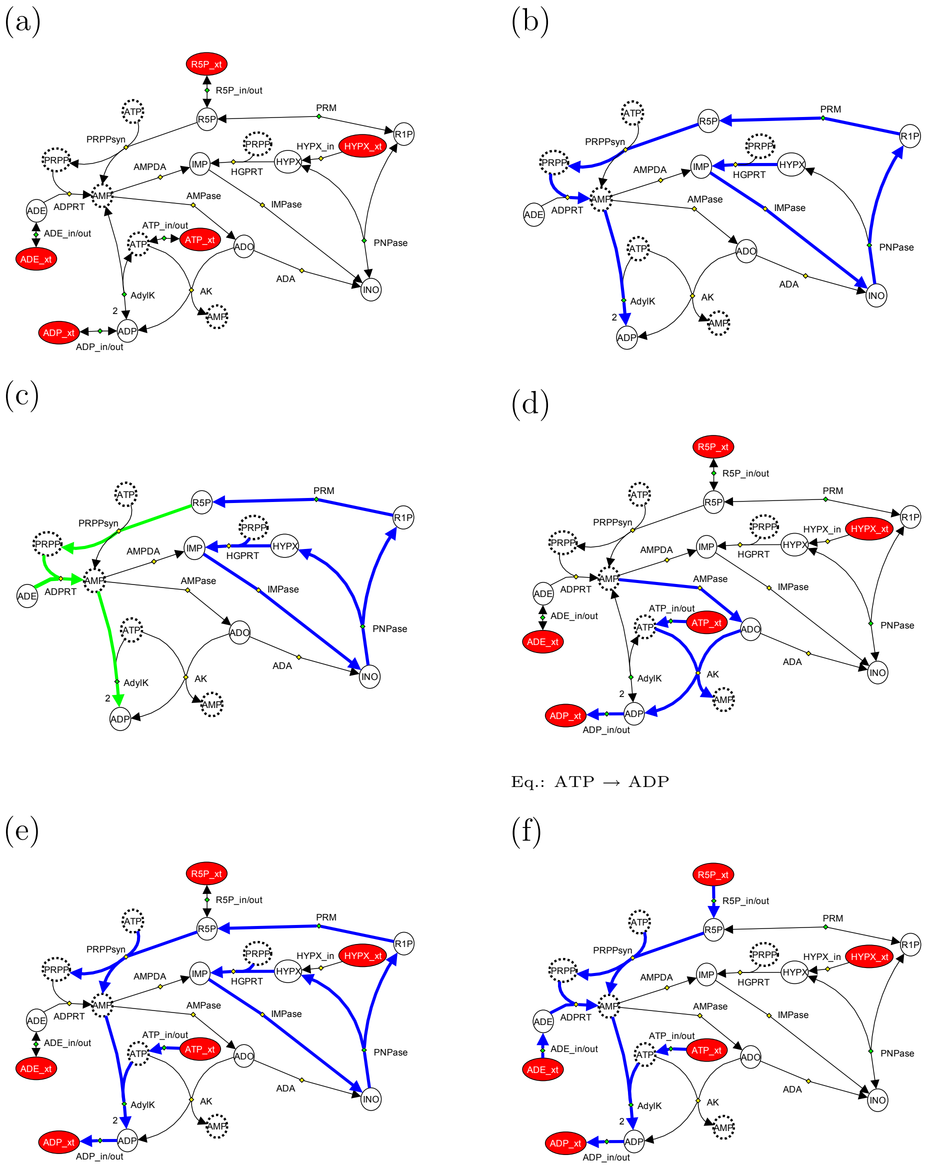
The three benchmark problems presented in Chapter 2 correspond to biological questions that were intensively discussed and for which there are abundant literature and experimental evidence to support their answer. The conversion of fatty acids to carbohydrates was discussed during the first half of the 20^{th} century, precisely at the time where pathways like glycolysis and the tricarboxylic acid cycle (TCA cycle) had been discovered (see Introduction). To correctly answer this question researchers brought together experimental results and algebra (Weinman *et al.*, 1957; Exton and Park, 1967; Heath, 1968). Another benchmark problem deals with the conversion of glucose to pyruvate in the absence of phosphofructokinase (EC 2.7.1.11), from glycolysis, and glucose 6-phosphate dehydrogenase (EC 1.1.1.49), from the pentose phosphate pathway. A route converting glucose to pyruvate in such a system was hypothesized by Pollack *et al.* (1997). However, we have shown that such pathway is not possible at steady state although there is a connected route between glucose and pyruvate (de Figueiredo *et al.*, 2009b). Moreover, this problem shows how relevant are constraint-based methods in the analysis of incomplete or recently sequenced genomes, helping in the annotation process.

The last benchmark problem deals with the metabolism of human erythrocytes for which elementary-flux mode analysis can be carried at the cell scale. In spite of the reduced metabolism of human erythrocytes, interesting questions can be formulated. For example, in erythrocytes the adenosine monophosphate (AMP) is irreversibly converted to hypoxantine leading to a decrease in the adenylate pool (AMP/ADP/ATP). In order to maintain this pool at a physiological required level, an inflow of adenine from the medium is required. Saying that, we can raise the following question, *Is it possible to convert hypoxantine back to the adenylate form?* This question underlies several experimental works (Salerno and Giacomello, 1985; Heptinstall *et al.*, 2005). Using elementary-flux mode analysis to study the system depicted in Figure 6.1 (a) we can confirm that it is impossible to convert hypoxanthine to adenosine diphosphate (ADP) (Figure 6.1 (d)-(f)). However, even the most recent version of PathFinding fails to answer such a question (Figure 6.1 (b)). The reason for this failure relies in the fact that the atom mapping information (i.e., the RPAIR information), is only used to establish the edges between metabolites and consequently the system does not have any memory recalling which metabolite provides the carbon backbone. For example, in the query for paths linking hypoxanthine to ADP (Figure 6.1 (b)), inosine shares a carbon backbone with hypoxantine but all the carbon atoms of ribose 1-phosphate (R1P) and present in inosine come from 5-phospho-α-D-ribose-1-diphosphate (PRPP). Thus, even the last version of PathFinding cannot be used for atom tracing. From elementary flux mode 2 (Figure 6.1 (e)) it is evident that the path predicted by PathFinding would only convert ATP to ADP, corresponding to a futile cycle. Additionally, the analysis of this subsystem shows that the oxypurine cycle is important for the salvage of hypoxanthine to inosine monophosphate (IMP) and that the PRPP pool controls this cycle (Berman and Human, 1990) (Figure 6.1 (e)).

An alternative method, also based in graph theory, is the recently introduced ReTrace tool (Pitkänen *et al.*, 2009). The results obtained with this tool (Figure 6.1 (c)) are very similar to the ones of elementary flux mode analysis. First, there is no path found from hypoxanthine to ADP. By querying this tool with other metabolites as source compounds, it is clear that adenine is required for ADP synthesis using a ribose ring, either from PRPP and consequently reducing the PRPP pool, or from ribose 5-phosphate (R5P), which can be obtained from central carbon metabolism. For more details about these simulations, see the Supplementary material (pages 139 ff). The success of this method relies in the atom-level representation of the metabolic network and in the recursive method to find branching pathways. Contrary to Pathway Hunter Tool and PathFinding approaches, ReTrace searches for pathways at the atom-level instead of the conventional reaction-metabolite graph. In this representation, nodes correspond to the atoms in a metabolite and the edges link the atoms between metabolites. Then, the search algorithm finds branching pathways transferring as many atoms as possible from the source to the target metabolite. Nevertheless, it would be interesting to test this tool with problems where indeed there is a carbon route from the source to the target metabolite but no net flux, such as the benchmark problem dealing with the conversion of even-chain fatty acids into carbohydrates.

The work presented here supports the recent paradigm shift in the understanding of the topology of metabolic networks. The representation of metabolic networks as graphs, without any further chemical constraints, may lead to incorrect interpretation of their structural properties. The initial studies on the topology of metabolic networks, have pointed out that these networks are scale-free and can be seen as a small-world (see Introduction). However, a careful look on the paths computed in such analysis, shows that some of the paths are not biologically relevant because they correspond to shortcuts through cofactors (Küffner *et al.*, 2000; Arita, 2004; Rahman *et al.*, 2005; de Figueiredo *et al.*, 2009b), or to paths where no carbon flux from the source to the target takes place (Arita, 2004; de Figueiredo *et al.*, 2009b). The small-world hypothesis is weakened when the chemical information embedded in metabolic networks is included in the path prediction method (Arita, 2004; Lima-Mendez and van Helden, 2009; Montañez *et al.*, 2010).

Tanaka (2005) suggested that metabolic networks are rather scale-rich instead of scale-free, based on the stoichiometric analysis of metabolic networks. The representation of metabolism through its stoichiometry is indeed an important



Eq.: ATP → ADP

Eq.: 2 ATP → 2 ADP

Eq.: ADE + R5P + 3 ATP → 4 ADP

Figure 6.1: Benchmark problem concerning the conversion of hypoxanthine to ADP in human erythrocytes as briefly outlined in de Figueiredo *et al.* (2009c) (a). Duplicated nodes are dashed and red nodes correspond to external metabolites. In the graph based approaches it is not required to define the system's boundary. The 2^{nd} shortest path computed with the improved version of PathFinding (Faust *et al.*, 2009b), using a directed RPAIR graph (b). The path from PRPP to R5P (blue arrows) together with the path from ADE and PRPP to ADP (green arrows) were computed with ReTrace (Pitkänen *et al.*, 2009). Indeed, the blue arrow corresponds to the exit of the ribose ring from the subsystem whereas the green arrow shows the sequestration of this ring in the adenylate pool. The system defined in (a) has only three elementary flux modes. Elementary flux modes (d) and (e) burn ATP to ADP and are usually named futile (or substrate) cycles. There is only one elementary flux mode producing ADP (f). See the Supplementary material (pages 139 ff.) for metabolite and reaction abbreviations.

alternative. This representation has been used by several tools in the field of systems biology (cf. Price *et al.*, 2004; Feist and Palsson, 2008). The major advantage of this approach is that the stoichiometry of a reaction network is easily accessible, whereas the similarity between chemical structures or atom maps have to be computed *a priori*. The stoichiometry together with the law of mass conservation form the core of constraint-based methods, making them powerful tools to query the metabolic capabilities and to assess the functional modes of biological systems. Therefore, it is not surprising that elementary-flux mode analysis can correctly answer questions such as whether the conversion of metabolite A into metabolite B is possible (de Figueiredo *et al.*, 2009b; Kaleta *et al.*, 2009b). Moreover, we expect that other methods for convex analysis like extreme pathways and methods based on Petri nets theory like the minimal T-invariants are also capable of answering these questions correctly.

Nevertheless, the main disadvantage of methods based on convex analysis is the impossibility of enumerating all the metabolic pathways present in large networks, as explained in the Introduction and in Chapter 2. The question that we addressed in Chapter 3 is: *Can we compute a subset of elementary flux modes in a genome-scale network?* Indeed, the K-shortest EFMs effectively computes a subset of elementary flux modes in such large networks (de Figueiredo *et al.*, 2009a). With this new method we show, once more, the potential of optimization problems in the study of metabolism. Other methods that use optimization frameworks, like FBA, Flux Coupling Analysis (Burgard *et al.*, 2004), Minimization Of Metabolic Adjustment (MOMA) (Segrè *et al.*, 2002) or OptKnock (Burgard *et al.*, 2003), just to cite some, are important for the study and rational design of microorganisms, and have been playing an important role in systems biology. The K-shortest EFM method is a milestone in metabolic pathway analysis because it allows the scale-up of elementary-flux modes analysis to genome-scale networks.

The key point in the K-shortest EFM method is to avoid full enumeration of elementary flux modes by focusing on solutions that produce/consume a given metabolite or contain a specific reaction. This is achieved by the inclusion of additional constraints forcing certain flux(es) to be non-null. Moreover, the enumeration starts with the shortest elementary flux mode (i.e., the solution containing less reactions) and proceeds toward longer elementary flux modes. This ranking of the solutions has a practical outcome. Starting from the elementary flux mode containing less reactions allows to evaluate the solution faster and eventually spot any modelling inconsistency faster.

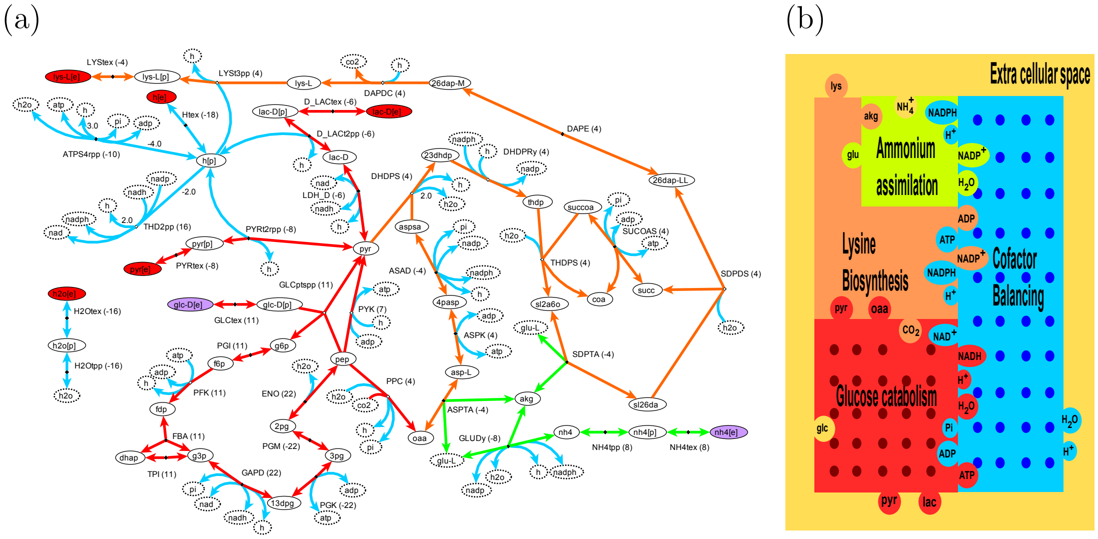
There are also methodological reasons for starting with the shortest elementary flux mode. First, the minimization of the number of active reactions assures that the computed solution is an elementary flux mode. Second, in order to enumerate a new elementary flux mode it is required to exclude the possibility that previous solutions are computed again in combination with other reaction sets. Thus, starting from the shortest solution is a good strategy for enumerating elementary flux modes in genome-scale networks.

The biological arguments for computing shorter metabolic pathways focus on two main points, one concerning experimental issues and the other one, a theoretical aspect of metabolism. First, genetic manipulation of metabolic pathways is a laborious work and it is convenient to perform few changes as possible. Moreover, it was experimentally shown that higher pathway fluxes can be obtained when all the enzymes or (at least) a considerable number of enzymes in the pathway are simultaneously and coordinately over-expressed (Niederberger *et al.*, 1992). Later, this experimental evidence was explained in the light of MCA (Kacser and Acerenza, 1993; Fell and Thomas, 1995). Thus, shorter pathways are more suitable for genetic manipulations towards the production improvement of target metabolites. Second, the theoretical work by Meléndez-Hevia *et al.* (1994) shows that when the kinetic constants are the same, shorter pathways can carry higher fluxes. For example, a comparison of lactic acid fermentation and respiration shows that the former is the shorter pathway and carries a higher flux.

Each of the 10-shortest elementary flux modes producing lysine can be subdivided into four subsystems: glucose catabolism, cofactor balancing, lysine biosynthesis and ammonium assimilation. The definition of these subsystems, in particular their limits, is sometimes fuzzy, but it is very helpful for reasoning on the pathways they represent. In *E. coli*, the shortest elementary flux mode producing lysine (Figure 6.2 (a)) contains the fermentation of glucose to pyruvate and lactate as the catabolic subsystem. Indeed, fermentation is not very efficient in the synthesis of ATP but can have very high rates. Moreover, fermentation is the major catabolic route when *E. coli* is growing anaerobically (Tempest and Neijssel, 1987; Clark, 1989). In these growing conditions, the end products of glucose catabolism are acetate, ethanol, lactate and formate. However, acetate is usually the major end product of fermentation but its synthesis requires more reactions steps. In order to obtain solutions excreting acetate, one has to increase K in order to enumerate longer solutions. Nevertheless, studies on the metabolic response of *E. coli* to changes in glucose supply show that pyruvate and lactate, together with methylglyxoal, are indeed excreted in the first place followed, with some delay, by acetate (Weber *et al.*, 2005). Thus, the catabolic subsystem of the shortest elementary flux mode corresponds to a physiological state of the cell.

Fermentation generates the precursors of lysine, pyruvate and oxaloacetate, and produces ATP required for lysine synthesis. The lysine biosynthesis subsystem contains all the essential reactions for lysine synthesis. In *E. coli* there is only one pathway for lysine biosynthesis (Schrumpf *et al.*, 1991) and it could be shown, using also the K-shortest EFM method, that all reactions except the one catalyzed by succinyl-CoA synthetase (SUCOAS; EC 6.2.1.5) are essential for lysine synthesis. Thus, it is expected that the lysine biosynthesis subsystem does not change in all the elementary flux modes producing lysine, with exception of the conversion of succinate to succinyl-CoA.

The cofactor balancing subsystem, is responsible for balancing cofactor demand and supply from all the other three subsystems. This subsystem, like the lysine biosynthesis subsystem, interfaces three other subsystems, showing a central



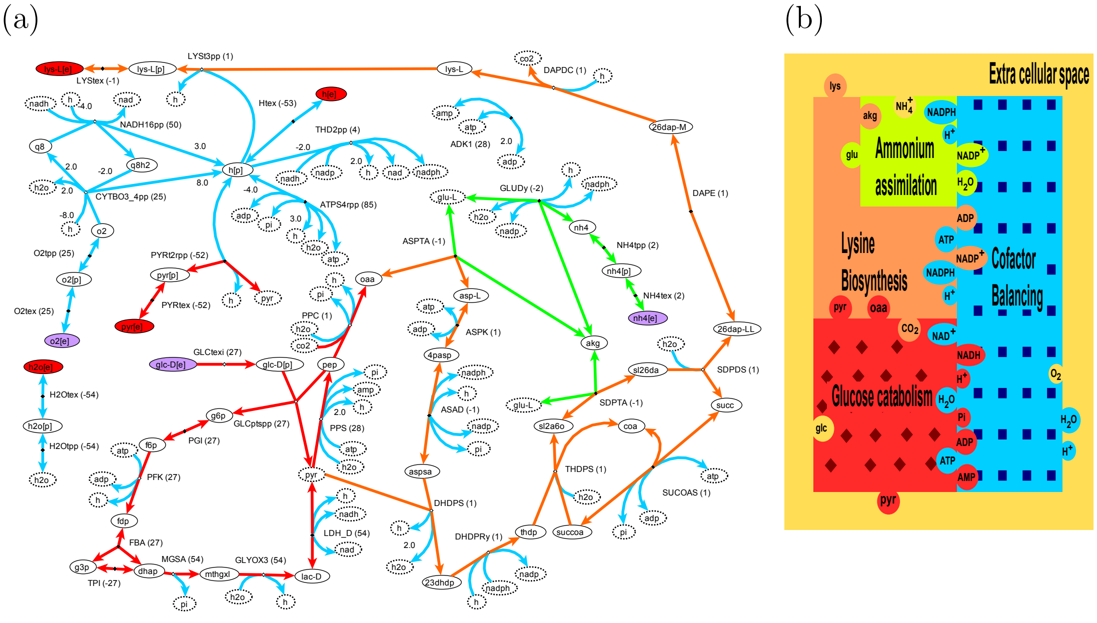
Eq.: 11 glc-D[e] + 8 nh4[e] → 16 h2o[e] + 18 h[e] + 6 lac-D[e] + 4 lys-L[e] + 8 pyr[e]

Figure 6.2: The shortest elementary flux mode producing lysine from glucose in the genome-scale network of *E. coli* (a). Red nodes are metabolites in the extra cellular space, violet nodes correspond to metabolites available in the *in silico* growth medium and in dash are the duplicated nodes used for simplifying the visualization. Reaction fluxes are in brackets following the reaction abbreviation. For metabolite and reaction abbreviations and directionality of reaction see Feist *et al.* (2007). Schematic representation of the metabolite interchange between the observed subsystems in the shortest elementary flux mode (b). The arrow coloring in (a) follows the color code of the subsystem division in (b).

role in lysine synthesis. There are two points worth mentioning in the cofactor balancing subsystem. First, the ATP synthase (ATPS4rpp; EC 3.6.3.14) has a negative flux meaning that it is pumping protons to the periplasm. Usually, ATP synthase performs ATP phosphorylation using the proton gradient as driving force. The inverse flux of ATP synthase corresponds to a mechanism of energy spilling (Russell and Cook, 1995; Trchounian, 2004) and, it shows that there is an imbalance between the catabolism and the anabolism subsystems. Second, the membrane transhydrogenase present in *E. coli* plays an important role converting NADH produced in the catabolism of glucose to reduced NADPH which is required for lysine synthesis.

The ammonium assimilation, responsible for the incorporation of nitrogen atoms in the carbon backbone, is only connected to the lysine biosynthesis subsystem and the cofactor balancing subsystem. In *E. coli* there are two routes for ammonium assimilation (Reitzer and Magasanik, 1987). The *K*-shortest method clearly favors the shortest one.

