# Hierarchical decomposition of metabolic networks using k-modules

Arne C. Reimers[1]

August 13, 2015

**Abstract**

The optimal solutions obtained by flux balance analysis (FBA) are typically not unique. Flux modules have recently been shown to be a very useful tool to simplify and decompose the space of FBA-optimal solutions. Since yield-maximization is sometimes not the primary objective encountered in vivo, we are also interested in understanding the space of sub-optimal solutions. Unfortunately, the flux modules are too restrictive and not suited for this task.

We present a generalization, called k-module, that compensates the limited applicability of flux modules to the space of sub-optimal solutions. Intuitively, a k-module is a subnetwork with low connectivity to the rest of the network. Recursive application of k-modules yields a hierarchical decomposition of the metabolic network, which is also known as a branch-decomposition in matroid-theory. In particular, decompositions computed by existing methods like the nullspace-based approach introduced by Poolman and coworkers can be interpreted as branch-decompositions.

With k-modules we can now compare alternative decompositions of metabolic networks to the classical subsystems of glycolysis, TCA-cycle, etc. They can be used to speed up algorithmic problems (theoretically shown for EFM enumeration) and have the potential to present computational solutions in a more intuitive way independently from the classical subsystems.

**Keywords:** metabolic network, flux module, k-module, branch-decomposition, connectivity, matroid

## 1 Introduction

Constraint based methods have proven to be very successful in the analysis of metabolic networks [15, 17], which are used to model metabolic capabilities and predict behaviors of organisms. In contrast to kinetic models, constraint based metabolic network models do not aim to predict a single phenotype, but a space of biologically possible phenotypes. This is achieved

[1] Centre for Mathematics and Computer Science (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands

by excluding unrealistic phenotypes using constraints. This reduces the data requirements enormously such that also large models with thousands of reactions can be built.

Because of the size of the networks, however, even the interplay of very simple constraints can yield very complex and high-dimensional solution spaces that are very hard to comprehensively analyze. This is already the case for networks solely based on the steady-state assumption and irreversibility constraints, which are the basic assumptions for methods like Flux Balance Analysis (FBA) [12, 23] and related methods. The steady-state assumption states that every metabolite must be produced at the same rate as it is consumed. Formally, a vector of reaction rates (flux vector) $v \in \mathbb{R}^{\mathcal{R}}$ is in steady-state if it satisfies

$$Sv = 0,$$

where $S$ is the stoichiometric matrix. We use $\mathcal{R}$ to denote the set of all reactions and $\mathcal{M}$ to denote the set of all metabolites. With a set $\texttt{Irrev} \subseteq \mathcal{R}$ of reactions that are only allowed to operate in forward direction, the *full steady-state flux space*

$$\{v \in \mathbb{R}^{\mathcal{R}} : Sv = 0, v_{\texttt{Irrev}} \geq 0\}$$

is obtained. While extreme pathways [20] or elementary flux modes (EFM) [21, 22] can comprehensively characterize the solution space based on easily understandable pathways, the number of pathways explodes with the size of the network. This makes these approaches only applicable to small networks.

Therefore, many methods try to determine only special properties of the network. For example, FBA asks for the maximal biomass yield for a given uptake of nutrients [12, 23]. Although the space of optimal yield fluxes (optimal yield flux space) also contains many solutions [9, 6], Kelk et al. [5] discovered a method that allows a comprehensive pathway-based description for the optimal yield flux space of many genome-scale networks. They observed that the optimal yield flux space can be decomposed into *flux modules*. For a flux space $P \subseteq \mathbb{R}^{\mathcal{R}}$, a *P-module* is a set of reactions $A \subseteq \mathcal{R}$ for which there exists a vector $d \in \mathbb{R}^{\mathcal{M}}$ with

$$S_A v_A = d \text{ for all } v \in P,$$

where $S_A$ denotes the submatrix of $S$ with only columns corresponding to the reactions in $A$. Similarly, $v_A$ is the subvector of $v$ with only entries corresponding to reactions in $A$. With this definition, originally introduced in [10], the flux modules can be efficiently computed [18]. By computing the pathways through each module, a comprehensive pathway-based description can be obtained efficiently [8]. However, this unfortunately only works for the optimal yield space, because for the full steady-state flux space (without yield-optimality condition) no interesting flux modules can typically be found.

