



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Network Analysis and Modeling in Systems Biology

PhD Dissertation

Author: Gabriel Bosque Chacón

Supervisor: Jesús Picó i Marco

February 2017

Instituto Universitario de Automática e Informática Industrial
Departamento de Ingeniería de Sistemas y Automática
Universidad Politécnica de Valencia

Acknowledgements

I do have to thank a lot of people for the help, guidance and support in the process of carrying out this thesis. First and foremost, I would like to thank my thesis supervisor Professor Jesús Picó i Marco for his endless support, intelligence, generosity and guidance. In four years working together I can only say good things about him. Always kind, generous and focused, this thesis is a direct result of his vision. I can only hope I have not let you down. Previous Professor Picó PhDs students Marta Tortajada and specially Francisco Llaneras and Alejandro Vignoni have been immensely helpful and they have always been there to answer my questions and doubts.

I would like to thank Professor Alberto Ferrer Riquelme, from the Departamento de Estadística for all his support and all that I have learn about multivariate statistics. A huge thank you to one of his PhD students, Abel Folch-Fortuny for being always insightful and very sharp, hard-working and very generous with his time and for being my partner in crime during the first couple of years of the PhD. I owe a lot as well to Research Professor Santiago Elena, from the Instituto de Biología Molecular y Celular de Plantas for his support, direction and focus. The first section of this thesis, devoted to the study of viral PPINs, is based in his previous experimental work.

I was privileged enough to enjoy three research visits overseas during my PhD. Many thanks to Professor Masami Hirai and Postdoctoral Researcher Kansuporn Syrudsak for making my stay in the RIKEN Center in Yokohama an absolute delight. I have to thank Professor Mauricio Barahona from the Mathematics Department at Imperial College London for his patience, time and insight and for welcome me twice at ICL. Many thanks to Junior Research Fellow Diego Oyarzún from ICL and Research Fellow Mariano Beguerisse from Oxford University.

A special thanks to the other PhD students I shared these years with: Ale, Yadira, Laguna, Fátima, Gilberto, Diego, Jopipe, Ana, Jesús, Mikel, Alberto, Manuel and Vanessa. For the laughs, the great moments and the memories. I couldn't have hoped for a better group of workmates.

Finally, I would like to thank my uncle Juan and my aunt Amparo for their massive and life-changing support for many years. What I owe you can't be described in words. It really can't. Thanks to my brilliant and crazy cousin Esteban for being always a source of happiness, joy and knowledge for me. Lastly, thank you abuela for showing me how to live a life full of compassion, forgiveness and generosity. I hope this humble thesis brings you some pride and joy, wherever you are. This thesis is dedicated to you.

Abstract

This thesis is dedicated to the study and comprehension of biological networks at the molecular level. The objectives were to analyse their topology, integrate it in a genotype-phenotype analysis, develop richer mathematical descriptions for them, study their community structure and compare different methodologies for estimating their internal fluxes.

The work presented in this document moves around three main axes. The first one is the biological. Which organisms were studied in this thesis? They range from the simplest biological agents, the viruses, in this case the *Potyvirus* genus to prokariotes such as *Escherichia coli* and complex eukariotes (*Arabidopsis thaliana*, *Nicotiana benthamiana*). The second axis refers to which biological networks were studied. Those are protein-protein interaction (PPIN) and metabolic networks (MN). The final axis relates to the mathematical and modelling tools used to generate knowledge from those networks. These tools can be classify in three main branches: graph theory, constraint-based modelling and multivariate statistics.

The document is structured in six parts. The first part states the justification for the thesis, exposes a general thesis roadmap and enumerates its main contributions. In the second part important literature is reviewed, summarized and integrated. From the birth and development of Systems Biology to one of its most popular branches: biological network analysis. Particular focus is put on PPIN and MN and their structure, representations and features. Finally a general overview of the mathematical tools used is presented. The third, fourth and fifth parts represent the central work of this thesis. They deal respectively with genotype-phenotype interaction and classical network analysis, constraint-based modelling methods comparison and modelling metabolic networks and community structure. Finally, in the sixth part the main conclusions of the thesis are summarized and enumerated.

This thesis highlights the vital importance of studying biological entities as systems and how powerful and promising this integrated analysis is. Particularly, network analysis becomes a fundamental avenue of research to gain insight into

those biological systems and to extract, integrate and display this new information. It generates knowledge from just data.

Resumen

Esta tesis está dedicada al estudio y comprensión de redes biológicas a nivel molecular. Los objetivos fueron analizar su topología, integrar esta en un análisis de genotipo-fenotipo, desarrollar descripciones matemáticas propias más completas, estudiar su estructura de comunidades y comparar diferentes metodologías para estimar sus flujos internos.

El trabajo presentado en este documento gira entorno a tres ejes principales. El primero es el biológico. ¿Qué organismos han sido estudiados en esta tesis? Estos van desde los agentes biológicos más simples, los virus, en este caso el género *Potyvirus*, hasta procariotas como *Escherichia coli* y eucariotas complejos (*Arabidopsis thaliana*, *Nicotiana benthamiana*). El segundo eje hace referencia a las redes biológicas estudiadas, que fueron redes de interacción de proteínas (PPIN) y redes metabólicas (MN). El eje final es el de las herramientas matemáticas y de modelización empleadas para interrogar esas redes. Estas herramientas pueden clasificarse en tres grandes grupos: teoría de grafos, modelización basada en restricciones y estadística multivariante.

Este documento está estructurado en seis partes. La primera expone la justificación para la tesis, muestra un mapa visual de la misma y enumera sus contribuciones principales. En la segunda parte, la bibliografía relevante es revisada y resumida. Desde el nacimiento y desarrollo de la Biología de Sistemas hasta una de sus ramas más populares: el análisis de redes biomoleculares. Especial interés es puesto en PPIN y MN: su estructura, representación y características. Finalmente, un resumen general de las herramientas matemáticas usadas es presentado. Los capítulos tercero, cuarto y quinto representan el cuerpo central de esta tesis. Estos tratan respectivamente sobre la interacción de genotipo-fenotipo y análisis topológico clásico de redes, modelos basados en restricciones y modelización de redes metabólicas y su estructura de comunidades. Finalmente, en la sexta parte las principales conclusiones de la tesis son resumidas y expuestas.

Esta tesis pone énfasis en la vital importancia de estudiar los fenómenos biológicos como sistemas y en la potencia y prometedor futuro de este análisis integrativo. En concreto el análisis de redes supone un camino de investigación fundamental

para obtener conocimiento sobre estos sistemas biológicos y para extraer y mostrar información sobre los mismos. Este análisis genera conocimiento partiendo únicamente desde datos.

Resum

Aquesta tesi està dedicada a l'estudi i comprensió de xarxes biològiques a nivell molecular. Els objectius van ser analitzar la seva topologia, integrar aquesta en una anàlisi de genotip-fenotip, desenvolupar descripcions matemàtiques més completes per a elles, estudiar la seva estructura de comunitats o modularitat i comparar diferents metodologies per estimar els fluxos interns.

El treball presentat en aquest document gira entorn de tres eixos principals. El primer és el biològic. ¿Què organismes han estat estudiats en aquesta tesi? Aquests van des dels agents biològics més simples, els virus, en aquest cas el gènere *Potyvirus*, fins procariotes com *Escherichia coli* i eucariotes complexos (*Arabidopsis thaliana*, *Nicotiana benthamiana*). El segon eix fa referència a les xarxes biològiques estudiades, que van ser les xarxes d'interacció de proteïnes (PPIN) i les xarxes metabòliques (MN). L'eix final és el de les eines matemàtiques i de modelització emprades per interrogar aquestes xarxes. Aquestes eines poden classificar-se en tres grans grups: teoria de grafs, modelització basada en restriccions i estadística multivariant.

Aquest document està estructurat en sis parts. La primera exposa la justificació per a la tesi, mostra un mapa visual de la mateixa i enumera les seves contribucions principals. A la segona part, la bibliografia rellevant és revisada i resumida. Des del naixement i desenvolupament de la Biologia de Sistemes fins a una de les seves branques més populars: l'anàlisi de xarxes moleculars. Especial interès és posat en PPIN i MN: la seva estructura, representació i característiques. Finalment, un resum general de les eines matemàtiques utilitzades és presentat. Els capítols tercer, quart i cinquè representen el cos central d'aquesta tesi. Aquests tracten respectivament sobre la interacció de genotip-fenotip i anàlisi topològic clàssic de xarxes, models basats en restriccions i modelització de xarxes metabòliques i la seva estructura de comunitats. Finalment, en la sisena part les principals conclusions de la tesi són resumides i exposades.

Aquesta tesi posa èmfasi en la vital importància d'estudiar els fenòmens biològics com sistemes i en la potència i prometedor futur d'aquesta anàlisi integratiu. En concret l'anàlisi de xarxes suposa un camí d'investigació fonamental per obtenir

coneixement sobre aquests sistemes biològics i per extreure i mostrar informació sobre els mateixos. Aquest anàlisi genera coneixement partint únicament des de dades.

Contents

Acknowledgements	iii
Abstract	v
Contents	xi
1 Justification, outline and contributions	1
2 State of the Art	7
2.1 Molecular systems biology	7
2.1.1 Systems biology	7
From molecules to systems	8
Components vs. systems	9
Top-down vs Bottom-up	9
Characteristics	11
Applications and challenges	11
2.1.2 Molecular Biology	12
Biology and cells	12
Central dogma and molecular agents	14
Biological and system hierarchy	15
Dual causation: physics and evolution	16
2.1.3 Modelling biology	17
What is a model? Definition and purposes	17
Cells as modelling environment	18
Classification of biological models	18

Modelling questions	19
2.1.4 Experimental data integration	19
2.2 Biological networks	20
2.2.1 Networks	21
Macroscopic networks	21
Microscopic networks	22
2.2.2 Metabolic networks	23
2.2.3 Protein-protein interaction networks	27
2.3 Modelling tools	29
2.3.1 Graph theory	29
Nodes and edges	29
Visual and mathematical representations	31
Network properties	31
Functional modules	32
Community detection based on Markov Stability	34
2.3.2 Constraint-based modelling	37
Kinetics models	38
Main features of constraint-based models	38
The stoichiometric matrix	39
Main principles in constraint-based models	40
Constraint-based approach	41
The biomass reaction	43
Brief history, uses and applications of CBMs	43
2.3.3 Multivariate statistics	45
3 Potyvirus network analysis	51
3.1 Introduction	52
3.2 Potyvirus- <i>Arabidopsis thaliana</i> protein-protein interaction network	53
3.2.1 Summary	53
3.2.2 Background	54
3.2.3 Methodology	57
Data collecting	57
Data integration: interactions, matrices and networks	59
Network topology	60
Virus-host interactome	61

3.2.4 Results and discussion	62
VVPI network analysis	62
VHPI network analysis	70
3.3 Potyvirus data fusion	79
3.3.1 Summary	79
3.3.2 Background	79
3.3.3 Methodology	82
Amino acid substitution matrix	82
Partial Least Squares regression (PLS)	83
Cross-validation and Jackknife confidence intervals	83
3.3.4 Results and discussion	84
Protein-protein interaction network reconstruction	84
Mutation and fitness	84
Mathematical modelling	87
Statistical modelling	88
Functional modules	92
3.4 Conclusions	95
4 Constraint-based approaches for metabolic analysis of <i>E. coli</i>	99
4.1 Introduction	100
4.2 Summary	101
4.3 Background	101
4.4 Materials	103
4.4.1 <i>E. coli</i> metabolic model	103
<i>E. coli</i> constraint-based models	104
Main features of the core <i>E. coli</i> metabolic model	104
Mathematical formulation of the core <i>E. coli</i> metabolic model	107
Computational form of the core <i>E. coli</i> metabolic model	108
4.4.2 Experimental <i>E. coli</i> data	109
4.5 Methodology	110
4.5.1 Metabolic Flux Analysis	110
Mathematical formulation	110
Dealing with uncertainty: Interval MFA	111
4.5.2 Flux Balance Analysis	113
4.5.3 Flux Variability Analysis	115
4.5.4 Comparison of methods	117

4.5.5 Statistical description and analysis	118
4.6 Results and discussion	121
4.6.1 Data vs. model agreement	121
4.6.2 Growth rate estimation	122
4.6.3 Flux distributions.	125
4.6.4 Pathway flow profiles	129
4.6.5 Statistical description and comparison	132
4.6.6 Variability analysis	134
Original variables variability: Principal Component Analysis.	134
Outcome variables variability: Partial Least Squares regression	137
4.7 Conclusions	138
5 Metabolic graph analysis of <i>E. coli</i>	141
5.1 Introduction	142
5.2 Summary	143
5.3 Background	143
5.4 Methodology	146
5.5 Results and discussion.	147
5.5.1 Graphs of metabolic networks: constructing graphs that incorporate di- rectionality and context	147
Directed Probabilistic Flux Reaction Graphs of metabolism in the absence of context.	151
Flux-Balance Graphs	154
5.5.2 Graph analysis of the core <i>E. coli</i> metabolic model	156
Probabilistic flux reaction graph of the <i>E. coli</i> core metabolic model	159
Flux-Balance Graphs of the <i>E. coli</i> core metabolic model	160
5.5.3 Structure of FBGs at multiple resolutions	165
5.6 Conclusions	169
6 Conclusions	171
7 Annexes	175
7.1 Annex I: Mutations performed on TEV, distances registered, and fitness measured.	176
7.2 Annex II: collected <i>E. coli</i> experimental data	178

7.3 Annex III: <i>MFA</i> , <i>MFAg</i> and <i>FVA</i> interval estimates for the experimental scenarios.	181
7.4 Annex IV: Toy model example matrices.. . . .	203
7.5 Annex V: Additional comparison between graphs A and D_p	205
7.6 Annex VI: communities from Markov Stability	208
7.6.1 Graph A	211
7.6.2 Graph D_p	212
7.6.3 Graph M_{glc}	213
7.6.4 Graph M_{etoh}	214
7.6.5 Graph M_{anaero}	215
7.6.6 Graph M_{lim}	216
 Bibliography	 217

Chapter 1

Justification, outline and contributions

- The whole is greater than the sum of its parts.

Artistotle

Justification

In a research project such as a PhD thesis, it is usual to follow a linear path. Observation, hypothesis, predictions, experimentation, conclusions. Then you make new observations, think of new hypothesis and so on and so forth. We call that scientific method. Its elegance and power is virtually unbounded. Reality, however, is usually messier. Research is not a path, it's a maze. You have some ideas, you make some experiments and you obtain some findings. Most of the time nature pushes you around, blocks your path, laughs at your preconceived ideas and challenges you relentlessly. This thesis does not fight that apparent chaos, it embraces it. Famous scientist Uri Alon wrote in Alon 2007 that sometimes it is more interesting to follow the path lead by research itself which is convoluted, not linear. Hopefully, at the end of it, we will be able to connect the dots of the work done, just like Steve Jobs said in his famous commencement speech in Stanford University in 2005. Since one can only connect the dots backwards, this will be done at the end of this thesis, in the Conclusions section.

As it has been previously mentioned in the Abstract, this thesis is devoted to the study of biological networks at the molecular level. The work presented here is the result of the collusion of two large fields of knowledge: complex systems and molecular biology. The network is the quintessential complex system: large amount of components, very intricate structure and highly dynamic. Network analysis has been so successful in the last two decades because is one of the disciplines that is better suited to explain the reality that surround us: highly interconnected, ever-changing, global, de-localized and apparently unpredictable. On the other hand, molecular biology is with no doubt the most promising branch of science. Its possibilities are endless: personalized health-care, bioenergetic production, bioremediation, industrial biotechnology and many more world-changing applications can be drawn from manipulation of living organisms. However, to achieve those goals we first have to understand these very complex living systems. This thesis aims to contribute at that exact point. To develop a series of tools and novel analyses to improve our understanding of those organisms, always revolving around the concept of network.

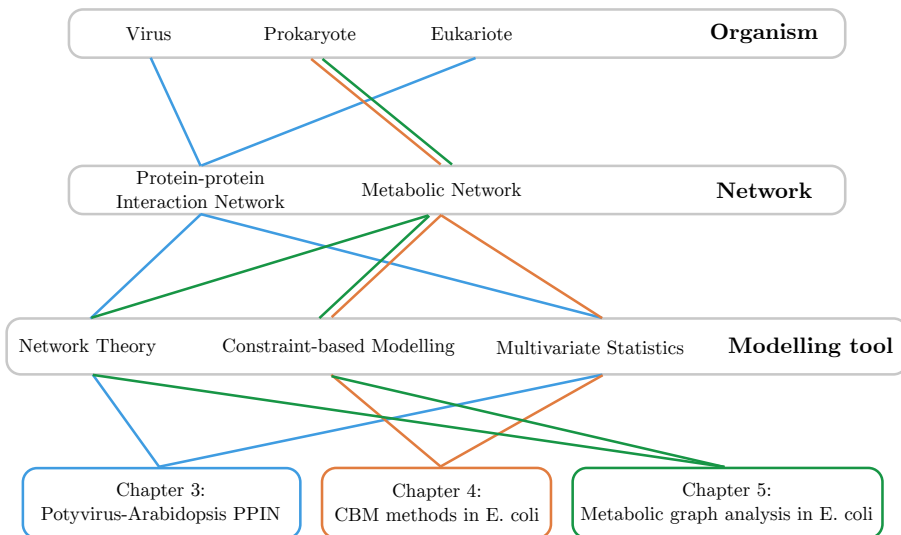


Figure 1.1. Thesis roadmap.

Thesis outline

As it was briefly mentioned in the Abstract, this thesis is divided in five main parts. An extensive literature review about systems biology and network theory is carried out in the first chapter, Chapter 2. Chapters 3, 4 and 5 represent the core of the work done. Since this thesis is devoted to network analysis, it seems appropriate to illustrate the structure of these main chapters with a network representation (Figure 1.1). This figure is structured representing in each level one of the axes defined in the abstract: the biological one at the top, the network one in the middle and the mathematical one at the bottom. The work done in each chapter is a combination of different elements in those axes.

Chapter 3 analyses the topology of protein-protein intereaction networks of viruses of the *Potyvirus* genus with some of their main plant hosts. The interactions between both organisms are studied in detail. Besides, the topology of the virus is used as a channel to explain the relationship between genotype and phenotype through organismal fitness. Chapter 4 is focused on the comparison of different constraint-based methods of metabolic flux determination at steady state in *E. coli*. Its main characteristics are evaluated and the validity of their assumptions is address and quantified. Chapter 5 deals with the mathematical description of metabolic networks and how to make them much more flexible and richer. The community structure of *E. coli* is analysed as well.

Contributions

The main contributions of this work are the following:

- **(Chapter 3)** Novel analysis of protein-protein interaction networks in viral-plant dual systems. Topology, connectivity, similarity, perturbation and effect propagation analyses are proposed to better understand the system.
- **(Chapter 3)** An approach for using molecular network topology as a connection between genotype alterations (mutants) and phenotype measurements (fitness). This produces a triple system (genotype-topology-phenotype) that is able to explain the variability of the organism performance.
- **(Chapter 4)** An exhaustive analysis of the most popular constraint-based methods for obtaining metabolic flux distributions. Data vs. model agreement, biological plausibility, pathway profiles, solution sizes, growth rate prediction capability, robustness and biological premise quantification are proposed to address the quality and potency of the methods.
- **(Chapter 5)** Novel graph description of metabolic networks including topology, stoichiometry, directionality and reaction flux information.

- **(Chapter 5)** Determination of the coarse-grained structure of the *E. coli* metabolic network under different environmental conditions. Robustness, topological analysis and community vs. pathway mixing are used to infer useful biological information.

Publications

The results of this thesis have been published in:

Refereed Journal Papers

- G. Bosque et al. (2014). “Topology analysis and visualization of Potyvirus protein–protein interaction network”. In: *BMC Systems Biology* 8.1, p. 129
- A. Folch-Fortuny et al. (2016). “Fusion of genomic, proteomic and phenotypic data: the case of potyviruses”. In: *Molecular BioSystems* 12.1, pp. 253–261
- Y. Morales et al. (2016). “PFA toolbox: a MATLAB tool for Metabolic Flux Analysis”. In: *BMC Systems Biology* 10, p. 46
- M. Beguerisse-Díaz et al. (2016). “Context-dependent metabolic networks”. In: *arXiv:1605.01639 [physics, q-bio]*

Conference Presentation and Posters

- Bosque, G. and Picó, J. (2013). Redes de Interacción de Proteínas. *XI Simposio CEA de Ingeniería de Control. Automática y Biología celular: una combinación emergente.*
- Vignoni, A., Bosque, G., Tabor, J., and Picó, J. (2013). How to tell bistable cells in which state they should be? On modeling of population fraction control using light. *6th International Meeting on Synthetic Biology (SB6.0).*
- Bosque, G., Picó, J., Folch-Fortuny, A., Ferrer, A., and Elena, S. (2014). Topological analysis and visualization of Potyvirus protein-protein interaction network. *Advanced Lecture Course on Systems Biology.*
- Bosque, G., Picó, J., Folch-Fortuny, A., Ferrer, A., and Elena, S. (2014). Latent Structures-based Modeling of Mutated Protein-Protein Interaction Networks. *12th International Conference on Computational Methods in Systems Biology (CMSB 2014).*
- Bosque, G., Picó, J., Folch-Fortuny, A., Ferrer, A., and Elena, S. (2015). Genomic, proteomic and phenotypic data fusion in potyviruses. *III Reunión*

*de la Red Española Interdisciplinar de Biofísica de Virus (BioFiViNet 3).
Physical Virology: From Structure to Evolution.*

- Bosque, G., Picó, J., Beguerisse-Díaz, M., Oyarzún, D., and Barahona, M. (2015). Community detection in *E. coli* metabolic network using Markov Stability and constrain-based modeling. *2nd Symposium on Complex Biodynamics and Networks*.

Chapter 2

State of the Art

- We are drowning in a sea of data and thirsting for knowledge.

Sydney Brenner

2.1 Molecular systems biology

Systems analysis and study has been applied to many fields in biology such as ecology, systematic biology and evolutionary biology. In the last two decades, the explosion of high-throughput techniques has placed the traditional molecular biology in the reach of a system approach given origin to a new discipline usually called molecular systems biology. Although this systemic approach is not able yet to tackle problems for higher organisms it has seen a lot of success when dealing with unicellular organisms. In this section, the origin, roots and main features and characteristics of this new discipline are reviewed and addressed.

2.1.1 Systems biology

The simultaneous measurement of the majority of molecular agents (metabolites, proteins, mRNA) in cells is nowadays possible thanks to the development of a series of high-throughput techniques. Consequently, complete datasets that define very precisely the state and composition of cells under specific conditions are available for study. Even more than information about the particular elements,

interactions (enzyme-substrate or protein-protein interactions) among those are already becoming well known.

Systems biology (Kitano 2002; Klipp et al. 2005; Palsson 2006) is more aimed to the study of the interactions or links that connect the cellular elements than the elements themselves. Those links spread across the cell connecting all the components and producing a system extremely complex and highly interconnected. Being able to connect, integrate and use all those networks to obtain useful biological information is the fundamental promise of systems biology.

From molecules to systems

The actual etymology of the discipline sheds light to the two fundamental roots in its development: molecular biology and systemic or integrative thinking. Molecular systems biology as it is known now appeared as a viable and promising approach only a couple of decades ago with the proliferation of high-throughput technologies (first the massive parallel sequencing and the other soon after).