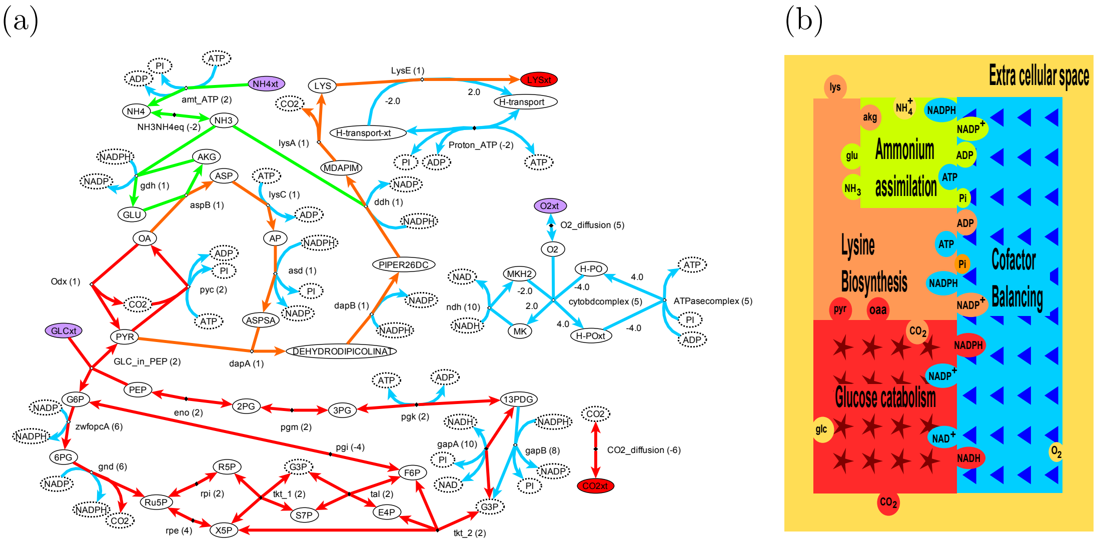
The main difference between the shortest elementary flux mode and the 10^{th} shortest elementary flux mode producing lysine in *E. coli* is in the glucose catabolism and cofactor balancing subsystems. In the former, the methylglyoxylate bypass plays an important role, decoupling the ATP synthesis from the glucose catabolism (Figure 6.3). Experimentally, the methylglyoxylate bypass was also identified as a spilling mechanism used by *E. coli* when growing in glucose excess (Tempest and Neijssel, 1987; Weber *et al.*, 2005). The cofactor balancing subsystem, is responsible for converting the NADH produced during glucose catabolism into the periplasmatic proton gradient which is then used for ATP and NADPH production.



Eq.: 27 glc-D[e] + 2 nh4[e] + 25 o2[e] → 54 h2o[e] + 53 h[e] + 1 lys-L[e] + 52 pyr[e]

Figure 6.3: The 10^{th} shortest elementary flux mode producing lysine from glucose in the genome-scale network of *E. coli* (a). Schematic representation of the metabolite interchange between the observed subsystems in the 10^{th} shortest elementary flux mode (b).

In *C. glutamicum*, each elementary flux mode can also be subdivided into the same four subsystems. The main difference here is that the pentose phosphate pathway is now central to glucose catabolism because it is the main route for NADPH production (Figure 6.4). It is worth mentioning that *C. glutamicum* does not have a transhydrogenase like *E. coli* but, some reactions in the central carbon metabolism can be coupled together to function as transhydrogenase (Wittmann and Becker, 2007). Interestingly, in the genome-scale model of *C. glutamicum* there is a coupled reaction functioning as transhydrogenase that actually is only converting NADPH to NADH, increasing the amount of NADH produced. This coupling between the reactions catalyzed by GapA and GapB, is responsible for the balancing between catabolism and anabolism in *C. glutamicum*. Note again that a clear subdivision into subsystems is not always exact, in particular between the catabolic and the cofactor balancing subsystems, but the objective of this exercise will be explained later.



Eq.: 2 GLCxt + 2 NH4xt + 5 O2xt → 6 CO2xt + LYSxt

Figure 6.4: The shortest elementary flux mode producing lysine from glucose in the genome-scale network of *C. glutamicum* (a). For metabolite and reaction abbreviations and directionality of reaction see Kjeldsen and Nielsen (2008). Schematic representation of the metabolite interchange between the observed subsystems in the shortest elementary flux mode (b). The arrow coloring in (a) follows the color code of the subsystem division in (b).

Few words on the computational complexity of *K*-shortest EFM method follow. The integer linear program used for computing the *K*-shortest EFMs, like any other optimization problem is not a decision (or recognition) problem and therefore, it is outside NP (Papadimitriou and Steiglitz, 1998; Garey and Johnson, 2000; Cormen *et al.*, 2001). However, decision problems can be polynomial-time transformed into optimization problems and consequently, the corresponding optimization problem is at least as hard as the corresponding decision problem (Papadimitriou and Steiglitz, 1998; Garey and Johnson, 2000; Cormen *et al.*, 2001). Thus, from the polynomial-time transformation of the satisfiability problem to an integer linear program, it follows that solving a integer linear program is NP-hard (Papadimitriou and Steiglitz, 1998). Moreover, as mentioned in the Introduction, Acuña *et al.* (2009) showed that computing the shortest elementary flux mode is also NP-hard. More recently, Liu *et al.* (2009) have proved that finding the shortest elementary flux mode with fixed parameter algorithms is W[1]-hard. The W[1] is a complexity class for parameterized algorithms that captures the property of fixed parameter intractability, being the W[1]-hard class analogous to the NP-hard in classical NP-completeness framework (Niedermeier, 2006). Nevertheless, these results from complexity analysis do not preclude algorithms able to compute the shortest elementary flux mode in a reasonable computation time for specific problem instances.

Indeed, the K-shortest EFM method was able to enumerate a subset of the shortest elementary flux modes producing lysine in one of the largest genome-scale networks (cf. Table 1.2). In de Figueiredo *et al.* (2009a), it is evident that the size of the network influences the computation time. Thus, the enumeration of the 10-shortest elementary flux modes for the *E. coli* network require in general more time than for the *C. glutamicum* network. Also the length of the target elementary flux modes plays a role in the computation time. In Figure 6.5, it is shown how the running time for the K-shortest EFM varies for two different biological problems, such as the 100-shortest elementary flux modes producing lysine from glucose and the 100-shortest futile cycles. These results cannot be used to deduce the computational complexity of finding the shortest elementary flux mode because this can only be done by mathematical proof as performed by Acuña *et al.* (2009) and by Liu *et al.* (2009). Nevertheless, given the existence of such a proof, it is always challenging to search problem formulations for which the K-shortest EFM method cannot compute a solution.

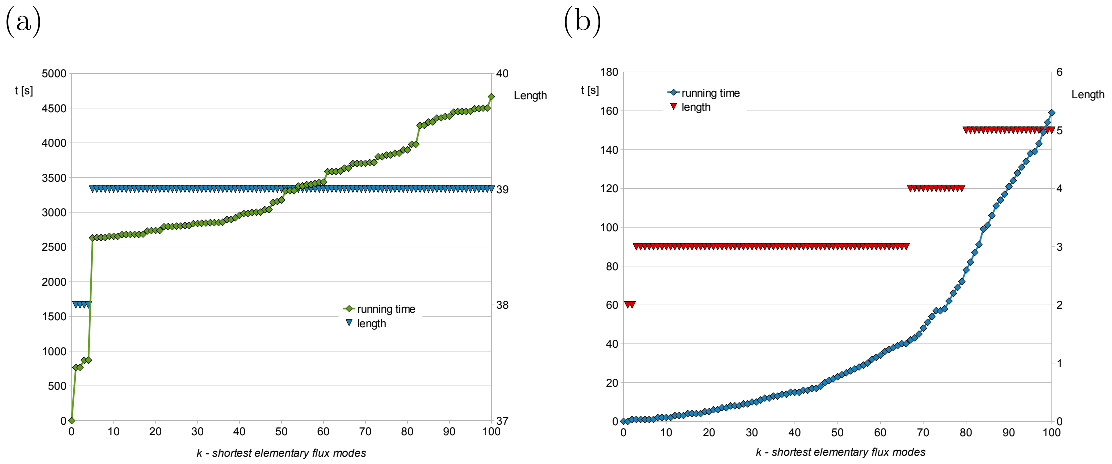


Figure 6.5: Running time and elementary flux mode length for two different biological problems, using the genome-scale network of *E. coli* from Feist *et al.* (2007). (a) The computation of the 100-shortest elementary flux mode producing lysine from glucose. (b) The computation of 100-shortest futile cycles involving the ATP synthase (ATPS4rpp).

In Chapter 3, we show that optimization problems can be used to compute elementary flux modes. In order to develop the method presented in Chapter 4, I analyzed the limitations of the K-shortest EFM method and consequently, pushed the capabilities of the optimization framework to an extreme. What are the main limitations of the K-shortest EFM method? First, this method is an iterative process that in each iteration adds a new constraint to the previous integer linear program. Thus, for larger values of K the optimization problem will become more complex which will increase the time required to solve each new iteration step (see Figure 6.5). Moreover, it is impossible to perform the enumeration of elementary flux modes with the K-shortest EFM method using parallel computing because of the iterative nature of this method. Second, the computed elementary flux modes tend to be very similar to each other, sometimes differing only in one single reaction.

What properties of the model can we use to solve these two initial limitations? One possibility is to set more restrictive removal constraints after each iteration. Thus, Eq. (8) from de Figueiredo *et al.* (2009a) can be re-written as:

$$\sum_{r=1}^{R} Z_r^k z_r \leq \left(\sum_{r=1}^{R} Z_j^k \right) - d \qquad k = 1, ..., K-1 \tag{6.1}$$

where d is a positive natural number such that $d \geq 1$. The bigger d is, the higher is the number of reactions that have to differ between each elementary flux mode. For example, in the case of lysine synthesis in *C. glutamicum*, the 10-shortest elementary flux modes when $d = 12$, reach now the reactions from the TCA cycle and can capture the two alternative pathways for lysine synthesis existing in this organism (Figure 6.6). The inclusion of more restrictive elimination constraints is a systematic way of enumerating distinct elementary flux modes. However, taking into account that the number of elementary flux modes is very high and that the K-shortest EFM method is bias towards shorter solutions, we may not have a good overview of all the elementary flux modes in the solution space.

Nevertheless, the previous approach enables us to see a very important property of elementary flux mode computation with optimization methods, which is the following. By blocking a reaction in a network, either by setting a specific reaction flux to zero or by adding a constraint applied to a set of reactions, such as in the Eq. 6.1, we can limit the access to a subset of elementary flux modes containing that reaction. Thus, we just have to block at least one reaction from the metabolic network, for example, a reaction that is present in an initial elementary flux mode and recompute the optimization problem again, to go from that initial elementary flux mode to a new one. This way, not only the constraints that are added to the new optimization problem are limited by the number of reactions in the metabolic network but they also simplify the computation by removing columns to the stoichiometric matrix. Moreover, this process can be reverted by removing the flux constraint of a blocked reaction, which enables to access again the elementary flux modes containing that reaction.

In the EFMEvolver we used a genetic algorithm for exploring this property.

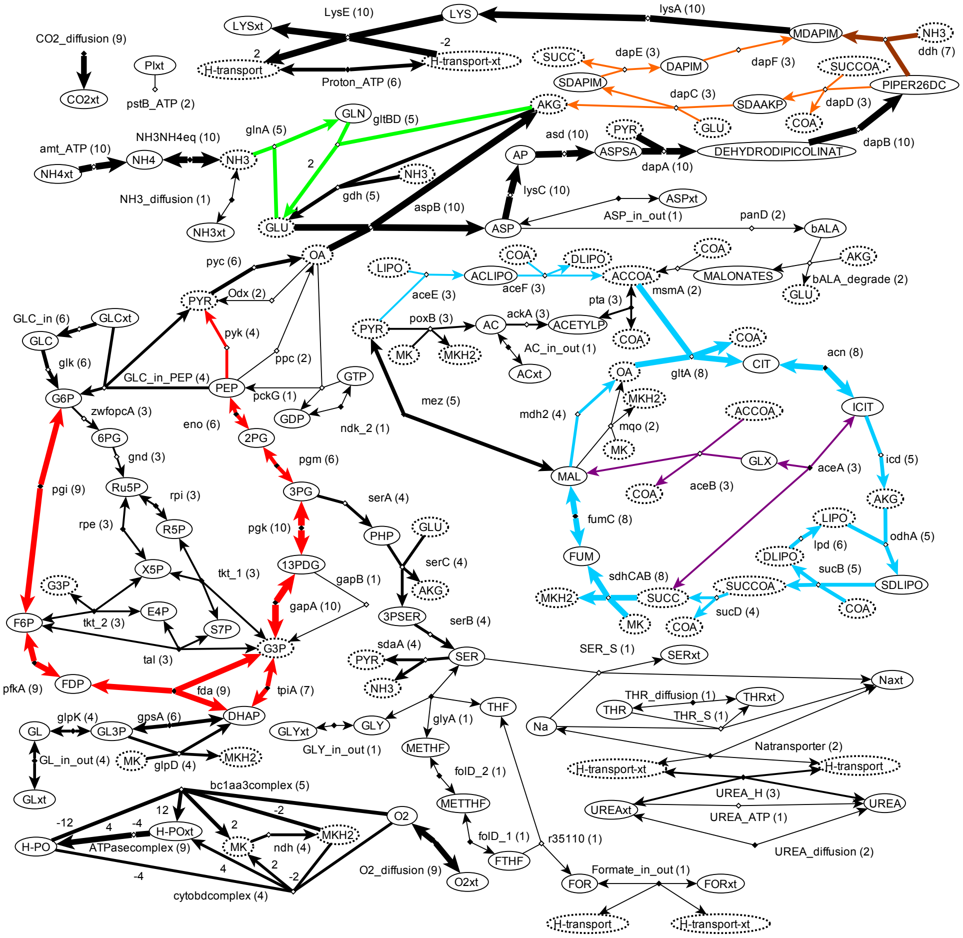


Figure 6.6: Merged representation of the 10-shortest elementary flux modes producing lysine from glucose in the genome-scale network of *C. glutamicum* using a more restrictive removal constraint (d=12). For metabolite and reaction abbreviations, and directionality of reaction see Kjeldsen and Nielsen (2008). Duplicated nodes are dashed, numbers in brackets after the reaction abbreviation correspond to the number of elementary flux modes in which the reaction takes place (arrow thickness is also proportional to this number). Glycolysis is in red, the TCA cycle is in blue, the glyoxylate shunt in violet, the active ammonium assimilation pathway is in green, the shared part between *E. coli* and *C. glutamicum* of the lysine biosynthesis is in orange and the part specific to *C. glutamicum* is in brown. The nodes corresponding to ATP, ADP, AMP, NAD, NADH, NADP, NADPH, PI, PII and CO2 were removed to simplify the visualization.

Genetic algorithms are a type of evolutionary algorithms that combine Darwin's theory of evolution with molecular genetics in order to solve combinatiorial problems. These algorithms aim at exploiting the historical information of previous solutions to speculate on new search points (Goldberg, 1989). There are distinct phases of search that can be categorized in terms of exploration (i.e., the generation of new individuals in unexplored regions of the search space) and exploitation (i.e., the concentration of individual in a specific part of the search space that is known to have good solutions) (Eiben and Smith, 2003).

Another limitation of the K-shortest EFMs method is the time required to compute a single elementary flux mode. Following what was said about the computational complexity of an integer linear program, we complement this information by saying that the simplex algorithm, in spite exponential in the worst case, on average solves linear programs in polynomial-time (Schrijver, 2000; Papadimitriou and Steiglitz, 1998). Consequently, a linear program formulation for computing elementary flux modes is expected to run faster than an integer linear program formulation.

The two key points of the linear program formulation present in the EFMEvolver method are the decoupling of reversible reactions into two irreversible reactions and the fact that the sum of fluxes is minimized. The former point will change the geometry of the solution space so that all the extreme rays of the polyhedral cone are elementary flux modes. The second key point is that the objective function, contrary to what happens often in FBA, does not maximize any reaction flux. Instead, the minimization of the sum of the fluxes is a reasonable objective function. With such an objective function one can compute a single extreme ray of the solution space and consequently an elementary flux mode. Note that, the same cannot be assured when maximizing the sum of the fluxes. First, there is no upper bound on the fluxes, which could imply some numerical issues, and second, by maximizing reaction fluxes the optimal solution can correspond to a combination of extreme rays, similarly to what happens in FBA. Another alternative, is to perform only a feasibility check without using any objective function.

The EFMEvolver, presented in Chapter 4, takes advantage of the fact that genetic algorithms are very good exploring the search space rather than fine tuning the solutions and of the computational efficiency of the algorithms used for solving linear programs. Additionally, the integration of these two frameworks allows to run the process on several processors simultaneously and benefit from multi-core CPU architectures. The computational power and capability of exploring the solution space from the EFMEvolver are clearly distinguishable from that of the K-shortest EFM, even though the latter allows a systematic ranking of elementary flux modes in increasing order of number of reactions.

These two methods, the K-shortest EFMs and the EFMEvolver, allow us to compute elementary flux modes in genome-scale networks, which was so far impossible with the current methods, like efmtool or METATOOL (see Table 1.1), at least in their conventional mode of use. Moreover, almost simultaneously, the method of elementary flux patterns was suggested to study metabolic subsytems in the context of flux distributions at genome scale (Kaleta *et al.*, 2009b). We have, for the first time, the possibility of studying the properties of these large-scale metabolic networks and of the embedded pathways. We are now prepared to answer questions such as the one posed in Chapter 2 even for genome-scale networks. Can humans convert fatty acids into sugars? Well, the elementary-flux mode analysis performed with the K-shortest EFM method on the human genome-scale metabolic network shows that this conversion is stoichiometrically possible (see Figure 6.7). This answer complements what was said in Chapter 2, that the TCA cycle is not involved in such conversion. In fact, this is just an initial answer. Indeed, elementary-flux mode analysis allows us to compute a metabolic pathway given the information that is stored in the organisms genome. This does not mean that all the solutions obtained with this method are feasible or active in a given condition. There are other constraints of thermodynamic or regulatory type limiting the number of feasible metabolic pathways. Consequently, the elementary flux modes can be used in the iterative process of hypothesis formulation and further validation through experiment or by doing a more detailed modelling in terms of kinetics.

The challenge of high-throughput data integration in reconstruction and analysis of metabolic networks is discussed in Chapter 5. The integration of this data,

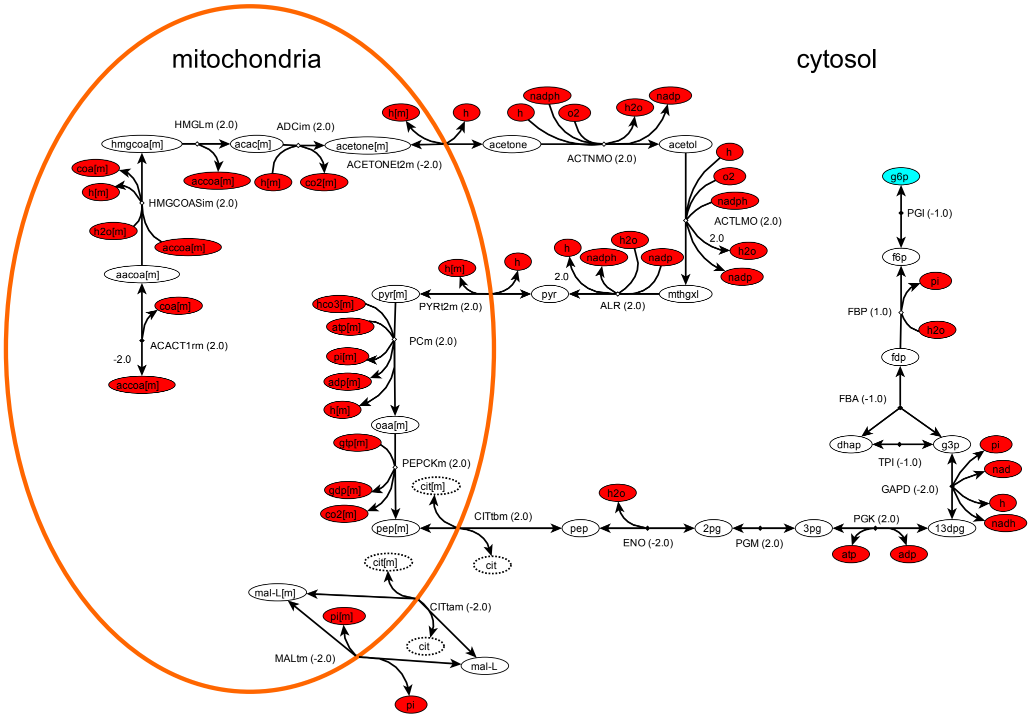


Figure 6.7: The shortest elementary flux mode converting acetyl-CoA to glucose 6-phosphate in the genome-scale network of *Homo sapiens*. The route through acetoacetone is used for the conversion to pyruvate and there is no enzyme from the TCA cycle involved in this pathway. Red nodes correspond to external metabolites and the blue node is the target metabolite. For reaction and metabolite abbreviations see Duarte *et al.* (2007)

also known as omics data, can also take place at the modelling level. For example, in the computation of genome-scale elementary flux modes that are associated with a given physiological state of an organism. In the K-shortest EFM approach, this integration cannot be done directly in the mathematical formulation of the method. The network has to be preprocessed to remove reactions that may not be active under certain conditions and the computed elementary flux modes have to be evaluated *a posteriori*. As discussed in Chapter 3, one cannot force the presence of more than one reaction otherwise, the non-decomposability condition cannot be assured. Thus, changing the mathematical model to cope with the integration of experimental data cannot be accomplished. On the other hand, the genetic algorithm from the EFMEvolver brings an extra flexibility to the elementary flux mode computation. A new fitness function can be defined, to evaluate the elementary flux modes that better describe the experimental data. The genetic algorithm can then explore this information keeping the elementary flux modes with better fitness in the population.