In [19] we introduced the concept of $k$-modules to overcome the limitations of flux modules. There, we followed a mathematical approach and considered the general problem of vertex enumeration of polyhedra. Here, we will now focus on the application to metabolic networks and the biological interpretation of $k$-modules, while keeping the mathematical overhead to a minimum.

With $k$-modules we can define a hierarchical decomposition of metabolic networks that is similar to tree-decompositions and tree-width in graph theory [19]. In parameterized complexity theory there exist many results that show that, if a graph has low tree-width, many otherwise NP-hard problems can be solved efficiently in polynomial time [1, 2]. This gives us the chance to also obtain similar results for metabolic networks.

In Sec. 2, we will introduce and define $k$-modules. They will then form the basis for the hierarchical decompositions discussed in Sec. 3. Finally, an application is given in Sec. 3.2 by considering the problem of EFM enumeration.

## 2 $k$-modules

Let us consider the network shown in Fig. 1 and the set $A = \{r_1, \dots, r_{10}\}$ of reactions. The metabolites $B = \{m_1, m_2, m_3\}$ are each involved in a reaction of $A$ and also in one of the other reactions. These metabolites form the boundary of $A$ and therefore connect $A$ to the rest of the network. Therefore, we call them the boundary metabolites $B(A)$ of $A$:

$$B(A) := \{m \in \mathcal{M} : S_{mr} \neq 0 \neq S_{ms} \; \exists r \in A, s \notin A\}$$

We observe that for any set of reactions $A$, we can compute how well $A$ is connected to the rest of the network using the formula

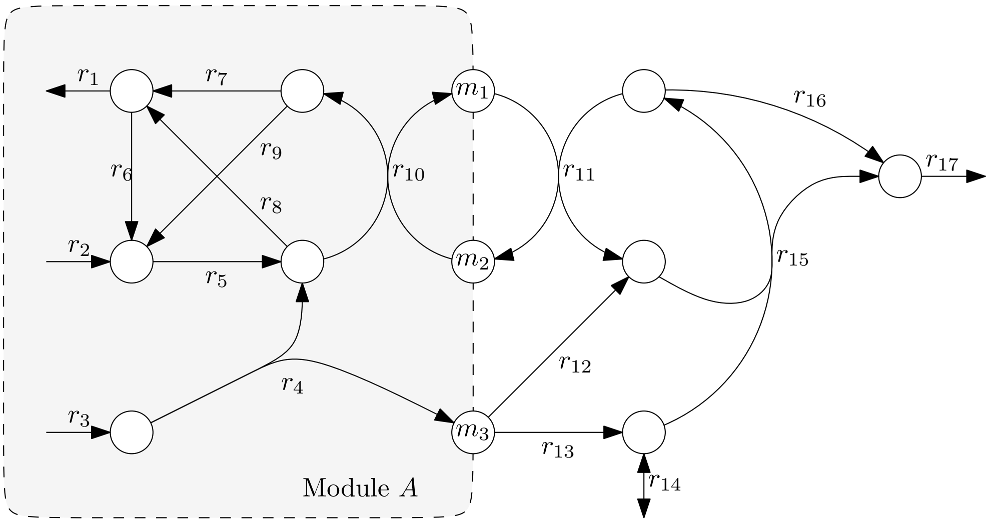

Figure 1: Example toy network. The reactions in $A = \{r_1, \dots, r_{10}\}$ have 3 boundary metabolites and form a 2-module.

$$\mu(A) := |B(A)|.$$

We argue that a set of reactions $A$ with low connectivity $\mu(A)$ should be easy to analyze by itself, because the interaction with the rest of the network that could influence the interpretation of $A$ is low.

However, we also observe for the set $A$ from Fig. 1 that $m_1$ is always produced at the rate by which $m_2$ is consumed (in real networks this can happen for example with currency metabolites like ATP and ADP). Hence, the interaction of $A$ through $m_2$ is already given by the interaction through $m_1$ with the rest of the network. To deal with such redundancies, we use the concept of $k$-modules.