The second half of the 20th century saw a series of ground-breaking discoveries that pushed biology into the molecular level: structure of DNA and proteins, the restriction enzymes and cloning. These advances established the biotechnology industry as it is known now. After that, the size and scope of some experimental techniques grew (e.g. PCR) and finally automated DNA sequencing techniques took the world by storm in the mid-1990s reaching the present 'omics' era. This led to the appearance of bioinformatics in order to extract useful information from the massive amount of data that it was being produced. Soon after that, a more nominal approach to the analysis, integration and modelling of this data appeared originating the molecular systems biology discipline.

Non-equilibrium thermodynamics can be considered the very beginning of the study of integrated processes in living systems (Westerhoff and Palsson 2004). From then on most of the system analysis dealt with the bioenergetics of the cell and coupling process principles (Mitchell 1961, 1966). Subsequently, metabolic models developed quickly in the last decades of the past century, either pathway-level kinetic (a brief introduction to kinetic modelling in biology and stochastic simulations can be found in Resat, Petzold, and Pettigrew 2009) or genome-scale constraint-based (Edwards and Palsson 1999).

Probably the most notorious attempt to create a theoretical system framework for biological entities was the general system theory (Bertalanffy 1968). Later on metabolic control analysis (see a general introduction in Fell 1992) and the close biochemical system theory (Savageau 1969) were noteworthy attempts of characterizing properties of networks of chemical reactions. These two methodologies showed that some of the properties of the molecular components escape the tra-

ditional reductionist approach. They depend on the structure of the network and therefore a network-wise kind of analysis is needed. A vast range of approaches are now available for researchers to deal with biological data: kinetic (stochastic and deterministic) and constraint-based modelling, network and graph analysis, multivariate statistics and data fusion.

The need for integrative thinking was always important in biological research. Nowadays, the raw amount of data we are able to produce makes this need stronger than ever. Ideally, both sides of a system approach in biology should feed back each other: experimental research fuelling modelling and mathematical description pointing experimental work in the most promising and useful directions.

Components vs. systems

Components come and go, however the system remains (Palsson 2015). As it has been mentioned, high-throughput technologies completely changed the landscape of biological research. Information was suddenly available not for one gene or protein but for thousands at the same time. However, this shift was not merely a scale-up of the subjects of study. This required a new approach much more suitable for large amounts of data.

Viewing components as part of a system and study them as such is the core idea behind the discipline of systems biology. Although traditional molecular biology research is still very potent and necessary, the systemic approach for solving multitude of biological problems is simply a necessity. Phenotype is the result of a large and very complex combination of elements and therefore to study the genotype-phenotype relationship an integrative approach is usually very useful. This approach is able to highlight the systemic or emergent properties of the system. These properties arise from the structure of the system itself and escape the traditional, reductionistic and more isolated approach. Figure 2.1 shows a simple diagram connecting those concepts.

Top-down vs Bottom-up

Metabolic networks are located at the heart of the present systems biology research. They have been very well studied for many decades and there is a lot of information available to develop and contrast models. When studying metabolic networks one faces a methodology crossroad that actually spreads to any type of molecular network and that defines the two main research streams in systems biology. This dichotomy is usually known as the top-down and the bottom-up approaches (Bruggeman and Westerhoff 2007).

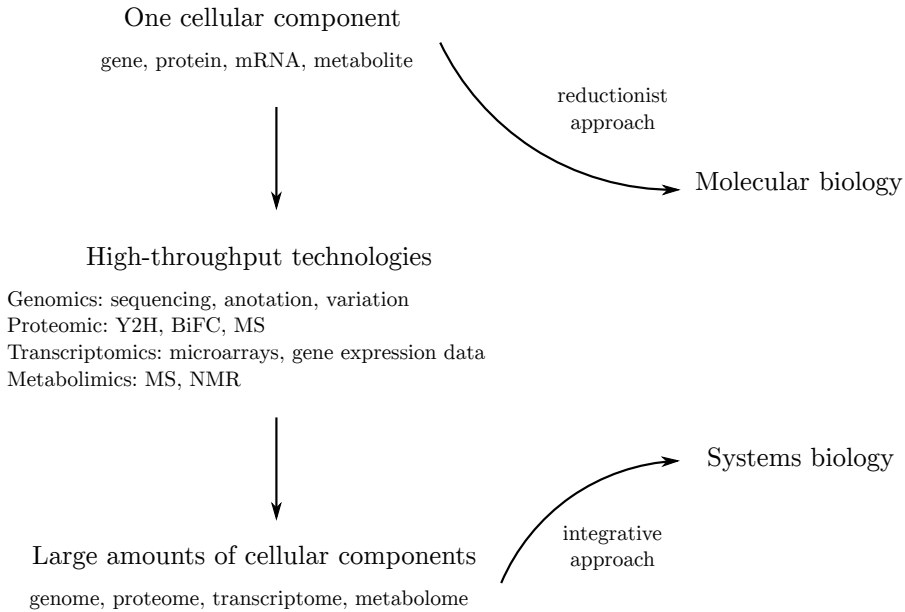


Figure 2.1. Different approaches and data sources of traditional molecular biology and systems biology.

The top-down approach starts with cell-wide experimental data and aims to unravel and characterize new molecular mechanisms closer to the components and their interactions. Therefore it goes from top (genome, proteome, transcriptome data) to bottom (genes, proteins interactions, metabolic reactions). The work-flow starts with the experimental data, which is analysed and integrated to find out correlations between particular elements and concludes with testable hypothesis about the behaviour and relationship among those components. Methodologically it uses inference techniques to build phenomenological models (not based on known mechanisms and biological knowledge) to reverse-engineering from solely cell data.

On the other hand, the bottom-up approach tries to elucidate the functions and states of a subsystem that has been studied and characterized with detail. It begins with the components or constitutive parts (bottom) and then formulates and uses the interactions (usually reactions) among those elements to predict the system behaviour. The models constructed from this perspective are mechanistic rather than purely phenomenological.

Characteristics

Systems biology aims to understand, explain and characterize biological entities. Its ultimate objective is no different from traditional molecular biology or even other branches of the biological sciences like ecology or zoology. It is the system approach to achieve those goals that sets it apart. In the near future this integrative approach will be so embedded in the stream of research that the "system" label will probably disappear and systems biology will be called just biology.

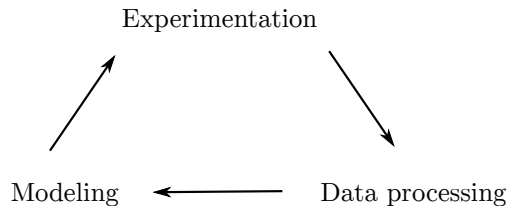


Figure 2.2. Traditional workflow in systems biology.

Systems biology is an inter-disciplinary field with basically three main legs: biological objects of study, experimental techniques and purpose, data gathering, integration and fusion and mathematical modelling and description. The stream from one leg to another is pretty straightforward (Figure 2.2). At the beginning, there is a particular biological system (in this usually a cell or a cellular subsystem) of interest. Experimental techniques available nowadays allow us to generate massive amounts of data for that particular subsystem. This data is integrated if it comes from different techniques or if it partially overlaps with existing data in the literature. Then, mathematical tools are used to construct a model that explains the data gathered and that is able to make useful predictions (Klipp et al. 2005). Finally, these predictions are tested in the laboratory and the process can start again (refining or scaling-up the model or changing the subsystem to study).

Applications and challenges

The present applications of systems biology are significant. Discovery of new metabolic functions (Guzmán et al. 2015), dynamic estimation of metabolic fluxes (Vercaemmen, Logist, and Impe 2014), industrial biotechnology (González-Martínez et al. 2014), gene similarity (Fuxman Bass et al. 2013), biological optimality (Schuetz et al. 2012) and even bacterial computing (Amos et al. 2015). The system approach will become another available toolbox for biological research as molecular biology did at the end of the last century.

In its fast growth systems biology faces important challenges. The way we manage to confront these challenges will surely determine the shape of the biological

research in this century. First and maybe most importantly is the connection between experiments and modelling. Still much needs to be done to close the circle of the Figure 2.2 and generate a real, stable and fluid feedback between data gathering and mathematical modelling. Secondly, data integration is probably the biggest bottleneck in systems biology right now. The ability to find, use and share biological models, networks and equations needs to reach the level of coherence found in genome sequences or protein structure. For this, scientific communication and database development is of the utmost importance. And finally, we need to start developing models that integrate different biological sources of information. Metabolite and protein concentrations, expression data, genome sequence must come together eventually to achieve a true whole-cell model.

Particular fields of application such as medicine, drug development, food production, industrial biotechnology, environmental sustainability benefit already from a more systemic approach of molecular biology. This trend will continue to develop along with other promising sides of the multi-layered biological research.

2.1.2 Molecular Biology

In this section a brief overview of molecular biology is provided. Since biological agents at molecular level (molecules, interactions, pathways, networks) are the main object of study of systems biology, summarizing at least the main features and characteristics of molecular biology is necessary. Among these are the concept of life and biology as discipline, structure and composition of cells, the central dogma of molecular biology, biological hierarchy, evolution and complexity. For a much more detailed introduction see Alberts et al. 2014.

Biology and cells

Life is probably the most complex event ever known to occur. Biology is the science that studies life and living matter in all its forms and phenomena. It tries to explain the origin of life, growth processes, reproduction events, diversity of organisms, structure of living beings and how they relate with their environment, adapt and evolve. This vast field is usually divided in many disciplines such as physiology, morphology, cytology, ecology, biochemistry, molecular biology, genetics and many more. Since the scope of systems biology reaches only the molecular level at this point, I highlight here cell and molecular biology.

The cell is the structural and functional unit of all living organisms. It can replicate in an independent way and constitutes the irreducible building block of life. It has the ability to grow, differentiate and reproduce. Molecular interactions within the cell determine its structure and functions. These interactions take place through several kinds of chemical bonds and forces; ionic, covalent, hydrogen bonds, non polar

bonds and van der Waals forces. The four main classes of molecules bond by those interactions are carbohydrates, lipids, nucleic acids and proteins.

Carbohydrate basic function is energy storage. The individual blocks of all carbohydrates are monosaccharides. They are formed by a chain of three to seven carbon atoms. Glucose is maybe the most important and frequent carbohydrate, being involved in many cellular processes. It is metabolized during glycolysis into ATP and reducing equivalents such as NADH or NADPH. On the other hand, lipids are a very heterogeneous group. They are formed by non polar groups and therefore are highly hydrophobic. Due to this fact, they form hydrophobic compartments essential for some biochemical reactions which need to occur in absence of water. Depending of the kind of lipid their functions range from fat and oil storage, membrane constituents and hormones. Two main nucleic acids exist in cells: deoxyribonucleic acid (DNA) in charge of storing the essential hereditary information (the genes) and ribonucleic acid (RNA) that participate in a much larger number of processes. Its main biological function however is translate the information contained in the DNA into proteins. Both nucleic acids are polymers formed by nucleotides, each containing a nitrogen base, a pentose and one or more phosphate groups.

The final class of molecular elements or components in the cell is proteins. They perform many indispensable functions in the cell. They are the principal cellular actuators. They build up the cytoskeletal framework, form the extracellular matrix, participate in signal transduction and above all they function as catalytic enzymes allowing for many chemical reactions to happen in rates that sustain the cell's life. Proteins are made up of one or more polypeptides. Each one consists of a linear chain of amino acids linked by covalent bonds usually called peptide bonds. Protein structure, which defines its functions, comes from the amino acid distribution along the chain.

Structure is well known for both prokaryotic and eukaryotic cells. The interior of any cell is surrounded by a semipermeable membrane that separates it from the environment. Eukaryotic cells are divided in different compartments; mainly two: the nucleus, where the genetic information is stored and the cytoplasm that contains numerous structures called organelles which carry out different tasks. Endoplasmatic reticulum, mitochondria, Golgi system, transport vesicles and peroxisomes are the most important. Plant cells contain chloroplasts as well. Prokaryotic cells are simpler in general: do not contain nucleus, no subcellular compartmentalization, no cytoskeleton and form mostly single cell organisms.

Central dogma and molecular agents

There is a framework for understanding the transfer of genetic information at the core of molecular biology. In very simple terms it states that in most organisms DNA produces RNA (transcription) and RNA produces proteins (translation). These proteins are later modified and the carry out most of the cellular functions. This flow of information is called central dogma of molecular biology (Crick 1970).

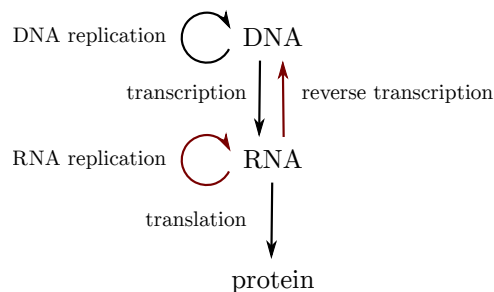


Figure 2.3. Basic diagram of the central dogma of molecular biology. Red arrows represent processes occurring in particular organisms.

This framework (Figure 2.3) becomes the center of the very complex process that the cell uses to regulate itself. Internal and external shifting conditions make the cell to express different proteins to adapt, survive and grow. A series of interconnected networks of interactions and reactions are established among those agents.

Figure 2.4 shows some typical interactions among those agents. Genes $g1$, $g2$ and $g3$ codify proteins $p1$, $p2$ and $p3$. Protein $p1$ inhibits the transcription of $g2$ into $p2$. Proteins $p2$ and $p3$ catalyse the conversion of metabolites $m1$ into $m2$ and $m2$ into $m3$ respectively. Metabolite $m3$ inhibits the catalytic action of $p3$ in a very common feature usually called feedback inhibition by downstream product. Metabolite $m4$ represents an external substance that activates the transcription of $g3$ into $p3$. Regulatory, protein-protein, signalling and metabolic networks although studied usually independently form a whole system in the cell. In this particular example $g1 - p1 - g2$ and $g3 - m4$ form two small regulatory networks and $m1 - p2 - m2 - p3 - m3$ represents a short metabolic pathway.

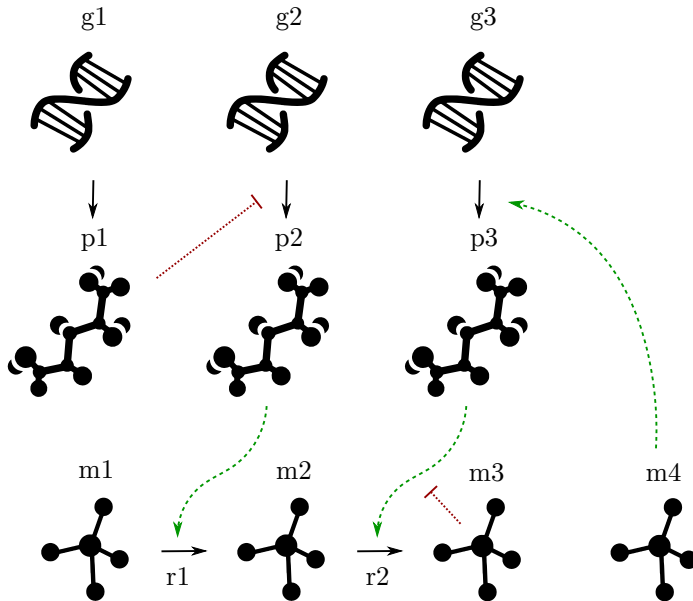


Figure 2.4. Interactions among genes, proteins and metabolites.

Biological and system hierarchy

As it is shown in Figure 2.4 particular agents (genes, mRNAs, proteins) in the cell interact with each other to form more complex structures that can achieve particular cellular goals. This interacting process defines a system and biological hierarchy (Palsson 2015) that starts with particular elements containing the genetic information and ends up with a measurable, defined cellular physiology and behaviour (Figure 2.5). It goes from the genotype to the phenotype. Genes produce proteins through transcription and translation; these proteins catalyse metabolic reactions; these reactions link to more reactions forming massive networks that produce a physical shift in the cell.

Ideally systems biology should be able to develop models that precisely represent that hierarchy from genotype to phenotype. However, sometimes only some levels are represented in a particular model (for instance metabolic constraint-based models that ignore expression data). In other occasions all levels are represented but the scale is small (for example in some kinetic pathway level metabolic models). Although some very valuable attempts have been recently made towards a whole-cell computational model (Karr et al. 2012; Lee et al. 2008), the predictive and explicative capabilities of these models are still very limited.

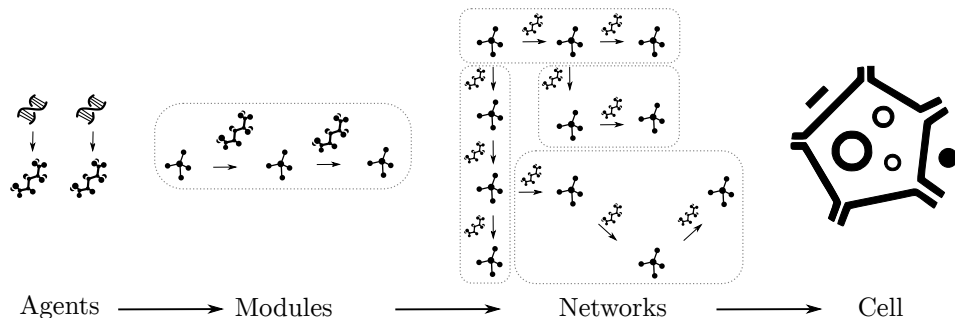


Figure 2.5. Biological hierarchy from genotype to phenotype.

Dual causation: physics and evolution

The concept of dual causation (Mayr 1998) represents the double causality that biological systems exist under. Biological system like any other obey the laws of physics. These can take the form of thermodynamics, diffusion process, mass conservation, chemical kinetics and many others. However, these laws are not enough to explain the particular phenotypes that are present in nature. The second axis of causation originates from the evolutionary process fuelled by natural selection mechanism.

Evolution (Darwin 1859) as we understand it now is an iterative process that elegantly explain the vast amount of diversity (different phenotypes) that occur in nature. First, a genotype defines a particular phenotype. Then natural selection adjust the survival and reproduction chances of an individual containing that phenotype. If mating is successful processes such as mutation and recombination produce a slightly different genotype and the process starts again. Therefore, only a very small subgroup of possible phenotypes (possible in physical terms) will occur due to this evolutionary process.

Systems biologists often think about this iterative selection mechanism as an optimization process. Environmental conditions and physics and chemistry constrains bound the allowable space of possible phenotypes and natural selection finds a temporary optimal solution for the problem. Using typical mathematical terminology genetic variability ensures exploration and natural selection provides the exploitation. Although this idea underlies many different models in systems biology is in constraint-based modelling of metabolic networks where it is represented in the purest way. How to mathematically express this optimization function (the search for the fittest individual) is one of the key aspects to keep in mind when working with these kind of models.

2.1.3 Modelling biology

The biological processes that surround us possess a complexity that can not be explained and figured out only from pure observation and deduction. Mathematical modelling allows us to understand the nature, dynamics and development of biological systems and to make plausible and useful predictions about how those system will behave in the future. The task of modelling is indispensable for research and for the scientific method as a tool to understand the physical reality of our world.

What is a model? Definition and purposes

A model is a simplified representation of reality able to describe and explain an existing phenomenon. The key concept in that definition is "simplified" since all models are simplifications. They only try to represent a particular aspect or area of a real process. Considering this, modelling becomes a subjective procedure. A model needs to be developed with a specific purpose, which sets what magnitudes and variables are important for that purpose and what can be disregarded or at least de-emphasized. Therefore a model explains only some aspects of a process, under certain environmental conditions and only for a particular degree of accuracy.

A model that aims at predicting a system's output can be based and developed around a precise description of input-output variables. However, if finding out the function of an object or the mechanisms of a system is the purpose of the model, then the relations between its parts must be included and described carefully. Some models in biology are quite general such as Michaelis-Menten kinetics (that aims to explain the dynamics of many different enzymes) and some can be very specific (for instance the structure of a protein, the sequence of a gene or the bio-energetics of mitochondria).

Models obtain the data to fill its structure from experiments. Therefore, a model can only be as precise and potent as the experimental data is based on. However, the process of modelling has some advantages when done in parallel or in an iterative way along experiments. First, modelling helps to clarify concepts. Detecting data inconsistencies and knowledge gaps are typical outcomes of good models. Secondly, it is cheaper, easier and faster to try out some test or hypothesis with a good model than to perform experiments. Good models save money and time. Finally, and this is probably the key point, models assist and guide experimentation. Different perturbations, long time courses, virtually endless repetitions can lead research to the most promising possibilities.

Cells as modelling environment

Biological models are as diverse as biology itself. Planet-level ecosystems, complex organisms, particular organs, tissues, cells, sub-cellular systems, pathways, molecular interactions, molecular sequence and structure. The size and scope of the models used in biology can vary dramatically. As it was mentioned before, systems biology deals usually with individual cells. Furthermore, this thesis focuses mainly on cellular metabolism and cellular protein interactome analysis.

There are two main reasons to circumscribe systems biology to cell-level modelling. First, there is still massive amount of information that we do not know yet. Although high-throughput techniques have multiplied the cellular data available to test and study, much needs to be done to improve our knowledge of cell functions and structure. Information about some metabolic or signalling pathways or gene and protein functions is still partial and fragmented and sometimes even contradictory.

The second reason arises from the inherent complexity of biological system. Genomes contain thousands of genes, poly-peptides have thousands of residues, protein-protein interaction networks have tens of thousands of interactions. Assuming the information needed to construct models for those elements is known, the sheer complexity of the system makes the model implementation and interpretation very difficult. There is a general understanding that at this moment cells are complex enough for the experimental data and the modelling tools available.

Classification of biological models

There are different classes of models depending on the purpose of the model and the amount and nature of the experimental data available.

Theory-based or data driven. Theory-based models draw their mathematical structure from first principle knowledge from the system. Data-driven models are based solely on the relation and dependency of the data. This thesis shows examples of both types.

Parametric or non parametric. Parametric models need the presence of parameters fitted by the experimental data. If the model does not need parameter fitting then it is non parametric.

Static or dynamic. Static models represent the system at specific time instants or snapshots and assume steady state. Dynamic models display the evolution of the system variables over time.

Black box or white box. A black box model or approximation represents a system which can be viewed in terms of its inputs and outputs, without any information

of its internal mechanisms. On the contrary, a white box is a system where the inner elements are available for study and testing.

Homogeneous or heterogeneous. Homogeneous models consider cell populations homogeneous in its functions, mechanisms and structures. Heterogeneous models distinguish between different cell sub-population that show different features.

Modelling questions

Building a model is a subjective endeavour. However, the process of modelling, in particular modelling in biology, can be broken down in a series of simple steps that is relevant to point out here. The next questions serve more of a series of issues that is generally useful to address when modelling than an exhaustive roadmap.

Problem and purpose. What problem or process do we want to explain or study? What particular aspect of that problem should we focus on?

Available data. What kind of available information we do have to build the model? Is there some pre-existing structural knowledge about the problem? Are the knowledge gaps? Do we know all the components and interactions that form the system?

Model structure. Which structure should the model have? Dynamic or static? Deterministic or stochastic? How many variables, parameters and equations should have?