Nevertheless, one can also think about other alternatives to these two methods that in the future could be explored. Faced with the difficulty of performing convex analysis in large metabolic networks, some groups have chosen to reduce the complexity of these networks by dividing them into smaller subsystems (see Introduction). The approach by Schuster *et al.* (2002b) is particularly interesting in light of the work presented in this thesis. Schuster *et al.* (2002b) have defined a threshold in the metabolite connectivity, setting a given metabolite to external whenever its connectivity is higher than that threshold. The definition of this threshold was based on the topological study of metabolic networks showing that these metabolic *hubs* have important functions (Jeong *et al.*, 2000; Wagner and Fell, 2001), for example, as currency compounds or as precursors of other metabolites in the network (see also Table 1.3). The consequence of setting these *hubs* to external is the fragmentation of the metabolic network into subsystems in which the elementary-flux mode analysis can be carried out. Moreover, taking the list of *hubs* presented in Schuster *et al.* (2002b) and comparing with the schemes in Figures 6.2 to 6.4 (b) we can see that they often match the interfaces between the puzzle pieces. This means that an elementary flux mode of a subsystem would correspond to a part of a bigger puzzle (Figure 6.8) that, at the end, we would like to put together in order to have a clearer view of the metabolic capabilities of an organism.

Being a puzzle enthusiast, I have to remember the reader the kind of exercise that usually one has to do when solving puzzles. The brute force approach is, of course, to test if each and every piece fit together. This is clearly a very difficult approach and will not bring us anywhere. The often used approach is to classify the pieces according to colors and shapes. At this point, we are not sure if the piece will fit exactly in the place where we thought it should fit, specially if there

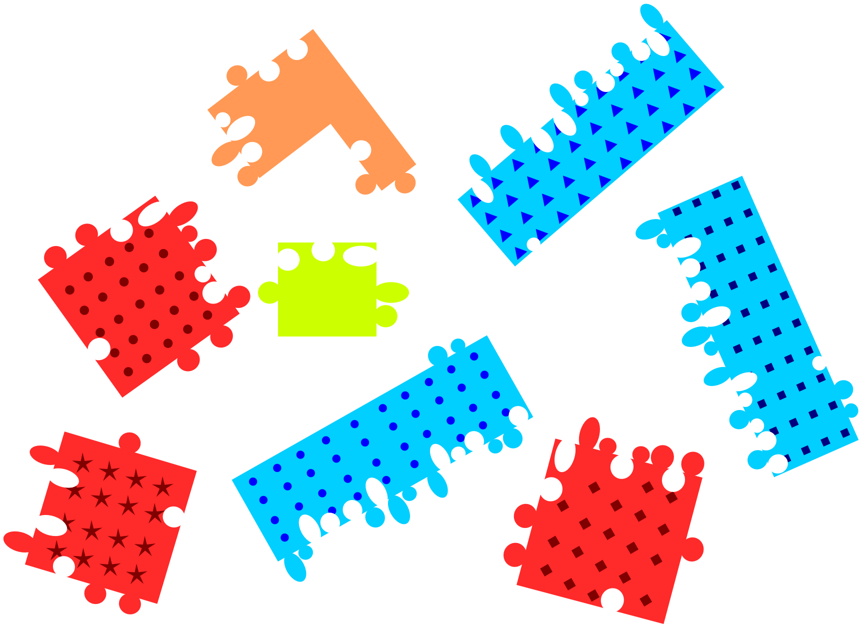


Figure 6.8: Schematic representation of some of the pieces of the puzzle that can be used to study lysine biosynthesis (c.f. with Figures 6.2 to 6.4 (b)). These pieces can be sorted according to their color and shape. For example red pieces play a role in central carbon metabolism, whereas blue in the energy production.

are several other pieces that seem to have the same color. But, we will for sure be able to put some of the trivial pieces together. Then, we start to increase the effort in testing more often whether the pieces fit in a specific place and making more assumptions where exactly the pieces will be placed. We reach then the point where we can start to see parts of the portrait in the puzzle and bigger parts of the puzzle can be now put together.

We can see each single reaction in a metabolic network as piece of this big puzzle. We know that the pieces have to fit at some place in the puzzle. We can try to see if every two pieces pass together in a brute force approach, similar to what was done in the beginning of the 20th century. At the end of the 20th century we were able already to put some of the trivial parts of the puzzle together and compute metabolic pathways at the subsystem level. Nowadays, we have access to even more pieces allowing us to put larger parts of the puzzle together using the methods present in this thesis and compute a subset of elementary flux modes at the scale of the genome.

Nevertheless, it may be wise to first characterize these parts of the puzzle in other words, to analyze the subsystems in the first place and then, assemble them to reach a larger scale. Schwartz *et al.* (2007) have shown the potential of this approach by integrating transcriptional data with elementary flux modes computed from the KEGG pathway maps and then searching for pairs of elementary flux modes from different maps that could be connected by boundary metabolites. Consequently, we can compute these smaller parts of the puzzle by removing the metabolic *hubs* of genome-scale networks from the balancing constraints and consequently, decomposing the network into subsystems (Zhao *et al.*, 2006). Then, we can put some effort in characterizing the pathways in these subsystems with the experimental data and select the ones that we want to further extend to genome-scale pathways. Later, the assembly of the larger parts of the puzzle, or more clearly speaking, the recovery of the genome-scale elementary flux modes, can be done by merging the most relevant solutions of each of the subsystems into a larger one, set the metabolic *hubs* back to internal metabolites and recompute the elementary flux modes. In the work presented in this thesis, it is evident that setting some metabolites to external can be used as a modelling simplification reducing the computational time required for elementary flux mode enumeration (de Figueiredo *et al.*, 2009a). This modelling simplification is reversible and consequently, by combining these simplified solutions with the cofactor subsystem and setting currency metabolites to internal, more complete solutions are reached.

This method is a good example of an alternative way of computing elementary flux modes in genome-scale networks, that combines graph theoretical with convex analysis methods. Moreover, the integration of omics data can be done in a preliminary stage and consequently, reduce the number of subsystem solutions that are used to generate genome-scale elementary flux modes. Of course, the main difference between building puzzles and studying biology is the fact that in the former we often know the picture *a priori*, whereas in the study of biology we only are able to understand the picture at the end, when we have already acquired enough knowledge to fit the reaming pieces.

Chapter 7

Conclusion and prospects

In this thesis, I show that the chemical information encoded in the stoichiometry of metabolic networks is a valuable resource for the prediction of metabolic pathways. In general, methods based in graph theory do not use this information and therefore, they fail in the prediction of relevant metabolic pathways. These incorrect predictions are now becoming evident in the field of metabolic engineering. Moreover, this issue is also very important when studying the topology of metabolic networks and has consequences in the conclusions made on the evolution of these networks.

Atom mapping rules are an alternative to the stoichiometry of metabolic networks. Approaches based on atom mapping and graph theory have been recently developed in order to improve pathway prediction. In particular, ReTrace is capable of solving some of the benchmark problems presented in this thesis. However, the presence of atoms in the target metabolite originated from the source metabolite do not assure that there is a net flux between these metabolites. Additional benchmark problems can be designed and used to assert the accuracy of metabolic pathway analysis tools.

On the other hand, the new methods presented in Chapters 3 and 4 are milestones in the field of metabolic pathway analysis. They allow us, for the first time, to compute a subset of elementary flux modes in genome-scale networks using only stoichiometric information. The scaling up of elementary flux mode analysis to larger networks is possible due to the use of optimization frameworks. These methods are a great asset to systems biology as they increase the spectrum of analysis that can be carried on genome-scale networks.

The current challenge in systems biology is the integration of disparate high-throughput data in the reconstruction of metabolic networks and in the modelling frameworks. The new methods presented here can be improved to fulfill this challenge. EFMEvolver is more suitable for this purpose due to the additional flexibility associated with the genetic algorithm. Nevertheless, other kinds of metaheuristics can be explored in the future. In particular, the decomposition of a genome-scale metabolic network into smaller subsystems where the full enumeration of elementary flux modes can be performed, is of special interest. Some initial work in this direction has been already performed by Schuster *et al.* (2002b). Such an approach would allow the characterization of elementary flux modes at the subsystem level using experimental data (Schwartz *et al.*, 2007). The elementary flux modes of interest at the subsystem level, can be used to build elementary flux modes at the genome-scale level.

In the long term, it will also be very exciting to follow the developments in genetic engineering or its extension to synthetic biology. In particular, the integration of heterologous pathways in new host organisms allowing us to evaluate *in vivo* the consequences of metabolism rewiring as well as the study of pathway evolution, in addition of course to all the biotechnological applications that are made possible. Thus, modelling techniques provide an excellent basis for guiding the experimental work and the iterative process between experiment and model improvement gives an important contribute to the emergence of new knowledge.

Bibliography

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.

T. Akutsu. Efficient extraction of mapping rules of atoms from enzymatic reaction data. *J Comput Biol*, 11(2-3):449--462, 2004.

M. Arita. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci U S A*, 101(6):1543--1547, 2004.

P. C. Babbitt. Definitions of enzyme function for the structural genomics era. *Curr Opin Chem Biol*, 7(2):230--237, 2003.

J. E. Bailey. Toward a science of metabolic engineering. *Science*, 252(5013):1668--1675, 1991.

J. A. Barnett. Glucose catabolism in yeast and muscle. In A. T. G. Semenza (editor), *Selected Topics in the History of Biochemistry: Personal Recollections IX*, volume 44 of *Comprehensive Biochemistry*, pages 1 -- 132. Elsevier, 2005.

J. Behre, T. Wilhelm, A. von Kamp, E. Ruppin, and S. Schuster. Structural robustness of metabolic networks with respect to multiple knockouts. *J theor Biol*, 252(7):433--441, 2008.

P. A. Berman and L. Human. Regulation of 5-phosphoribosyl 1-pyrophosphate and of hypoxanthine uptake and release in human erythrocytes by oxypurine cycling. *J Biol Chem*, 265(12):6562--6568, 1990.

F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453--1474, 1997.

T. Blum and O. Kohlbacher. MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics*, 24(18):2108--2109, 2008a.

T. Blum and O. Kohlbacher. Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J Comput Biol*, 15(6):565--576, 2008b.

B. A. Boghigian, H. Shi, K. Lee, and B. A. Pfeifer. Utilizing elementary mode analysis, pathway thermodynamics, and a genetic algorithm for metabolic flux determination and optimal metabolic network design. *BMC Syst Biol*, 4(1):49, 2010.

S. Brohée, K. Faust, G. Lima-Mendez, O. Sand, R. Janky, G. Vanderstocken, Y. Deville, and J. van Helden. NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res*, 36(Web Server issue):W444--W451, 2008.

A. P. Burgard, P. Pharkya, and C. D. Maranas. OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647--657, 2003.

A. P. Burgard, E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res*, 14(2): 301--312, 2004.

T. Cakir, C. S. Tacer, and K. O. Ulgen. Metabolic pathway analysis of enzyme-deficient human red blood cells. *BioSystems*, 78(1-3):49--67, 2004.

D. P. Clark. The fermentation pathways of *Escherichia coli*. *FEMS Microbiol Rev*, 5 (3):223--234, 1989.

B. L. Clarke. Stoichiometric network analysis. *Cell Biophys*, 12:237--253, 1988.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press, London, 2^{nd} edition, 2001.

D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res*, 33(Web Server issue): W326--W330, 2005.

D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol*, 356(1):222--236, 2006.

L. F. de Figueiredo, S. Schuster, C. Kaleta, and D. A. Fell. Can sugars be produced from fatty acids? a test case for pathway analysis tools. *Bioinformatics*, 24(22):2615--2621, 2008.

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, 2009a.

L. F. de Figueiredo, S. Schuster, C. Kaleta, and D. A. Fell. Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, 25(1):152--158, 2009b.

L. F. de Figueiredo, S. Schuster, C. Kaleta, and D. A. Fell. Response to comment on 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'. *Bioinformatics*, 25(24):3330--3331, 2009c.

N. C. Duarte, M. J. Herrgård, and B. Ø. Palsson. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298--1309, 2004.

N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6):1777--1782, 2007.

J. S. Edwards and B. Ø. Palsson. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J Biol Chem*, 274(25):17410--17416, 1999.

J. S. Edwards and B. Ø. Palsson. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, 97(10):5528--5533, 2000.

V. Egelhofer, I. Schomburg, and D. Schomburg. Automatic assignment of EC numbers. *PLoS Comput Biol*, 6(1):e1000661, 2010.

A. E. Eiben and J. E. Smith. *Introduction to evolutionary computing.* Springer, Berlin, 2003.

J. H. Exton and C. R. Park. Control of gluconeogenesis in liver. I. general features of gluconeogenesis in the perfused livers of rats. *J Biol Chem*, 242(11):2622--2636, 1967.

K. Faust, D. Croes, and J. van Helden. In response to "Can sugars be produced from fatty acids? A test case for pathway analysis tools". *Bioinformatics*, 25(23):3202--3205, 2009a.

K. Faust, D. Croes, and J. van Helden. Metabolic pathfinding using RPAIR annotation. *J Mol Biol*, 388(2):390--414, 2009b.

A. M. Feist and B. Ø. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol*, 26(6):659--667, 2008.

A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, 7(2):129--143, 2009.

D. A. Fell. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J*, 286 (Pt 2):313--330, 1992.

D. A. Fell. *Understanding the control of metabolism.* Portland Press, London, 2^{nd} edition, 2003.

D. A. Fell. Metabolic control analysis. In L. Alberghina and H. V. Westerhoff (editors), *Systems Biology: Definitions and Perspectives*, volume 13 of *Topics in Current Genetics*, pages 69--80, Berlin, 2005. Springer.

D. A. Fell and J. R. Small. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem J*, 238(3):781--786, 1986.

D. A. Fell and S. Thomas. Physiological control of metabolic flux: the requirement for multisite modulation. *Biochem J*, 311 (Pt 1):35--39, 1995.

E. Fischer and U. Sauer. A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*. *J Biol Chem*, 278(47):46446--46451, 2003.

R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223): 496--512, 1995.

J. Förster, I. Famili, P. Fu, B. Ø. Palsson, and J. Nielsen. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res*, 13(2):244--253, 2003.

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

M. R. Garey and D. S. Johnson. *Computers and intractability : a guide to the theory of NP-completeness*. Freeman, New York, 2000.

E. P. Gianchandani, A. K. Chavali, and J. A. Papin. The application of flux balance analysis in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2:372--382, 2010.

A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287): 546--567, 1996.

D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley, New York, 1989.

S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, K. Sato, and M. Kanehisa. Organizing and computing metabolic pathway data in terms of binary relations. In *Pacific Symposium on Biocomputing*, pages 175--186, 1997.

M. L. Green and P. D. Karp. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res*, 33(13):4035--4039, 2005.

M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc*, 125(39):11853--11865, 2003.

D. F. Heath. The redistribution of carbon label by the reactions involved in glycolysis, gluconeogenesis and the tricarboxylic acid cycle in rat liver. *Biochem J*, 110(2): 313--335, 1968.

R. Heinrich and T. A. Rapoport. A linear steady-state treatment of enzymatic chains. general properties, control and effector strength. *Eur J Biochem*, 42(1):89--95, 1974a.

R. Heinrich and T. A. Rapoport. A linear steady-state treatment of enzymatic chains. critique of the crossover theorem and a general procedure to identify interaction sites with an effector. *Eur J Biochem*, 42(1):97--105, 1974b.

R. Heinrich and S. Schuster. *The regulation of cellular systems.* Chapman & Hall, New York, 1996.

R. Heinrich, S. M. Rapoport, and T. A. Rapoport. Metabolic regulation and mathematical models. *Prog Biophys Mol Biol*, 32(1):1--82, 1977.

S. Heptinstall, A. Johnson, J. R. Glenn, and A. E. White. Adenine nucleotide metabolism in human blood -- important roles for leukocytes and erythrocytes. *J Thromb Haemost*, 3(10):2331--2339, 2005.

R. Hofestädt. A Petri net application to model metabolic processes. *Systems Analysis Modelling Simulation*, 16(2):113--122, 1994.

S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI--a COmplex PAthway SImulator. *Bioinformatics*, 22(24): 3067--3074, 2006.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651--654, 2000.

H. Kacser and L. Acerenza. A universal method for achieving increases in metabolite production. *Eur J Biochem*, 216(2):361--367, 1993.

H. Kacser and J. A. Burns. The control of flux. *Symposia of the Society for Experimental Biology*, 27:65104, 1973.

C. Kaleta, L. F. de Figueiredo, J. Behre, and S. Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, and P. Stadler (editors), *Lecture Notes in Informatics*, volume P-157, pages 179--189, Bonn, 2009a. Gesellschaft für Informatik. ISBN 978-3-88579-251-2 ; ISSN 1617-5468.

C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res*, 19:1872--1883, 2009b.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27--30, 2000.

M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database issue):D354--D357, 2006.

P. D. Karp. Pathway databases: a case study in computational symbolic theories. *Science*, 293(5537):2040--2044, 2001.

P. D. Karp and M. L. Mavrovouniotis. Representing, analyzing, and synthesizing biochemical pathways. *IEEE Expert: Intelligent Systems and Their Applications*, 9: 11 -- 21, 1994.

P. D. Karp, M. Riley, S. M. Paley, and A. Pelligrini-Toole. EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res*, 24(1):32--40, 1996.

N. Kashtan, E. Noor, and U. Alon. Varying environments can speed up evolution. *Proc Natl Acad Sci U S A*, 104(34):13711--13716, 2007.

K. R. Kjeldsen and J. Nielsen. In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol Bioeng*, 102(2):583 -- 597, 2008.

S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233--236, 2002.

S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends Biotechnol*, 21(2):64--69, 2003.

S. Klamt, J. Gagneur, and A. von Kamp. Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *IEE Proc Syst Biol*, 152(4): 249--255, 2005.

S. Klamt, J. Saez-Rodriguez, and E. D. Gilles. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol*, 1:2, 2007.

S. Klamt, U.-U. Haus, and F. Theis. Hypergraphs and cellular networks. *PLoS Comput Biol*, 5(5):e1000385, 2009.

E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwing. *Systems Biology: a textbook*. Wiley-Blackwell, Weinheim, 2009.

I. Koch, B. H. Junker, and M. Heiner. Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, 21 (7):1219--1226, 2005.

M. Kotera, Y. Okuno, M. Hattori, S. Goto, and M. Kanehisa. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J Am Chem Soc*, 126(50):16487--16498, 2004.

H. A. Krebs. *Otto Warburg: Zellphysiologe, Biochemiker, Mediziner 1883-1970*. WVG, Stuttgart, 1979.

H. A. Krebs and W. A. Johnson. The role of citric acid in intermediate metabolism in animal tissues. *FEBS Lett*, 117 Suppl:K1--10, 1980.

R. Küffner, R. Zimmer, and T. Lengauer. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, 16(9):825--836, 2000.

V. Lacroix, L. Cottret, P. Thébault, and M.-F. Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans Comput Biol Bioinform*, 5 (4):594--617, 2008.

A. Larhlimi and A. Bockmayr. A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, 157(10):2257--2266, 2009.

K. Lautenbach. *Exakte Bedingungen der Lebendigkeit für eine Klasse von Petri-Netzen*. PhD thesis, Gesellschaft für Mathematik und Datenverarbeitung Bonn, 1973.

S. Lee, C. Palakornkule, M. M. Domach, and I. E. Grossmann. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput Chem Eng*, 24(2-7):711--716, 2000.

J. C. Liao, S. Y. Hou, and Y. P. Chao. Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnol Bioeng*, 52(1):129--140, 1996.

G. Lima-Mendez and J. van Helden. The powerful law of the power law and other myths in network biology. *Mol Biosyst*, 5(12):1482--1493, 2009.

H. Liu, H. Feng, and D. Zhu. Parameterized complexity of finding elementary modes in metabolic networks. In *Proc. International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing IJCBS '09*, pages 479--482, 2009.

R. Mahadevan and C. H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*, 5(4):264--276, 2003.

R. A. Majewski and M. M. Domach. Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnol Bioeng*, 35(7):732--738, 1990.

M. L. Mavrovouniotis, G. Stephanopoulos, and G. Stephanopoulos. Computer-aided synthesis of biochemical pathways. *Biotechnol Bioeng*, 36(11):1119--1132, 1990.

E. Meléndez-Hevia, T. G. Waddell, and F. Montero. Optimization of metabolism: The evolution of metabolic pathways toward simplicity through the game of the pentose phosphate cycle. *J theor Biol*, 166(2):201--220, 1994.

G. Michal. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*. Spektrum Akademischer Verlag, Heidelberg, Berlin, 1999.

R. Montañez, M. A. Medina, R. V. Solé, and C. Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *BioEssays*, 32(3): 246--256, 2010.

H. J. Morowitz. *Beginnings of cellular life: Metabolism recapitulates biogenesis*. Yale University Press, London, 1992.

P. Niederberger, R. Prasad, G. Miozzari, and H. Kacser. A strategy for increasing an in vivo flux by genetic manipulations. The tryptophan system of yeast. *Biochem J*, 287 (Pt 2):473--479, 1992.

R. Niedermeier. *Invitation to fixed-parameter algorithms*. Oxford Univ. Press, Oxford, 2006.

J. H. Northrop and R. M. Herriott. Chemistry of the crystalline enzymes. *Annual Review of Biochemistry*, 7:37--50, 1938.

H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 27(1):29--34, 1999.

M. Oh, T. Yamada, M. Hattori, S. Goto, and M. Kanehisa. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J Chem Inf Model*, 47(4):1702 -- 1712, 2007.

C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: algorithms and complexity*. Dover Publications, Mineola, 1998.

J. A. Papin, J. Stelling, N. D. Price, S. Klamt, S. Schuster, and B. Ø. Palsson. Comparison of network-based pathway analysis methods. *Trends Biotechnol*, 22(8):400--405, 2004.