**Definition 1** A set of reactions $A \subseteq \mathcal{R}$ is a *$P$-$k$-module*, if there exists a $d \in \mathbb{R}^k$ and a matrix $D \in \mathbb{R}^{\mathcal{M} \times k}$ (interface) such that for every $v \in P$ there exists an $\alpha \in \mathbb{R}^k$ with

$$S_A v_A = d + D\alpha.$$

We simply write *$k$-module* if $P$ is the full steady-state flux space and the network contains no blocked reactions.

The set $A$ from the example of Fig. 1 is a 2-module, because we can choose

$$D = \begin{array}{c} m_1 \\ m_2 \\ m_3 \end{array} \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The notion of $k$-module allows us now to define an alternative connectivity function that does not only consider network topology but also stoichiometries:

$$\lambda(A) := \min\{k : A \text{ is a } k\text{-module } \}$$

To improve the reader's understanding of $k$-modules and their relation to the previously defined $P$-modules, we here list a few properties that hold in general:

- $\mu(A) = \mu(\mathcal{R} \setminus A)$
- $\lambda(A) = \lambda(\mathcal{R} \setminus A)$
- $\lambda(A) \leq \mu(A)$
- $\lambda(A) \leq |A|$ (note that $\mu(A) \leq |A|$ does not hold in general)
- $\lambda(A) \leq |\mathcal{R} \setminus A|$
- $A$ is a $P$-module if and only if $A$ is a $P$-0-module (i.e., A is a 0-module if $P$ is the full steady-state flux space without blocked reactions).

Furthermore, we want to remark that the simplifications (see [18]) that make an efficient computation of flux modules possible, also apply to $k$-modules. This means that as soon as all reactions with fixed reaction rate have been identified, the connectivity function $\lambda$ can be computed based on linear algebra alone [19], i.e. we have

$$\lambda(A) = \lambda(A \cap V),$$

where $V$ is the set of reactions with variable flux. Furthermore, if all reactions can carry variable flux, then $\lambda$ depends on the stoichiometric matrix alone. For $V = \mathcal{R}$, this leads to the surprising result that (see [19])

$$\lambda(A) = \text{rank}(S_A) + \text{rank}(S_{\mathcal{R} \setminus A}) - \text{rank}(S_{\mathcal{R}}).$$

Since the rank of a matrix can be efficiently computed, we can also compute $\lambda$ efficiently. For example, in the case of the *E. coli* iAF1260 network and its subsystem annotations, we computed that Glycolysis/Gluconeogenesis has a connectivity of 15 and the Citric Acid Cycle has a connectivity of 12. A complete list can be found in the supplementary material.

# 3 Branch-decomposition

While we can compute $\lambda(A)$ efficiently for a given $A$, we are still left with the problem of finding "interesting" sets of reactions $A$ with low $\lambda(A)$. In particular, it is not really clear what an "interesting" set of reactions is. We observe by the properties mentioned above that there are many sets of reactions, where $\lambda(A)$ is low (for example if $A$ contains only one reaction, or if $A$ contains all but one reaction), but which are clearly not interesting.

We conclude that we want to find sets of reactions $A$ where $A$ is large, the complement $\mathcal{R} \setminus A$ is large, and $\lambda(A)$ is low. However, if $A$ contains many reactions, we will be interested to understand $A$ more deeply. Hence, we want to be able to split $A$ recursively into smaller $k$-modules. This leads us to the concept of branch-decompositions and branch-width [3, 4]. Branch-width is related to the more popular concept of tree-width, which is used to measure how tree-like a graph is. In contrast to tree-width, branch-width has a natural extension to matroids and thus, also to metabolic networks.

A branch-decomposition is a subcubic tree, i.e. all nodes have either degree 3 or they are leaves. Every leaf is uniquely associated to a reaction of the network. An example is shown in Fig. 2. We observe that if we remove an edge $e$ of the tree, we get two connected components. Let $A$ be the set of reactions associated to the leaves of one of the connected components. We observe that $\mathcal{R} \setminus A$ is the set of reactions associated to the leaves of the other connected component. Hence, we can annotate the edge $e$ with the value of the connectivity function $\lambda(A) = \lambda(\mathcal{R} \setminus A)$. Note, that alternatively, we can do the same with the connectivity function $\mu$. Furthermore, we observe that by deleting the edge $e$, we get two rooted binary trees, each rooted at a vertex that was incident to $e$. These rooted binary trees give a straightforward rule on how to recursively split the $k$-module $A$ (resp. the $k$-module $\mathcal{R} \setminus A$) into two smaller $k$-modules. In the example of Fig. 2 the reaction set $A = \{r_1, \ldots, r_{10}\}$ would be split up into the reaction set $\{r_1, r_2, r_5, \ldots, r_{10}\}$ with connectivity 2 and the fully coupled reactions $\{r_3, r_4\}$ with connectivity 1.