Testing. Do the model agrees with the experimental data? Is it able to predict results from new experiments? Do the model make easy predictions to test?

Refinement. Which parts of the models should we adapt to better represent the phenomenon? Does it need more parameters? Different variables? Bigger or smaller scope?

Modelling, like any process in research, highly benefits from preparation and thinking ahead. It is a fine balance between having a solid approach to the problem and the aim of the model and the flexibility of expand it or limit it in scope and shape.

2.1.4 Experimental data integration

As it was commented before, high-throughput technologies gave birth to the systemic approach in the molecular biology field. More often than not the shape and content of the experimental data is not directly comparable. This is common problem when new experimental techniques appear. In this particular scenario the problem becomes even more significant due to the massive amount of information this methods can produce.

Data integration is important for any biological research but it is particularly relevant in systems biology. This discipline is based on the idea some properties of a biological system can only be studied when such system is considered as a whole. Therefore it is of vital importance that the data coming from different parts of this system fit together as easily and seamlessly as possible.

There are a series of problems that difficult the data integration in systems biology (Hwang et al. 2005). First, the types of data can be very different in shape and range from discrete to continuous. Second, each method has a different degree of reliability and comes with particular uncertainties, errors and biases. And third, very useful available information can be found outside the high-throughput techniques in small-scope experiments, computational predictions and high-quality curated databases.

Klipp et al. 2011 proposed a division of data integration in several levels of complexity. The first level refers to common standards for information storage, representation and transfer. The second level focuses on developing shared schemes for biological models and pathways. A good example of this is the Systems Biology Markup Language (SBML) (Hucka et al. 2003). The final level revolves around data correlation and fusion. In other words, to develop analysis and modelling tools that allow researchers to integrate very different information to learn and explain biological processes.

2.2 Biological networks

Biology deals with a wide range of subjects from molecules to vast ecosystems. DNA, RNA, proteins and metabolites combine and interact to form cells. These cells aggregate to form tissues on complex organisms. Different tissues and organs form organisms, that vary from very simple to the most sophisticated and complex systems known. Finally, these organisms (millions and millions of different species) form ecosystems. All these layers of complexity and connection can be (and usually are) represented by networks. These have become so fundamental to represent and understand biological beings that the concept itself has become somewhat indivisible from biology. The network representation and concept has become the core of the systemic approach to biology as we understand it nowadays.

2.2.1 Networks

Why is the network the most successful representation of large biological systems? And why is it the perfect fit for a systemic approach of biology? The answer to those questions is multiple. Networks are abstractions of much more complex systems and they emphasises the interactions between elements more than the nature or characteristics of each isolated element (Klipp et al. 2011). This produces a series of advantages that prove to be useful.

Networks are a good first step into systems for which there is not a lot of information available. They are very understandable when compare to some mathematical descriptions. Network topology is a quick way to compare whether two systems are similar. Furthermore, when compare to random networks, structural features appear highlighting the most informative sections of the network. Moreover, some processes are especially more suited for a network-type description (disease spread, food chains, neural systems) than a more traditional reductionist approach. Finally, when dealing with massive amounts of data (i.e. high-throughput technologies) networks may point to potential functions of particular branches. It is the same idea behind the famous principle that *structure defines function* in molecular biology only that in a network scale: network structure defines network function. Biological networks can be divided according to scale in macroscopic and microscopic (Junker 2008).

Macroscopic networks

The very beginning of network analysis and representation in biology is found in ecology. It is defined as the science that studies the interactions between organisms and their environment. The interactions between different species are known as ecological networks. Probably the most famous example of this type of network is food webs, especially predator-prey. These describe who is affected and in which way in a feeding interaction. This knowledge is very important for understanding the mechanisms that govern populations and entire ecosystems. Other common ecological network is plant-pollinator interaction network.

The second most frequent example of macroscopic network is phylogenetic networks. These represent the evolutionary relationships between organisms. Usually, these relations were illustrated by tress where branching points represented the of two species during evolution. However, recent discoveries such as multi-separation and reticulate relationships picture a much more complex landscape.

Microscopic networks

Microscopic or biochemical networks have been studied for many years. Traditional molecular biology approaches focused mainly on the particularities of each component: activity of one enzyme, targets of a transcription factor, interactions of one protein. The process was a solid but slow build-up from components to higher order structures such as pathways and complete networks. High-throughput technologies massively increased the scope and speed of network-based molecular biology (Covert et al. 2004). Gene regulatory networks, metabolic networks and protein interaction networks are the basic three types of biochemical networks.

Transcription regulatory networks control which genes are expressed in each moment in the cell (Carrera, Elena, and Jaramillo 2012; Carrera et al. 2009; Even, Lindley, and Coccia 2003; Perrenoud and Sauer 2005; Thiele et al. 2009). The expression of one gene is controlled by the product (usually proteins called transcription factors) of another. Gaining insight into how the cell controls the expression of its genes depending on internal or external conditions is the next big step in systems biology. Small motifs such as feed-forward and feed-back loops are very common in these networks. On the other hand, signalling networks are a particular kind of regulatory networks that contain signalling cascades or chains usually associated with phosphorylation events. These networks link intracellular processes to extracellular environments and adjust cellular functions according to those external conditions.

Proteins operate mostly interacting with almost all kinds of molecules in the cell: small metabolites, lipids, nucleic acids or other proteins. Protein interaction networks illustrate how proteins associate with other proteins to carry out their functions (De Las Rivas and Fontanillo 2010; Fossum et al. 2009; Rajagopala et al. 2014). There are several experimental and computational methods to determine these networks. Metabolic networks display metabolites being transformed into each other due to the presence of enzymes. Metabolic networks are by far the most studied biochemical network (Chassagnole et al. 2002; Link, Christodoulou, and Sauer 2014; Vercauteren, Logist, and Impe 2014). Biochemists have been studying metabolic reactions and linking them with other reactions for many decades. A lot of metabolic constraint-based static models and pathway-level dynamic models have been developed in the last decade with great success. These past years a lot of effort has been invested in integrating metabolic and transcriptional networks (Herrgard et al. 2006; Simeonidis, Chandrasekaran, and Price 2013). Since the analysis carried out in this thesis used protein interaction and metabolic networks as objects of study, the next section will describe them in more detail.

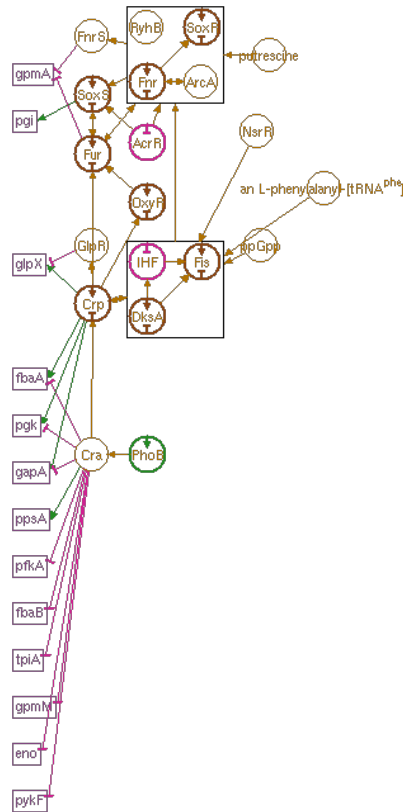


Figure 2.6. Gene regulatory network associated to the glycolysis pathway in *Escherichia coli* K12 MG1655 from the *EcoCyc* database.

2.2.2 Metabolic networks

Cellular metabolism is the set of chemical reactions that take place in the cell and ensure its life. The concept of metabolism is essential to understand life itself. It converts available substances to the cell into energy and molecular blocks to build cellular components. It obeys the laws of chemistry and thermodynamics (like any set of chemical reactions) and it has the ability to shift and modulate its structure to adapt to a wide range of external environments and ensure the cell's survival.

Metabolic networks (Silva et al. 2008) are composed by metabolites and the reactions that relate them. Metabolites are normally small molecules such as monosaccharides, amino acids or inorganic ions. The biochemical reactions are often catalysed by enzymes, a type of protein that works lowering the activation energy of

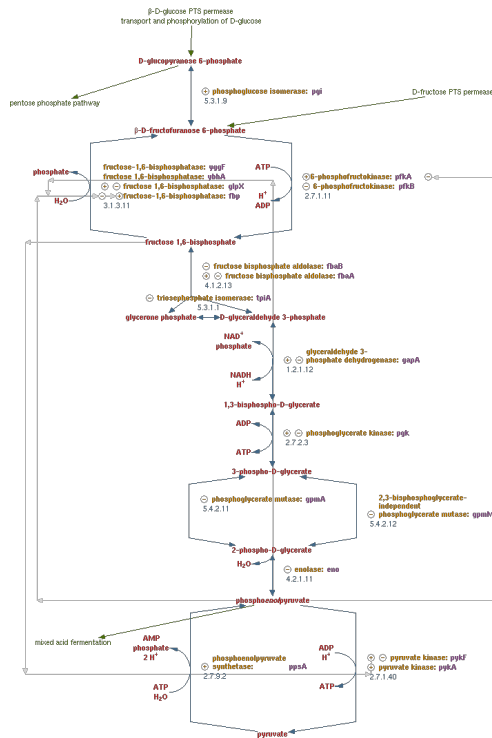


Figure 2.7. Metabolic network representing the glycolysis pathway in *Escherichia coli* K12 MG1655 from the *EcoCyc*.

those reactions making them occur at much faster rates. Metabolites and reactions have been traditionally organized in metabolic pathways. Those are defined as a group or successive metabolic reactions that carry out a specific cellular function as a group. The genome-scale metabolic network of an organism includes all the different pathways that represent sections or branches of the network.

How to divide a network in functional modules is one of the core problems this thesis tackles. Biochemical pathways is the traditional metabolic network division. It progressively originated when biochemists started to determine the functions of particular enzymes. If the product of one reaction was the substrate of another then those reactions were linked. This manually curated method eventually formed a network of metabolites and reactions. Pathways are simply the groups of reactions that made most biochemical sense to group together. Boehringer Mannheim chart (<http://web.expasy.org/pathways/>) is the quintessential division of cellular metabolism. However, recently some efforts have been made to analyse the par-

tioning of the metabolic networks from a more network-level approach (Papin et al. 2003).

Visualizing and displaying metabolic networks is much harder than it seems. In the most popular metabolic databases, such as *EcoCyc* (Keseler et al. 2013) and *KEGG* (Kanehisa and Goto 2000), metabolites are presented by vertices and reactions by edges joining them. Reactions usually carry the names of the enzymes that catalyse them. Sometimes the EC number (Enzyme Commission number) can be present too. EC numbers contain four digits in the form x.x.x.x that represent the type of reaction being catalysed by one particular enzyme. Reaction name, enzyme and EC number are used as synonyms in these representations. Figures 2.7 and 2.8 show those features clearly.

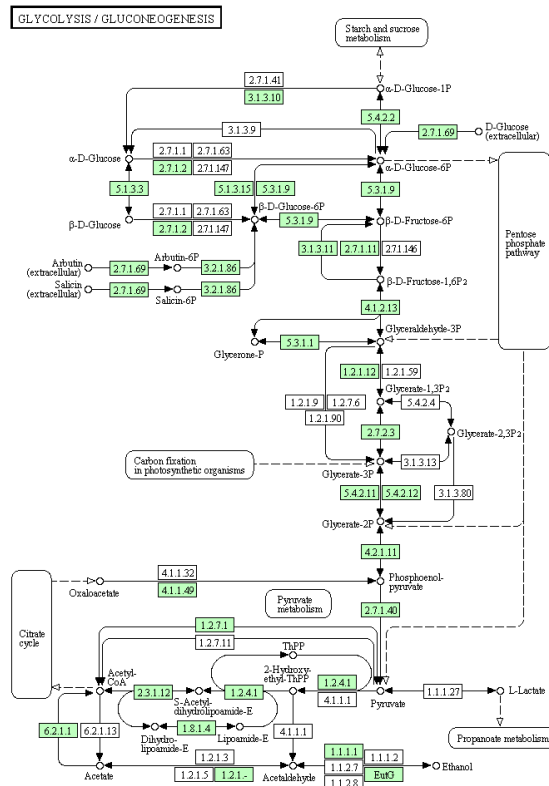


Figure 2.8. Metabolic network representing the glycolysis pathway in *Escherichia coli* K12 MG1655 from the *KEGG* database.

This approach is very intuitive and useful for understanding the basic layout of the system in a quick look. However, it lacks a rigorous mathematical description that can be used to study the network much further. This issue has significant implications for metabolic network analysis and will be addressed in detail in Chapter 5. Some metabolites, usually called pool metabolites (Ma and Zeng 2003b) such as H_2O , CO_2 or adenosin triphosphate (ATP), appear in these representations many times, distorting the real connectivity of the network. These pool metabolites are usually overlooked when studying metabolic networks; sometimes are down-weighted to be less important or even removed completely from the network. There are valid reasons to support these simplifications but each one distance us from a rigorous mathematical description of the network. Other example is the clumping together of the edges representing a concrete reaction. That means that usually a reaction with two substrates and two products is represented by only one link instead of four. Figure 2.9 shows the traditional and the direct unipartite graph representation. Once again, this makes the representation simpler but complicates the network analysis.

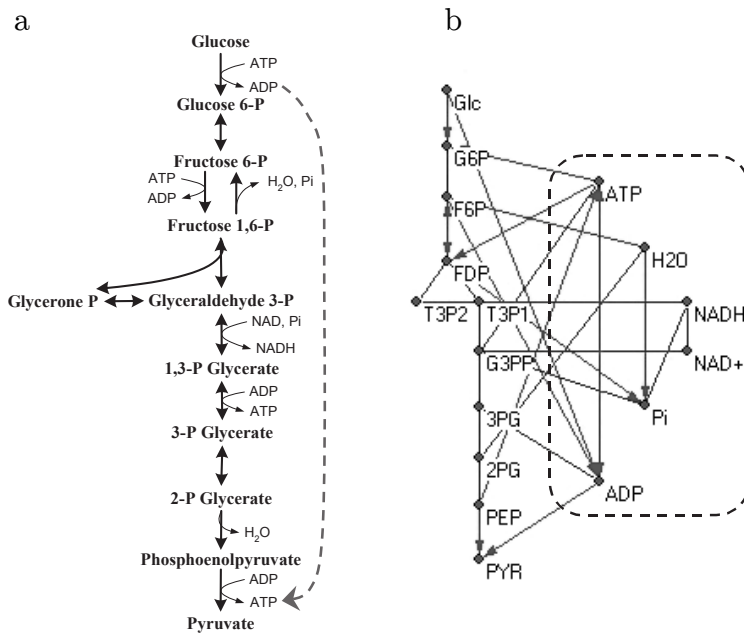


Figure 2.9. Metabolic network representing the general glycolysis pathway (from Ma and Zeng 2003b). (a) Traditional representation of glycolysis. (b) Unipartite graph representation of glycolysis.

2.2.3 Protein-protein interaction networks

Proteins play a major role in the majority of cellular processes. Catalysts, storage, transport, signalling, cellular structure, immune system, cell growth are main examples of their importance but there are many more. All cellular building blocks (DNA, RNA, proteins, lipids, carbohydrates) interact with each other but the protein interactions are particularly important for their functions. Proteins are the main actuators in the cell and the basic way they have to carry out their functions is to interact. The concept of interaction is built in the role of proteins much more than for instance in DNA, which main function is information storage.

A protein-protein interaction (PPI) (De Las Rivas and Fontanillo 2010) obviously occurs when at least two proteins interact. This interaction requires some kind of molecular docking event, has a cellular purpose and occurs in a specific biological background or context. Generic interactions such as the ones that occur when a protein is being made, folded or degraded should be excluded. PPIs can be static, for instance when they form part of a protein complex such as ATP synthase or even a ribosome, or they can be temporary or transient, for example in signalling interactions or activation of gene expression via transcription factors.

Detecting PPIs can be achieved using a variety of experimental methods, all of them high-throughput techniques. The main three used nowadays to detect PPIs are yeast two-hybrid (Y2H) (Brückner et al. 2009; Suter, Kittanakom, and Stagljar 2008), bimolecular fluorescent complementation (BiFC) (Kerppola 2006, 2008) and affinity purification coupled to mass spectrometry (AP-MS) (Berggård, Linse, and James 2007). Figure 2.10 shows the PPIN of *Saccharomyces cerevisiae* detected by Y2H and AP-MS methods (from Yu et al. 2008). In Chapter 3, data from the literature obtained through Y2H and BiFC is used to build the PPI networks (PPIN), hence their molecular mechanisms are explained further next.

In Y2H method, one protein is fused to a DNA-binding domain and the other is fused to an activation domain. If both proteins interact the fused proteins work as a transcription factor that express some reporter gene. GAL4-binding domain is the most commonly used system. BiFC is based on the union of fluorescent protein subunits that are attached to elements of the same macromolecular complex. Proteins being tested are fused to unfolded complementary fragments of a fluorescent reporter protein and expressed *in vivo*. Interaction of these proteins pushes the fluorescent fragments closer, allowing the reporter protein to reform in its original structure and release its fluorescent signal.

These methods produce an immense amount of PPIs data, which is stored in online databases. The most common are STRING (Franceschini et al. 2013), IntAct (Kerrien et al. 2012) and BioGRID (Stark et al. 2006) although there are some more. In general, information about PPIs is spread in more databases than metabolic information which is centred in two or three main sources. Therefore it is

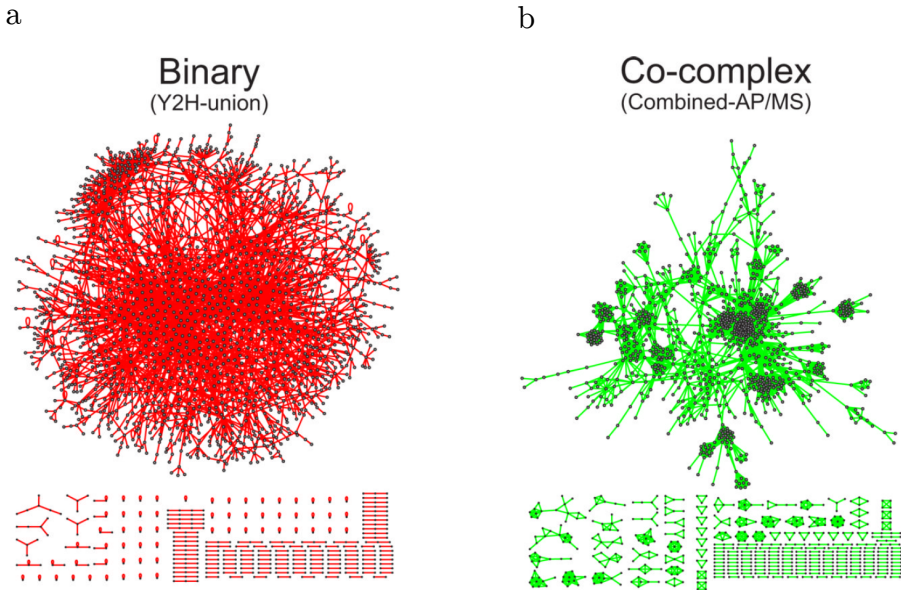


Figure 2.10. PPIN of *Saccharomyces cerevisiae* detected by (a) Y2H and (b) AP-MS methods (from Yu et al. 2008).

very frequent to find major discrepancies and contradictory information regarding PPIs even in well studied organisms such as *Escherichia coli* or *Saccharomyces cerevisiae*. This issue is dealt with in detail in Chapter 3.

PPINs are formed by thousands of proteins interacting with each other. They are involved in practically every aspect of the cell's life. A lot of attention has been given to these networks in the past decade, both for relatively simple organisms (Rajagopala et al. 2014; Uetz et al. 2000; Yu et al. 2008) and more complex (Hegele et al. 2012; Uetz et al. 2006). This system level proteomic analysis highlights very useful information such as discovery of functional modules or motifs, prediction of protein function (the *guilt by association* principle) and comparative analysis among species.

2.3 Modelling tools

The study of biological networks by its very nature requires an interdisciplinary approach. Different analytical methods should be used in different scenarios to obtain different conclusions. In this section, a brief introduction to the methods used to carry out this thesis is provided. The objective here is to get used to the general ideas and capabilities of each family of methods and their technical vocabulary. As it was commented in the Abstract, three main toolboxes were used to analyse biological networks and to extract useful biological information: graph theory, constraint-based modelling and multivariate statistics.

2.3.1 Graph theory

Complex dynamical systems are formed by many interacting components that generate the so called emergent properties, which exceed superposition of individual components. This is very clear in biology where the behaviour of cells (and much more complex systems) depends on the interaction of thousands of molecular elements. The analysis of these complex networks has become one of the main tools in molecular systems biology. Network theory (and its core of graph theory) allows the translation of the biological reality of the cell complexity into a mathematical description that can be modelled and analysed. This section highlights the basic concepts of graph theory and the main areas in network analysis.

Nodes and edges

The word networks refers to an informal representation of a group of elements that are connected to each other. The mathematical entity normally used to analyse networks is the graph. A graph is a mathematical abstraction that has nodes and edges, representing elements and interactions respectively. A graph $G = (N, E)$ is composed of a group of nodes or vertices N and a group of edges E . Each edge is assigned to two nodes. An edge e that connects the nodes a and b is denoted by a, b . Nodes a and b are then adjacent and the edge e is incident to them. The number of edges incident to a node is called degree. A edge connecting a node with itself (a, a) is called loop.

A sequence $(n_1, e_1, n_2, \dots, e_k, n_k)$ of connected and consecutive nodes and edges is called a walk or path, although some authors reserve the term path for a walk without repeating nodes or edges. A path that starts and ends in the same node is usually called cycle. Two nodes are connected if there is a path between them. The length of a path is usually given by its number of edges. The shortest path between two nodes is the path with minimal length. The distance between two nodes is the length of the shortest path between them. For instance, in Figure 2.11a $P_1 = ((1, 2), (2, 3), (3, 4))$ is a path from node 1 to node 4 of distance 3. However,

the shortest path between those two nodes is $P_2 = ((1,3), (3,4))$, which has a distance of 2. Additionally, the path $P_3 = ((3,2), (2,1), (1,3))$ is an example of cycle.

The simplest way to measure density of nodes in a network or in a specific area of it is the clustering coefficient (Watts and Strogatz 1998). This parameter shows how tightly a node is linked to its neighbours. For each node, this coefficient is usually defined as the proportion of links between the nodes within its neighbourhood divided by the number of links that could actually exist between them. It gives an idea of how probable is that if node A is connected to node B and B is connected to node C, then A is also connected to C.

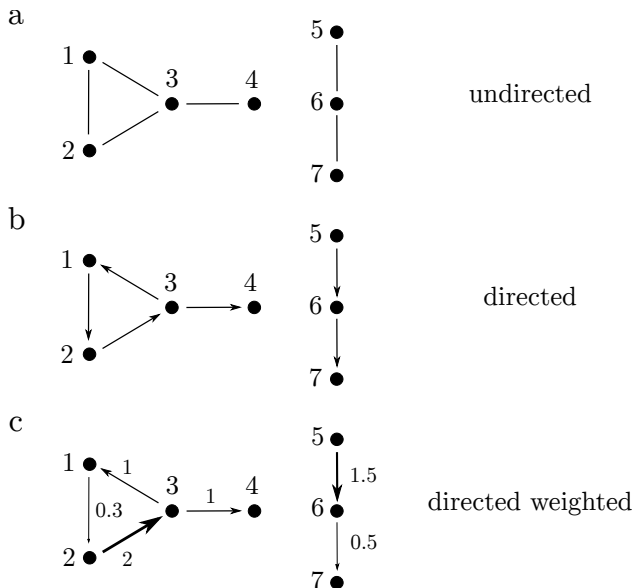


Figure 2.11. Undirected (a), directed (b) and directed weighted (c) variants of a graph.