E. T. Papoutsakis and C. L. Meyer. Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnol Bioeng*, 27(1):50--66, 1985.

K. R. Patil, I. Rocha, J. Frster, and J. Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6:308, 2005.

C. A. Petri. *Kommunikation mit Automaten*. PhD thesis, Universität Bonn, Bonn, 1962.

T. Pfeiffer, I. Sánchez-Valdenebro, J. C. Nuño, F. Montero, and S. Schuster. METATOOL: for studying metabolic networks. *Bioinformatics*, 15(3):251--257, 1999.

E. Pitkänen, P. Jouhten, and J. Rousu. Inferring branching pathways in genome-scale metabolic networks. *BMC Syst Biol*, 3(1):103, 2009.

J. D. Pollack, M. V. Williams, and R. N. McElhaney. The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit Rev Microbiol*, 23(4):269--354, 1997.

M. Poolman. ScrumPy: metabolic modelling with Python. *IEE Proceedings Systems Biology*, 153(5):375--378, 2006.

M. G. Poolman, D. A. Fell, and C. A. Raines. Elementary modes analysis of photosynthate metabolism in the chloroplast stroma. *Eur J Biochem*, 270(3):430--439, 2003.

M. G. Poolman, L. Miguet, L. J. Sweetlove, and D. A. Fell. A genome-scale metabolic model of *Arabidopsis thaliana* and some of its properties. *Plant Physiol*, 151:1570--1581, 2009.

N. D. Price, J. A. Papin, and B. Ø. Palsson. Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res*, 12(5):760--769, 2002.

N. D. Price, J. L. Reed, and B. Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, 2(11):886--897, 2004.

S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, 21 (7):1189--1193, 2005.

K. Raman and N. Chandra. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform*, 10(4):435--449, 2009.

S. Ranganathan and C. D. Maranas. Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions. *Biotechnol J*, 5 (7):716 -- 725, 2010.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551--1555, 2002.

J. W. Raymond, E. J. Gardiner, and P. Willett. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J Chem Inf Comput Sci*, 42(2):305--316, 2002.

V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman. Petri net representations in metabolic pathways. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 1:328--336, 1993.

C. Reder. Metabolic control theory: a structural approach. *J theor Biol*, 135(2):175--201, 1988.

L. J. Reitzer and B. Magasanik. Ammonia assimilation and the biosynthesis of glutamine, glutamate, aspartate, asparagine, L-alanine, and D-alanine. In F. C. Neidhardt (editor), *Escherichia coli and Salmonella typhimurium: cellular and molecular biology*, pages 302--320, Washington D. C., 1987. ASM.

E. Ruppin, J. A. Papin, L. F. de Figueiredo, and S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr Opin Biotechnol*, 21(4):502--510, 2010.

J. B. Russell and G. M. Cook. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev*, 59(1):48--62, 1995.

C. Salerno and A. Giacomello. Hypoxanthine-guanine exchange by intact human erythrocytes. *Biochemistry*, 24(6):1306--1309, 1985.

J. M. Savinell and B. Ø. Palsson. Network analysis of intermediary metabolism using linear optimization. I. development of mathematical formalism. *J theor Biol*, 154(4): 421--454, 1992a.

J. M. Savinell and B. Ø. Palsson. Network analysis of intermediary metabolism using linear optimization. II. interpretation of hybridoma cell metabolism. *J. theor. Biol.*, 154(4):455--473, 1992b.

C. H. Schilling and B. Ø. Palsson. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J theor Biol*, 203(3):249--283, 2000.

C. H. Schilling, D. Letscher, and B. Ø. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J theor Biol*, 203(3):229--248, 2000.

I. Schomburg, A. Chang, O. Hofmann, C. Ebeling, F. Ehrentreich, and D. Schomburg. BRENDA: a resource for enzyme data and metabolic information. *Trends Biochem Sci*, 27(1):54--56, 2002.

A. Schrijver. *Theory of linear and integer programming.* Wiley, Chichester, 3^{rd} edition, 2000.

B. Schrumpf, A. Schwarzer, J. Kalinowski, A. Pühler, L. Eggeling, and H. Sahm. A functionally split pathway for lysine synthesis in *Corynebacterium glutamicium. J Bacteriol*, 173(14):4510--4516, 1991.

R. Schuetz, L. Kuepfer, and U. Sauer. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli. Mol Syst Biol*, 3:119, 2007.

R. Schuster and S. Schuster. Refined algorithm and computer program for calculating all non-negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. *Comput Appl Biosci*, 9(1):79--85, 1993.

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

S. Schuster and D. Kenanov. Adenine and adenosine salvage pathways in erythrocytes and the role of S-adenosylhomocysteine hydrolase. a theoretical study using elementary flux modes. *FEBS J*, 272(20):5278--5290, 2005.

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 Biotechnol*, 17(2):53--60, 1999.

S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol*, 18(3):326--332, 2000.

S. Schuster, C. Hilgetag, J. H. Woods, and D. A. Fell. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J Math Biol*, 45(2):153--181, 2002a.

S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae. Bioinformatics*, 18(2):351--361, 2002b.

S. Schuster, T. Pfeiffer, and D. A. Fell. Is maximization of molar yield in metabolic networks favoured by evolution? *J theor Biol*, 252(3):497--504, 2008.

J.-M. Schwartz, C. Gaugain, J. Nacher, A. de Daruvar, and M. Kanehisa. Observing metabolic functions at the genome scale. *Genome Biol*, 8(6):R123, 2007.

R. Schwarz, C. Liang, C. Kaleta, M. Khnel, E. Hoffmann, S. Kuznetsov, M. Hecker, G. Griffiths, S. Schuster, and T. Dandekar. Integrated network reconstruction, visualization and analysis using yanasquare. *BMC Bioinformatics*, 8:313, 2007.

D. Segrè, D. Vitkup, and G. M. Church. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23):15112--15117, 2002.

A. Seressiotis and J. E. Bailey. MPS: An algorithm and data base for metabolic pathway synthesis. *Biotechnol Lett*, 8(12):837--842, 1986.

A. Seressiotis and J. E. Bailey. MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnol Bioeng*, 31(6):587--602, 1988.

J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912): 190--193, 2002.

L. Stryer. *Biochemistry*. W.H. Freeman and Company, New York, 4^{th} edition, 1995.

J. B. Sumner. The isolation and crystallization of the enzyme urease: Preliminary paper. *J Biol Chem*, 69:435--441, 1926.

J. B. Sumner. The story of urease. *J Chem Educ*, 14(6):255--259, 1937.

R. Tanaka. Scale-rich metabolic networks. *Phys Rev Lett*, 94(16):168101, 2005.

D. W. Tempest and O. M. Neijssel. Growth yield and energy distribution. In F. C. Neidhardt (editor), *Escherichia coli and Salmonella typhimurium: cellular and molecular biology*, pages 797--805, Washington D. C., 1987. ASM.

M. Terzer and J. Stelling. Accelerating the computation of elementary modes using pattern trees. In P. Bücher and B. M. Moret (editors), *Algorithms in Bioinformatics*, volume 4175 of *Lecture Notes in Computer Science*, pages 333--343, Berlin, 2006. Springer.

M. Terzer and J. Stelling. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229--35, 2008.

B. Teusink, A. Wiersma, L. Jacobs, R. A. Notebaart, and E. J. Smid. Understanding the adaptive growth strategy of *Lactobacillus plantarum* by in silico optimisation. *PLoS Comput Biol*, 5(6):e1000410, 2009.

I. Thiele and B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93--121, 2010.

A. Trchounian. *Escherichia coli* proton-translocating F0F1-ATP synthase and its association with solute secondary transporters and/or enzymes of anaerobic oxidation-reduction under fermentation. *Biochem Biophys Res Commun*, 315(4):1051--1057, 2004.

C. T. Trinh and F. Srienc. Metabolic engineering of *Escherichia coli* for efficient conversion of glycerol into ethanol. *Appl Environ Microbiol*, 75(21):6696--6705, 2009.

C. T. Trinh, P. Unrean, and F. Srienc. Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Appl Environ Microbiol*, 74(12):3634--3643, 2008.

R. Urbanczik. SNA--a toolbox for the stoichiometric analysis of metabolic networks. *BMC Bioinformatics*, 7:129, 2006.

R. Urbanczik and C. Wagner. An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, 21(7):1203--1210, 2005.

L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410--421, 1979a.

L. G. Valiant. The complexity of computing the permanent. *Theor Comput Sci*, 8(2): 189 -- 201, 1979b.

D. Voet and J. Voet. *Biochemistry.* Wiley, USA, 3^{rd} edition, 2004.

C. A. Voigt, C. Martinez, Z.-G. Wang, S. L. Mayo, and F. H. Arnold. Protein building blocks preserved by recombination. *Nat Struct Biol*, 9(7):553--558, 2002.

A. von Kamp and S. Schuster. Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, 22(15):1930--1931, 2006.

A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478):1803--1810, 2001.

C. Wagner. Nullspace approach to determine the elementary modes of chemical reaction systems. *J Phys Chem B*, 108(7):2425--2431, 2004.

C. Wagner and R. Urbanczik. The geometry of the flux cone of a metabolic network. *Biophys J*, 89(6):3837--3845, 2005.

M. R. Watson. A discrete model of bacterial metabolism. *Comput Appl Biosci*, 2(1): 23--27, 1986.

E. C. Webb. *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.* Academic Press, San Diego, 6^{th} edition, 1992.

J. Weber, A. Kayser, and U. Rinas. Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. II. dynamic response to famine and feast, activation of the methylglyoxal pathway and oscillatory behaviour. *Microbiology*, 151(Pt 3):707--716, 2005.

E. O. Weinman, E. H. Strisower, and I. L. Chaikoff. Conversion of fatty acids to carbohydrate: application of isotopes to this problem and role of the Krebs cycle as a synthetic pathway. *Physiol Rev*, 37(2):252--272, 1957.

H. V. Westerhoff and L. Alberghina. Systems biology: Did we know it all along? In L. Alberghina and H. V. Westerhoff (editors), *Systems Biology: Definitions and Perspectives*, volume 13 of *Topics in Current Genetics*, pages 3--9, Berlin, 2005. Springer.

H. V. Westerhoff and J.-H. S. Hofmeyr. What is systems biology? From genes to function and back. In L. Alberghina and H. V. Westerhoff (editors), *Systems Biology: Definitions and Perspectives*, volume 13 of *Topics in Current Genetics*, pages 119--141, Berlin, 2005. Springer.

H. V. Westerhoff and B. Ø. Palsson. The evolution of molecular biology into systems biology. *Nat Biotechnol*, 22(10):1249--1252, 2004.

S. J. Wiback and B. Ø. Palsson. Extreme pathway analysis of human red blood cell metabolism. *Biophys J*, 83(2):808--818, 2002.

S. J. Wiback, R. Mahadevan, and B. Ø. Palsson. Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum. *J theor Biol*, 224(3):313--324, 2003.

T. Wilhelm, J. Behre, and S. Schuster. Analysis of structural robustness of metabolic networks. *Syst Biol (Stevenage)*, 1(1):114--120, 2004.

C. Wittmann and J. Becker. The L-lysine story: from metabolic pathways to industrial production. In V. F. Wendisch (editor), *Amino acid biosynthesis - pathways, regulation and metabolic engineering*, pages 39--70, Heidelberg, 2007. Springer.

M. Yeung, I. Thiele, and B. Ø. Palsson. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics*, 8(1):363, 2007.

J. Zhao, H. Yu, J.-H. Luo, Z.-W. Cao, and Y.-X. Li. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics*, 7:386, 2006.

Supplementary material

Supplementary Material

1 List of Abbreviations

Table 1 contains the list of abbreviations used in the article and in the networks.

Table 1: List of Abbreviations*

Abbreviation	Name
2PG	2-Phospho-D-glycerate
3PG	3-Phospho-D-glycerate
3PGP	3-Phospho-D-glyceroyl phosphate
AcCoA	Acetyl coenzyme A
Acet	Acetate
Ala	L-Alanine
Asp	L-Aspartate
bF6P	β-D-Fructose 6-phosphate
Cit	Citrate
CO2	Carbon dioxide
CoA	Coenzyme A
dATP	2'-Deoxyadenosine 5'-triphosphate
dADP	2'-Deoxyadenosine 5'-diphosphate
Dihydroxyethyl-TPP	alpha,beta-Dihydroxyethyl-TPP
EM	Elementary mode
EMA	Elementary mode analysis
Ery4P	D-Erythrose 4-phosphate
F1,6PP	β-D-Fructose 1,6-bisphosphate
F6P	D-Fructose 6-phosphate
Fum	Fumarate
G3P	D-Glyceraldehyde 3-phosphate
G6P	α-D-Glucose-6-phosphate
bG6P	β-D-Glucose-6-phosphate
GlcN	D-Glucosamine

GlcN6P	D-Glucosamine 6-phosphate
GlcNAc6P	N-Acetyl-D-glucosamine 6-phosphate
Gln	L-Glutamine
Glu	L-Glutamate
Gly	Glyoxylate
GP	Glycerone phosphate
HCO3-	Bicarbonate
IDP	Inosine 5'-diphosphate
ITP	Inosine 5'-triphosphate
Isocit	Isocitrate
Mal	(S)-Malate
Man	D-Mannose
Man6P	D-Mannose 6-phosphate
NH3	Ammonia
OAA	Oxaloacetate
OG	2-Oxoglutarate
PHT	Pathway Hunter Tool
PEP	Phosphoenolpyruvate
Pi	Orthophosphate
PPi	Pyrophosphate
Protein N-P-histidine	Protein N(pi)-phospho-L-histidine
Pyr	Pyruvate
R5P	D-Ribose 5-phosphate
Ru5P	D-Ribulose 5-phosphate
Sed7P	Sedoheptulose 7-phosphate
DSed7P	D-Sedoheptulose 7-phosphate
Succ	Succinate
SucCoA	Succinyl coenzyme A
ThPP	Thiamin diphosphate
Xyl5P	D-Xylulose 5-phosphate

* External metabolites in external reactions are indicated with the suffix "ex". Well known abbreviations in biochemistry, such as ATP, are not included in the list.

2 Reaction list of the reconstructed models

2.1 Human metabolism

Table 2 contains the list of reactions from the model used to study the human metabolism. Note that the Gene symbols are written according to the official symbols present in NCBI gene database, in capital letters.

Table 2: Reaction list containing all the reactions present in the model of human metabolism studied by EMA.

Gene Symbol	Reaction	KEGG Reaction	Ref.
PGI	bF6P ↔ G6P	R02740	[1, 2]
FBP1	F1,6PP + H2O → bF6P + Pi	R04780	[1, 2]
PFKL	ATP + bF6P → ADP + F1,6PP	R04779	[1, 2]
ALDOA, ALDOB	G3P + GP ↔ F1,6PP	R01070	[1, 2]
TPI1	G3P ↔ GP	R01015	[1, 2]
GAPDH, PGK1, PGAM1, ENO1 [†]	G3P + Pi + NAD + ADP ↔ PEP + ATP + NADH + H	R01061, R01512, R01518, R00658	[1, 2]
PKLR	ATP + Pyr ← ADP + PEP	R00200	[1, 2]
PDC [‡]	Pyr + CoA + NAD → NADH + H + AcCoA + CO2	—	[1, 2]
PC	ATP + Pyr + HCO3- → ADP + Pi + OAA	R00344	[1, 2]
PCK1	OAA + GTP → PEP + GDP + CO2	R00431	[1, 2]
ME1	Mal + NADP ↔ Pyr + CO2 + NADPH + H	R00216	[1, 2]
CS	AcCoA + OAA + H2O → Cit + CoA	R00351	[1, 2]
ACO1, ACO2	Cit ↔ Isocit	R01324	[1, 2]
IDH3A, IDH3B, IDH3G	Isocit + NAD → OG + CO2 + NADH + H	R00709	[1, 2]
OGDH/ DLST/ DLD [⋆]	OG + NAD + CoA → SucCoA + CO2 + NADH + H	—	[1, 2]
SUCLG2/ SUCLG1, SUCLA2 [◦]	GTP/ATP + Succ + CoA ↔ GDP/ADP + Pi + SucCoA	R00432, R00405	[1, 2]
MDH1, MDH2	Mal + NAD ↔ OAA + NADH + H	R00342	[1, 2]
GLUD1 [•]	OG + NH3 + NAD(P)H + H ↔ Glu + NAD(P)	R00243, R00248	[1, 2]
GOT1	OAA + Glu ↔ Asp + OG	R00355	[1, 2]
GPT	Ala + OG ↔ Pyr + Glu	R00258	[1, 2]
ICL	Isocit + H2O → Succ + Gly	R00479	[1, 2]
MAS	Mal + CoA ← AcCoA + H2O + Gly	R00472	[1, 2]
ex_AcCoA [♮]	AcCoAex → AcCoA	—	—
ex_G6P [♮]	G6P → G6Pex	—	—
ex_Glu [♮]	Gluex → Glu	—	—
ex_Asp [♮]	Aspex → Asp	—	—
ex_Ala [♮]	Alaex → Ala	—	—

[†] The reactions catalyzed by these enzymes were lumped.

[‡] Multi-enzyme complex pyruvate dehydrogenase (lipoamide).

[⋆] Multi-enzyme complex 2-oxoglutarate dehydrogenase.

[◦] The reactions catalyzed by the complex SUCLG2/SUCLG1 and SUCLA2 were lumped in order to reduce the number of EMs and, the nomenclature GTP/ATP is used because the complex SUCLG2/SUCLG1 is GTP specific and the enzyme SUCLA2 is ATP specific.

[•] The enzyme GLUD1 can use both NADH and NADPH and, therefore, the two reactions (R00243, R00248) were lumped and the nomenclature NAD(P)H is used.

[♮] External reactions used to control the influx and outflux of metabolites in the system.

2.2 *Bordetella* metabolism

Table 3 contains the additional reactions from the pentose phosphate pathway that together with reactions from glycolysis/gluconeogenesis pathways make the model of *Bordetella* metabolism.

Table 3: Additional reactions from the pentose phosphate pathway present in *Bordetella pertussis* present in the model of *Bordetella* metabolism studied by EMA.

Gene Symbol	Reaction	KEGG Reaction	Ref.
talB	Sed7P + G3P ↔ Ery4P + F6P	R07378	KEGG, BioCyc, [3]
tktA	Sed7P + G3P ↔ R5P + Xyl5P	R07246	KEGG, BioCyc, [3]
tktA[†]	F6P + G3P ↔ Ery4P + Xyl5P	R01067	KEGG, BioCyc, [3]
rpe	Ru5P ↔ Xyl5P	R01529	KEGG, BioCyc, [3]
rpiA	R5P ↔ Ru5P	R01056	KEGG, BioCyc, [3]
ppsA	ATP + Pyr + H2O → AMP + PEP + Pi	R00199	KEGG, BioCyc, [3]
ex_Pyr [‡]	Pyr → Pyrex	—	—
ex_G6P [‡]	G6Pex → G6P	—	—

[†] The name is represented with an additional "_b" in the graphical representation of the network, Figure 2.

[‡] External reactions used to control the influx and outflux of metabolites in the system.

3 Elementary modes

3.1 Human metabolism

The elementary modes of the model without glyoxylate cycle enabling the input of glutamate, aspartate and alanine and with the synthesis of G6P are represented in Figure 1.

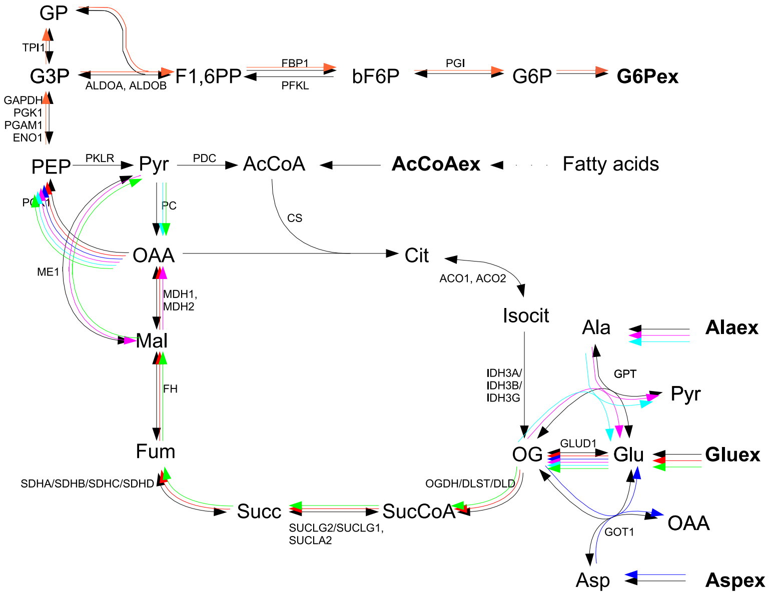


Figure 1: Schematic representation of the EMs with influx of glutamate, aspartate or alanine and synthesis of G6P. The common part of the modes, connection of PEP to G6P, is represented in orange, the different parts are represented with different colors. The two EMs requiring influx of glutamate are complemented, in addition to the parts shown in orange, with the paths in red and green. The two EMs requiring influx of alanine are complemented with the paths in magenta and cyan. The EM requiring influx of aspartate is complemented with the path in blue.

3.2 *Bordetella* metabolism

The model used to calculate the EMs of *B. pertussis*, consisting of gluconeogenesis and the non-oxidative part of the pentose-phosphate pathway is represented in Figure 2. In the figure, gene symbols from reactions that are present in Table 2 and belong also to this model, were rewritten in the official nomenclature for *B. pertussis*. The reactions from glycolysis are not lumped in this network.

The network was automatically generated and the red ellipses are the external metabolites, white ellipses are internal metabolites, yellow diamonds are irreversible reactions and green diamonds are reversible reactions. This symbolism is applied to the following networks as well.