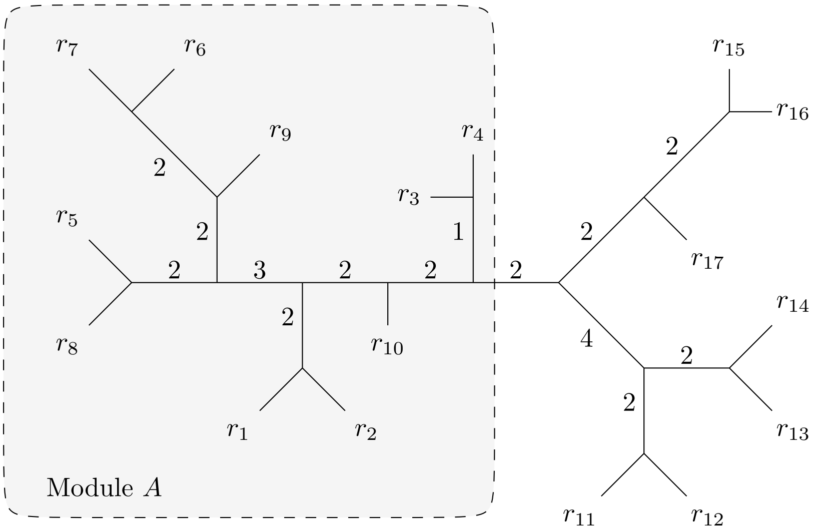

Figure 2: A branch-decomposition of the example network in Fig. 1. The edges are annotated with the corresponding value of the connectivity function, except edges incident to leaves. Edges incident to leaves correspond to $k$-modules containing only one reaction. Hence, they have connectivity 1.

The largest value with which an edge of a branch-decomposition is annotated is called the branch-width of the branch-decomposition. For the example in Fig. 1, we see in Fig. 2 a branch-decomposition with branch-width 4, because $A' := \{r_{11}, r_{12}, r_{13}, r_{14}\}$ is only a 4-module, i.e., $\lambda(A') = 4$. We can now turn the question of finding an interesting $k$-module into the question of finding a branch-decomposition with low branch-width. In particular, we can consider a metabolic network modular if we can find a branch-decomposition with low branch-width.

## 3.1 Computing branch-decompositions for metabolic networks

Unfortunately, it is NP-hard to find a branch-decomposition with minimal branch-width. There exist theoretical results on how to solve this problem in polynomial time for low branch-width instances [13, 14]. However, these algorithms are not likely to be practical, which is why we need to use heuristics. Heuristics have been developed by Ma et al.[7]. Core idea is to compute a similarity matrix by computing the similarity for each pair of reactions. Interestingly, Ma et al. [7] use a similarity measure, which is closely related to the one introduced by Poolman et al. [16]. Indeed, the decomposition computed by Poolman et al. [16] is a branch-decomposition. In Fig. 3 an excerpt of the branch-decomposition computed for an E. coli core network [11] is shown. The full version can be found in the supplementary material.