If every node in the graph is connected then the graph is called also connected. The largest connected component is the largest subgraph of the original graph that is connected. In Figure 2.11a, for instance, the node set $N_1 = (1, 2, 3, 4)$ is the biggest connected component (it is larger than $N_2 = (5, 6, 7)$). Graphs can be undirected, if the nodes are linked by edges without any directional information or directed. In directed graphs, the out-degree of a node is defined as the number of edges going out of it while the in-degree is the number of edges coming in. The degree of a node is the sum of both. In addition, in weighted graphs, each edge carries a value that represents the strength or flow of such interaction. Figure 2.11 shows three graphs with the same layout but with increasing information: Figure 2.11a shows an undirected graph, Figure 2.11b displays a directed graph and finally Figure 2.11c

illustrates a directed weighted graph. It is possible to have undirected weighted graphs but they are less common, since a magnitude such as flow or cost usually carries also a particular direction.

Visual and mathematical representations

Displaying graphs to truthfully represent real connected systems is a tough task. Real phenomena can be very complex and therefore their representation and mathematical description is not straight forward. In molecular biology many interactions occur among more than two elements, for instance the substrates and products of a metabolic reaction or the formation of a protein complex. Hypergraphs (Klamt, Haus, and Theis 2009; Zhou and Nakhleh 2011) are used to represent such interactions. Hypergraphs are defined as normal graphs in which edges link together more than two nodes. As it was commented in Section 2.2.2 this approach displays the system in a simpler and more intuitive way but has a mayor disadvantage. Hypergraphs are not commonly used in graph theory and most of the methods and algorithms developed can not be used on them. The classic way to dodge this issue is using bipartite graphs. Figure 2.12 shows the difference between a hypergraph and a bipartite graph for a metabolic reaction (figures built using the *Escher* visualization tool from King et al. 2015b). Precise and truthful network representations is one of the core issues dealt with in Chapter 5.

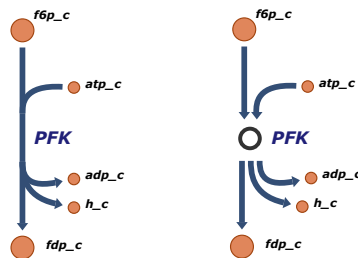


Figure 2.12. Phosphofruktokinase reaction that catalyses the conversion of D-fructose-6-phosphate to D-fructose-1,6-biphosphate. (a) shows the hypergraph and (b) the bipartite graph.

Network properties

Graph theory has developed a series of parameters to better describe and understand a network's structure. The most simple parameters are the distribution or mean across the network of the node parameters already defined: degree distribution, shortest path and clustering coefficient. The degree distribution gives the probability that a node has an exact number of nodes. This representation is one

of the most common ways of classifying networks. The degree distribution of most biological networks approximates a power law (with an exponent usually between two and three). These are commonly called scale-free networks (Barabási and Albert 1999). In scale-free networks a few well connected nodes called hubs keep bind together most of the network generating an low average shortest path. These hubs are a common feature in every biochemical network: very connected metabolites such as ATP or pyruvate, transcription factors influencing the expression of many genes or proteins interacting with many other proteins.

Coupled with a scale-free behaviour, many biochemical networks exhibit a high clustering coefficient (Wagner 2001; Wagner and Fell 2001; Wuchty 2001). The clustering coefficient of a network is the average of the coefficients of every node. Both characteristics together (scale-free behaviour and high clustering coefficient) define the "small world" effect (originally proposed in Watts and Strogatz 1998). Typically, in a small world network the shortest path between two random nodes grows proportionally to the logarithm of the number the nodes in the network. It is important to mention that for stating that some network parameter is high or low, a network benchmark is needed. Traditionally, the Erdős-Rényi model for random networks (Erdős 1959) is used to compare the structure and features of other networks.

Assortativity is another frequent property in network description and analysis. It refers to the preference of the network's nodes to link to other that are similar to them in some way. Node degree is normally used to define that similarity among nodes. In general terms a network is said to be assortative if hubs tend to attach to other hubs more than to normal nodes. The opposite is true for dissortative networks. Most biological networks seem to show a dissortative behaviour, which may be related to how the network was formed and evolved. Finally, centrality measures how "important" a node is in the network structure. Importance here refers to the cohesiveness of the network and its flow structure. Two main centrality measures exist: closeness and betweenness centrality. All these properties are studied in detail for a protein-protein interaction network in Chapter 3.

Functional modules

It is frequent that networks have communities of highly interconnected nodes that are less connected to nodes in other communities. This modular structure has been reported in all kinds of networks, being biochemical networks no exception (Guimera and Nunes Amaral 2005; Hartwell et al. 1999; Holme, Huss, and Jeong 2003; Papin, Reed, and Palsson 2004; Ravasz et al. 2002). It is broadly accepted that the modular structure of complex networks plays a essential role in their functionality, just as protein structure plays a critical role in its function. Consequently there is a need to develop methods that are able to search and identify

those modules. This problem takes many names in the literature from module identification to community detection.

The idea of functional modules in biological networks appears in many shapes and forms across recent bibliography (Girvan and Newman 2002; Hartwell et al. 1999). It is also given a lot of attention in this thesis. The core premise is simple: biological networks are not structure uniformly, they have sections that are densely connected with themselves but sparsely connected with the rest of the network (see Figure 2.13). This core idea has been extensively studied in several biomolecular networks, using different data and following different methodologies. Mering et al. 2003 used genomic data to cluster metabolic enzymes and then compared them with metabolic pathways. Spirin and Mirny 2003 determined the multibody structure of protein-protein interaction networks from yeast proteomic data. In the seminal paper Ideker et al. 2002, authors integrated protein-protein and protein-DNA interaction networks with gene expression data to find regulatory pathways. Guimera and Nunes Amaral 2005 used modularity to find functional modules in the metabolic networks of twelve organisms. In Dittrich et al. 2008, authors used an heuristic approach to calculate maximum scoring network regions in protein-protein interaction networks.

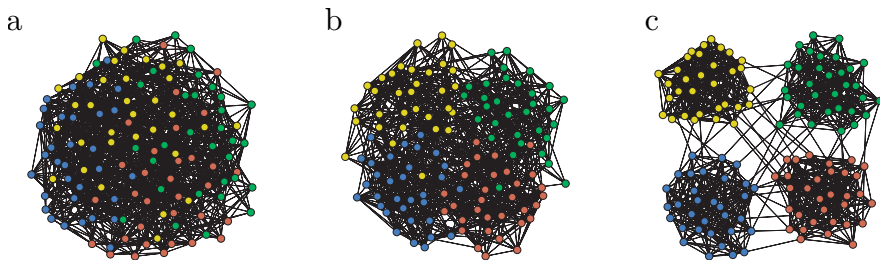


Figure 2.13. Community detection process in a high density network (edited from Guimera and Nunes Amaral 2005). (a) Original representation. (b) Communities start to appear. (c) Identified communities.

Many modularity-based methods (Newman 2006; Newman and Girvan 2004) have been successfully applied to identify functional modules or communities in protein interaction networks by searching for high density connectivity subnetworks. However, it has been recently recently shown that optimizing modularity can over- or under partition the network, failing to find the most natural community structure (Fortunato and Barthélemy 2007). To overcome this issue, a wave of approaches based on the modularity metric was developed including ad hoc parameters that can be tuned to focus on specific-size modules or communities (Lancichinetti, Fortunato, and Kertesz 2009; Reichardt and Bornholdt 2006). On the other hand, community detection methods based on Markov random walks on graphs started to be developed and successfully applied to the problem at

hand (Satuluri, Parthasarathy, and Ucar 2010; Voevodski, Teng, and Xia 2009). Combining these two trends (resolution capability to find which communities are relevant in each time scale and application of Markov random walks of graphs) the Markov Stability methodology was developed (Delvenne, Yaliraki, and Barahona 2010; Delvenne et al. 2013). The stability of a partition (defining partition has a particular network structured in different communities) measures its quality as a community structure based on the clustered autocovariance of a dynamic Markov process taking place on the network. This methodology is explained in more detail in the next subsection and applied to the metabolic network of *E. coli* in Chapter 5.

Another branch of research that deals with community detection, in this case restricted to metabolic network, is known as elementary modes, extreme pathways and minimal generators (Ferreira et al. 2011; Llaneras and Picó 2010). They are not used in this thesis but it is worth it to mention them because they have gathered some attention these last few years. The three groups aim to identify the relevant network-based pathways in a metabolic network. Essentially, there are two properties that these groups of pathways (which is just another name for communities of reactions) can hold: they can generate the flux space or they can comprise all the nondecomposable pathways in the network. Another tool for identifying functional modules is used in this thesis in conjunction with graph theory: multivariate statistics. Since these techniques come from a slightly different mathematical background, they are exposed in another section (Section 2.3.3). The exact mathematical formulation for Multivariate Statistics techniques is shown in Chapter 3.

Community detection based on Markov Stability

The communities were extracted in each network using the Markov Stability (MS) community detection framework (Delvenne, Yaliraki, and Barahona 2010; Delvenne et al. 2013). This framework uses diffusion processes on the network to find groups of nodes (i.e., communities or functional modules) that retain flows for longer than one would expect on a comparable random network; in addition, MS incorporates directed flows seamlessly into the analysis (Beguirisse-Díaz et al. 2014; Lambiotte, Delvenne, and Barahona 2014). The inclusion of direction in the flows is vital for an accurate representation of metabolism.

The diffusion process used is a continuous-time Markov process on the network. From the adjacency matrix \mathbf{A} a rate matrix is constructed for the process: $\mathbf{M} = \mathbf{K}_{\text{out}}^{-1} \mathbf{A}$, where \mathbf{K}_{out} is the diagonal matrix of out-strengths, $k_{\text{out},i} = \sum_j a_{i,j}$. When a node has no outgoing edges then simply $k_{\text{out},i} = 1$. In general, a directed network will not be strongly-connected and thus a Markov process on \mathbf{M} will not have a unique steady state. To ensure the uniqueness of the steady state a *teleportation* component must be added to the dynamics by which a random walker visiting a

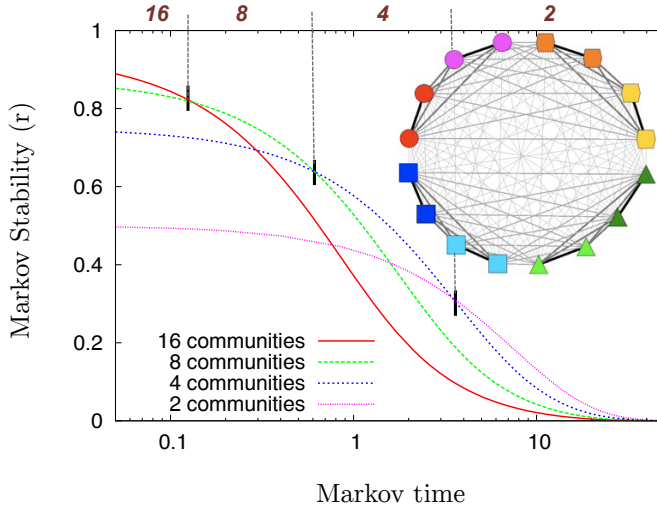


Figure 2.14. Evolution of the multiscale community structure of a simple hierarchical network (from Lambiotte, Delvenne, and Barahona 2014). The number of communities decrease with increasing Markov time. At certain time points the partition with maximum stability jumps from a higher number of communities to a lower number. For instance, until Markov time around 0.12 the stablest partition contains 16 communities (red curve) but at that point the partition with 8 communities (green line) overtakes its predecessor and becomes the stablest partition. Evaluation of the Markov Stability shows that, as time grows, the optimal partition goes from 16 communities to 8 to 4 to 2 over different time intervals. The figure shows a network with $2^4 = 16$ nodes with edges shaded according to their strength. By symmetry, the natural partitions are into 16 single nodes, 8 pairs (colours), 4 tetrads (shapes) and 2 groups of 8 nodes (upper and lower hemispheres).

node can follow an outgoing edge with probability λ or jump (teleport) uniformly to any other node in the network with probability $1 - \lambda$ (Page et al. 1999). The rate matrix of a Markov process with teleportation is:

$$\mathbf{B} = \lambda \mathbf{M} + \frac{1}{N} [(1 - \lambda) \mathbf{I}_N + \lambda \text{diag}(\mathbf{a})] \mathbf{1} \mathbf{1}^T, \quad (2.1)$$

where the $N \times 1$ vector \mathbf{a} is an indicator for dangling nodes: if node i has no outgoing edges then $a_i = 1$, and $a_i = 0$ otherwise. Here $\lambda = 0.85$ is used. The Markov process is described by the ODE:

$$\dot{\mathbf{x}} = -\mathbf{L}^T \mathbf{x}, \quad (2.2)$$

where $\mathbf{L} = \mathbf{I}_N - \mathbf{B}$. The solution of (2.2) is $\mathbf{x}(t) = e^{-t\mathbf{L}^T} \mathbf{x}(0)$ and its stationary state (i.e., $\dot{\mathbf{x}} = 0$) is $\mathbf{x} = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is the leading left eigenvector of \mathbf{B} .

A hard partition of the network into C communities can be encoded into the $N \times C$ matrix \mathbf{H} , where $h_{ic} = 1$ if node i belongs to community c and zero otherwise. The $C \times C$ clustered autocovariance of (2.2) is

$$\mathbf{R}(t, \mathbf{H}) = \mathbf{H}^T \left(\boldsymbol{\Pi} e^{-t\mathbf{L}^T} - \boldsymbol{\pi} \boldsymbol{\pi}^T \right) \mathbf{H}. \quad (2.3)$$

The entry (c, s) of $\mathbf{R}(t, \mathbf{H})$ is how likely it is that a random walker that started the process in community c at finds itself in community s at time t . Crucially, the diagonal elements of $\mathbf{R}(t, \mathbf{H})$ show how good are the communities in \mathbf{H} at retaining flows. The *stability of the partition* is then

$$r(t, \mathbf{H}) = \text{trace} \mathbf{R}(t, \mathbf{H}). \quad (2.4)$$

The communities are found in the network by optimising (2.4) over the space of partitions for a given time t using the Louvain greedy optimisation heuristic (Blondel et al. 2008). The Louvain algorithm does not guarantee a globally optimal partition of the network into communities; for this reason (2.4) is optimized 100 times for each value of t . The consistency of the resulting partition is assessed using the Variation of Information (VI) metric (Meila 2007), as described in Lambiotte, Delvenne, and Barahona 2014; Schaub et al. 2012. A low value of the VI implies that the 100 partitions obtained from Louvain are similar; a VI of exactly zero means that *all* the partitions in each of the 100 optimisations are identical.

The value of the Markov time t , i.e. the duration of the Markov process, is also a resolution parameter for the partition of the network into communities (Delvenne, Yaliraki, and Barahona 2010; Schaub et al. 2012). In the limit $t \rightarrow 0$, Markov stability will assign each node to its own community; as t grows, larger communities are obtained because the random walkers have more time to explore the network. Finally, when $t \rightarrow \infty$ all nodes merge into a single community comprising the entire network (Delvenne et al. 2013). A range of values of t is scanned to explore the multiscale community structure of the network.

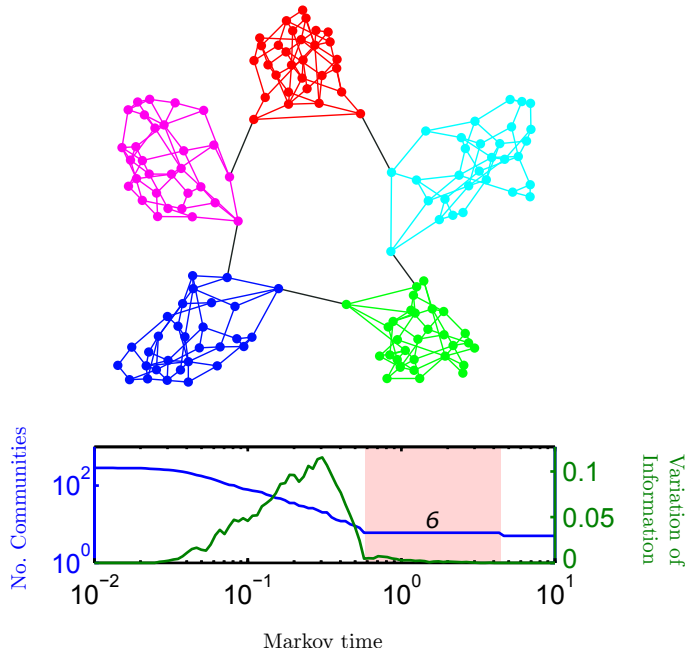


Figure 2.15. Network analysed with edges coloured according to the community structure found in the flow-redistribution matrix using the Markov stability method (from Schaub et al. 2014). The partition into 6 communities is stable over a long span of the Markov times with vanishing variation of information, thus signalling its robustness.

2.3.2 Constraint-based modelling

Constraint-based models (CBMs) are used in this thesis in Chapters 4 and 5. They possess a series of characteristics that allow for a network-based analysis of metabolism. On one hand, they have been very successful in practical applications in the last decade and on the other, they keep a light mathematical structure that facilitates their integration with other mathematical tools such as network theory (Chapter 5) or multivariate statistics (Chapter 4). They are usually opposed to metabolic kinetic models, which are not used in this thesis. However, it is useful to at least describe them in general terms and comment their main drawbacks, which make the constraint-based approach more attractive for certain applications.

Kinetics models

The simplest kinetic models are often called macroscopic models (Dunn 2003), because they do not consider the internal structure of the cells. They are black-box models that transform substrates into products and a certain biomass production. Frequently, biomass production is coupled with extracellular species (both substrates and products) through a series of macroreactions. These large reactions lump together many real metabolic reactions. Monod, for instance, is a typical kinetic expression for these macroreactions. Obviously, these models simplify the cell's complexity but in some applications such as bioprocess engineering (Niu et al. 2013) are quite popular. They are simple and do not require a lot of experimental data to be fitted. However when biomass production is not the only main variable to explain the biological activity of the cell, their predictions tend to fail. A common example of this would be the over-expression or repression of a target gene of interest. This may have a huge impact on the cell but the model is unable to take it into account.

Structured kinetics models (Nielsen and Villadsen 1992) are the natural evolution of macromolecular models. Typically, the cell is divided into several intracellular substances that are connected with each other and with the environment. A series of ordinary differential equations describe the relationships among the compounds, including reaction rates and other kinetic parameters. In general terms these structured kinetics models are more realistic, more precise and more flexible than macromolecular models. However there are a number of serious drawbacks to these models: they require more information, information of particular reaction, kinetic mechanisms and kinetic parameters is often lacking. This final disadvantage (many kinetic parameters are still unknown) is the practical bottleneck in the development of these models. To avoid this issue kinetic models are usually restricted to particular pathways (Costa et al. 2014) or specially studied groups of reactions such as the central carbon metabolism (Chassagnole et al. 2002). In recent years, however, promising attempts of complete cell models have been developed (Karr et al. 2012).

Main features of constraint-based models

Constraint-based models (Figure 2.16) are representations of the cellular metabolism and they all share two fundamental properties:

- The metabolic network of the organism represents the core of the model. The topology (which reactions consume and produce which compounds), the stoichiometry (the molar relationship among the compounds involved in the reactions) and the directionality (which reactions are irreversible under normal biological conditions) are the basic information that the model contains. Regarding the directionality, it is important to state that constraining posi-

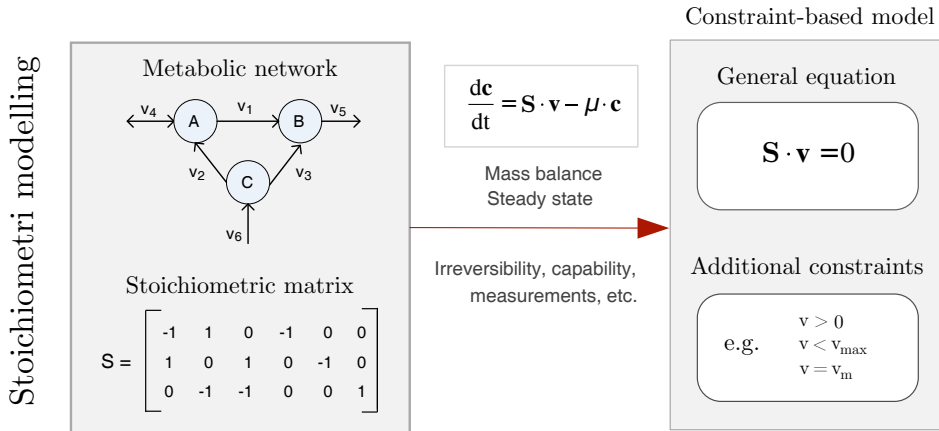


Figure 2.16. Basic principles of the stoichiometric modelling approach. An ODE represents the mass balance of the metabolite pool. Under the assumption of steady state it simplifies to the general equation. Typical of constraint-based models are additional constraints that reduce the possible flux space.

tive values for a given flux does not necessarily mean that the corresponding reaction is irreversible. It can still be considered reversible, with direct and reverse reactions simultaneously occurring, but with the assumption that the net flux is positive.

- They ignore the intracellular dynamics, assuming a steady state for the internal metabolites (Stephanopoulos, Aristidou, and Nielsen 1998).

The stoichiometric matrix

The metabolic network in an organism can be represented in the form a stoichiometric matrix (Figure 2.17). It lists the metabolites and the reactions occurring among them. It is the main feature of the constraint-based models. Reactions include intracellular (both substrates and produces are metabolites within the cell) and exchange (some metabolite involved is out of the cell). Exchange reactions, therefore, represent the uptake of necessary nutrients and the production of by-products. All this information is contained in the stoichiometric matrix, which has m metabolites and n reactions and takes the form of a $m \times n$ matrix S , in which rows match metabolites and columns reactions.

The stoichiometric matrix S is formed by the stoichiometric coefficients or the reactions that form a metabolic network. Each column contains the elements that participate in the corresponding reaction (reaction participation) and their

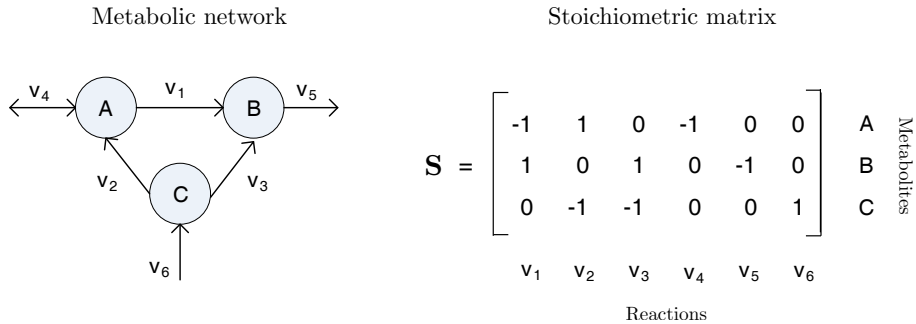


Figure 2.17. A simple example metabolic network. Nodes represent metabolites, links correspond to fluxes of metabolic reactions \mathbf{v} and arrowheads the reaction directionality. Reaction v_4 is reversible. Fluxes v_4 , v_5 and v_6 exchange mass with the environment.

stoichiometric coefficients. Besides, they must obey the rules of chemistry, such as mass and charge balance. Every row describes all the reaction in which the metabolites participate (usually called metabolite connectivity), and therefore how the reactions are linked (which reaction produce metabolites that are substrates for other reactions). For large networks, the stoichiometric matrix has zero in most of its elements. Being a sparse matrix has implications in the computational procedures, specially when reaching genome-scale levels. For a detailed mathematical description of \mathbf{S} see Chapter 9 in Palsson 2015.