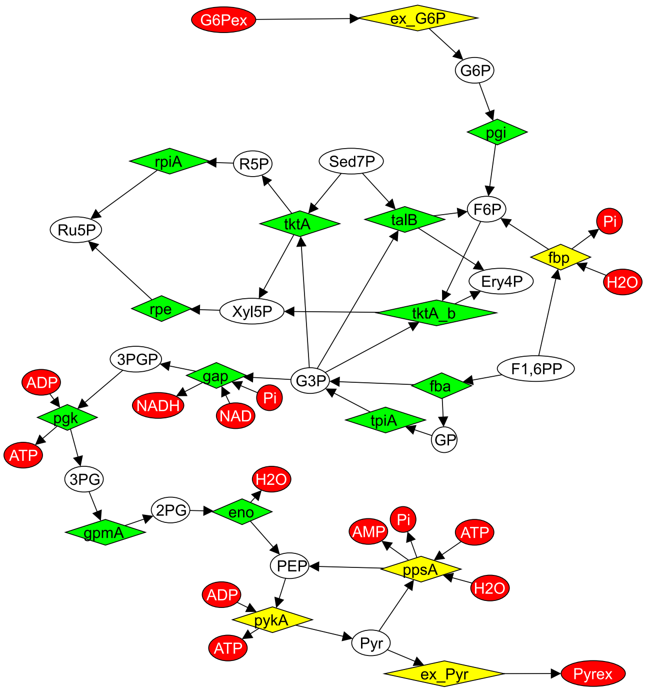


Figure 2: Schematic representation of the model used to study the assumption of steady-state synthesis of pyruvate from G6P in *B. pertussis*.

The network represented in Figure 2 does not have any EM consuming G6P and producing pyruvate, therefore there is no steady-state synthesis of pyruvate from G6P, in agreement with experimental observations.

4 Paths automatically generated by PathFinding and Pathway Hunter Tool algorithms

4.1 Human metabolism

In this subsection, the paths generated by graph-theory based tools for the study of G6P synthesis from AcCoA are shown.

4.1.1 Paths from PathFinding

Figures 3 and 4 contain the paths that connect AcCoA to PEP in humans.

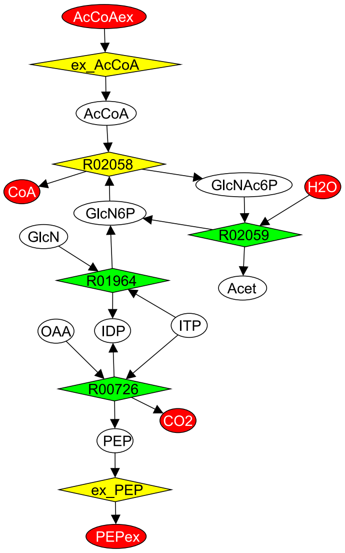


Figure 3: Path with rank 16. Weight=206.0

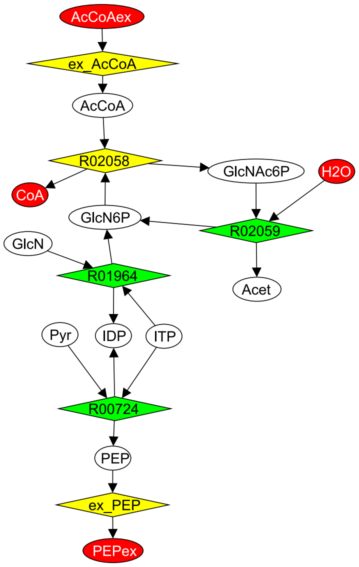


Figure 4: Path with rank 17. Weight=206.0

Figures 5 to 7 contain the paths that connect PEP to G6P in humans.

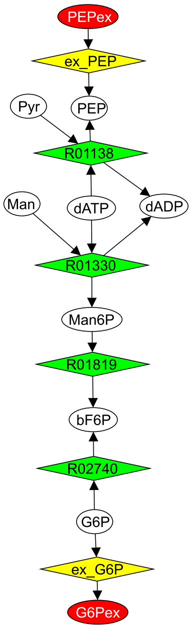


Figure 5: Path with rank 6. Weight=78.0

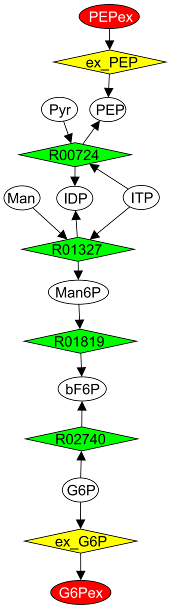


Figure 6: Path with rank 17. Weight=85.0

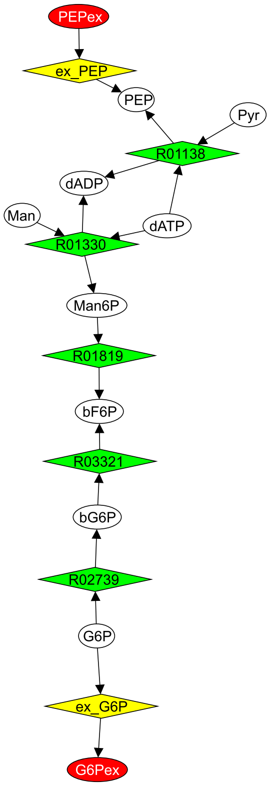


Figure 7: Path with rank 35. Weight=87.0

4.1.2 Pathway Hunter Tool

The PHT was queried using AcCoA as source and G6P as sink. Different combinations of molecular similarity options were used.

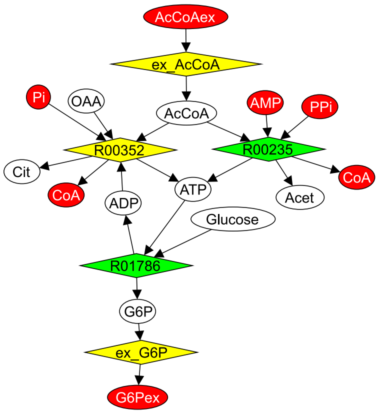


Figure 8: Result from PHT. Program options in PHT: Select one or more organisms as model=Homo sapiens (human); Atom Mapper=Off; Atom Tracer=On; Source Metabolite=C00024; Destination Metabolite=C00668

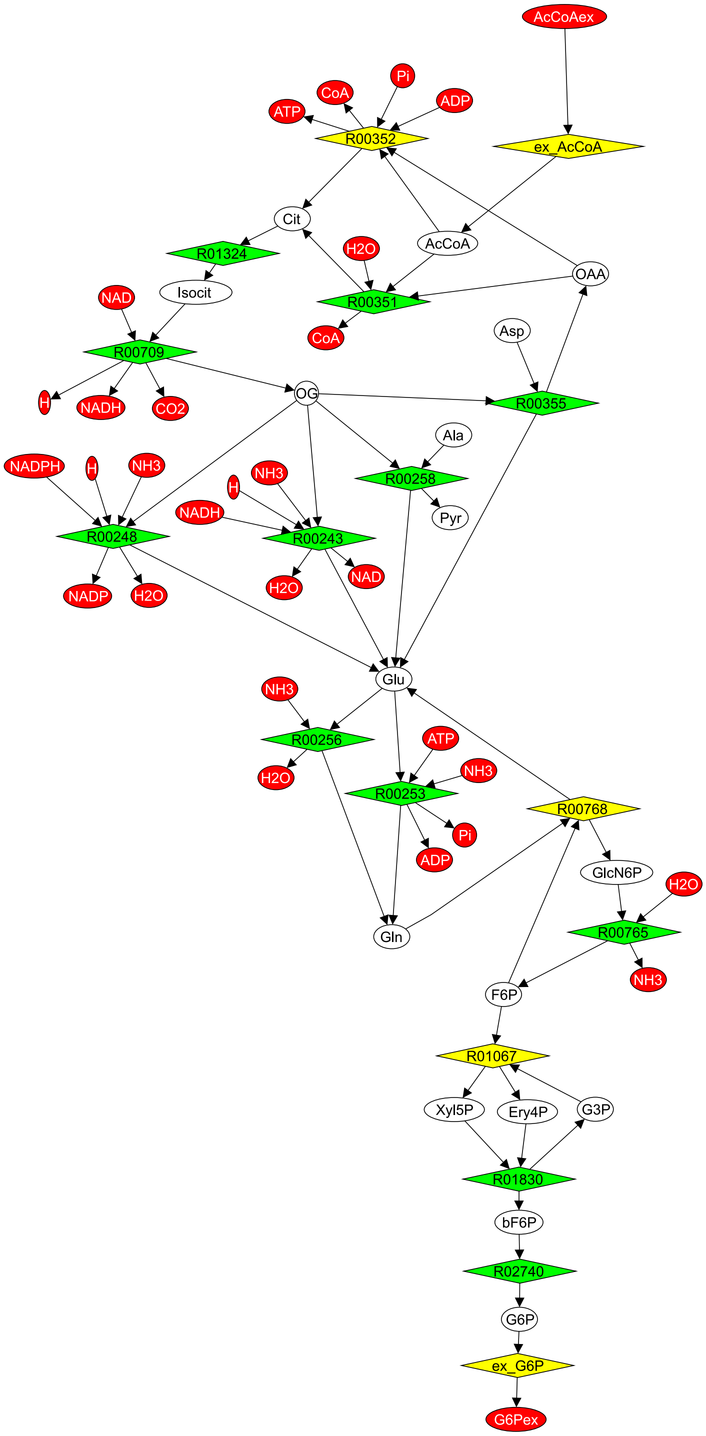


Figure 9: Paths 2 and 3 from the results of PHT. Options: Select one or more organisms as model=Homo sapiens (human); Atom Mapper=Off; Atom Tracer=Off; Source Metabolite=C00024; Destination Metabolite=C00668

4.2 *Bordetella* Metabolism

In this subsection the paths generated by graph-theory based tools for the study of pyruvate synthesis from G6P in *B. pertussis* are shown.

4.2.1 Paths from PathFinding

Figure 10 shows the network generated from merging all the paths in the output of PathFinding when queried for consumption of G6P and synthesis of G3P. The link from G3P to pyruvate was omitted to avoid more complex networks and due to the trivial path linking G3P to pyruvate.

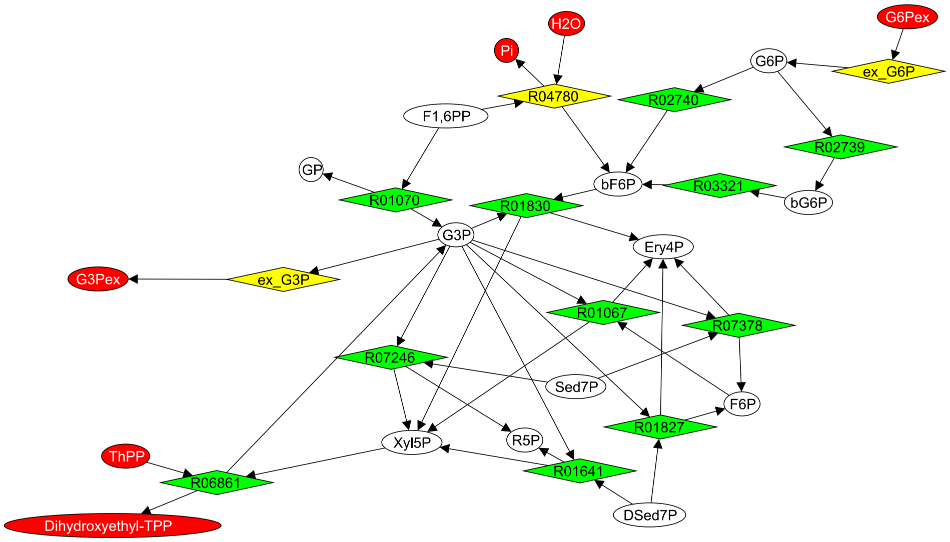


Figure 10: 16 paths found from G6P to G3P, all merged in a single network. Weight range between 60.0 and 74.0

The EMs for the network in Figure 10 were computed. No EM enabling the steady state synthesis of G3P from G6P was obtained.

4.2.2 Pathway Hunter Tool

For PHT a different query was used because the query where G6P and G3P were a source and sink, respectively, did not result in any path. Therefore, we did a query using pyruvate as a sink, instead of G3P. The output of this query is shown in Figure 11. As the query with molecular similarity features *on* was successful we did not perform any further query testing different molecular similarity options. Strangely, G3P is an intermediary metabolite.

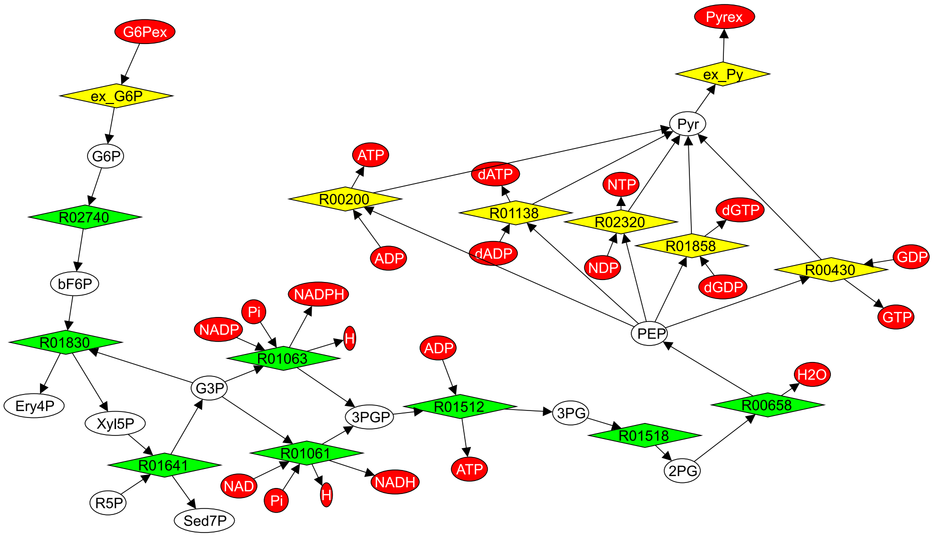


Figure 11: Path from the results of PHT. Options: Select one or more organisms as model=Bordetella pertussis Tohama I; Atom Mapper=On; Atom Tracer=On; Source Metabolite=C00668; Destination Metabolite=C00022

The calculation of EMs in the network presented in Figure 11 revealed that no steady-state flux was possible between G6P and pyruvate.

References

[1] Michal, G. (1999) *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology*, Spektrum Akademischer Verlag, Heidelberg, Berlin, 1st edition.

[2] Voet, D. and Voet, J. (2004) *Biochemistry*, Wiley, USA, 3rd edition.

[3] Armstrong, S. K. and Gross, R. (2007) Primary Metabolism and Physiology of *Bordetella* Species. In Locht, C., (ed.), *Bordetella: Molecular Microbiology*, Taylor & Francis, 1st edition, pp. 165–190.

Supplementary material

Computing the shortest elementary flux modes in genome-scale metabolic networks

Luis F. de Figueiredo, Adam Podhorski, Angel Rubio, Christoph Kaleta, John E. Beasley, Stefan Schuster and Francisco J. Planes

Simulation details

The mathematical model was implemented in a Intel Core® Duo Processor T2400 machine with 2GB RAM . We used CPLEX® 11.0 to solve the model in single thread. The computation of the EFMs in the small model was carried in the same machine using METATOOL 5.1 running on Matlab® R2008b. The visual representation of the EFMs was done using yEd® 3.1.2.

The metabolic model in Schuster *et al.* 1999 was used to enumerate all EFMs. We also tested further modeling strategies. Results are shown below.

The metabolic models of *Escherichia coli* K-12 MG1655 (Feist *et al.* 2007) and *Corynebacterium glutamicum* ATCC 13032 (Kjeldsen and Nielsen, 2009) were used as input network to compute the 10-shortest EFMs. Details as to EFMs are presented below. We also describe minor changes done in the metabolic models in the application of our procedure.

Testing K-shortest EFMs in small scale model

We used the metabolic network from Schuster *et al.*, 1999, to test our mathematical model. Reactions and metabolites names are the same. In Schuster *et al.*, 1999, cofactors were considered as external and consequently they did not figure in the stoichiometric matrix. We simplified the metabolic model by not considering cofactors at all. We derived our network from Figure 2 of the above mentioned work. Details as to the biochemical reactions used here are shown in Table 1. Note that reversible reactions have the sign == separating both sides of the reaction equation, while irreversible reactions have the sign =>. In the computation of the EFMs, the following metabolites were considered external: PG, Alaxt, Gluxt, Aspxt and SucCoAxt. Once the full set of EFMs was enumerated, we computed the *K*-shortest EFMs producing SucCoAxt. Results are shown in Table 2.

Table 1: Metabolic model derived from Schuster *et al*. (1999).

Enzyme abbreviation	Simplified reaction equation
Eno	PG == PEP
Pyk	PEP => Pyr
Pps	Pyr => PEP
AceEF	Pyr => AcCoA
GltA	AcCoA + OAA => Cit
Acn	Cit == IsoCit
Icd	IsoCit => OG
SucAB	OG => SucCoA
SucCD	SucCoA == Succ
Sdh	Succ == Fum
Fum	Fum == Mal
Mdh	Mal == OAA
Icl	IsoCit => Gly + Succ
Mas	Gly + AcCoA => Mal
Ppc	PEP => OAA
Pck	OAA => PEP
AspC	Glu + OAA == Asp + OG
AspA	Asp => Fum
Gdh	OG == Glu
IlvE_AvtA	Pyr + Glu == Ala + OG
GluCon	Glu => Gluxt
AlaCon	Ala => Alaxt
AspCon	Asp => Aspxt
SucCoAcon	SucCoA => SucCoAxt

Table 2. Enumeration of all the EFMs with respective overall equations and enzyme sets and *K*-shortest EFMs producing SucCoAxt

K	*L*	Overall Equation	Enzyme Set	Order of the EFMs in the Schuster et al. (1999)	*K*-shortest EFMs producing SucCoAxt
1*	2	-->	Pck; Ppc	1	--
2*	2	-->	Pps; Pyk	2	--
3	5	PG --> Alaxt	AlaCon; Eno; Gdh; IlvE_AvtA; Pyk	4	--
4	5	PG --> Aspxt	AspC; AspCon; Eno; Gdh; Ppc	3	--
5*	5	-->	AspA; AspC; Fum; Gdh; Mdh	5	--
6	7	PG --> SucCoAxt	Eno; Ppc; SucCoAcon; - Fum; - Mdh; - Sdh; - SucCD	10	1
7	8	PG --> SucCoAxt	AspA; AspC; Eno; Gdh; Ppc; SucCoAcon; - Sdh; - SucCD	8	2
8	9	2 PG --> SucCoAxt	AceEF; Acn; 2 Eno; GltA; Icd; Ppc; Pyk; SucAB; SucCoAcon	13	3
9	9	2 PG --> Gluxt	AceEF; Acn; 2 Eno; Gdh; GltA; GluCon; Icd; Ppc; Pyk	12	--
10	10	2 PG --> SucCoAxt	2 AceEF; Acn; 2 Eno; GltA; Icl; Mas; Mdh; 2 Pyk; SucCoAcon; - SucCD	9	4
11*	11	PG -->	AceEF; Acn; Eno; Fum; GltA; Icd; Mdh; Pyk; Sdh; SucAB; SucCD	15	--
12*	11	PG -->	2 AceEF; Acn; Eno; Fum; GltA; Icl; Mas; 2 Mdh; Pck; 2 Pyk; Sdh	6	--
13	12	3 PG --> 2 SucCoAxt	2 AceEF; Acn; 3 Eno; GltA; Icl; Mas; Ppc; 2 Pyk; 2 SucCoAcon; - Fum; - Sdh; -2 SucCD	11	5
14	13	3 PG --> SucCoAxt	3 AceEF; 2 Acn; 3 Eno; Fum; 2 GltA; Icd; Icl; Mas; 2 Mdh; 3 Pyk; Sdh; SucAB; SucCoAcon	16	6
15	13	3 PG --> Gluxt	3 AceEF; 2 Acn; 3 Eno; Fum; Gdh; 2 GltA; GluCon; Icd; Icl; Mas; 2 Mdh; 3 Pyk; Sdh	14	--
16	13	2 PG --> Aspxt	2 AceEF; Acn; AspC; AspCon; 2 Eno; Fum; Gdh; GltA; Icl; Mas; 2 Mdh; 2 Pyk; Sdh	7	--

* these EFMs have overall equations with one or both sides missing because in our derived model cofactors were discarded. These EFMs represent futile cycles or the conversion of PG in cofactors. More details can be found in Schuster *et al*. 1999.

Computing EFMs in Escherichia coli in genome-scale model

The *Escherichia coli* iAF1260 genome-scale model was used for the computation of EFMs. This can be obtained from BIGG database (Feist *et al.*, 2007). The abbreviations were kept. This model contains three main compartments, cytosol, periplasm and extra-cellular compartment. In the extra-cellular there are several reactions carrying the in/out flux of metabolites from the system. In cytosol there are five reactions called sink which are similar to the previous, namely: DM_4HBA, DM_5DRIB, DM_AACALD, DM_HMFURN and DM_OXAM. All these reactions were removed from the model because they are not required for calculation of EFMs and they do not have biological meaning.

The mathematical model to compute the *K*-shortest EFMS uses Integer Linear Programming, ILP. For this reason, biomass reaction was removed from the model. In Flux Balance Analysis (FBA), the biomass equation plays an important role since it represents the conversion of macromolecules precursors into biomass, which is one of the experimental variables that can be measured. To some extent, the biomass reaction represents the phenotype of the *in silico* organism. In this work we are interested in the computation of pathways synthesizing a given metabolite starting from a set of precursors. Therefore, we are not interested in computing cell growth so that the biomass equation does not have the same relevance as in FBA. The other reactions containing non-integer stoichiometric coefficients were multiplied by a factor of two so as to obtain integer stoichiometric coefficients. These reactions are associated with the respiratory chain, where stoichiometric coefficients associated with oxygen are usually fractional. Biologically, this would mean that the flux calculated should also be multiplied by two in the original reaction form. The reactions multiplied by a factor are the following: CYTBD2pp, CYTBDpp, CYTBO3_4pp, OMMBLHX, OMPHHX, OPHHX and PPPGO.