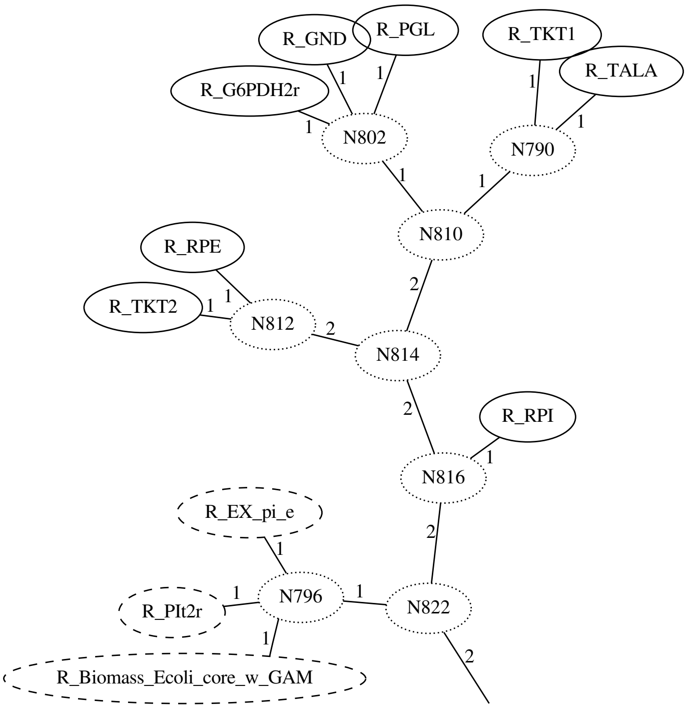

Figure 3: Excerpt of branch-decomposition computed using a variant of Poolman's method [16] for the E. coli core network. All non-dashed leaves belong to the pentose phosphate pathway.

In Tab. 1 we have listed upper bounds on the branch-widths for a set of genome-scale networks computed using variants of Poolman's method. The methods for computing the branch-decompositions are described in the supplementary material and they are implemented in the `cbmpy` toolbox (`cbmpy.sourceforge.net`).

Table 1: Upper bounds on branch-width for some genome-scale metabolic networks.

| Network | reactions | branch-width |
|---|---:|---:|
| *E. coli* core | 95 | 13 |
| *E. coli* iJR904 | 1075 | 40 |
| *E. coli* iAF1260 | 2382 | 59 |
| *H. pylori* iIT341 | 554 | 26 |
| *H. sapiens* recon 1 | 3742 | 99 |
| *H. sapiens* recon 2 | 7440 | 146 |
| *M. barkeri* iAF692 | 690 | 29 |
| *M. tuberculosis* iNJ661 | 1025 | 35 |
| *S. aureus* iSB619 | 743 | 39 |
| *S. cerevisiae* iND750 | 1266 | 53 |

### 3.2 Application: EFM enumeration

In [19] we have shown that the set of EFMs can theoretically be efficiently enumerated if the branch-width of the network is low. To be more precise, we show that given a branch-decomposition of branch-width $k$, we can enumerate all EFMs in time

$$O(|\mathcal{R}||\mathrm{EFM}|^{2k+2}t),$$

where $t$ is the time needed to solve a linear program (LP) and $|\mathrm{EFM}|$ is the number of EFMs [19]. While this is the first result that shows that elementary flux modes can be enumerated in total polynomial time, it unfortunately is not very practical due to the $|\mathrm{EFM}|^{2k+2}$ term.

Without going into details (we refer to [19] for that), the idea is to enumerate recursively pathways that correspond to EFMs through the k-modules in the branch-decomposition. I.e., the pathways of a $k$-module $C$, which is split into two $k$-modules $A$ and $B$, can be computed from the pathways through $A$ and $B$. The bad runtime bound arises from the fact that it is hard to bound the number of pathways through each $k$-module well enough.

## 4 Conclusion

In this article we have shown how we can use $k$-modules to measure how well connected metabolic subsystems are to the rest of the network. While this measure is similar to counting the number of metabolites on the boundary, it is stoichiometry-based and hence it smoothly deals with redundancies due to coupled metabolites.

By recursively decomposing a network using branch-decompositions, $k$-modules give us a measure of modularity. Unfortunately, it is very hard to compute the best branch-decomposition. Therefore, we use heuristics, such as the method by Poolman et al. [16]. While this method gives us a branch-decomposition from which we can recognize familiar subnetworks, its branch-width is not very small. In addition, the connectivity of many subsystems as annotated in the *E. coli* iAF1260 model is also very large compared to their size. Therefore, we conclude that the high branch-width computed by our algorithm is probably not due to its lack of a quality guarantee, but because metabolic networks are not very modular (in the k-modules sense).

# 5 Acknowledgements

I thank Timo Maarleveld, Frank Bruggeman, Brett Olivier, Marie-France Sagot and Leen Stougie for insightful discussions.