Main principles in constraint-based models

Once the stoichiometric matrix has been defined, the mass balances involving intracellular metabolites can be mathematically described by a set of ordinary differential equations (Llaneras and Picó 2008):

$$\frac{dc}{dt} = \mathbf{S} \cdot \mathbf{v} - \mu \cdot \mathbf{c} \quad (2.5)$$

where $\mathbf{c} = (c_1, c_2, \dots, c_m)$ is the vector of intracellular metabolite concentrations, $\mathbf{v} = (v_1, v_2, \dots, v_n)$ the vector of fluxes and μ the specific growth rate of the cells. This equation represents the dynamic mass balance, and therefore describes how the concentration of each metabolite c_i changes over time. To solve this equation information about stoichiometry (\mathbf{S}), reaction fluxes (\mathbf{v}) and cellular growth (μ) is necessary.

As it was previously stated, in stoichiometric models the dynamics of intracellular metabolites are ignored under the assumption that there is an intracellular global

steady state. This assumption is corroborated by the observation that intracellular dynamics are much faster than extracellular so it is reasonable to assume that intracellular metabolites reach a steady state very quickly. Furthermore, the dilution term $\mu \cdot \mathbf{c}$ is also disregarded it is commonly much smaller than the intracellular fluxes affecting the corresponding metabolites. Under those two assumptions, the total mass balance of the cell can be mathematically represented by Equation 2.6, routinely called general equation.

$$\mathbf{S} \cdot \mathbf{v} = 0 \quad (2.6)$$

Equation 2.6 defines the space of possible flux distributions \mathbf{v} . Although it reduces significantly the number of metabolic states that feasible, it does not predict the one that is actually occurring. Obviously, different carbon sources or oxygen availability produce metabolic states wildly different from each other. This matrix can be translated into a system of equations with m independent equations. Being n normally larger than m the system is undetermined with $n-m$ degrees of freedom. Additional information is needed to reduce even more the space of feasible solutions to a biologically useful set.

Constraint-based approach

The basic idea behind constraint-based models is that cells are subject to a series of constraints that limit their possible behaviour. Reaching a complete knowledge of all the constraints that control the cell's features is a long term goal. However, nowadays it is possible to enumerate a set of physical, chemical and biological constraints that reduce a great deal the possible flux space of a metabolic network. These constraints, added to classical stoichiometric constraints of Equation 2.6 start to define a very limited space of feasible metabolic flux distributions. From that point of view, stoichiometric modelling based on the general equation may be viewed as a particular branch of constraint-based modelling that only considers stoichiometric constraints. Since the flux distributions define the metabolic phenotypes of the cell, this space contains all the feasible phenotypes (Edwards, Covert, and Palsson 2002).

Typical constraints range from thermodynamics (e.g. irreversibility of fluxes) to enzyme capacities (which limits a maximum flux). It is very common to have at least some reaction fluxes measured, which constrains the flux space even more. These constraints are mathematically represented in Figure 2.16 under *Additional constraints*. Additional regulatory constraints (Covert, Schilling, and Palsson 2001; Lerman et al. 2012; O'Brien et al. 2013) may be included as well but are disregarded in this thesis.

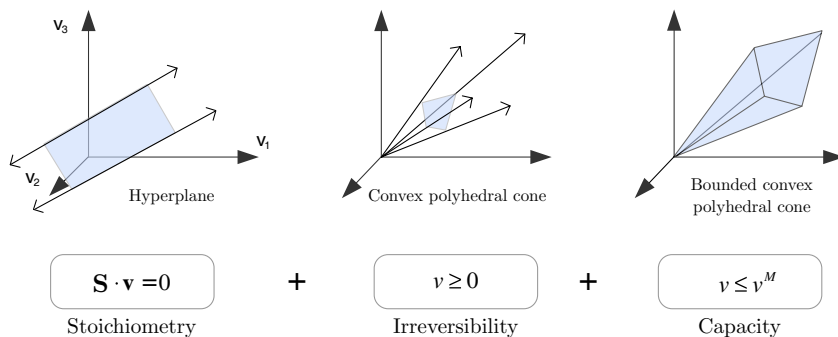


Figure 2.18. Space of possible steady-state flux distributions. Each figure is bounded by consecutively more restricting constraints.

The general equation, Equation 2.6, procures a set of purely stoichiometric constraints that relate the fluxes with each other and reduce the space of possible flux distributions to a hyperplane (see Figure 2.18), subspace of \mathbb{R}^n where each axis represents a particular reaction flux in the network. After that, irreversibility constraints are usually added. These are codified as reaction which are assumed to flow in only one direction, thus making the flux through those reactions always positive (as in Equation 2.7).

$$v_i \geq 0 \quad (2.7)$$

Ultimately, maximum flux values derived from enzyme or transporter capacities can be defined as well (Equation 2.8).

$$\mathbf{v} < \mathbf{v}_{\max} \quad (2.8)$$

If this data is available for the flux of every reaction in the network, then the flux space becomes bounded. In mathematical terms, the space of solutions is shaped as a bounded convex polyhedral cone (see the last panel in Figure 2.18). Equations 2.6, 2.7 and 2.8 embody the most common constraints. They are the only constraints used in this thesis. These three types of constraints define a space of solutions that the complete flux distribution of the metabolic network of a cell always inhabits. Therefore, together they form what is popularly called a *constraint-based model*, which will be one of the basic object of study in this work. The three methodologies that were used to study these models are known as FBA, FVA and MFA. These are explained in detail in Chapter 4.

The biomass reaction

Arguably the most important reaction in many CBMs is known as the *biomass reaction* (Feist and Palsson 2010). This reaction represents the cellular growth and works draining precursor metabolites from the network at stoichiometrically fixed relative rates while producing some by-product metabolites. These precursors are used to produce lipids, proteins, nucleic acids and other macromolecules necessary for the cell's growth. Besides generating macromolecules, significant energetic requirements exist for ensuring cellular replication and growth. These requirements are usually divided in growth associated and non-growth associated. To represent the former, ATP is converted to ADP as a part of the biomass reaction. Non-growth associated maintenance is not part of the biomass reaction in most CBMs. Instead, it is represented as the lower bound of another drain reaction, the ATP maintenance reaction (ATPM), that simulates the consumption of energy the cell undergoes only for staying alive. Biomass and ATPM reactions are codified in the model as additional columns in the stoichiometric matrix \mathbf{S} . Information about the biomass reaction for the specific *E. coli* model used in this thesis is given in Chapter 4.

Brief history, uses and applications of CBMs

CBMs have their origin as metabolic network reconstructions. Early reconstructions were small and they only included the most basic reactions of the central carbon of the organisms they were representing. Probably their actual shape as CBMs has its origins in a series of papers dated more than twenty years ago, specially in Varma and Palsson 1993a. Nowadays there are genome-scale CBMs for hundreds of different organisms. Most of them are organized in the BIGG database (King et al. 2015a). Popular examples are the RECON 1 reconstruction of the *Homo Sapiens* metabolism (Duarte et al. 2007) and the latest iteration of the *E. coli* (str. K-12 substr. MG1655) genome-scale CBM, called iJ1366 and published in Orth et al. 2011. As it was stated before, these models contain information about the reaction stoichiometry, reversibility and capabilities but they also include the relationships between genes, proteins and reactions.

Constructing a genome-scale CBM is a long, time-consuming process. This process is usually divided in four consecutive steps (Orth, Fleming, and Palsson 2010):

- First, the organism's annotated genome is used to build a draft reconstruction.
- Second, this draft reconstruction is curated through a long process that comprises the analysis of many specific experimental data.

- Third, the reconstruction is translated into a proper mathematical model, usually called constraint-based model (CBM). At this point model simulations can already be compared with real phenotypic data.
- Finally, in a fourth step high-throughput data such as fluxomics, metabolomics, proteomics or transcriptomics can be used to refine the model even further.

There have been many practical uses to constraint-based models. Some of these studies encompass:

- Bacterial evolution
- Gene deletion and horizontal gene transfer
- Adaptation to new environments
- Evolution to minimal genomes
- Identification of optimal network states
- Determination of groups of coupled reactions
- States of regulatory networks
- Linking phenotypes and genotypes, mainly through the prediction of cellular growth
- Discovery of unknown biological features
- Applications in metabolic and bioprocess engineering

A comprehensive enumeration of the most popular applications of CBMs in *E. coli* can be found in Feist and Palsson 2008 (see Figure 2.19). In general terms, CBMs have reach such a spectacular success because of their ability to make good predictions with small amounts of experimental data. The turning point in the scientific community occurred in the first years of the present century when CBMs started to be regarded as a useful tool and not an overly simplified model of metabolism. Around that time the seminal paper Edwards, Ibarra, and Palsson 2001 was published and served as validation of the whole approach.

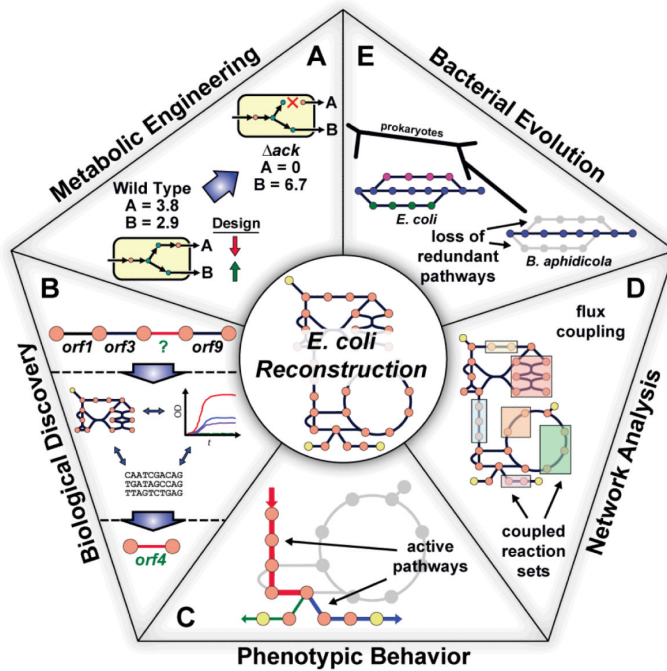


Figure 2.19. Main applications of constraint-based modelling (from Feist and Palsson 2008).

2.3.3 Multivariate statistics

Multivariate statistics is a subdivision of statistics that includes the simultaneous observation and analysis of more than one outcome variable. This branch of techniques are used to carry out trade studies across multiple dimensions while taking into account the effects of all variables on the system's response. Some types of very common data analysis such as linear and multiple regression are not multivariate analysis in the sense that they consider only the univariate conditional distribution of a single outcome variable given the other variables. Some techniques in multivariate analysis include multivariate analysis of variance (MANOVA), multivariate regression and factor analysis among others.

The methods used in this thesis are maybe the most successful and widely used examples of multivariate analysis: the Principal Component Analysis (PCA) and the Partial Least-Squares regression (PLS). These have been extensively used in systems biology before in a wide range of applications: from data fusion (Folch-Fortuny et al. 2016), pathway determination (Ferreira et al. 2011; Folch-Fortuny et al. 2015), network decomposition (Barrett, Herrgard, and Palsson 2009) and

metabolic flux analysis (González-Martínez et al. 2014). PCA will be used in Chapter 4 for constraint-based metabolic model characterization and PLS will also be used in Chapter 4 for analysing the robustness of the conclusions and in Chapter 3 for finding functional modules in a viral PPIN.

Principal Component Analysis

PCA is probably the most widespread multivariate statistical method being used in virtually all scientific fields. Its modern formulation comes from Hotelling 1933. An excellent recent introductory review of the method can be found in Abdi and Williams 2010. PCA analyses a data table representing observations described by some correlation variables that often are inter-correlated. Its main objective is to extract the most relevant information from the data and to display it as a set of new orthogonal variables known as *principal components*. The main goals of PCA can be summarized in:

- Gathering the most relevant information from the data.
- Condensing the size of the data set.
- Break down the description of the data set.
- Dissecting the structure of the observations and variables.

The data table to be studied by PCA contains I observations or individuals described by J variables and it is represented by the $I \times J$ matrix \mathbf{X} . This matrix \mathbf{X} has rank L where $L \leq \min\{I, J\}$. There is a standard pre-processing of the data known as centering and re-scaling. Centering means that the mean of each column is equal to 0 so each column is centered around this value. Re-scaling refers to setting the variance of each variable to 1 in order to make possible the comparison (see Figure 2.20).

The first principal component is required to explain the largest variability possible. The second component determined must be orthogonal to the first component and it has to explain the largest possible variability. Following components (if necessary) are calculated this way. The values of these new variables are known as factor scores and represent the projection of the observations onto the principal components. Finding the components comes from the SVD of the data table \mathbf{X} :

$$\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T \tag{2.9}$$

where \mathbf{P} is the $I \times L$ matrix of left singular vectors, \mathbf{Q}^T is the $L \times J$ matrix of right singular vectors and $\mathbf{\Delta}$ is the diagonal matrix of singular values. In PCA the $I \times L$ matrix of factor scores, \mathbf{F} is defined as

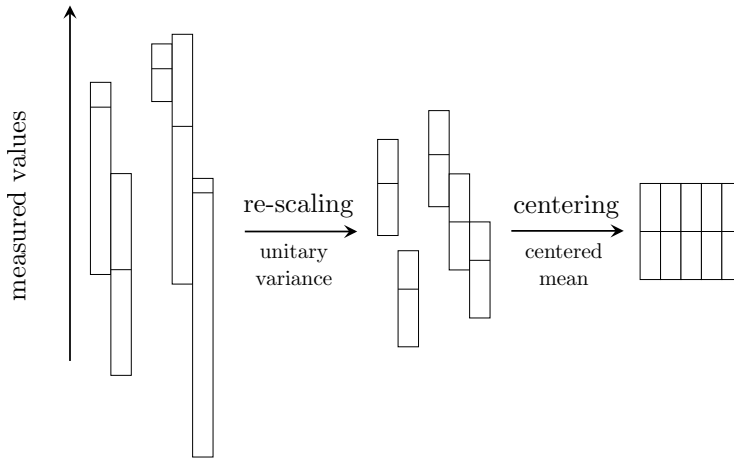


Figure 2.20. Two-step pre-processing in PCA and PLS methods.

$$\mathbf{F} = \mathbf{P}\mathbf{\Delta} \quad (2.10)$$

The matrix \mathbf{Q} , known as the loading matrix, returns the coefficients of the linear combinations used to compute the factor scores. The principal components can be represented geometrically by the rotation of the original axes. The factor scores give the length of the projections of the observations on the new components. The loadings are then interpreted as direction cosines of the new components from the original variables. From this point of view the matrix \mathbf{X} can be interpreted as a bilinear decomposition that produces the factor score matrix \mathbf{F} by the loading matrix \mathbf{Q} :

$$\mathbf{X} = \mathbf{F}\mathbf{Q}^T \quad (2.11)$$

Figure 2.21 shows a simple example of the PCA methodology. The data matrix \mathbf{X} has three variables: \mathbf{x}_1 , \mathbf{x}_2 and \mathbf{x}_3 . PCA found two principal components: \mathbf{t}_1 and \mathbf{t}_2 . The first component \mathbf{t}_1 is the straight line that better approximate the data (solving a least square problem to minimize the distance between the data and the new line). The direction of \mathbf{t}_1 is determined by the loading vector \mathbf{p}_1 . The new coordinate for observation i is t_{i1} . The second principal component \mathbf{t}_2 is another straight, orthogonal to \mathbf{t}_1 that better approximate the data. The direction of \mathbf{t}_2 is determined by the loading vector \mathbf{p}_2 . Both principal components define a hyperplane in the space defined by The direction of \mathbf{t}_1 is determined by the loading vector \mathbf{X} . The projections of the observations onto this hyperplane is the *scores* vectors for the first \mathbf{t}_1 and the second \mathbf{t}_2 component.

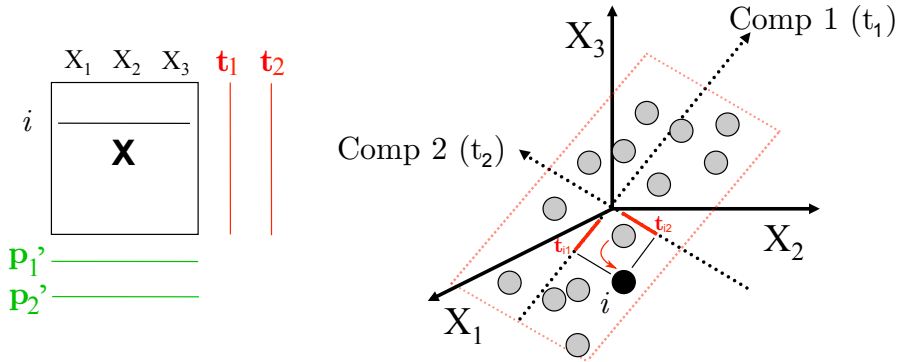


Figure 2.21. Simple example of the PCA methodology.

The maximum number of components is equal to L , the rank of \mathbf{X} . This corresponds to the number of non-zero singular values present in $\mathbf{\Delta}$. These singular values are ranked from higher to lower in the diagonal matrix $\mathbf{\Delta}$. The actual value of each singular value is precisely the variance that is explained by each component. Therefore, only the first components with high singular values are actually useful. In most cases a few components explain most of the variance of the data and accordingly there is a significant compression and clarification of the data. It is possible to approximately reconstruct the original data matrix \mathbf{X} from the scores \mathbf{t}_a and the loadings \mathbf{p}_a , being a the number of selected components ($a \leq L$) (see Equation 2.12).

$$\mathbf{X} = t_1 p_1^T + \dots + t_a p_a^T + E \quad (2.12)$$

where \mathbf{E} is the residual matrix that contains the part of each observation not explained by the principal components. When $a = L$, Equations 2.11 and 2.12 are equivalent. On the other hand, there are a few graphic representations of the PCA analysis available to explore the results:

- T^2 -Hotelling and SPE
- Scores t/t
- Loadings p/p
- Scores t / observations (or experiments)
- Loadings p / variable \mathbf{X}

They will be explained in detail when used with the experimental data in Chapter 4.

Partial Least-Squares regression

Also known as *projection to latent structures* combines ideas from PCA and multiple linear regression (Abdi 2010). Its main goal is to analyse a set of dependent variables from a set of independent variables. This analysis or prediction is carried out by extracting from the original variables a set of orthogonal latent variables which represent the best predictive power. Following the notation defined previously with PCA, I observations described by K dependent variables are stored in an $I \times K$ matrix called \mathbf{Y} . The values of the independent variables J on these I observations are collected in the $I \times J$ matrix \mathbf{X} .

The goal of PLS is to describe or predict \mathbf{Y} from \mathbf{X} and analyse their common structure. By contrast, PCA decomposes \mathbf{X} in order to obtain components that explain \mathbf{X} better. PLS, however, finds components from \mathbf{X} that best predict \mathbf{Y} . The method searches for a set of latent variables and carries out a simultaneous decomposition of \mathbf{X} and \mathbf{Y} with the constraint that these components must explain as much covariance as possible between \mathbf{X} and \mathbf{Y} .

PLS regression decomposes both \mathbf{X} and \mathbf{Y} as a product of a common set of orthogonal factors and a set of specific loadings. The independent variables are broken up in:

$$\mathbf{X} = \mathbf{TP}^T \quad (2.13)$$

By analogy with PCA, \mathbf{T} is the score matrix and \mathbf{P} is the loading matrix. \mathbf{Y} is estimated as

$$\hat{\mathbf{Y}} = \mathbf{TBC}^T \quad (2.14)$$

where \mathbf{B} is a diagonal matrix with the 'regression weights' as diagonal elements and \mathbf{C} is the 'weight matrix' of the dependent variables. The columns of \mathbf{T} are the latent vectors (just as the columns in \mathbf{F} were the scores in PCA). When their number is equal to the rank of \mathbf{X} , they represent an exact decomposition of \mathbf{X} . Additional constraints are needed to define \mathbf{T} . In the case of PLS this means finding two sets of weights \mathbf{w} and \mathbf{c} in order to create a linear combination of the columns of \mathbf{X} and \mathbf{Y} such that these two linear combinations have maximum covariance. First the pair of weight vectors are obtained:

$$\mathbf{t} = \mathbf{Xw} \quad \text{and} \quad \mathbf{u} = \mathbf{Yc} \quad (2.15)$$

meeting always the constraints that $\mathbf{w}^\top \mathbf{w} = 1$, $\mathbf{t}^\top \mathbf{t} = 1$ and $\mathbf{t}^\top \mathbf{u}$ is maximum. When the first latent vector is found, it is subtracted from both \mathbf{X} and \mathbf{Y} and the process starts all over again until \mathbf{X} becomes a null matrix.

In general terms the quality of the prediction does not always increase with the number of latent variables used in the approximation. Usually, the quality increases first and then decreases. If and when the quality of the prediction decreases when the number of latent variables increases, an event of data overfitting is occurring. Therefore it is very relevant to construct a model with the optimal amount of latent variables. The most common approach is the ratio Q_l^2 which relates the residual sum of squares and the predicted residual sum of squares. Concrete examples of this criteria and the graphic representations of PLS (which are analogous to those in PCA) are shown in Chapters 3 and 4.

Chapter 3

Potyvirus network analysis

- You move to an area and you multiply and multiply until every natural resource is consumed and the only way you can survive is to spread to another area. There is another organism on this planet that follows the same pattern. Do you know what it is? A virus.

Agent Smith. "The Matrix", 1999.

Part of the contents of this chapter appeared in the following journal articles:

- Bosque, G. et al. (2014). "Topology analysis and visualization of Potyvirus protein-protein interaction network". In *BMC Systems Biology* 8.1, p. 129.
- Folch-Fortuny, A. et al. (2016). "Fusion of genomic, proteomic and phenotypic data: the case of potyviruses". In: *Molecular BioSystems* 12.1, pp. 253-261.

3.1 Introduction

This chapter revolves around the study and analysis of the protein-protein interaction network (PPIN) of potyviruses and some of their main plant hosts. The combination of organisms studied, type of network analysed and mathematical tools used is shown in Figure 3.1.