Table 3 shows the metabolite sets used in the the computation of 10-shortest EFMs producing L-Lysine. The letter in brackets refers the compartment: [e] extra-cellular; [p] periplasm; [c] cytosol.

Table 3. Sets of metabolites used in the 10-shortest EFM modes producing L-lysine

Metabolites	
Medium	External
k[e]	nadp[c]
h2o[e]	co2[c]
glc-D[e]	adp[c]
so4[e]	nadph[c]
na1[e]	h[p]
fe2[e]	h2o[c]
nh4[e]	coa[c]
pi[e]	atp[c]
o2[e]	amp[c]
	nad[c]
	h[c]
	nadh[c]
	pi[c]

LF de Figueiredo, A Podhorski, A Rubio, C Kaleta, JE Beasley, S Schuster and FJ Planes.
Bioinformatics, 25(23):3158-3165, Dec 2009.
-Supplemental Material-

The 10-shortest EFMs were computed for *M*= 10, 100, 1000 and 10000 when cofactors were considered internal and external metabolites, respectively. Results are shown in Table 4 and Table 5. In brackets, correspondent compartment for metabolites are shown, namely [e] - extra-cellular; [p] - periplasm; nothing or [c] - cytosol. Reactions abbreviations are preceded by the stoichiometric coefficient when higher than 1. The minus sign, whick is only present in reversible reactions, means that the reaction runs in the opposite direction. K represents the enumeration order, whilst L is the number of reactions involved in the EFMs.

Table 4. 10-shortest EFMs producing lysine with cofactors set internal and values of *M* ranging from 10 to 10000.

K				*L*	Overall reaction	Reaction set
M= 10	100	1000	10000			
	4	1	1	38	9 glc-D[e] + 4 nh4[e] --> 8 h2o[e] + 16 h[e] + 10 lac-D[e] + 2 lys-L[e] + 4 pyr[e]	2 ASPK; 2 DAPDC; 2 DAPE; 2 DHDPRy; 2 DHDPS; 18 ENO; 9 FBA; 18 GAPD; 9 GLCptspp; 9 GLCtexi; 2 LYSt3pp; 4 NH4tex; 4 NH4tpp; 7 OAADC; 9 PFK; 9 PGI; 9 PPC; 2 SDPDS; 2 SUCOAS; 8 THD2pp; 2 THDPS; 9 TPI; -2 ASAD; -2 ASPTA; -5 ATPS4rpp; -10 D_LACt2pp; -10 D_LACtex; -4 GLUDy; -8 H2Otex; -8 H2Otpp; -16 Htex; -10 LDH_D; -2 LYStex; -18 PGK; -18 PGM; -4 PYRt2rpp; -4 PYRtex; -2 SDPTA
	2	2	2	38	11 glc-D[e] + 8 nh4[e] --> 16 h2o[e] + 18 h[e] + 6 lac-D[e] + 4 lys-L[e] + 8 pyr[e]	4 ASPK; 4 DAPDC; 4 DAPE; 4 DHDPRy; 4 DHDPS; 22 ENO; 11 FBA; 22 GAPD; 11 GLCptspp; 11 GLCtexi; 4 LYSt3pp; 8 NH4tex; 8 NH4tpp; 11 PFK; 11 PGI; 4 PPC; 7 PYK; 4 SDPDS; 4 SUCOAS; 16 THD2pp; 4 THDPS; 11 TPI; -4 ASAD; -4 ASPTA; -10 ATPS4rpp; -6 D_LACt2pp; -6 D_LACtex; -8 GLUDy; -16 H2Otex; -16 H2Otpp; -18 Htex; -6 LDH_D; -4 LYStex; -22 PGK; -22 PGM; -8 PYRt2rpp; -8 PYRtex; -4 SDPTA
	1	3	3	38	11 glc-D[e] + 8 nh4[e] --> 16 h2o[e] + 18 h[e] + 6 lac-D[e] + 4 lys-L[e] + 8 pyr[e]	4 ASPK; 4 DAPDC; 4 DAPE; 4 DHDPRy; 4 DHDPS; 22 ENO; 11 FBA; 22 GAPD; 11 GLCptspp; 11 GLCtex; 4 LYSt3pp; 8 NH4tex; 8 NH4tpp; 11 PFK; 11 PGI; 4 PPC; 7 PYK; 4 SDPDS; 4 SUCOAS; 16 THD2pp; 4 THDPS; 11 TPI; -4 ASAD; -4 ASPTA; -10 ATPS4rpp; -6 D_LACt2pp; -6 D_LACtex; -8 GLUDy; -16 H2Otex; -16 H2Otpp; -18 Htex; -6 LDH_D; -4 LYStex; -22 PGK; -22 PGM; -8 PYRt2rpp; -8 PYRtex; -4 SDPTA
	3	4	4	38	9 glc-D[e] + 4 nh4[e] --> 8 h2o[e] + 16 h[e] + 10 lac-D[e] + 2 lys-L[e] + 4 pyr[e]	2 ASPK; 2 DAPDC; 2 DAPE; 2 DHDPRy; 2 DHDPS; 18 ENO; 9 FBA; 18 GAPD; 9 GLCptspp; 9 GLCtex; 2 LYSt3pp; 4 NH4tex; 4 NH4tpp; 7 OAADC; 9 PFK; 9 PGI; 9 PPC; 2 SDPDS; 2 SUCOAS; 8 THD2pp; 2 THDPS; 9 TPI; -2 ASAD; -2 ASPTA; -5 ATPS4rpp; -10 D_LACt2pp; -10 D_LACtex; -4 GLUDy; -8 H2Otex; -8 H2Otpp; -16 Htex; -10 LDH_D; -2 LYStex; -18 PGK; -18 PGM; -4 PYRt2rpp; -4 PYRtex; -2 SDPTA

	5	5	5	39	27 glc-D[e] + 2 nh4[e] + 25 o2[e] --> 54 h2o[e] + 53 h[e] + lys-L[e] + 52 pyr[e]	28 ADK1; ASPK; 85 ATPS4rpp; 25 CYTBO3_4pp; DAPDC; DAPE; DHDPRy; DHDPS; 27 FBA; 27 GLCptspp; 27 GLCtex; 54 GLYOX3; 54 LDH_D; LYSt3pp; 54 MGSA; 50 NADH16pp; 2 NH4tex; 2 NH4tpp; 25 O2tex; 25 O2tpp; 27 PFK; 27 PGI; PPC; 28 PPS; SDPDS; SUCOAS; 4 THD2pp; THDPS; - ASAD; - ASPTA; -2 GLUDy; -54 H2Otex; -54 H2Otpp; -53 Htex; - LYStex; -52 PYRt2rpp; -52 PYRtex; - SDPTA; -27 TPI
		6	6	39	27 glc-D[e] + 2 nh4[e] + 25 o2[e] --> 54 h2o[e] + 53 h[e] + lys-L[e] + 52 pyr[e]	28 ADK1; ASPK; 85 ATPS4rpp; 25 CYTBO3_4pp; DAPDC; DAPE; DHDPRy; DHDPS; 27 FBA; 27 GLCptspp; 27 GLCtexi; 54 GLYOX3; 54 LDH_D; LYSt3pp; 54 MGSA; 50 NADH16pp; 2 NH4tex; 2 NH4tpp; 25 O2tex; 25 O2tpp; 27 PFK; 27 PGI; PPC; 28 PPS; SDPDS; SUCOAS; 4 THD2pp; THDPS; - ASAD; - ASPTA; -2 GLUDy; -54 H2Otex; -54 H2Otpp; -53 Htex; - LYStex; -52 PYRt2rpp; -52 PYRtex; - SDPTA; -27 TPI
		7	7	39	6 glc-D[e] + 2 nh4[e] --> 4 h2o[e] + 11 h[e] + 8 lac-D[e] + lys-L[e] + 2 pyr[e]	ASPK; DAPDC; DAPE; 6 DHAPT; DHDPRy; DHDPS; 12 ENO; 6 F6PA; 12 GAPD; 6 GLCt2pp; 6 GLCtex; 6 HEX7; LYSt3pp; 2 NH4tex; 2 NH4tpp; 5 OAADC; 6 PPC; SDPDS; SUCOAS; 4 THD2pp; THDPS; 6 TPI; 6 XYLI2; - ASAD; - ASPTA; -4 ATPS4rpp; -8 D_LACt2pp; -8 D_LACtex; -2 GLUDy; -4 H2Otex; -4 H2Otpp; -11 Htex; -8 LDH_D; - LYStex; -12 PGK; -12 PGM; -2 PYRt2rpp; -2 PYRtex; - SDPTA
		8	8	39	6 glc-D[e] + 2 nh4[e] --> 4 h2o[e] + 11 h[e] + 8 lac-D[e] + lys-L[e] + 2 pyr[e]	ASPK; DAPDC; DAPE; 6 DHAPT; DHDPRy; DHDPS; 12 ENO; 6 F6PA; 12 GAPD; 6 GLCt2pp; 6 GLCtexi; 6 HEX1; LYSt3pp; 2 NH4tex; 2 NH4tpp; 5 OAADC; 6 PGI; 6 PPC; SDPDS; SUCOAS; 4 THD2pp; THDPS; 6 TPI; - ASAD; - ASPTA; -4 ATPS4rpp; -8 D_LACt2pp; -8 D_LACtex; -2 GLUDy; -4 H2Otex; -4 H2Otpp; -11 Htex; -8 LDH_D; - LYStex; -12 PGK; -12 PGM; -2 PYRt2rpp; -2 PYRtex; - SDPTA
		9	9	39	3 glc-D[e] + 2 nh4[e] + o2[e] --> 6 h2o[e] + 5 h[e] + lys-L[e] + 4 pyr[e]	ASPK; CYTBD2pp; DAPDC; DAPE; DHDPRy; DHDPS; 6 ENO; 3 FBA; 6 GAPD; 3 GLCptspp; 3 GLCtex; LYSt3pp; 2 NADH17pp; 2 NH4tex; 2 NH4tpp; O2tex; O2tpp; 3 PFK; 3 PGI; 4 PPC; 3 PPCK; 2 PYK; SDPDS; SUCOAS; 4 THD2pp; THDPS; 3 TPI; - ASAD; - ASPTA; -2 GLUDy; -6 H2Otex; -6 H2Otpp; -5 Htex; - LYStex; -6 PGK; -6 PGM; -4 PYRt2rpp; -4 PYRtex; - SDPTA
7	8	10	10	39	4 glc-D[e] + 2 nh4[e] + 2 o2[e] --> 3 5dglcn[e] + 5 h2o[e] + 4 h[e] + lys-L[e]	ASPK; 3 ATPS4rpp; 2 CYTBO3_4pp; DAPDC; DAPE; DHDPRy; DHDPS; EDA; EDD; ENO; GAPD; 4 GLCDpp; 4 GLCNt2rpp; 4 GLCtex; GNK; LYSt3pp; 2 NH4tex; 2 NH4tpp; 2 O2tex; 2 O2tpp; 2 PPC; PPCK; SDPDS; SUCOAS; THD2pp; THDPS; -3 5DGLCNR; -3 5DGLCNt2rpp; -3 5DGLCNtex; - ASAD; - ASPTA; -2 GLUDy; -5 H2Otex; -9 H2Otpp; -4 Htex; - LYStex; - PGK; - PGM; - SDPTA
3	6			39	4 glc-D[e] + 2 nh4[e] + 2 o2[e] --> 3 5dglcn[e] + 5 h2o[e] + 4 h[e] + lys-L[e]	ASPK; 2 ATPS4rpp; 2 CYTBO3_4pp; DAPDC; DAPE; DHDPRy; DHDPS; EDA; EDD; ENO; GAPD; 4 GLCDpp; 4 GLCNt2rpp; 4 GLCtexi; GNK; LYSt3pp; 2 NADTRHD; 2 NH4tex; 2 NH4tpp; 2 O2tex; 2 O2tpp; PPC; SDPDS; SUCOAS; 3 THD2pp; THDPS; -3 5DGLCNR; -3 5DGLCNt2rpp; -3 5DGLCNtex; - ASAD; - ASPTA; -2 GLUDy; -5 H2Otex; -9 H2Otpp; -4 Htex; - LYStex; - PGK; - PGM; - SDPTA

1	7	39	4 glc-D[e] + 2 nh4[e] + 2 o2[e] --> 3 5dglcn[e] + 5 h2o[e] + 4 h[e] + lys-L[e]	ASPK; 2 ATPS4rpp; 2 CYTBO3_4pp; DAPDC; DAPE; DHDPRy; DHDPS; EDA; EDD; ENO; GAPD; 4 GLCDpp; 4 GLCNt2rpp; 4 GLCtex; GNK; LYSt3pp; 2 NADTRHD; 2 NH4tex; 2 NH4tpp; 2 O2tex; 2 O2tpp; PPC; SDPDS; SUCOAS; 3 THD2pp; THDPS; -3 5DGLCNR; -3 5DGLCNt2rpp; -3 5DGLCNtex; - ASAD; - ASPTA; -2 GLUDy; -5 H2Otex; -9 H2Otpp; -4 Htex; - LYStex; - PGK; - PGM; - SDPTA
5	9	39	4 glc-D[e] + 2 nh4[e] --> 4 h2o[e] + 7 h[e] + 4 lac-D[e] + lys-L[e] + 2 pyr[e]	ASPK; DAPDC; DAPE; DHDPRy; DHDPS; 4 EDA; 4 EDD; 4 ENO; 4 G6PDH2r; 4 GAPD; 2 GLCptspp; 2 GLCt2pp; 4 GLCtex; 2 HEX1; LYSt3pp; 2 NH4tex; 2 NH4tpp; 4 PGL; PPC; PYK; SDPDS; SUCOAS; THDPS; - ASAD; - ASPTA; - ATPS4rpp; -4 D_LACt2pp; -4 D_LACtex; -2 GLUDy; -4 H2Otex; -4 H2Otpp; -7 Htex; -4 LDH_D; - LYStex; -4 PGK; -4 PGM; -2 PYRt2rpp; -2 PYRtex; - SDPTA
2	10	39	4 glc-D[e] + 2 nh4[e] --> 4 h2o[e] + 7 h[e] + 4 lac-D[e] + lys-L[e] + 2 pyr[e]	ASPK; DAPDC; DAPE; DHDPRy; DHDPS; 4 EDA; 4 EDD; 4 ENO; 4 G6PDH2r; 4 GAPD; 2 GLCptspp; 2 GLCt2pp; 4 GLCtexi; 2 HEX1; LYSt3pp; 2 NH4tex; 2 NH4tpp; 4 PGL; PPC; PYK; SDPDS; SUCOAS; THDPS; - ASAD; - ASPTA; - ATPS4rpp; -4 D_LACt2pp; -4 D_LACtex; -2 GLUDy; -4 H2Otex; -4 H2Otpp; -7 Htex; -4 LDH_D; - LYStex; -4 PGK; -4 PGM; -2 PYRt2rpp; -2 PYRtex; - SDPTA
4		39	4 glc-D[e] + 2 nh4[e] + 2 o2[e] --> 3 5dglcn[e] + 5 h2o[e] + 4 h[e] + lys-L[e]	ASPK; 3 ATPS4rpp; 2 CYTBO3_4pp; DAPDC; DAPE; DHDPRy; DHDPS; EDA; EDD; ENO; GAPD; 4 GLCDpp; 4 GLCNt2rpp; 4 GLCtexi; GNK; LYSt3pp; 2 NH4tex; 2 NH4tpp; 2 O2tex; 2 O2tpp; 2 PPC; PPCK; SDPDS; SUCOAS; THD2pp; THDPS; -3 5DGLCNR; -3 5DGLCNt2rpp; -3 5DGLCNtex; - ASAD; - ASPTA; -2 GLUDy; -5 H2Otex; -9 H2Otpp; -4 Htex; - LYStex; - PGK; - PGM; - SDPTA
6		39	3 glc-D[e] + 2 nh4[e] + o2[e] --> 6 h2o[e] + 5 h[e] + lys-L[e] + 4 pyr[e]	ASPK; CYTBD2pp; DAPDC; DAPE; 3 DHAPT; DHDPRy; DHDPS; 6 ENO; 3 F6PA; 6 GAPD; 3 GLCabcpp; 3 GLCtex; 3 HEX1; LYSt3pp; 2 NADH17pp; 2 NH4tex; 2 NH4tpp; O2tex; O2tpp; 3 PGI; PPC; 2 PYK; SDPDS; SUCOAS; 4 THD2pp; THDPS; 3 TPI; - ASAD; - ASPTA; -2 GLUDy; -6 H2Otex; -6 H2Otpp; -5 Htex; - LYStex; -6 PGK; -6 PGM; -4 PYRt2rpp; -4 PYRtex; - SDPTA
8		39	3 glc-D[e] + 2 nh4[e] + o2[e] --> 6 h2o[e] + 5 h[e] + lys-L[e] + 4 pyr[e]	ASPK; CYTBD2pp; DAPDC; DAPE; 3 DHAPT; DHDPRy; DHDPS; 6 ENO; 3 F6PA; 6 GAPD; 3 GLCabcpp; 3 GLCtexi; 3 HEX1; LYSt3pp; 2 NADH17pp; 2 NH4tex; 2 NH4tpp; O2tex; O2tpp; 3 PGI; PPC; 2 PYK; SDPDS; SUCOAS; 4 THD2pp; THDPS; 3 TPI; - ASAD; - ASPTA; -2 GLUDy; -6 H2Otex; -6 H2Otpp; -5 Htex; - LYStex; -6 PGK; -6 PGM; -4 PYRt2rpp; -4 PYRtex; - SDPTA
9		39	3 glc-D[e] + 2 nh4[e] + o2[e] --> 6 h2o[e] + 5 h[e] + lys-L[e] + 4 pyr[e]	ASPK; CYTBD2pp; DAPDC; DAPE; 3 DHAPT; DHDPRy; DHDPS; 6 ENO; 3 F6PA; 6 GAPD; 3 GLCabcpp; 3 GLCtex; 3 HEX7; LYSt3pp; 2 NADH17pp; 2 NH4tex; 2 NH4tpp; O2tex; O2tpp; PPC; 2 PYK; SDPDS; SUCOAS; 4 THD2pp; THDPS; 3 TPI; 3 XYLI2; - ASAD; - ASPTA; -2 GLUDy; -6 H2Otex; -6 H2Otpp; -5 Htex; - LYStex; -6 PGK; -6 PGM; -4 PYRt2rpp; -4 PYRtex; - SDPTA

10	39	3 glc-D[e] + 2 nh4[e] + o2[e] --> 6 h2o[e] + 5 h[e] + lys-L[e] + 4 pyr[e]	ASPK; CYTBD2pp; DAPDC; DAPE; DHDPRy; DHDPS; 6 ENO; 3 FBA; 6 GAPD; 3 GLCabcpp; 3 GLCtexi; 3 HEX1; LYSt3pp; 2 NADH17pp; 2 NH4tex; 2 NH4tpp; O2tex; O2tpp; 3 PFK; 3 PGI; PPC; 5 PYK; SDPDS; SUCOAS; 4 THD2pp; THDPS; 3 TPI; - ASAD; - ASPTA; -2 GLUDy; -6 H2Otex; -6 H2Otpp; -5 Htex; - LYStex; -6 PGK; -6 PGM; -4 PYRt2rpp; -4 PYRtex; - SDPTA

Table 5. 10-shortest EFMs producing lysine with cofactors set external and values of *M* ranging from 10 to 10000.