This work was carried out during the tenure of an ERCIM Alain Bensoussan Fellowship Programme.

# References

[1] Stefan Arnborg. Efficient algorithms for combinatorial problems on graphs with bounded decomposability - a survey. *BIT Numerical Mathematics*, 25(1):1–23, 1985.

[2] William Cook and Paul Seymour. Tour merging via branch-decomposition. *INFORMS Journal on Computing*, 15(3):233–248, 2003.

[3] Illya V. Hicks, Arie M. C. A. Koster, and Elif Kolotoğlu. *Tutorials in Operations Research*, chapter Branch and Tree Decomposition Techniques for Discrete Optimization. INFORMS, 2005.

[4] Illya V. Hicks and Sang-Il Oum. *Wiley Encyclopedia of Operations Research and Management Science*, chapter Branch-Width and Tangles. Wiley, 2011.

[5] Steven M. Kelk, Brett G. Olivier, Leen Stougie, and Frank J. Bruggeman. Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Scientific Reports*, 2:580, 2012.

[6] Chiraphan Khannapho, Hongjuan Zhao, Bhushan L. Bonde, Andrzej M. Kierzek, Claudio A. Avignone-Rossa, and Michael E. Bushell. Selection of objective function in genome scale flux balance analysis for process feed development in antibiotic production. *Metabolic Engineering*, 10(5):227–233, 2008.

[7] Jing Ma, Susan Margulies, Illya V. Hicks, and Edray Goins. Branch decomposition heuristics for linear matroids. *Discrete Optimization*, 10:102–119, 2013.

[8] Timo R. Maarleveld, Meike Wortel, Brett G. Olivier, Bas Teusink, and Frank J. Bruggeman. Interplay between constraints, objectives, and optimality for genome-scale stoichiometric models. *PLoS Computational Biology*, 11(4):e1004166, 2015.

[9] R. Mahadevan and C.H. Schilling. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5:264–276, 2003.

[10] Arne C. Müller and Alexander Bockmayr. Flux modules in metabolic networks. *Journal of Mathematical Biology*, 69(5):1151–1179. (AC Müller is now called AC Reimers).

[11] Jeffrey D. Orth, Ronan M.T. Fleming, and Bernhard Ø Palsson. Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide. *EcoSal Plus*, Chapter 10.2.1, 2010.

[12] Jeffrey D. Orth, Ines Thiele, and Bernhard Ø. Palsson. What is flux balance analysis? *Nature Biotechnology*, 28:245–248, 2010.

[13] Sang-il Oum and Paul Seymour. Approximating clique-width and branch-width. *Journal of Combinatorial Theory*, Series B 96:514–528, 2006.

[14] Sang-il Oum and Paul Seymour. Testing branch-width. *Journal of Combinatorial Theory*, Series B 97:385–393, 2007.

[15] A. Jason Papin, Joerg Stelling, Nathan D. Price, Steffen Klamt, Stefan Schuster, and Bernhard Ø. Palsson. Comparison of network-based pathway analysis methods. *TRENDS in Biotechnology*, 22(8):400–405, 2004.

[16] Mark G. Poolman, Cristiana Sebu, Michael K. Pidcock, and David A. Fell. Modular decomposition of metabolic systems via null-space analysis. *Journal of Theoretical Biology*, 249:691–705, 2007.

[17] Nathan D. Price, Jennifer L. Reed, and Bernhard Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2:886–897, 2004.

[18] Arne C. Reimers, Frank J. Bruggeman, Brett G. Olivier, and Leen Stougie. Fast flux module detection using matroid theory. *Journal of Computational Biology*, 22(5):414–424, 2015.

[19] Arne C. Reimers and Leen Stougie. A decomposition theory for vertex enumeration of convex polyhedra. arXiv:1404.5584 [cs.CG] `http://arxiv.org/abs/1404.5584`, 2014.

[20] Christophe H. Schilling, David Letscher, and Bernhard Ø. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203:229–248, 2000.

[21] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical systems at steady state. *J. Biol. Systems*, 2:165–182, 1994.

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

[23] Amit Varma and Bernhard Ø. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Nature Biotechnology*, 12:994–998, 1994.