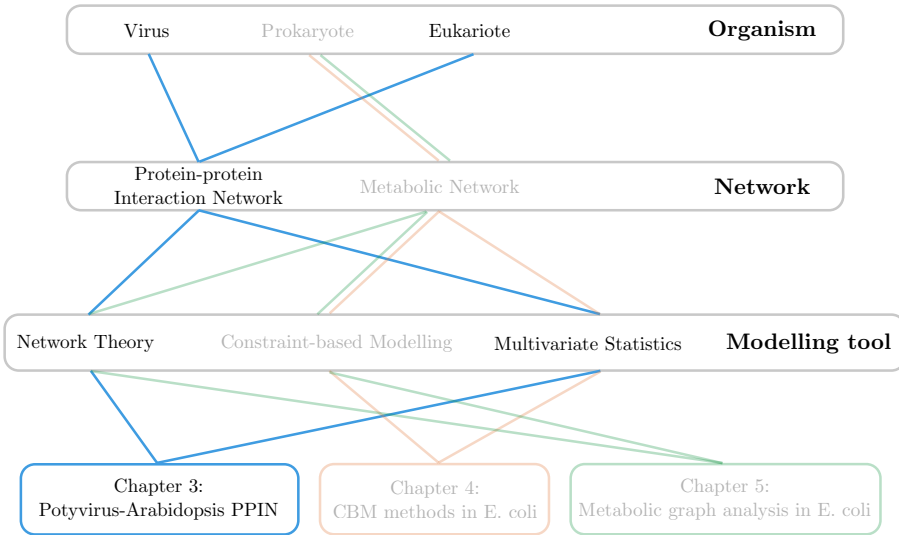


Figure 3.1. Chapter roadmap.

In the first section (Section 3.2) the integrated PPIN of virus and network is thoroughly studied. From data collected from bibliography and previous studies the binary system virus-host is analysed using tools from graph theory and network topology. In the second section (Section 3.3), the network structure of the virus is integrated with phenotypic measurements of its fitness through a set of mutants. In this section, multivariate statistics is used to unveil a set of functional modules that relate virus proteins, mutations and fitness. In the final section, Section 3.4, the main conclusion from the entire chapter are enunciated.

3.2 Potyvirus-*Arabidopsis thaliana* protein-protein interaction network

3.2.1 Summary

One of the central interests of Virology is the identification of host factors that contribute to virus infection. Despite tremendous efforts, the list of factors identified remains limited. With omics techniques, the focus has changed from identifying and thoroughly characterizing individual host factors to the simultaneous analysis of thousands of interactions, framing them on the context of protein-protein interaction networks and of transcriptional regulatory networks. This new perspective is allowing the identification of direct and indirect viral targets. Such information is available for several members of the Potyviridae family, one of the largest and more important families of plant viruses.

Information on virus protein-protein interactions from different potyviruses was collected, processed and used for inferring a protein-protein interaction network. All proteins are connected into a single network component. Some proteins show a high degree and are highly connected while others are much less connected, with the network showing a significant degree of dissortativeness. It was attempted to integrate this virus protein-protein interaction network into the largest protein-protein interaction network of *Arabidopsis thaliana*, a susceptible laboratory host. To make the interpretation of data and results easier, a new approach was developed for visualizing and analysing the dynamic spread across the host network of the local perturbations induced by viral proteins. It was found that local perturbations can reach the entire host protein-protein interaction network, although the efficiency of this spread depends on the particular viral protein. By comparing the spread dynamics among viral proteins, it was found that some proteins spread their effects fast and efficiently by attacking hubs in the host network while other proteins exert more local effects.

These findings confirm that potyvirus protein-protein interaction networks are highly connected, with some proteins playing the role of hubs. Several topological parameters depend linearly on the protein degree. Some viral proteins focus their effect in only host hubs while others diversify its effect among several proteins at the first step. Future new data will help to refine this model and to improve its predictions.

3.2.2 Background

Potyvirus is the mayor genus in the Potyviridae family, accounting for 30% of all known plant viruses, with more than 180 members. Many potyviruses are important pathogens of agricultural crops. They are able to infect a wide range of mono- and dicotyledonous plant species (Gibbs and Ohshima 2010), causing symptoms that severely reduce the yield and quality of crops. The economic impact of these viruses on agriculture is well-documented (Spence et al. 2007). Figure 3.2 shows the effect of several potyvirus on different hosts. Some examples of potyviruses are *Plum pox virus* (PPV), *Soybean mosaic virus* (SMV), *Turnip mosaic virus* (TuMV), and *Tobacco etch virus* (TEV) (Ward and Shukla 1991).

Potyvirus virions are flexuous and rod-shaped, 680 to 900 nm long and 11 to 15 nm wide (Riechmann, Laín, and García 1992). Potyviruses have a single-stranded, positive-sense RNA genome of approximately 10 kilobases (kb) (see Figure 3.3). They contain two open reading frameworks (ORF). The first one is a long ORF which is translated into a large polyprotein, which subsequently self-processes into 10 mature functional proteins: P1, a serine protease also involved in enhancement of polyprotein translation; HC-Pro, a protease with RNA silencing suppressor activity that also mediates aphid transmission; P3, which play a role in cell-to-cell movement; 6K1, a small peptide that links the replication complexes to ER membranes; CI, an RNA helicase with ATPase activity; 6K2, another small peptide of unknown function; VPg, linked to the 5' end of the genome; NIaPro, the mayor protease; NIB, the RNA-dependent RNA polymerase; and CP, the capsid protein (Elena and Rodrigo 2012). The second ORF is a small one embedded within the P3 coding region and results from + 2 frame-shift (Chung et al. 2008; Wei et al. 2010). This recently discovered ORF encodes the eleventh protein, P3N-PIPO, also involved in cell-to-cell movement. Much research in the last two decades has focused on understanding the functions of the different potyvirus proteins during the virus life cycle. Rapid rise of academic interest in this topic followed the complete sequencing of the first two potyviruses: TEV (Allison, Johnston, and Dougherty 1986) and *Tobacco vein mottling virus* (TVMV) (Domier et al. 1986). Many excellent reviews have been published since then (Revers et al. 1999; Riechmann, Laín, and García 1992); some addressing particular issues such as protein function (Urcuqui-Inchima, Haenni, and Bernardi 2001), polyprotein processing (Adams, Antoniw, and Beaudoin 2005; Merits et al. 2002), cellular localization (Zheng et al. 2011) and genome structure (Gibbs and Ohshima 2010).

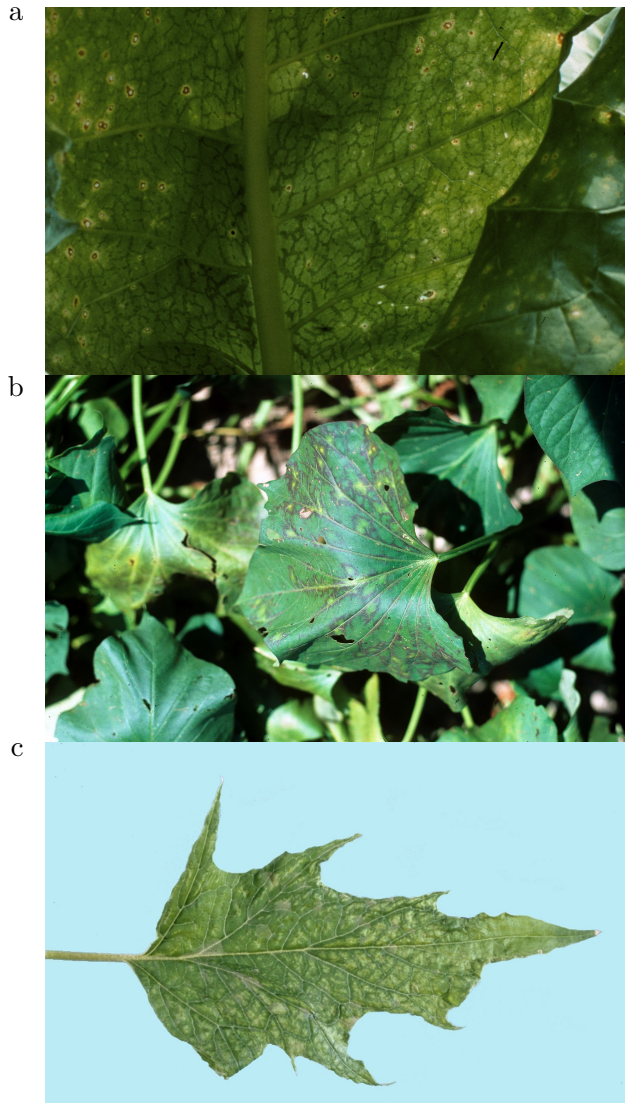


Figure 3.2. Potyvirus symptoms on plant's leaves. Vein clearing appears. Leaves develop a faint mottling and the characteristics vein banding: dark green bands along the veins with light green tissue around them. Veins may become necrotic on infected plants. (a) *Potato Virus Y* on *Nicotiana tabacum* (R.J. Reynolds Tobacco Company 1990). (b) *Sweet Potato Feathery Mottle Virus* on *Ipomoea batatas* (Charles Averre, North Carolina State University 2009). (c) *Tobacco Etch Virus* on *Nicotiana tabacum* (Florida Division of Plant Industry 2007).

During the last decade there has been an increasing number of studies of protein-protein interactions (PPIs) and the effect that these interactions cause on a wide range of biological processes (Culver and Padmanabhan 2007). PPIs are defined as physical contacts that take place in cells through molecular docking (De Las Rivas and Fontanillo 2010). Proteins work typically linked to other molecules including lipids, nucleic acids or other proteins (Börnke 2008). Biological activity usually arises from the association of several proteins, which form protein complexes. In viruses, interactions between proteins play vital roles in many processes during infection such as virus trafficking between the nucleus and the cytoplasm, formation of replication complexes, assembly of virions, or virus transmission to other cells. Traditionally, PPIs have been studied using methods such as coimmunoprecipitation or chromatography (Phizicky and Fields 1995). However, over the past decade two experimental strategies have been used to detect these interactions: yeast two-hybrid (Y2H) (Börnke 2008; Brückner et al. 2009; Fields and Song 1989) and affinity purification coupled with mass spectrometry (AP-MS) (Ho et al. 2002). Additionally, bimolecular fluorescence complementation (BiFC) (Hu, Chinenov, and Kerppola 2002; Kodama and Hu 2010) has grown in popularity during the last few years because it allows PPI visualization in living cells, which is a key aspect to understand their cellular functions.

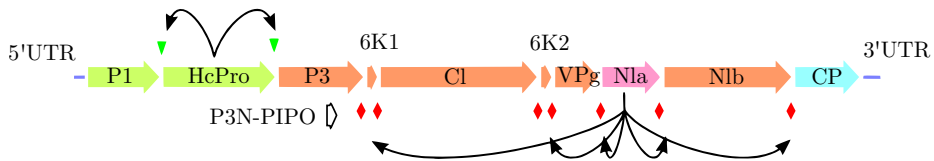


Figure 3.3. Potyvirus genome structure.

PPIs form networks of linked proteins which are called consequently protein-protein interaction networks (PPINs) (De Las Rivas and Fontanillo 2010). PPINs can be seen as a visual representation of the complete map of interactions that a system (pathway, cell, living organism) establishes in a particular moment and for a certain time window. Detection methods (specially Y2H) opened the possibility to tackle protein-protein interactions on a genome wide scale, producing complete PPINs, which have been called interactomes (Ito et al. 2001; Rual et al. 2005; Uetz et al. 2000; Venkatesan et al. 2008). Viral PPINs have also been developed (Fossum et al. 2009; Uetz et al. 2006), revealing quite useful biological information.

The analysis of viral PPINs presents interactions between two proteins of the virus (VVPIs) or interactions between viral proteins and host proteins (VHPIs). These PPINs illustrate a fundamental property of viral proteins: their multifunctionality. Viral proteins usually perform different functions at different stages of the infection cycle. Moreover, their role changes along with the infection process. Thus,

detecting VHPIs provides valuable insight into viral mechanisms and processes. VHPIs are responsible of channeling the effect of the virus into the plant. In addition, interactions between host proteins (HHPIs) are also fundamental in order to understand the interplay between virus and host, and the biological consequences once the virus effect starts to propagate across the host PPIN (Rodrigo et al. 2012).

PPINs, as any other network, may be described and studied from a complex systems point of view. Over the past fifteen years many researchers have focused on developing tools and frameworks to study, categorize and understand networks (Albert and Barabási 2002; Boccaletti et al. 2006; Newman 2003; Watts and Strogatz 1998). Some work has been done applying network theory to biological networks, developing a new discipline or approach called Network Biology (Albert et al. 2011; Barabási and Oltvai 2004; Cho, Kim, and Przytycka 2012; Russell and Aloy 2008). An excellent and updated review on topology of interaction networks may be found in (Winterbach et al. 2013). Some studies have dealt with the topological properties and features of PPINs (Albert and Barabási 2002; Pržulj, Wigle, and Jurisica 2004; Pržulj 2011; Yook, Oltvai, and Barabási 2004), however just a few have focused on viral PPINs (Elena, Carrera, and Rodrigo 2011; Elena and Rodrigo 2012; Fossum et al. 2009). Viral infection is a complex process and it requires a systems approach to be fully described. A more detailed and systematic understanding of how viral proteins interact with each other, and with host proteins, might allow developing new drugs and treatments that block the viral replication in a more efficient and durable manner. Unfortunately, there remains a need for a much deeper understanding of viral PPINs using the topological tools and methods developed by complex systems and network science.

Following this major current approach, in this section a topological analysis of the potyvirus PPIN constructed by integrating data from several different species of potyvirus is presented. VHPIs are studied as well using the complete *Arabidopsis thaliana* PPIN. Furthermore the effect that the viral network and each of its components has on the host interactome is described and quantified. Finally, new ways to visually represent the VHPI network (VHPIN) are proposed.

3.2.3 Methodology

Data collecting

All currently available potyvirus VVPI datasets were gathered as a first step. These data were obtained from six different articles published over the last decade (Guo et al. 2001; Kang, Lim, and Kim 2004; Lin et al. 2009; Shen et al. 2010; Yambao et al. 2003; Zilian and Maiss 2011). This initial dataset is the starting point of the subsequent analysis. An overview of the data is shown in TFigure 3.4.

681 PPIs were tested and 194 PPIs were detected among the 11 viral proteins from eight different viruses: *Plum pox virus* (PPV), *Soybean mosaic virus* (*Pinellia ternate* isolate, SMV-P), *Shallot yellow strip virus* (onion isolate, SYSV-O), *Potato virus A* (PVA), *Pea seed-borne mosaic virus* (PSbMV), *Soybean mosaic virus* (G7H strain, SMV-G7H) and *Clover yellow vein virus* (CIYVV). Some of the Y2H original studies included information about the relative intensity of each interaction, represented by a higher or lower number of colonies appearing after an incubation time. However, integrating the intensity data is not straightforward because it depends on some experimental variables such as sampling schemes, growth variables or environment conditions. Furthermore, differences in normalization methods, categorization and batch effects also contribute to make comparisons difficult. Especially problematic was the inclusion of the P3N-PIPO protein. This protein was discovered and characterized only recently and, therefore, it was not included in some of the studies in which this work was based. However, the statistical standardization of the data allows an appropriate representation of P3N-PIPO interactions (see Section 3.2.4, Interaction relevance subsection).

Reference	Virus	Interactions		Method
		Tested	Detected	
[44]	PPV	105	54	BFC
[45]	SMVP	100	39	Y2H
	SYSV-O	100	45	Y2H
[46]	PVA	80	16	Y2H
	PSbMV	56	10	Y2H
[47]	PRSV-P	100	16	Y2H
[48]	SMV-G7H	100	9	Y2H
[49]	CIYV	40	5	Y2H

Figure 3.4. Potyvirus interactions initial dataset. It contains data from six different studies and eight different viruses.

The second basic source of data was the *A. thaliana* interactome formed by 12654 interactions and 5127 proteins published in (Arabidopsis Interactome Mapping Consortium 2011) plus the most recently discovered HHPIs. Although some studies have analysed the changes produced by virus infection in natural hosts, *A. thaliana* is the standard model host used with viruses belonging to different taxonomic families (Elena and Rodrigo 2012). The final data source was the group of VHPIs detected between proteins from potyviruses and *A. thaliana* published originally in Elena and Rodrigo 2012 and later updated. Therefore the data covers all possible protein interactions: virus-virus (VVPI), virus-host (VHPI) and host-host (HHPI).

Data integration: interactions, matrices and networks

Integrating data from different sources in a common framework required of statistical standardization and preprocessing. First, each interaction tested in the original studies was collected. Some of them were able to test more interactions than others. In some studies it was not possible to produce enough quantity of a certain protein to test its interactions with the others. In other cases proteins had not been yet discovered when the studies took place so they are obviously absent. Additionally, not all interactions tests resulted in a positive interaction being detected. All detected interactions were collected as well. Tested and detected interactions across all sources were grouped in two common pools.

The molecular methods used to detect the interactions have an inherent directionality. Experimentally, it is common to swap the fused tags among the pair of proteins to avoid possible structural problems that may interfere with the detecting methods (*e.g.*, Y2H and BiFC). Original studies tested all interactions in two directions, for instance $P1 \leftrightarrow HC\text{-}Pro$ and $HC\text{-}Pro \leftrightarrow P1$. This produces a problem when only one direction was detected. Since the PPI itself has no directionality (it is a molecular docking phenomenon between two molecules) the disagreement comes from the molecular methods used. Some combinations of fused and viral proteins may be less stable or may block the docking of other proteins. To overcome this, it was assumed that an interaction was valid if it was detected in any of the two directions or in both. This produces symmetry in complementary interactions ($P1 \leftrightarrow HC\text{-}Pro$ and $HC\text{-}Pro \leftrightarrow P1$) representing the real process of interacting in a clearer and more truthful way.

The next step was to determine which interactions were relevant and which ones were fair representations of the *Potyvirus* genus topology. Given the variability among studies (*e.g.*, virus species and experimental conditions) it is not surprising that some interactions were detected only in one or few studies, while other were pervasive across the entire dataset. On the other hand, the relative scarcity of the data (only 194 interactions detected) made difficult and somewhat useless a more detailed statistical analysis. Even a confidence interval for each interaction with only eight independent values (corresponding to the eight viruses) is not reliable enough. Therefore, a relevance coefficient (RC) between the numbers of detected and tested interactions for each pair of proteins was defined. It is reasonable to assume that RC is a measure of biological importance. In other words, the more times an interaction has been detected, the higher the probability that this particular interaction is important for the virus to complete its infectious/replication cycle. However, considering the particularities of each method, percentages for Y2H and BiFC were weighted. The latter is closer and much more biologically coherent to natural conditions where potyvirus interactions take place. Therefore, it was decided to overweight the only study in which this method was used ((Zilian and Maiss 2011). Thus, RC takes the form $RC = 100 \cdot (2[BiFC] + [Y2H]) / (T + 1)$,

where T is the number of times that a particular interaction was tested (from 0 to 8), $[BiFC]$ is the number of times that a given interaction was detected using the BiFC method (from 0 to 1 because only one study used BiFC) and $[Y2H]$ corresponds to the number of times that an interaction was detected using the Y2H methodology (from 0 to 7). The factor of 2 multiplying the $[BiFC]$ term is a simple way to overweight this method against the Y2H. Doubling its importance was a compromise solution between being truthful to the particularities of each method and still gathering all the relevant information. RC can range then from 0% (the interaction was not detected in any of the studies) to 100% (was detected in every single study).

It was decided to establish the RC threshold for each interaction at the minimum value where all nodes were part of a single connected network, which occurred at $RC = 44\%$. This choice has biological meaning because is based on the fact that all Potyvirus genomes encode for the eleven proteins and that all these proteins have been reported to interact at least once with each other. Therefore, it is only possible to study this particular system assuming only one connected network, which appears at $RC = 44\%$. It was also decided to set the threshold at this value to include all information considered relevant from this approach. This threshold is data-dependent and therefore can change from network to network. Even with the same dataset it may be changed to satisfy a particular research objective. For instance, setting a higher RC makes the analysis focus on the most frequent interactions, which may be interesting in a specific situation. However, lower RC than 44% results in a disconnected network with various components. Using the relevant interactions an interaction matrix was constructed with the eleven viral proteins as rows and columns, and the RC values for the interactions in each position. Finally, this matrix is visually displayed in a PPIN.

Network topology

After integrating the data, an exhaustive topological analysis was carried out. First, the protein connectivity aspects of the network were studied: protein degree, RC relation with protein degree and assortativity. Then a group of topological parameters was calculated for the viral PPIN and its nodes:

- Clustering Coefficient is defined as the number of edges between the neighbours of a node divided by the maximum number of edges that could possibly exist between those neighbours.
- Closeness Centrality is the reciprocal of the average shortest path length between a particular node and all the rest.
- Betweenness Centrality is the number of shortest paths in which a particular node lies on.

- Topological Coefficient is the average of the number of neighbours shared between a particular node and all its neighbours. It is normalized by the degree of such node.

Finally an analysis of these topological parameters was carried out: their relation with the degree and their cumulative distributions.

The topological analysis of the viral PPIN and its nodes was repeated for those individual virus networks with enough interactions detected to form a complete topology (Table 3.4): PPV, SMV-P and SYSV-O. All the networks were constructed and the parameters calculated using the software Cytoscape (Shannon et al. 2003) and its network analyzer tool.

Virus-host interactome

The purpose of the analysis between the virus proteins and the host ones is to achieve an overall better understanding of their relationships and integration, which is pivotal to grasp the infection process. For this, an approach to quantify the importance that each viral protein has over the host network was used. The first order connectivity that each viral protein has with the host proteins can be extracted directly from the data. Starting from each viral protein, and following the host interactome, it was calculated how many steps (consecutive interactions) are needed to reach each host protein. At the end, it is possible to map the consecutive steps from the viral protein to the last host protein. This was repeated for all viral proteins and the propagation trajectories produced were plotted.

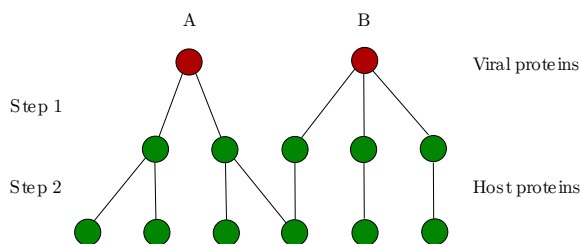


Figure 3.5. Examples of steps of interactions. Step is the measure used to define distance between proteins. In this example **A** would establish 2 interactions in step 1 and 4 in step 2. **B** has 3 in step 1 and 3 in step 2.

Several considerations are here in due, starting from the concept of *distance* in a graph. In this section it was used the simplest distance measure possible, which is the shortest path between two nodes, which comes directly from the adjacency matrix and the cross-interactions or VHPIs. The minimal measure of distance is called here step. The distance between two proteins interacting directly is one step.

The distance between two proteins that interact with another common protein is two steps (Figure 3.5). From this simple distance a metric was used to qualify the interaction-profile similarity of the viral proteins. Nonetheless, much more complex similarity coefficients (Fouss et al. 2012) can be used as kernels on graphs (e.g., exponential diffusion kernel, Laplacian exponential diffusion kernel, or the commute time kernel).

The similarity of the spreading trajectories was compared for every pair of viral proteins with a similarity coefficient or index (Fuxman Bass et al. 2013). The total amount of interactions is 66 (combinations of eleven proteins taken by pairs). The Simpson index (SI) was chosen, which is commonly used in systems biology and network science. It is defined as the proportion of shared nodes relative to the degree of the least connected node: $SI(A, B) = |N(A) \cap N(B)| / \min(|N(A)|, |N(B)|)$. SI changes in each step so the similarity evolves along the whole host interactome. This index offers a quick and insightful way of quantifying the similarity that two viral proteins show in their relationship with the host network.