K				*L*	Overall reaction	Reaction set
M=10	100	1000	10000			
6	5	2	1	27	5 atp + 1 glc-D[e] + 3 h2o + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 3 adp + 2 amp + 7 h + 1 lys-L[e] + 2 nadh + 4 nadp + 7 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 FBA; 1 GLCptspp; 1 GLCtex; 2 GLYOX3; 2 LDH_D; 1 LYSt3pp; 2 MGSA; 2 NH4tex; 2 NH4tpp; 1 PFK; 1 PGI; 1 PPC; 2 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 SDPTA; -1 TPI
4	2	5	2	27	1 atp + 1 glc-D[e] + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 1 adp + 3 h2o + 1 h + 1 lys-L[e] + 2 nadh + 4 nadp + 1 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 2 ENO; 1 FBA; 2 GAPD; 1 GLCptspp; 1 GLCtex; 1 LYSt3pp; 2 NH4tex; 2 NH4tpp; 1 PFK; 1 PGI; 1 PPC; 1 SDPDS; 1 SUCOAS; 1 THDPS; 1 TPI; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -2 PGK; -2 PGM; -1 SDPTA
5	4	1	3	27	5 atp + 1 glc-D[e] + 3 h2o + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 3 adp + 2 amp + 7 h + 1 lys-L[e] + 2 nadh + 4 nadp + 7 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 FBA; 1 GLCptspp; 1 GLCtexi; 2 GLYOX3; 2 LDH_D; 1 LYSt3pp; 2 MGSA; 2 NH4tex; 2 NH4tpp; 1 PFK; 1 PGI; 1 PPC; 2 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 SDPTA; -1 TPI
2	6	6	4	27	1 atp + 1 glc-D[e] + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 1 adp + 3 h2o + 1 h + 1 lys-L[e] + 2 nadh + 4 nadp + 1 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 2 ENO; 1 FBA; 2 GAPD; 1 GLCptspp; 1 GLCtexi; 1 LYSt3pp; 2 NH4tex; 2 NH4tpp; 1 PFK; 1 PGI; 1 PPC; 1 SDPDS; 1 SUCOAS; 1 THDPS; 1 TPI; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -2 PGK; -2 PGM; -1 SDPTA
3	1	3	5	27	5 atp + 1 glc-D[e] + 4 h2o + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 2 adp + 3 amp + 8 h + 1 lys-L[e] + 2 nadh + 4 nadp + 8 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHAPT; 1 DHDPRy; 1 DHDPS; 1 F6PA; 1 GLCptspp; 1 GLCtex; 2 GLYOX3; 2 LDH_D; 1 LYSt3pp; 2 MGSA; 2 NH4tex; 2 NH4tpp; 1 PGI; 1 PPC; 3 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 SDPTA; -1 TPI
1	3	4	6	27	5 atp + 1 glc-D[e] + 4 h2o + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 2 adp + 3 amp + 8 h + 1 lys-L[e] + 2 nadh + 4 nadp + 8 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHAPT; 1 DHDPRy; 1 DHDPS; 1 F6PA; 1 GLCptspp; 1 GLCtexi; 2 GLYOX3; 2 LDH_D; 1 LYSt3pp; 2 MGSA; 2 NH4tex; 2 NH4tpp; 1 PGI; 1 PPC; 3 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 SDPTA; -1 TPI
9	10	7	7	28	2 atp + 1 glc-D[e] + 1 h-p + 1 nad + 3 nadph + 2 nh4[e] --> 1 adp + 1 amp + 1 h2o + 3 h + 1 lys-L[e] + 1 nadh + 3 nadp + 3 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 EDA; 1 EDD; 1 ENO; 1 G6PDH2r; 1 GAPD; 1 GLCptspp; 1 GLCtex; 1 LYSt3pp; 2 NH4tex; 2 NH4tpp; 1 PGL; 1 PPC; 1 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 PGK; -1 PGM; -1 SDPTA
7	9	9	8	28	2 atp + 1 glc-D[e] + 1 h-p + 1 nad + 3 nadph + 2 nh4[e] --> 1 adp + 1 amp + 1 h2o + 3 h + 1 lys-L[e] + 1 nadh + 3 nadp + 3 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 EDA; 1 EDD; 1 ENO; 1 G6PDH2r; 1 GAPD; 1 GLCptspp; 1 GLCtexi; 1 LYSt3pp; 2 NH4tex; 2 NH4tpp; 1 PGL; 1 PPC; 1 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 PGK; -1 PGM; -1 SDPTA
10			9	28	3 atp + 1 glc-D[e] + 1 h-p + 1 nad + 3 nadph + 2 nh4[e] --> 3 adp + 1 h2o + 3 h + 1 lys-L[e] + 1 nadh + 3 nadp + 3 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 EDA; 1 EDD; 1 ENO; 1 G6PDH2r; 1 GAPD; 1 GLCabcpp; 1 GLCtexi; 1 HEX1; 1 LYSt3pp; 2 NH4tex; 2 NH4tpp; 1 PGL; 1 PPC; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 PGK; -1 PGM; -1 SDPTA

	7		10	28	4 atp + 1 glc-D[e] + 2 h2o + 1 h-p + 1 nad + 3 nadph + 2 nh4[e] --> 2 adp + 2 amp + 6 h + 1 lys-L[e] + 1 nadh + 3 nadp + 6 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 EDA; 1 EDD; 1 G6PDH2r; 1 GLCptspp; 1 GLCtex; 1 GLYOX3; 1 LDH_D; 1 LYSt3pp; 1 MGSA; 2 NH4tex; 2 NH4tpp; 1 PGL; 1 PPC; 2 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 SDPTA; -1 TPI
	8	8		28	7 atp + 1 glc-D[e] + 6 h2o + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 4 adp + 3 amp + 10 h + 1 lys-L[e] + 2 nadh + 4 nadp + 10 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHAPT; 1 DHDPRy; 1 DHDPS; 1 F6PA; 1 GLCptspp; 1 GLCtexi; 2 GLNS; 2 GLUSy; 2 GLYOX3; 2 LDH_D; 1 LYSt3pp; 2 MGSA; 2 NH4tex; 2 NH4tpp; 1 PGI; 1 PPC; 3 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -1 LYStex; -1 SDPTA; -1 TPI
		10		28	7 atp + 1 glc-D[e] + 6 h2o + 1 h-p + 2 nad + 4 nadph + 2 nh4[e] --> 4 adp + 3 amp + 10 h + 1 lys-L[e] + 2 nadh + 4 nadp + 10 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHAPT; 1 DHDPRy; 1 DHDPS; 1 F6PA; 1 GLCptspp; 1 GLCtex; 2 GLNS; 2 GLUSy; 2 GLYOX3; 2 LDH_D; 1 LYSt3pp; 2 MGSA; 2 NH4tex; 2 NH4tpp; 1 PGI; 1 PPC; 3 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -1 LYStex; -1 SDPTA; -1 TPI
8				28	4 atp + 1 glc-D[e] + 2 h2o + 1 h-p + 1 nad + 3 nadph + 2 nh4[e] --> 2 adp + 2 amp + 6 h + 1 lys-L[e] + 1 nadh + 3 nadp + 6 pi	1 ASPK; 1 DAPDC; 1 DAPE; 1 DHDPRy; 1 DHDPS; 1 EDA; 1 EDD; 1 G6PDH2r; 1 GLCptspp; 1 GLCtexi; 1 GLYOX3; 1 LDH_D; 1 LYSt3pp; 1 MGSA; 2 NH4tex; 2 NH4tpp; 1 PGL; 1 PPC; 2 PPS; 1 SDPDS; 1 SUCOAS; 1 THDPS; -1 ASAD; -1 ASPTA; -2 GLUDy; -1 LYStex; -1 SDPTA; -1 TPI

Computing EFMs in Corynebacterium glutamicumin genome-scale model

We also used the *Corynebacterium glutamicum* ATCC 13032 genome-scale model (Kjeldsen and Nielsen (2009)) to test our K-shortest EFMs procedure. The abbreviations were kept, aside from the exception of some special characters which were removed due to the use of sbml file format. This model contains only two compartments, cytosol and extra-cellular compartment. The metabolites belonging to the extra-cellular compartment are represented by a suffix "*xt*" in the name abbreviation.

Similarly to the *E. coli* model, some reactions containing non-integer stoichiometric coefficients were removed from the model and others adjusted. As opposed to the *E. coli* model, where the macromolecules assembly is represented by the biomass equation, the *C. glutamicum* model contains many reactions representing it. The following reactions were removed from the initial model: fas-IA_MA, Phospholipid-step, plsC, PROTEIN_Ass, DNA_Ass, RNA_Ass, ARABINOGALACTAN_Ass, PEPTIDOGLYCAN_Ass, FREEMYCOLICACID_Ass, MYCOLICACID_Ass, PHOSPHOLIPID_Ass and biomass_Ass.

Some reactions were multiplied by a factor of 2 in order to have integer stoichiometric coefficient: cyto-bd-complex, bc1-aa3-complex and FASC150.

In *C. glutamicum* the uptake of fructose can be done by a fructose and a mannose PTS system, which results in the conversion of fructose into fructose 1-phosphate and fructose 6-phosphate, respectively, Dominguez *et al*.,1998. The former was not present in the model and therefore a reaction was added to the model to better represent the uptake of fructose from the medium. Some reactions had typos. In particular, a mistake in reaction dapB caused a null synthesis of lysine. Table 6 shows the correct version of such reactions.

Table 6. Reactions added or changed in the model due to incorrectness.

Operation	Reaction name	Equation
Added	FRU_PTS	FRUxt + PEP => PYR + F1P
Changed	r3.1.5.1	DGTP => DEOXYGUANOSINE + 3 PI
	dapB	DEHYDRODIPICOLINAT + NADPH => PIPER26DC + NADP
	asd	AP + NADPH => ASPSA + PI + NADP
	mez	MAL + NADP == CO2 + NADPH + PYR
	NO3_H	NO3xt + H_transport_xt => NO3 + H_transport

Table 7 shows the medium and external metabolites defined for C. Glutamicum. Metabolites in extra-cellular compartment have the suffix “xt”.

Table 7. Medium and external metabolites used in the 10-shortest EFM modes producing L-lysine

Metabolites	
Medium	External
GLCxt	H-POxt
NH4xt	PI
PIxt	NADH
O2xt	NAD
Naxt	ADP
SLFxt	CO2
	AMP
	H-PO
	ATP
	NADP
	COA
	NADPH
	H-transport-xt
	H-transport

The 10-shortest EFMs were computed for *M*= 10, 100, 1000 and 10000 when cofactors were considered internal and external metabolites, respectively. Results are shown in Table 8 and Table 9. Note that in Table 8 and Table 9 metabolites in extra-cellular compartment have the suffix “xt”.

Table 8. 10-shortest EFMs producing lysine with cofactors set to internal metabolites and the value of *M* ranging from 10 to 10000

K				*L*	Overall reaction	Reaction set
M=10	100	1000	10000			
1	1	1	1	33	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 1 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 2 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
	3	2	2	33	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	15 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 11 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 bc1aa3complex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 12 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
2	6	6	3	33	2 GLCxt + 3 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt + 1 NH3xt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 3 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 1 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -3 NH3NH4eq; -1 NH3_diffusion; -2 Proton_ATP; -4 pgi
	5	3	4	33	2 GLCxt + 13 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt + 11 NH3xt	15 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 13 amt_ATP; 1 asd; 1 aspB; 5 bc1aa3complex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 1 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -13 NH3NH4eq; -11 NH3_diffusion; -2 Proton_ATP; -4 pgi
3	2	4	5	33	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 glnA; 1 gltBD; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 1 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
4	4	5	6	33	2 GLCxt + 2 NH4xt + 5 O2xt + 1 PIxt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 1 pstB_ATP; 1 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
	7	7	7	33	2 GLCxt + 2 NH4xt + 5 O2xt + 11 PIxt --> 6 CO2xt + 1 LYSxt	15 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 2 amt_ATP; 1 asd; 1 aspB; 5 bc1aa3complex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 11 pstB_ATP; 1 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
7	8	9	8	34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 9 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 gapA; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 8 mqo; 2 ndh; 2 pgk; 2 pgm; 2 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -8 mez; -4 pgi

	9	10	9	34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	15 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 19 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 bc1aa3complex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 gapA; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 8 mqo; 2 ndh; 2 pgk; 2 pgm; 12 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -8 mez; -4 pgi
8			10	34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 3 Odx; 2 UREA_diffusion; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 4 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 UREA_H; -4 pgi
		8		34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	15 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 19 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 bc1aa3complex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 gapA; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 8 mdh; 10 ndh; 2 pgk; 2 pgm; 12 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -8 mez; -4 pgi
9	10			34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 9 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 gapA; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 8 mdh; 10 ndh; 2 pgk; 2 pgm; 2 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -8 mez; -4 pgi
5				34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 1 Odx; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 gapA; 1 gdh; 6 gnd; 1 lysA; 1 lysC; 2 ndh; 2 pgk; 2 pgm; 8 proC; 8 putA; 2 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
6				34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 2 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 gapA; 1 glnA; 1 gltBD; 6 gnd; 1 lysA; 1 lysC; 2 ndh; 2 pgk; 2 pgm; 8 proC; 8 putA; 1 pyc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi
10				34	2 GLCxt + 2 NH4xt + 5 O2xt --> 6 CO2xt + 1 LYSxt	5 ATPasecomplex; 1 GLC_in; 1 GLC_in_PEP; 1 LysE; 5 O2_diffusion; 2 amt_ATP; 1 asd; 1 aspB; 5 cytobdcomplex; 1 dapA; 1 dapB; 1 ddh; 2 eno; 10 gapA; 8 gapB; 1 gdh; 1 glk; 6 gnd; 1 lysA; 1 lysC; 10 ndh; 2 pgk; 2 pgm; 1 ppc; 4 rpe; 2 rpi; 2 tal; 2 tkt_1; 2 tkt_2; 6 zwfopcA; -6 CO2_diffusion; -2 NH3NH4eq; -2 Proton_ATP; -4 pgi

Table 9. 10-shortest EFMs producing lysine with cofactors set to external metabolites and the value of *M* ranging from 10 to 10000.

K				*L*	Overall reaction	Reaction set
M=10	100	1000	10000			
1	1	1	1	21	2 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 2 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 2 PI	1 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 ppc; -2 NH3NH4eq; -1 tpiA; ;
2	2	2	2	22	2 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 2 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 2 PI	1 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 pyc; 1 pyk; -2 NH3NH4eq; -1 tpiA
3	3	3	3	22	3 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 3 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 3 PI	1 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 glnA; 1 gltBD; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 ppc; -2 NH3NH4eq; -1 tpiA
6	4	6	4	23	2 ATP + 2 GLCxt + 2 H-transport-xt + 4 NADPH + 2 NH4xt --> 2 ADP + 2 GLxt + 2 H-transport + 1 LYSxt + 4 NADP + 2 PI	2 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 fda; 2 gapA; 1 gdh; 2 gpsA; 1 lysA; 1 lysC; 2 pfkA; 2 pgi; 2 pgk; 2 pgm; 1 pyc; -2 GL_in_out; -2 NH3NH4eq; -2 glpK
4	8	4	5	23	4 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 4 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 4 PI	1 GLC_in; 1 LysE; 1 Odx; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 glk; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 2 ppc; -2 NH3NH4eq; -1 tpiA
9		10	6	23	3 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 3 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 3 PI	1 GLC_in; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 glk; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 pyc; 2 pyk; -2 NH3NH4eq; -1 tpiA
10	6		7	23	3 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 3 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 3 PI	1 GLC_in; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 glk; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 ppc; 1 pyk; -2 NH3NH4eq; -1 tpiA
	10	8	8	23	2 GLCxt + 2 H-transport-xt + 4 NAD + 4 NADPH + 2 NH4xt --> 2 GLYRxt + 2 H-transport + 1 LYSxt + 4 NADH + 4 NADP	2 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 2 fda; 4 gapA; 1 gdh; 1 lysA; 1 lysC; 2 pfkA; 2 pgi; 4 pgk; 2 pgm; 1 pyc; -2 GLYR_in_out; -2 NH3NH4eq; -2 glxK; -2 tpiA
5	5	7	9	23	3 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 3 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 3 PI	1 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 glnA; 1 gltBD; 1 lysA; 1 lysC; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 pyc; 1 pyk; -2 NH3NH4eq; -1 tpiA
8	9	5	10	23	1 ATP + 1 GLCxt + 2 H-transport-xt + 2 NAD + 4 NADPH + 2 NH4xt --> 1 ADP + 2 H-transport + 1 LYSxt + 2 NADH + 4 NADP + 1 PI	1 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 lysA; 1 lysC; 1 mdh2; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 pyk; -2 NH3NH4eq; -1 mez; -1 tpiA
7	7	9		23	1 ATP + 1 GLCxt + 2 H-transport-xt + 3 NAD + 5 NADPH + 2 NH4xt --> 1 ADP + 2 H-transport + 1 LYSxt + 3 NADH + 5 NADP + 1 PI	1 GLC_in_PEP; 1 LysE; 2 amt_ATP; 1 asd; 1 aspB; 1 dapA; 1 dapB; 1 ddh; 2 eno; 1 fda; 2 gapA; 1 gdh; 1 lysA; 1 lysC; 1 mdh; 1 pfkA; 1 pgi; 2 pgk; 2 pgm; 1 pyk; -2 NH3NH4eq; -1 mez; -1 tpiA

Overview of the genome-scale simulations

Table 10 is a summary of the simulations performed in this work given special attention to the computation time required for each different *M* value.

Table 10. Metabolic network size and computation time of the 10-shortest EFMs. Abbreviations: Met – metabolites; Reac. – reactions; internal – cofactors internal; external – cofactors external; time – time required to compute the 10-shortest EFMs; short. – size of the shortest EFMs.

	Number of			Cofactors			
				internal		external	
Model	Met.	Reac.	*M*	time	short.	time	short.
E. coli	1668	2077	10	36 min	39	4 min	27
			100	483 min	38	7 min	27
			1000	323 min	38	11 min	27
			10000	744 min	38	11 min	27
C. glutamicum	388	437	10	24 s	33	5 s	21
			100	113 s	33	8 s	21
			1000	85 s	33	7 s	21
			10000	88 s	33	3 s	21

References:

Dominguez H, Rollin C, Guyonvarch A, Guerquin-Kern JL, Cocaign-Bousquet M, Lindley ND (**1998**). Carbon-flux distribution in the central metabolic pathways of Corynebacterium glutamicum during growth on fructose. *Eur J Biochem.*, **254**(1)**:**96-102.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ (**2007**) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol.*, **3:**121.

Kjeldsen KR, Nielsen J (**2009**) In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnol Bioeng.*, **102**(2)**:**583-97.

Schuster S, Dandekar T, Fell DA (**1999**) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**(2)**:**53-60.

Reassessment of recently released tools for pathway prediction: The erythrocyte benchmark problem

During the last year of the PhD, several improvements on graph theory approaches for metabolic pathway prediction were published. Here we tested two of these tools. The new version of the tool PathFinding was released right after the publication by de Figueiredo *et al.* (2009b). This one uses the information on RPAIR to improve the quality of the paths (Faust *et al.*, 2009b). A different approach is used in the recently published ReTrace, which relies in the atom mapping of metabolites (Pitkänen *et al.*, 2009). The results presented here are intended to extend the benchmark problem introduced in de Figueiredo *et al.* (2009c). This benchmark problem can be described by the subsystem in Figure 1 which represents the nucleotide metabolism of human erythrocytes. This subsystem is an adaptation of the network published by Schuster and Kenanov (2005) including almost all reaction and metabolite abbreviations. The analysis presented here could be carried at the cell scale (see Introduction). Nevertheless, the reduced metabolism of human erythrocytes allow us to simplify the network and therefore, focus on the specific question we want to answer without loss of generality.

In order to test the new version of PathFinding, we build a directed graph corresponding to the subsystem in Figure 1 using the RPAIR database and the instructions an examples found in the website of NeAT (Brohée *et al.*, 2008). The file can be found in the digital supplemental material of this thesis (see pages 144).

The result present in Table 1 show that PathFinding predicts paths between hypoxanthine to ADP. From the first shortest pathway we can say that the PahtFinding tool does not handle correctly directed RPAIR graphs because two reaction pairs, RP00175> and RP01809>, associated to the same reaction, R01863, appear consecutively in the same path. In a private communication with the developers of PathFinding, the main reason for this behavior is that for RPAIR

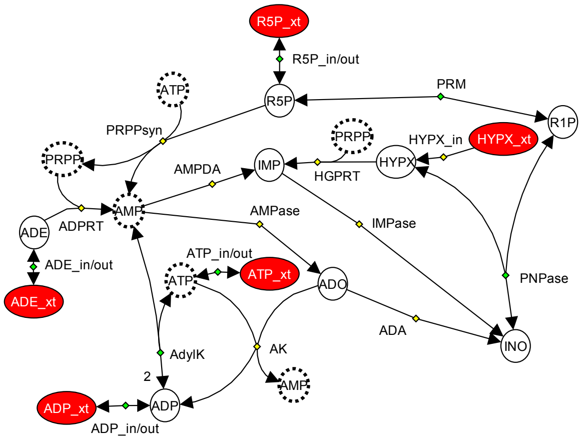


Figure 1: Benchmark problem concerning the conversion of hypoxanthine to ADP in human erythrocytes (de Figueiredo *et al.*, 2009c).

graphs the exclusion group are not directed and therefore, there is no way to avoid that two reaction pairs belonging to the same reaction can appear in the same path when analyzing directed RPAIR graphs.

Start node	End node	Path index	Rank	Weight	Step number	Path
Hypoxanthine	ADP	1	1	62.5	13	Hypoxanthine→ RP00175>→ Inosine→ RP01809>→ alpha-D-Ribose 1-phosphate→ RP01263>→ D-Ribose 5-phosphate→ RP01255>→ 5-Phospho-alpha-D-ribose 1-diphosphate→ RP04092>→ AMP→ RP00056<→ ADP
Hypoxanthine	ADP	2	2	71.5	15	Hypoxanthine→ RP00470>→ IMP→ RP00466>→ Inosine→ RP01809>→ alpha-D-Ribose 1-phosphate→ RP01263>→ D-Ribose 5-phosphate→ RP01255>→ 5-Phospho-alpha-D-ribose 1-diphosphate→ RP04092>→ AMP→ RP00056<→ ADP
Hypoxanthine	ADP	3	3	76.5	9	Hypoxanthine→ RP00470>→ IMP→ RP06442>→ Orthophosphate→ RP04035>→ AMP→ RP00056<→ ADP
Hypoxanthine	ADP	4	4	86.5	11	Hypoxanthine→ RP00175>→ Inosine→ RP01809>→ alpha-D-Ribose 1-phosphate→ RP05759<→ Orthophosphate→ RP04035>→ AMP→ RP00056<→ ADP
Hypoxanthine	ADP	5	5	95.5	13	Hypoxanthine→ RP00470>→ IMP→ RP00466>→ Inosine→ RP01809>→ alpha-D-Ribose 1-phosphate→ RP05759<→ Orthophosphate→ RP04035>→ AMP→ RP00056<→ ADP
Hypoxanthine	ADP	6	6	109.5	17	Hypoxanthine→ RP00175>→ Inosine→ RP01809>→ alpha-D-Ribose 1-phosphate→ RP01263>→ D-Ribose 5-phosphate→ RP01255>→ 5-Phospho-alpha-D-ribose 1-diphosphate→ RP01322>→ IMP→ RP06442>→ Orthophosphate→ RP04035>→ AMP→ RP00056<→ ADP
Hypoxanthine	ADP	7	7	145.5	15	Hypoxanthine→ RP00470>→ IMP→ RP06442>→ Orthophosphate→ RP05759>→ alpha-D-Ribose 1-phosphate→ RP01263>→ D-Ribose 5-phosphate→ RP01255>→ 5-Phospho-alpha-D-ribose 1-diphosphate→ RP04092>→ AMP→ RP00056<→ ADP

Table 1: The 7 shortest paths linking hypoxanthine to ADP computed with PathFinding using the graph based in the RPAIR database.