3.2.4 Results and discussion

As outlined in the Section 3.2.2, the aim of this study is to describe and characterize the PPIN of potyviruses using tools and techniques from network science. As mentioned earlier, the study starts from three different datasets: VVPis, VHPis and HHPIs. VVPis allowed the study of the topology of the network, composed exclusively by 11 viral proteins. Next, VHPis and HHPIs were evaluated and used to describe and quantify the integration of the viral PPIN within the larger interactome of the host plant.

VVPI network analysis

In this subsection different aspects of the topology of the network were studied in detail. Y2H and BiFC analysis and Y2H intensity subsections deal with the differences between the detection methods and the nature of the information they provide and the possible consequences for the study. VVPI network construction and visualization subsection shows how the network was visually defined and the last three (Interaction relevance, Protein connectivity and Topological analysis) focus on several aspects of the topological properties of the network.

Y2H and BiFC analysis

In this subsection, the results inferred from data generated using the two detection methods were compared. The aim of this comparison was to find out whether a method tends to detect some interactions but not others or, on the contrary, the main interactions were evenly detected by both methods. Interactions detected by both methods will be more reliable than those detected by only one method. The

number of observed interactions was classified according to the detection method (Figure 3.4). Some direct remarks can be made just from this simple classification. First, there are 5.4 times more data available from Y2H than from BiFC, which reflects the more recent technological development of BiFC but also introduces a bias towards Y2H-based studies. Despite the lower number of interactions studied using BiFC, the number of positive cases is significantly larger for this technique than for Y2H (Fisher's exact test p -value < 0.001), thus proving that BiFC is a more sensitive method. Moreover, BiFC preserves the biological relevance of the interactions detected, since this technique seeks for interactions in plant rather than detect heterologous expression of proteins in yeast cells.

Y2H is an older method, widely used because of its simplicity, speed and its ability to generate interactions at genome level. Y2H also provides a rough measure of interaction intensity given by the number of colonies that grow in each experiment and usually distributed in several ranges (from 1 to 5, from 5 to 10, etc.). Alternatively, BiFC does not provide a quantitative value. Some particularities arise when they are compared. The interaction between CI and P3N-PIPO was only tested and detected by BiFC (due to the recent discovery of the P3N-PIPO protein). Interestingly, the most common interactions are detected by both methods and appear in both networks; out of the 26 most relevant interactions (displayed in Table 2) only three were detected by Y2H but not by BiFC (HC-Pro \leftrightarrow HC-Pro, HC-Pro \leftrightarrow NIaPro and HC-Pro \leftrightarrow VPg). This implies that both methods, although different in scope and sensitivity, offer highly consistent results. This consistency validates the approach of integrating data from both techniques into a single dataset.

Y2H intensity

Intensity data was used (whenever available) to try to correlate it with the frequency of each interaction. All the data from Y2H studies was grouped together and the intensity was plotted against the overall frequency of all interactions (data not shown). No correlation was found ($r = 0.249$, 45 d.f., p -value = 0.172) between intensity and frequency for any of the seven potyvirus studied with Y2H. This leads to the conclusion that the biological importance of an interaction (related with the frequency with which it is detected) is not function of its intensity. In other words, interactions with lower intensity can be as vital to virus development as the more intense.

VVPI network construction and visualization

As it was explained in the Section 3.2.3, a threshold of 44% was set in the RC to separate relevant interactions from the rest. With this constraint, only 26 out of the 66 possible interactions were considered as relevant. With those interactions the global interaction matrix (GLIM) was built (Figure 3.6). The network defined by GLIM shows the proteins as nodes and the interactions as edges. It represents

	P1	HC-Pro	P3	6 K1	CI	6 K2	VPg	NIaPro	NIb	CP	P3N-PIPO
P1					57%		63%				
HC-Pro		78%			44%		44%	44%			
P3								56%	67%		
6 K1								44%			
CI					57%		56%	56%		50%	100%
6 K2							44%	44%			
VPg							89%	56%	56%	44%	
NIaPro								78%	78%	44%	
NIb									44%	56%	
CP										88%	
P3N-PIPO											

All interactions with a $RC > 44\%$ are displayed in a matrix form.

Figure 3.6. Global interaction matrix.

the VVPs detected in the studies with a $RC > 44\%$. Additionally, to increase the visual information the width of the edges was made proportional to the RC of the interactions. The resulting network (Figure 3.7) is the global interaction network (GLIN).

Interaction relevance

The starting point for the topological analysis is the computation of the RC for every interaction (with $RC > 44\%$) experimentally detected (Figure 3.8). Some interesting information arises from this representation. The most common interactions have a RC in the range 80% - 90% (with exception of the $CI \leftrightarrow P3N-PIPO$). $P3N-PIPO$ was tested only in one of the studies (Zilian and Maiss 2011) and only against three proteins: itself, CP and CI. The positive hit of the $CI \leftrightarrow P3N-PIPO$ interaction produces a $RC = 100\%$ for this particular interaction. However, it is reasonable to assume that after $P3N-PIPO$ is tested against all viral proteins in future studies, this RC value will decrease. Core interactions involve proteins CI, VPg, NIaPro and NIb. Out of the 66 possible interactions, 26 were considered relevant representing a striking 39.3%. This shows clearly that the intraviral network is highly connected. It is generally accepted that viral proteins are multi-functional, so this high connectivity was expected. Another interesting conclusion drawn from Figure 3.8 is that there is no specific RC threshold dividing the interactions between the most common and the rarest. In other words, there are interactions detected across all the RC range (from 100% to the established limit of 44%).

Protein connectivity

In a PPIN, the degree of each node matches the number of different interactions in which each protein is involved but only if there is no self-interaction. If there is, the protein degree equals the number of interactions plus one. Supporting the

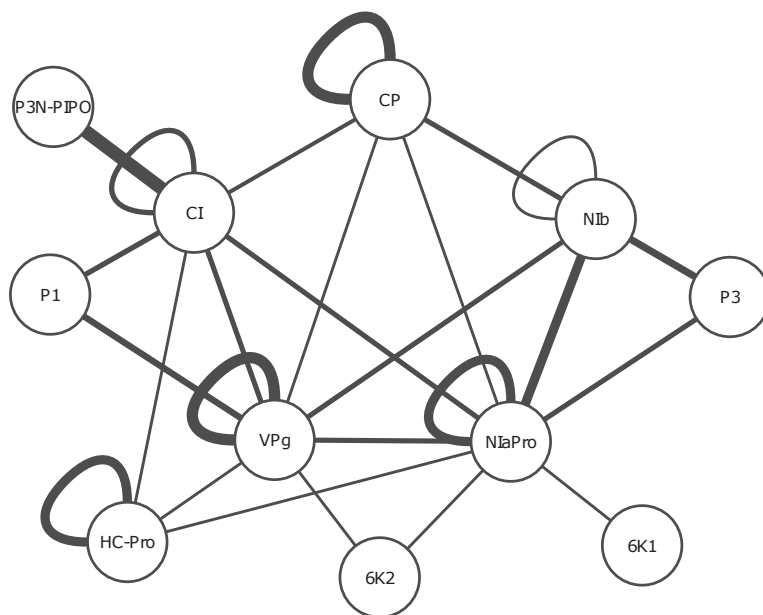


Figure 3.7. Global interaction network.

idea that viral PPINs are highly connected, Figure 3.9a shows that the degree of most proteins is in a narrow range (2-10). However, a clear distinction can be made between high and low connected proteins. Low connected proteins are P1, P3, 6K1, 6K2, and P3N-PIPO, and they have a degree in the low range of 1-2. Highly connected ones are HC-Pro, CI, VPg, NIAPro, NIB, and CP, with a degree of 5-10.

Furthermore, it was investigated if there is some relation between interactions relevance and protein degree. It seemed that interactions with the highest RC were formed by proteins with a high degree. To check this a correlation study was performed, and it found no relation between RC and degree ($r = -0.034$, 24 d.f., $p - value = 0.871$). In spite of that, it is noteworthy that the five most relevant interactions (VPg \leftrightarrow VPg, CP \leftrightarrow CP, NIAPro \leftrightarrow NIB, NIAPro \leftrightarrow NIAPro and HC-Pro \leftrightarrow HC-Pro) are formed by proteins with a high degree (without considering the CI \leftrightarrow P3N-PIPO interaction).

It is also interesting to study the assortativity (Newman 2002) of the network. Assortative mixing is the preference for the nodes of a network to attach to others that are similar. This is commonly examined in terms of a node degree. In PPINs, it consists of studying whether high degree proteins tend to establish interactions with other high degree proteins. One way to capture the assortative behaviour of

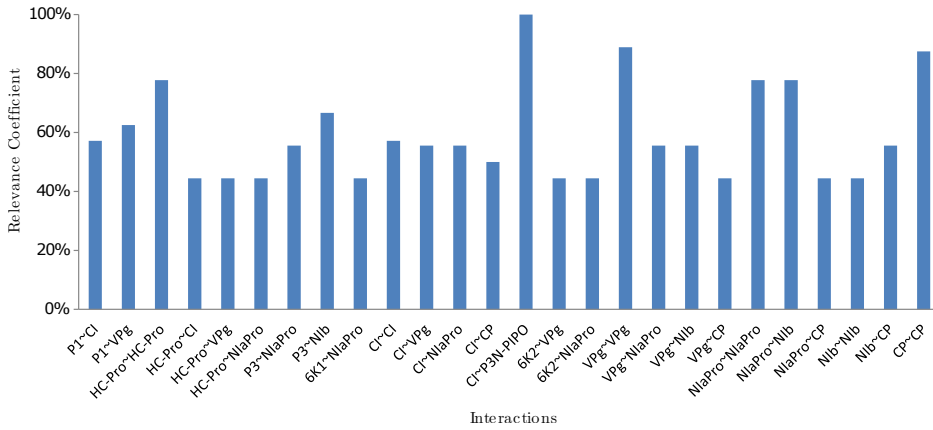


Figure 3.8. Relevance coefficient (RC) of all interactions of the global interaction network.

a network is to examine the average neighbour connectivity. The connectivity of a node is the number of its neighbours. The neighbourhood connectivity of a node is defined as the average connectivity of all its neighbours. The neighbourhood connectivity distribution gives the average of the neighbourhood connectivities of all nodes with k neighbours for $k = 0, 1, \dots$. If this function is increasing, the network is assortative, since it shows that nodes of high degree connect, on average, to nodes of high degree. On the other hand, if the function is decreasing, the network is disassortative, since nodes of high degree tend to connect to nodes of lower degree. Average neighbour connectivity distribution for the GLIN is shown in Figure 3.9b. The values of the parameter decrease with the number of neighbours, therefore the GLIN shows a disassortative behaviour. This agrees with previous studies that stated the disassortative nature of biological networks (Newman 2002). However, biological interpretation of this fact remains unclear. Hierarchical structures in biological networks may result in disassortativity. Regulatory genes or transcription factors influence many particular genes or proteins with specific biological functions. Therefore, hubs correspond to regulators and less connected nodes to actuators, dividing the network in several hierarchical levels. Among the 11 nodes in the PPIN, HC-Pro is the most highly connected component, interacting with all other nodes. Therefore, disassortativity in this network emerges as a simple consequence of the limited number of nodes and that the most connected one interacts with all other nodes, regardless their specific connectivity.

Topological analysis

As it was mentioned in the Section 3.2.3, a complete topological analysis of the GLIN and all its nodes was carried out. First, a set of general topological param-

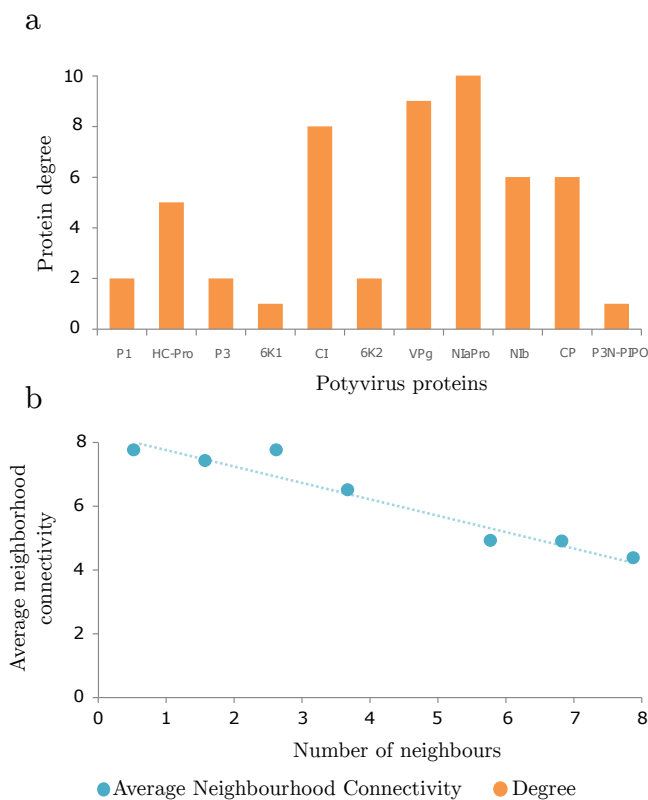


Figure 3.9. Protein degree and connectivity of the global interaction network (GLIN). (a) Degree of each potyvirus protein. (b) Average neighbourhood connectivity distribution.

eters was calculated for the entire GLIN (Figure 3.10). The clustering coefficient is high and the characteristic path length is lower than two, emphasizing the fact that GLIN is highly connected. The number of self-loops is quite high (six out of 11 possible), meaning that most proteins interact with themselves for carrying out some of the biological functions.

In addition, four topological parameters were computed for each protein in the GLIN. This topological information is displayed in Figure 3.11a. Some parameters contain related information such as centralities and the clustering and topological coefficients. NIaPro, VPg and CI have the highest centralities and the lowest clustering and topological coefficients. A similar conclusion can be drawn from the low clustering and topological coefficients of 6K1 and P3N-PIPO because they do not form any 3-loop in the network. P3N-PIPO is only linked to CI and 6K1 only to NIaPro. Therefore their topological parameters are quite different from

Clustering coefficient	0.605
Connected components	1
Network diameter	3
Network radius	2
Network centralization	0.533
Characteristic path length	1.745
Average number of neighbors	3.636
Number of nodes	11
Network density	0.364
Network heterogeneity	0.634
Number of self-loops	6

Figure 3.10. General topological parameters of the global interaction network.

the highly connected rest of proteins (especially the clustering and topological coefficients, which are based on common neighbors). An identical analysis was performed for PPV, SMV-P and SYSV-O, since they were the only ones with enough interactions detected to construct a complete topology.

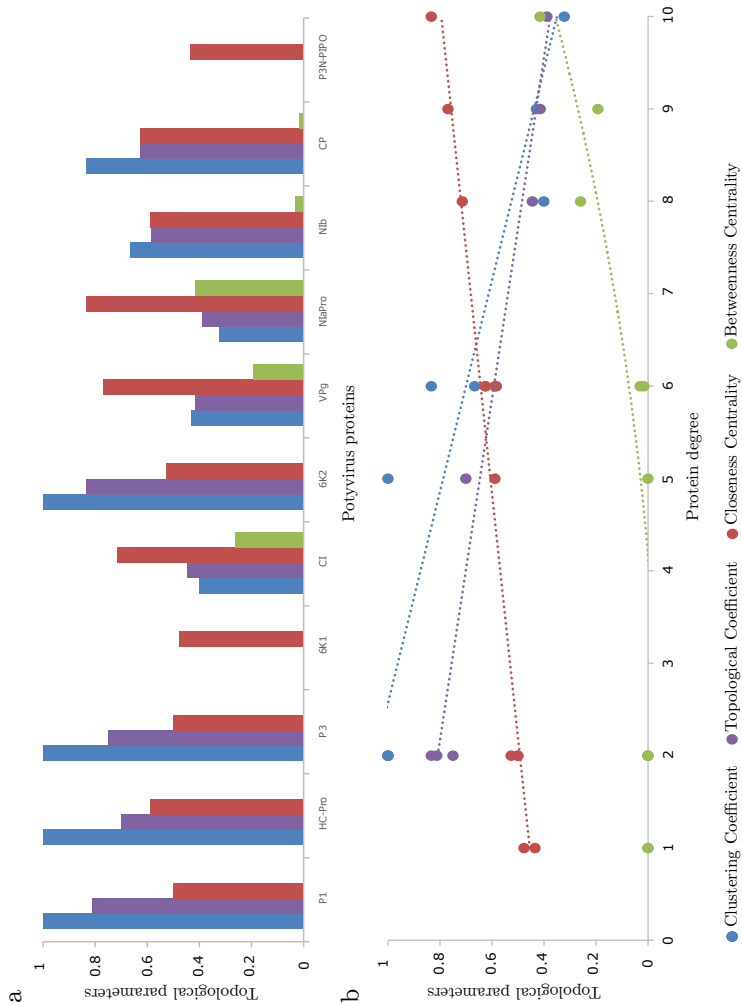


Figure 3.11. Topological analysis of the viral proteins. (a) Topological parameters of each protein. (b) Topological parameters of proteins related with their degree. 6K1 and P3N-PIPO data for the clustering and topological coefficients were removed from the representation (commented in text).

It is important to remark that these parameters are in part influenced by the degree of each protein (Figure 3.9a). In general, the clustering and topological coefficients increase with degree while closeness and between centrality decrease (Figure 3.11b). The least connected proteins have an extreme clustering coefficient (0 or 1) while the most connected ones have intermediate values. Both centralities

are higher for high degree proteins, which is to be expected. HC-Pro is located somewhere in the middle. It has a high degree but its centralities are low and its topological coefficient is high. It also has an extreme clustering coefficient. Clustering and topological coefficients have the worst fitting to a linear regression due to the low degree of 6K1 and P3N-PIPO, which was already discussed. Complete statistical description of the regressions (p -value, d.f. and r^2) can be found in Figure 3.12. It is worth noting that non-linear models have a better fit in the betweenness centrality data.

Y	X	Linear regression results			
		R2	g.l.	p-value	coefficient
Clustering coefficient	Degree	0.869	8	<0,001	-0.141
Topological coefficient	Degree	0.957	8	<0,001	-0.087
Closeness centrality	Degree	0.954	10	<0,001	0.038
Average neighborhood connectivity	Number of neighbors	0.93	6	<0,001	-0.541

Y	X	Non linear regression (quadratic)			
		R2	g.l.	p-value	coefficient
Betweenness centrality	Degree	0.837	10	<0,001	0.004

Figure 3.12. Statistical description of the regression for the topological parameters of the viral proteins.

Finally, the topological distributions of the different parameters were determined, displayed and studied. Topological distributions compute the probability that a node in a network presents a particular value in some parameter. For instance, the probability of a node to have a degree of three. Although informative, they are more useful when computed as cumulative distributions. Following the example, the probability of a node to have degree lower than or equal to three. Cumulative distributions of degree and other topological parameters were calculated for the GLIN (Figure 3.13a and b). The cumulative degree distribution for the GLIN shows a quasi-linear behaviour. Obviously, the probability increases with the degree. The other cumulative distributions also tend to be linear.

VHPI network analysis

In this subsection, integration of the virus network and the host network (through VHPIs and HHPIs) was studied. VHPI network construction and visualization subsection focuses on the difficulty of the faithfully representation of networks of this size. Effect propagation deals with the effect of specific viral proteins along the HHPIN and Similarity analysis focuses on the comparing the patterns of propagation of pairs of viral proteins.

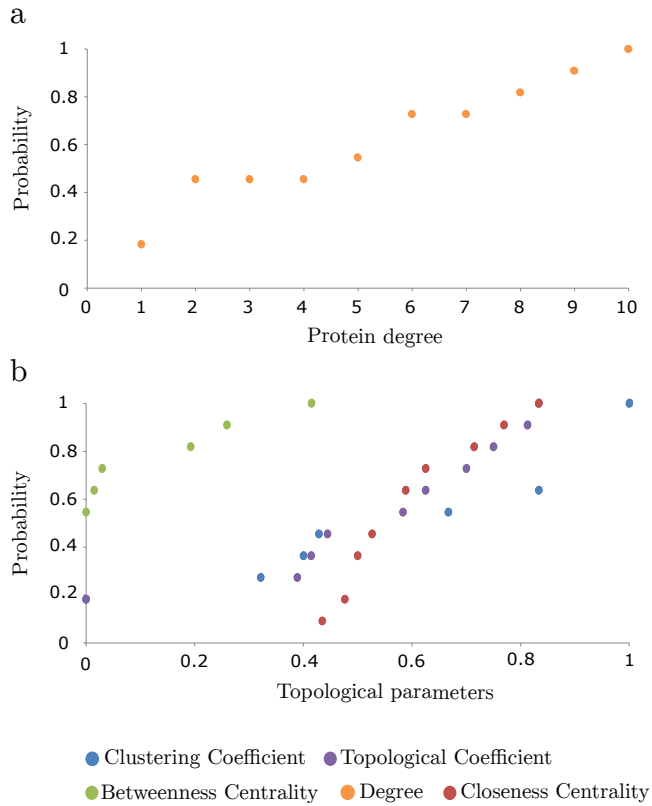


Figure 3.13. Cumulative distributions of topological parameters of viral proteins. (a) Degree cumulative probability distribution. It shows the probability that a protein has a determined degree or lower. (b) Cumulative probability distribution of several topological parameters.

VHPI network construction and visualization

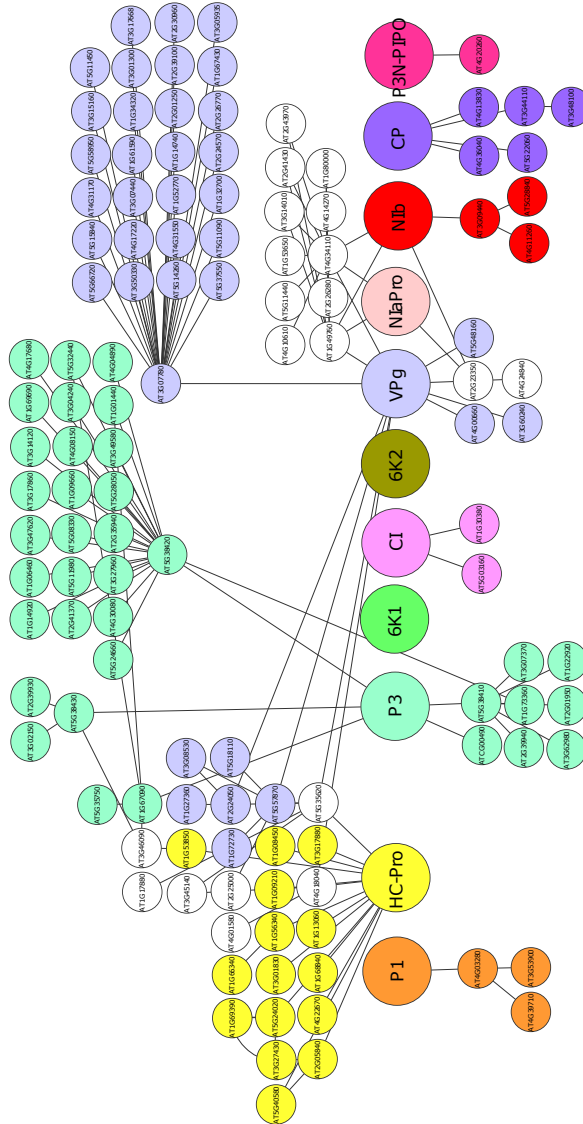


Figure 3.14. Potyvirus-*A. thaliana* VHPI network (VHPIN). Proteins and their host neighbors are grouped by colors. White color is assigned to host proteins connected to several viral proteins during the same step. For instance, host protein At2G23350 (located just below VPg protein) is coloured white because is linked directly to two different viral proteins: VPg and NIb).