In order to use ReTrace, this software tool was downloaded and installed locally. We remove many of the entries from LIGAND database in order to reduce the search problem to the subsystem in Figure 1. However, while running ReTrace there were some complications while generating the output. The reason for this behavior is that ReTrace tries to access some reactions that were not present in the reduced LIGAND database. The information about these entries was removed in order to comprise only the reactions, metabolites and reaction pairs present in the subsystem understudy. The input data as well as other data used in this analysis can be found in the digital supplemental material of this thesis (see pages 144).

One additional file was created to force ReTrace to follow the directionality of reactions. A bug in the file retrace.py from ReTrace version 1.20 was found between lines 575 and 581 that did not allow to use the information of reaction directionality. A suggestion to the developers of ReTrace was made for replacing the Python source code between these lines with the one in Listing 1.

```python
if reactionDirConstraints[r]:
  if sub in re.substrates and pro in re.products:
    dir = ">"
  elif sub in re.products and pro in re.substrates:
    dir = "<"
  if dir != None:
    constrained.add(dir)
else:
  if sub in re.substrates and pro in re.products:
    dir = "<"
  elif sub in re.products and pro in re.substrates:
    dir = ">"
  if dir != None:
    constrained.add(dir)
```

Listing 1: Suggested source code to correct the bug found in ReTrace version 1.20.

ReTrace did not find any route linking hypoxantine to ADP. Instead, two routes linking PRPP to ribose 5-phosphate were computed Figures 2 (a) and (b).

Moreover, when computing routes from PRPP to ADP the program only found one incomplete route. When setting Adenine also as source metabolite the status

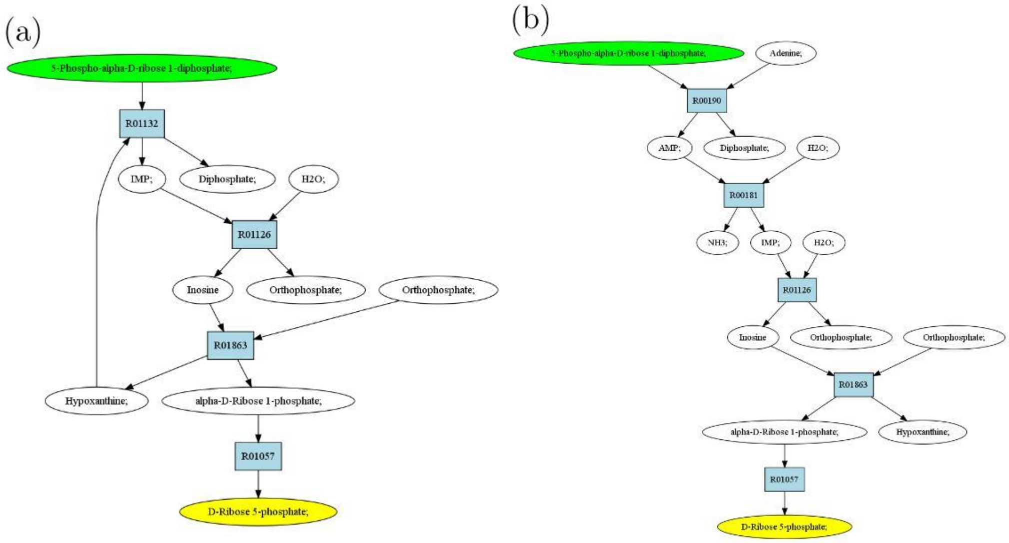


Figure 2: Routes calculated with ReTrace, converting PRPP to ribose 5-phosphate.

of that route changed to complete (Figure 3).

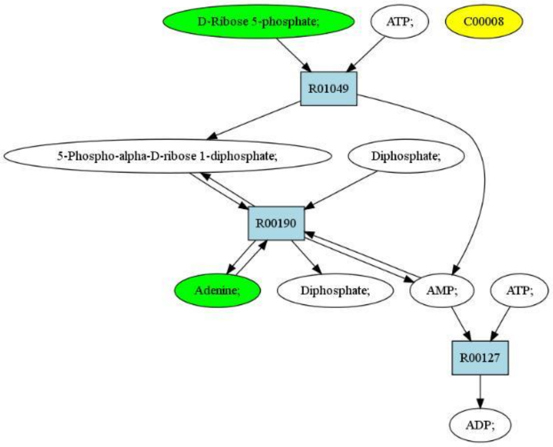


Figure 3: Routes calculated with ReTrace, converting PRPP and Adenine to ADP.

The subsystem in Figure 1 was used to compute the elementary flux modes. The nodes in red correspond to the external metabolites and duplicated nodes are dashed. The three elementary flux modes are depicted in the Discussion, Figures 6.1 (d) to (f).

Digital supplementary material

Beitragende Autoren

Angaben zum Eigenanteil

Titel	Literaturangabe	Autoren	Arbeitsanteil
Can sugars be produced from fatty acids? A test case for pathway analysis tools.	*Bioinformatics*, 25(1):152-158, Jan 2009.	**LF de Figueiredo**,	40%
		S Schuster,	30%
		C Kaleta,	10%
		DA Fell.	20%
Response to comment on 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'.	*Bioinformatics*, 25(24):3330-3331, Dec 2009.	**LF de Figueiredo**,	60%
		S Schuster,	20%
		C Kaleta,	5%
		DA Fell.	15%
Computing the shortest elementary flux mode in genome-scale networks.	*Bioinformatics*, 25(23):3158-3165, Dec 2009.	**LF de Figueiredo**,	55%
		A Podhorski,	5%
		A Rubio,	5%
		C Kaleta,	5%
		JE Beasley,	10%
		S Schuster,	5%
		FJ Planes.	15%
EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks.	In *Lecture Notes in Informatics*, vol P-157, pp 179-189, Bonn 2009. (ISBN: 978-3-88579-251-2)	C Kaleta,	45%
		LF de Figueiredo,	45%
		J Behre,	5%
		S Schuster.	5%
Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks.	*Curr Opin Biotechnol*, 21(4):502-510, Aug 2010.	E Ruppin,	30%
		JA Papin,	20%
		LF de Figueiredo,	20%
		S Schuster.	30%

....

bestätigt Prof. Dr. Stefan Schuster

About the author

Curriculum vitae

Personal data

First name: Luís Filipe

Surname: Domingos Pereira de Figueiredo

Date of birth: the 12th of March of 1982

Place of birth: Lisbon, Portugal

Nationality: Portuguese

Marital status: single

Education

since 2006: PhD studies in Computational Biology at Department of Bioinformatics, under the supervision of Prof. Dr. Stefan Schuster, Friedrich-Schiller-Universität Jena, Jena, Germany.

since 2005: PhD studies in Computational Biology at the PhD Program in Computational Biology from the Instituto Gulbenkian de Ciência, Oeiras, Portugal.

2007: Intensive course in German language and culture from January to March 2007 at the Institut für Interkulturelle Kommunikation e.V. in Jena, Germany.

2005: Diploma work at the BioCentrum under the supervision of Prof. Dr. Jens Nielsen, Danmarks Tekniske Universitet from March to August 2005 in Lyngby, Denmark.

2000-2005: Undergraduate studies in Biological Engineering at the Instituto Superior Técnico, Technical University of Lisbon, Portugal.

1992-2000: Escola Salesiana de Manique, Cascais, Portugal

Working Experience

2010: Member of the organizing committee of the workshop *Integration of OMICS Datasets into Metabolic Pathway Analysis* at the 11th International Conference on Systems Biology the 15th of October in Edinburgh, Scotland, UK.

2009: Member of the organizing committee of the 1st Portuguese Forum on Computational Biology from the 10th to 12th of July 2008 at the Instituto Gulbenkian de Ciência in Oeiras, Portugal.

2005: Trainnee at the Fluxome Sciences A/S under the supervision of Dr. Jochen Föster from March to August 2005 in Lyngby, Denmark.

Teaching

2007-2009: Supervision of the *Proseminar* “Recherche in molekularbiologischen Datenbanken”[¶], study course Bioinformatics of the Friedrich-Schiller-Universität Jena, summer terms.

Publications

- S. Schuster, L. F. de Figueiredo and C. Kaleta. Predicting novel pathways in genome-scale metabolic networks. *Biochem. Soc. Transactions*, 38:1202-1205, 2010.

- K. Bohl[‖], L. F. de Figueiredo[‖], O. Hädicke, S. Klamt, C. Kost, S. Schuster

[¶] together with colleagues

[‖] both authors contributed equally

and C. Kaleta. CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks. In D. Schomburg, A. Grote (Ed.), *Lecture Notes in Informatics - Proceedings*, vol. P-173, Gesellschaft für Informatik, Bonn 2010, pp 71-80. (ISBN:978-3-88579-267-3)

- E. Ruppin, J. A. Papin, L. F. de Figueiredo and S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.*, 21:502-510, 2010.

- L. F. de Figueiredo, S. Schuster, C. Kaleta and D. A. Fell. Response to comment on 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'. *Bioinformatics*, 25:3330-3331, 2009.

- C. Kaleta[**], L. F. de Figueiredo[**], J. Behre and S. Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, P. Stadler (Ed.), *Lecture Notes in Informatics - Proceedings*, vol. P-157, Gesellschaft für Informatik, Bonn 2009, pp 179-189. (ISBN:978-3-88579-251-2)

- 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*, 25:3158-3165, 2009.

- C. Kaleta, L. F. de Figueiredo and S. Schuster. Can the Whole Be Less Than the Sum of its Parts? Pathway Analysis in Genome-Scale Metabolic Networks Using Elementary Flux Patterns. *Genome Research*, 19: 1872-1883, 2009.

- L. F. de Figueiredo, A. Podhorski, A. Rubio, J. E. Beasley, S. Schuster and F. J. Planes. Calculating the K-shortest elementary flux modes in metabolic networks.In I. Troch and F. Breitenecker (Ed.), *Proceedings of the Vienna*

[**] both authors contributed equally

Conference on Mathematical Modelling, vol. 2 of ARGESIM Reports, pp. 736-747, 2009. (ISBN:978-3-901608-35-3)

- C. Kaleta, L. F. de Figueiredo and S. Schuster. Detecting Metabolic Conversions in Genome-Scale Metabolic Networks on the Basis of Elementary Flux Patterns in Subnetworks. In I. Troch and F. Breitenecker (Ed.), *Proceedings of the Vienna Conference on Mathematical Modelling*, vol. 2 of ARGESIM Reports, pp. 748-759, 2009. (ISBN:978-3-901608-35-3)

- L. F. de Figueiredo, S. Schuster, C. Kaleta and D. A. Fell. Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, 25:152-158, 2009. (Full Article as Erratum for de Figueiredo *et al.* , Bioinformatics, 24: 2615-2621, 2008.)

Presentations

- L. F. de Figueiredo, S. Schuster and C. Kaleta. *Metabolic pathway analysis in genome-scale networks and beyond.* Talk at the 2010 International Workshop on Integrative Biological Pathway Analysis and Simulation, Bielefeld, Germany, 18th May 2010.

- L. F. de Figueiredo. *Scaling up elementary flux mode analysis towards genome-scale metabolic networks.* Talk at the JCB seminar, Jena, Germany, 8th April 2010.

- L. F. de Figueiredo, *Computing metabolic pathways in Systems Biology.* Invited talk at the Gulbenkian Alumni Meetings (GAMeets), Oeiras, Portugal, 28th December 2009.

- L. F. de Figueiredo, C. Kaleta, S. Schuster and D. A. Fell. *Benchmarking tools in Metabolic Pathway Analysis.* Highlights track (Talk and Poster) at the ISMB/ECCB 2009, Stockholm, Sweden, 1st July 2009.

- L. F. de Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. Behre, J. E. Beasley, S. Schuster and F. J. Planes. *Computation of elementary flux modes in systems biology.* Talk at the 10th BioPathways meeting, Stockholm, Sweden, 27th June 2009.

- L. F. de Figueiredo, A. Podhorski, A. Rubio, J. E. Beasley, S. Schuster and F. J. Planes. *Calculating the K-shortest Elementary Flux Modes in Metabolic Networks.* Talk at the 6th Vienna International Conference on Mathematical Modelling (MATHMOD 2009), Vienna, Austria, 11th February 2009.

- L. F. de Figueiredo, S. Schuster,C. Kaleta, D.A. Fell. *Benchmarking new tools in Metabolic Pathway Analysis by didactic examples.* Talk at 13th Workshop of the International Study Group for Systems Biology (ISGSB), Elsinore, Denmark, 18th August 2008.

- L.F. de Figueiredo, F.J. Planes, S. Schuster. *Optimization as a Framework for Metabolic Reconstruction - Analysis of Suboptimal Solutions.* Poster at the German Conference on Bioinformatics, Dresden, Germany, 9th to 12th of September of 2008.

- L.F. de Figueiredo , F.J. Planes, S. Schuster. *Optimization as a Framework for Metabolic Reconstruction.* Poster at the 1st Portuguese Forum on Computational Biology, Oeiras, Portugal, 10th to 12th of July of 2008.

- L.F. de Figueiredo , D.A. Fell, S. Schuster. *Metabolic Pathway Analysis as a Tool in Synthetic Biology - Illustration by the example of the conversion of lipids into sugars.* Poster at :the 9th Functional Genomics meeting, Gothenburg, Sweden, 27th to 28th of August of 2007; the German Conference on Bioinformatics, Potsdam, Germany, 26th to 28th September 2007; 2nd Annual Meeting of gulbenkian PhD students, Lagoa, Portugal, 14th to 16th December 2007.

Über den Autor

Lebenslauf

Persönliche Daten

Vorname: Luís Filipe

Nachname: Domingos Pereira de Figueiredo

Geburtsdatum: 12. März 1982

Geburtsort: Lissabon, Portugal

Nationalität: portugiesich

Familienstand: ledig

Ausbildung

seit 2006: Promotion am Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena bei Prof. Dr. Stefan Schuster, Jena, Deutschland.

seit 2005: Promotion im Rahmen des Promotions-Programms über *Computational Biology*, am Instituto Gulbenkian de Ciência, Oeiras, Portugal.

2007: Sprachintensivkurs Deutsch, von Januar bis März 2007, am Institut für Interkulturelle Kommunikation e.V., Jena, Deutschland.

2005: Diplomarbeit am BioCentrum der Technischen Universität Dänemark bei Prof. Dr. Jens Nielsen, von März bis August 2005, Lyngby, Dänemark.

2000-2005: Studium für Biologische Technik am Instituto Superior Técnico der Technischen Universität Lissabon, Lissabon, Portugal.

1992-2000: Escola Salesiana de Manique, Cascais, Portugal

Arbeitserfahrung

2010: Mitglied des Organisationskomitee des Workshops *Integration of OMICS Datasets into Metabolic Pathway Analysis* der 11th International Conference on Systems Biology, 15. Oktober, Edinburg, Schottland, Großbritannien.

2009: Mitglied des Organisationskomitee des 1st Portuguese Forum on Computational Biology], 10. bis 12. Juli 2009, am Instituto Gulbenkian de Ciência, Oeiras, Portugal.

2005: Praktikum bei Fluxome Sciences A/S mit Betreuung von Dr. Jochen Föster, Zeitraum von März bis August 2005, Lyngby, Dänemark.

Lehre

2007-2009: Betreuung der Proseminar “Recherche in molekularbiologischen Datenbanken”[††], für das Studium im Bioinformatik der Friedrich-Schiller-Universität Jena, in dem Sommersemester.

Publikationen

- S. Schuster, L. F. de Figueiredo und C. Kaleta. Predicting novel pathways in genome-scale metabolic networks. *Biochem. Soc. Transactions*, 38:1202-1205, 2010.

[††] zusammen mit Kollegen

- K. Bohl[‡‡], L. F. de Figueiredo[‡‡], O. Hädicke, S. Klamt, C. Kost, S. Schuster and C. Kaleta. CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks. In D. Schomburg, A. Grote (Ed.), *Lecture Notes in Informatics - Proceedings*, vol. P-173, Gesellschaft für Informatik, Bonn 2010, pp 71-80. (ISBN:978-3-88579-267-3)

- E. Ruppin, J. A. Papin, L. F. de Figueiredo und S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.*, 21:502-510, 2010.

- L. F. de Figueiredo, S. Schuster, C. Kaleta und D. A. Fell. Response to comment on 'Can sugars be produced from fatty acids? A test case for pathway analysis tools'. *Bioinformatics*, 25:3330-3331, 2009.

- C. Kaleta[*], L. F. de Figueiredo[*], J. Behre und S. Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, P. Stadler (Ed.), *Lecture Notes in Informatics - Proceedings*, vol. P-157, Gesellschaft für Informatik, Bonn 2009, pp 179-189. (ISBN:978-3-88579-251-2)

- 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*, 25:3158-3165, 2009.

- C. Kaleta, L. F. de Figueiredo und S. Schuster. Can the Whole Be Less Than the Sum of its Parts? Pathway Analysis in Genome-Scale Metabolic Networks Using Elementary Flux Patterns. *Genome Research*, 19: 1872-1883, 2009.

- L. F. de Figueiredo, A. Podhorski, A. Rubio, J. E. Beasley, S. Schuster und F. J. Planes. Calculating the K-shortest elementary flux modes in metabolic

[‡‡] Beide Autoren trugen gleichermaßen zu dieser Arbeit bei

[*] Beide Autoren trugen gleichermaßen zu dieser Arbeit bei

networks.In I. Troch and F. Breitenecker (Ed.), *Proceedings of the Vienna Conference on Mathematical Modelling*, vol. 2 of ARGESIM Reports, pp. 736-747, 2009. (ISBN:978-3-901608-35-3)

- C. Kaleta, L. F. de Figueiredo und S. Schuster. Detecting Metabolic Conversions in Genome-Scale Metabolic Networks on the Basis of Elementary Flux Patterns in Subnetworks. In I. Troch and F. Breitenecker (Ed.), *Proceedings of the Vienna Conference on Mathematical Modelling*, vol. 2 of ARGESIM Reports, pp. 748-759, 2009. (ISBN:978-3-901608-35-3)
- L. F. de Figueiredo, S. Schuster, C. Kaleta und D. A. Fell. Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, 25:152-158, 2009. (Als Erratum für de Figueiredo *et al.* , Bioinformatics, 24: 2615-2621, 2008.)

Vorträge und Poster

- L. F. de Figueiredo, S. Schuster und C. Kaleta. *Metabolic pathway analysis in genome-scale networks and beyond.* Vortrag beim International Workshop on Integrative Biological Pathway Analysis and Simulation 2010, Bielefeld, Deutschland, 18. Mai 2010.
- L. F. de Figueiredo. *Scaling up elementary flux mode analysis towards genome-scale metabolic networks.* Vortrag beim JCB-Seminar, Jena, Deutschland, 8. April 2010.
- L. F. de Figueiredo, *Computing metabolic pathways in Systems Biology.* Eingeladen Vortrag beim Gulbenkian Alumni Meeting (GAMeets), Oeiras, Portugal, 28. Dezember 2009.
- L. F. de Figueiredo, C. Kaleta, S. Schuster und D. A. Fell. *Benchmarking tools in Metabolic Pathway Analysis. Highlights track* (Vortrag und Poster) bei dem ISMB/ECCB 2009, Stockholm, Schweden, 1st July 2009.

- L. F. de Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. Behre, J. E. Beasley, S. Schuster und F. J. Planes. *Computation of elementary flux modes in systems biology.* Vortrage beim 10. BioPathways Treffen, Stockholm, Schweden, 27. Juni 2009.

- L. F. de Figueiredo, A. Podhorski, A. Rubio, J. E. Beasley, S. Schuster und F. J. Planes. *Calculating the K-shortest Elementary Flux Modes in Metabolic Networks.* Vortrag bei der 6. Vienna International Conference on Mathematical Modelling (MATHMOD 2009), Wien, Österreich, 11. Februar 2009.

- L. F. de Figueiredo, S. Schuster,C. Kaleta, D.A. Fell. *Benchmarking new tools in Metabolic Pathway Analysis by didactic examples.* Vortrag beim 13. Workshop of the International Study Group for Systems Biology (ISGSB), Helsingör, Dänemark, 18. August 2008.

- L.F. de Figueiredo, F.J. Planes, S. Schuster. *Optimization as a Framework for Metabolic Reconstruction - Analysis of Suboptimal Solutions.* Poster bei der German Conference on Bioinformatics, Dresden, Deutschland, 9. bis 12. September 2008.

- L.F. de Figueiredo , F.J. Planes, S. Schuster. *Optimization as a Framework for Metabolic Reconstruction.* Poster beim 1. Portuguese Forum on Computational Biology, Oeiras, Portugal, 10. bis 12. Juli 2008.

- L.F. de Figueiredo , D.A. Fell, S. Schuster. *Metabolic Pathway Analysis as a Tool in Synthetic Biology - Illustration by the example of the conversion of lipids into sugars.* Poster beim 9. Functional Genomics Treffen, Göteborg, Schweden, 27. bis 28. August 2007; the German Conference on Bioinformatics, Potsdam, Deutschland, 26. to 28. September 2007; 2. Annual Meeting of gulbenkian PhD students, Lagoa, Portugal, 14. bis 16. Dezember 2007.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe. Mir ist die geltende Promotionsordnung bekannt und ich habe weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte unmittelbare oder mittelbare geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die vorgelegte Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad doctor rerum naturalium (Dr. rer. nat.) beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des o.g. akademischen Grades an einer anderen Hochschule beantragt.

Bei der Auswahl und Auswertung des Materials, sowie bei der Herstellung des Manuskripts hat mir der Lehrstuhl für Bioinformatik der Biologisch-Pharmazeutischen-Fakultät der Friedrich-Schiller-Universität Jena unter der Leitung von Prof. Dr. Stefan Schuster unterstützt.

Jena, den 23. Februar 2011

....

(Luís Filipe Domingos Pereira de Figueiredo)