Potyvirus proteins establish interactions with a large unknown number of host factors, disrupting the normal development of the plant. These VHPIs channel the harmful effect of the virus and point to the vital nodes of the PPIN and transcriptional regulatory network of the host (Rodrigo et al. 2012). The effect propagates from those direct VHPIs through the entire network of HHPIs. Visualization of the *A. thaliana* interactome is impossible in practical terms. It has 5127 nodes (proteins) and 12624 edges (interactions) and therefore any attempt to visually represent the network as a whole is not going to provide useful information. Instead, the 11 potyvirus proteins were illustrated surrounded by two levels or steps of plant interactions (Elena and Rodrigo 2012). This Potyvirus-*A. thaliana* VHPI network (VHPIN) (Figure 3.14) provides a quick overview of the anchor points that the virus uses to hijack the plant network. It is clear that the virus hits many proteins in the first step. However, the interactions vary in number and connectivity. For instance, proteins P3 and VPg hit two host proteins that are network hubs while HC-Pro directly interacts with more than 10 different proteins and then diversifies its effect to all the interactions of these proteins. The VHPIN does not show any information of the interactions happening in successive next steps (step 3, 4 and so on).

Effect propagation

To study the potential effect that the viral proteins have on the network, the 11 viral proteins were taken as starting point and used the *A. thaliana* interactome as a map to draw the complete tree of interactions that appear until no more interactions are possible. The first two steps are represented in the VHPIN but beyond that it is not practical to visualize the interactome as a network illustration, so we have to rely on mathematical description. For instance, the protein P1 establishes only one interaction with a plant protein (step 1), then this protein establishes two interactions with other plant proteins (step 2, the VHPIN displays the protein relationships up to this point) but the network keeps growing; these two proteins link with 13 proteins (step 3), these 13 link with 110 (step 4) and so on. These calculations were repeated for the 11 viral proteins and the results are displayed in Figure 3.15 (the figure shows the cumulated number of interacting proteins).

Some information may be directly extracted from the illustration. Hence, 6K1, CI, 6K2, and P3N-PIPO establish virtually no interactions with the host. Three possible explanations for this lack of interactions are possible. (i) These proteins function only by interacting with other viral proteins but not with host factors. (ii) These proteins may interact with host proteins via other viral proteins or via other host elements such as RNA, DNA, lipids or carbohydrates. And (iii) the lack of reported interactions does not necessarily means these interactions do not exist, reflecting the need of additional work. This is the especially relevant for the recently described P3N-PIPO.

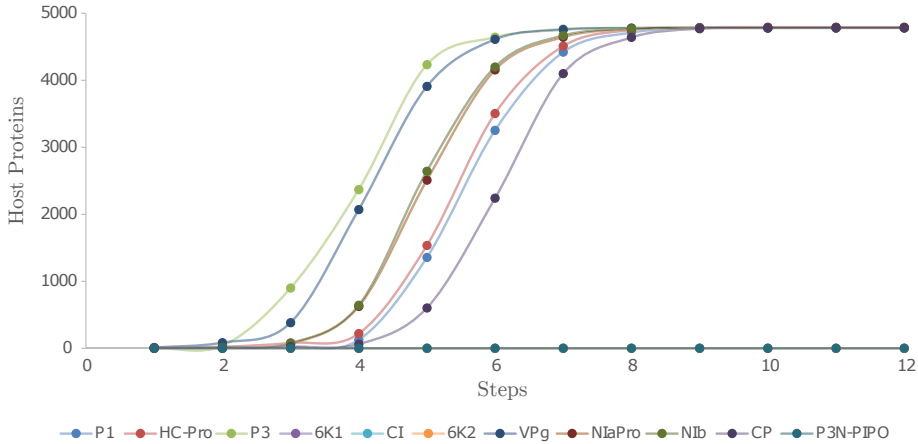


Figure 3.15. *A. thaliana* interactome coverage. It shows the protein-protein interactions occurring from each potyvirus protein and going across the whole plant network.

The other seven proteins are able to reach essentially the whole *A. thaliana* network (around 93%). Full speed propagation starts in step two and ends around step eight. Some small sections of the network are unreachable because they are not connected to the main module. Of course, this does not mean that the effect of those seven proteins is relevant and significant in the whole plant network. The effect may lose its biological importance after a few steps of interactions unless the affected proteins are transcription factors that may function as hubs in the global regulatory network. In such case, the perturbation will be efficiently transmitted along the entire network. In all other instances, viral proteins will affect the host network only to a certain extent and possibly circumscribe their action to specific branches or modules. However, a global analysis is still useful to compare the viral proteins with one another. Some proteins such as P3 or VPg propagate their action through the network remarkably faster than others like CP or P1. This may indicate the sequential order in which the effect of the proteins crosses the network during the virus cycle. This measure of steps can be seen as a temporal variable. The effect of one viral protein is likely to be noticed earlier in a host protein located two steps away than other located six steps away. It seems reasonable to assume that, in spite of the enormous diversity and relevance of host interactions, some viral proteins act earlier than others during the infection cycle and that this kind of propagation analysis is a reasonable approach to study them.

Protein	Interaction	Steps								Protein
		1	2	3	4	5	6	7	8	
P1	1	1	1	1	1	1	1	1	1	P1
	2	0	0	0	0.11	0.59	0.89	0.98	1	HCPPro
	3	0	0	0.44	0.94	1	1	1	1	P3
	7	0	0	0.31	0.92	1	1	1	1	VPg
	8	0	0	0	0.53	0.93	0.98	1	1	NlaPro
	9	0	0	0	0.53	0.93	0.99	1	1	Nlb
	10	0	0	0	0.02	0.5	0.86	0.98	0.99	CP
HC-Pro	12	0	0	0	0.11	0.59	0.89	0.98	1	P1
	13	1	1	1	1	1	1	1	1	HCPPro
	14	0	0.04	0.59	0.8	0.99	1	1	1	P3
	18	0.18	0.3	0.16	0.57	0.91	0.99	1	1	VPg
	19	0	0	0.01	0.4	0.77	0.97	0.99	1	NlaPro
	20	0	0	0.01	0.4	0.78	0.97	0.99	1	Nlb
	21	0	0	0.12	0.54	0.94	0.97	1	1	CP
P3	23	0	0	0.44	0.94	1	1	1	1	P1
	24	0	0.04	0.59	0.8	0.99	1	1	1	HCPPro
	25	1	1	1	1	1	1	1	1	P3
	29	0	0	0.36	0.76	0.96	0.99	1	1	VPg
	30	0	0	0.56	0.94	1	1	1	1	NlaPro
	31	0	0	0.51	0.93	0.99	1	1	1	Nlb
	32	0	0	0.28	0.71	0.99	1	1	1	CP
VPg	67	0	0	0.31	0.92	1	1	1	1	P1
	68	0.18	0.3	0.16	0.57	0.91	0.99	1	1	HCPPro
	69	0	0	0.36	0.76	0.96	0.99	1	1	P3
	73	1	1	1	1	1	1	1	1	VPg
	74	1	1	1	1	1	1	1	1	NlaPro
	75	0.75	0.81	0.88	0.98	0.99	1	1	1	Nlb
	76	0	0	0	0.3	0.87	1	1	1	CP
NlaPro	78	0	0	0	0.53	0.93	0.98	1	1	P1
	79	0	0	0.01	0.4	0.77	0.97	0.99	1	HCPPro
	80	0	0	0.56	0.94	1	1	1	1	P3
	84	1	1	1	1	1	1	1	1	VPg
	85	1	1	1	1	1	1	1	1	NlaPro
	86	1	1	1	1	1	1	1	1	Nlb
	87	0	0	0	0.13	0.77	0.98	1	1	CP
Nlb	89	0	0	0	0.53	0.93	0.99	1	1	P1
	90	0	0	0.01	0.4	0.78	0.97	0.99	1	HCPPro
	91	0	0	0.51	0.93	0.99	1	1	1	P3
	95	0.75	0.81	0.88	0.98	0.99	1	1	1	VPg
	96	1	1	1	1	1	1	1	1	NlaPro
	97	1	1	1	1	1	1	1	1	Nlb
	98	0	0	0	0.13	0.77	0.98	1	1	CP
CP	100	0	0	0	0.02	0.5	0.86	0.98	0.99	P1
	101	0	0	0.12	0.54	0.94	0.97	1	1	HCPPro
	102	0	0	0.28	0.71	0.99	1	1	1	P3
	106	0	0	0	0.3	0.87	1	1	1	VPg
	107	0	0	0	0.13	0.77	0.98	1	1	NlaPro
	108	0	0	0	0.13	0.77	0.98	1	1	Nlb
	109	1	1	1	1	1	1	1	1	CP

Figure 3.16. Simpson index evolution for all viral proteins.

Similarity analysis

Effect propagation analysis does not evaluate how similar two viral proteins are in their relationship with the host; whether they hit the same host proteins and in the same or similar number of steps. Some measure of similarity in effect propagation among viral proteins is thus needed. For example, let's assume that P1 reaches five host proteins while HC-Pro reaches 10 at a determined step, and that one of those host proteins (HP1) is common for both viral proteins. Two groups are formed: P1-group (with five members) and HC-Pro-group (with 10 members) having one member (HP1) belonging to both groups at the same time. It is possible to quantify the similarity of those two groups using a similarity coefficient such as the *SI*. It varies from 0 to 1 and expresses the similarity between two groups of proteins. The *SI* was calculated it for every pair of viral proteins (55) and for all the steps (12) (Figure 3.16). Note that proteins 6K1, CI, 6K2, and P3N-PIPO were removed from the table since they do not interact with the host proteins. The *SI* was calculated as an accumulative variable. This way each value gives an idea of similar behaviour up to that step. Plotting its evolution over the steps produces dynamical coinciding patterns. It tends to increase in the mid-steps because at that point the viral effects are propagating at full speed, and those interactions are usually common to most viral proteins.

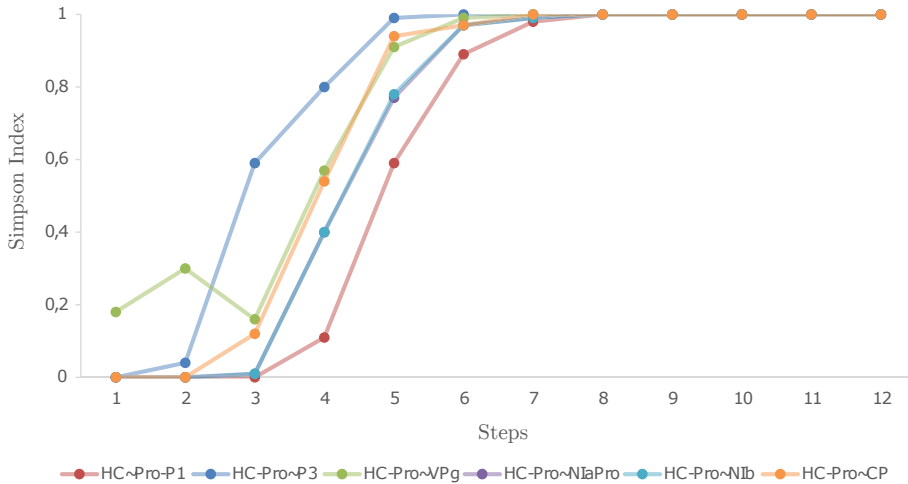


Figure 3.17. Simpson index evolution for HC-Pro. All possible combinations between HC-Pro and other viral proteins that propagate through the network. Differences in speed and shape of the spreading patterns for each pair can be easily observed. Straight lines link the values of the *SI* for each step representing how it varies while the protein pair effect propagates through the network.

Different features can be illustrated through *SI* graphs quite easily. Figure 3.17 shows the *SI* for all proteins paired with HC-Pro. The most common behaviour for a couple of proteins is that similarity starts at zero and begins to increase around step 2-3 until it reaches its maximum at step 7-9. The first and main difference is speed; some pairs reach a high *SI* much faster (e.g., HC-Pro↔P3) than others (HC-Pro↔P1). However, there are a few cases in which the *SI* for a pair of proteins decreases at some steps (HC-Pro↔VPg, steps 2-3). This is somehow surprising, since the index is calculated with accumulated proteins in each step. Therefore, the networks are always increasing their size in each step. However, in some interactions (and for some steps) the networks of both proteins increase but the common host proteins to both viral proteins in that steps does not increase proportionally. Consequently there is an absolute decrease in similarity. Nonetheless, *SI* always end up increasing until a value of almost one because the seven viral proteins that propagate their effect all reach the entire host network.

The information drawn from this similarity analysis complements the effect propagation study shown before. However, even for pairs of proteins, representing visually similarity is not trivial. Similarity evolution for a specific pair of proteins can be easily plotted but displaying all of them at the same time, while retrieving useful biological information, is much more difficult. To tackle this a voxel-based representation was used. A 3-dimensional matrix called voxel was constructed to visually represent the evolution of the *SI* over the host-host protein interaction network (HHPIN). The first two dimensions represent the eleven viral proteins; this creates a grid that assigns a pixel to each pair of viral proteins. The main diagonal has no biological meaning because the similarity of a protein with itself is always one. Furthermore, the information is repeated twice in the grid (P1↔HC-Pro pixel contains obviously the same information as the HC-Pro↔P1 pixel). The color of the pixel represents the value of the *SI* for that particular combination. The third dimension is the distance (measured in steps) from the original viral pair of proteins to any particular point in the HHPIN. This representation (Figure 3.18) allows any viewer to find quickly the spaces of interest: which viral proteins link with the host, in which steps the *SI* changes the most, which pairs of proteins follow a determined evolution, etc. Additionally the projection of each pixel over the steps (Figure 3.18d) reveals the particular evolution of the *SI* for that pair of proteins.

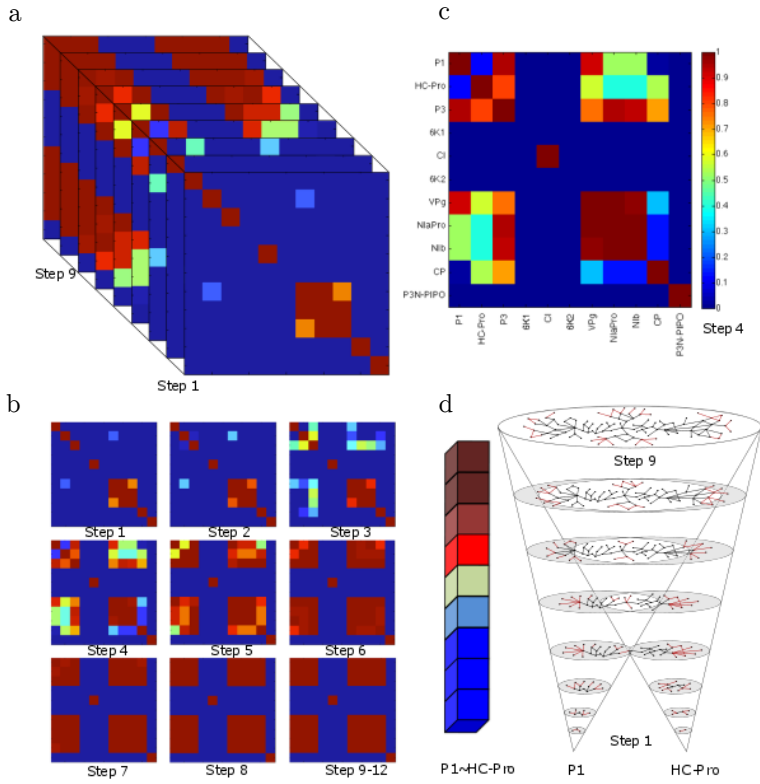


Figure 3.18. Voxel representation of the Simpson index. (a) Voxel representation of the Simpson index for the viral proteins across the HHPIN. (b) Consecutive pixel representations of the *SI* for the twelve steps that form the HHPIN. (c) Pixel representation for step 4. Viral proteins are shown in X and Y axes and relevance coefficient color legend is displayed on the right side vertical axis. (d) Evolution of the *SI* for the P1↔HC-Pro interaction across the entire HHPIN. A schematic cone of possible interactions is displayed as well to visually represent the networks growing from the viral proteins (step 1) until the end of the HHPIN.

3.3 Potyvirus data fusion

3.3.1 Summary

Data fusion has been widely applied to analyse different sources of information, combining all of them in a single multivariate model. This methodology is mandatory when different omic data sets must be integrated to fully understand an organism using a systems biology approach. In this second section of Chapter 3, a data fusion procedure is presented to combine genomic, proteomic and phenotypic data sets gathered for Tobacco etch virus (TEV). The genomic data correspond to random mutations inserted in most viral genes. The proteomic data represent both the effect of these mutations on the encoded proteins and the perturbation induced by the mutated proteins to their neighbours in the protein–protein interaction network (PPIN). This is the link with the previous section, where the potyvirus network topology is defined and thoroughly studied. Finally, the phenotypic trait evaluated for each mutant virus is replicative fitness. To analyse these three sources of information a Partial Least Squares (PLS) regression model is fitted in order to extract the latent variables from data that explain (and relate) the significant variables to the fitness of TEV. The final output of this methodology is a set of functional modules of the PPIN relating topology and mutations with fitness. Throughout the re-analysis of these diverse TEV data, valuable information is generated on the mechanism of action of certain mutations and how they translate into organismal fitness. Results show that the effect of some mutations goes beyond the protein they directly affect and spreads on the PPIN to neighbour proteins, thus defining functional modules.

3.3.2 Background

In this section, the genetic information are obtained from a collection of 20 Tobacco etch virus (TEV) single nucleotide substitution mutants and 53 double mutants resulting from the pairwise combination of the single ones (Lalić and Elena 2012). For each of these 73 mutant genotypes, absolute fitness was evaluated in its natural host *Nicotiana tabacum var Xanthi nc* during a single infection cycle (Lalić and Elena 2012). On the other hand, the PPIN inferred from empirical protein–protein interaction (PPI) data from several potyviruses analysed in Section 3.2 is used to relate the mutations and the organismal fitness. A mutation in a protein may change (slightly or dramatically) its ability to perform its biological functions correctly. The mutated TEV proteins establish interactions with other viral proteins according to the PPIN of potyviruses. Since viral proteins are multifunctional, and they carry out some of their functions as protein complexes, it is reasonable to assume that a part of the effect of the mutated protein on the fitness is channelled through its PPIs. In other words, mutations affect PPIs, which ultimately affect biological fitness. Some mutations are much more harmful while others have

no fitness effect. The PPIN of *Potyvirus* adds biological context to the mutation and allows for a deeper analysis of the importance of each protein in the virus infectious cycle.

Some assumptions are made in the present approach. The main one is that each mutation affects all the PPIs of a mutated protein in the same way. Probably the true modifications are subtler, depending also on other factors. Proteins are highly heterogeneous structures and modifications in different parts of their sequence may have different biological consequences for different interactions. However, the lack of available data and their nature constrained the present study. The problem revolves around two issues. On one hand, there are protein residues or domains that are much more sensitive to mutations than others. Mutations in some locations, such as the catalytic site of an enzyme, are potentially much more harmful to its function than mutations affecting other domains. In this study there is data available for only 73 mutants for a genome of 10 kb encoding for eleven proteins. Instead of relating mutants and fitness directly, the present approach relates mutants to fitness using proteins and interactions between them as a way to channel those effects and hopefully obtain useful information. Even with a relatively small pool of mutants it is possible to apply the proposed methodology and obtain valid results.

On the other hand, very scarce information is available for particular interactions. One way to include variability in the influence of a particular mutation on each interaction could be carrying out a docking study. Having structural information of two proteins it would be possible to estimate the influence that any change in their sequences has on a possible docking between them. Unfortunately, none of the TEV proteins have been crystallographically determined so far, hence this analysis is not possible yet. Therefore, until no new proteomic information arises, the influence of mutations is spread equally to all the interactions that the mutated protein establishes.

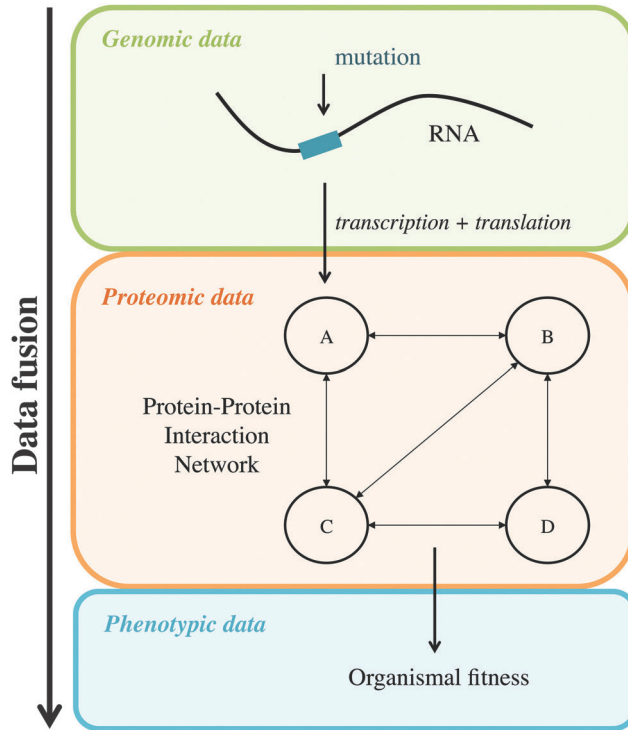


Figure 3.19. Schematic representation of the approach. The aim is to relate the mutations generated on the genome of TEV, their effect on the protein–protein interaction network, and the resulting phenotypic fitness of the virus *in vivo*.

The problem of relating different sources of data has been widely assessed in systems biology using data fusion. Data fusion can be defined as a statistical procedure to analyse simultaneously different sources of complex data sets (Van Mechelen and Smilde 2010). This methodology has been applied to identify genes related to specific diseases (Nie and Yu 2013), to PPINs and gene expression (Aerts et al. 2006) to fuse gene regulatory networks, transcriptional factors and amino acid sequences (Baumbach, Rahmann, and Tauch 2009), for metabolic profiling (Xu, Correa, and Goodacre 2013) and for biomarker search in proteomics (Forshed et al. 2008). One of the most used methods in data fusion (Conesa et al. 2010; Forshed et al. 2008; Lee et al. 2012; Rojas et al. 2004; Xu, Correa, and Goodacre 2013) is Partial Least Squares regression (PLS) (Wold, Sjöström, and Eriksson 2001), which pursues to relate a set of biological descriptors or process variables and a set of biological outputs or quality variables taking advantage of the existing correlations among them.

The aim of the present section is to fuse genomic, proteomic and phenotypic data of potyviruses in a single multivariate model to understand the relationships among the different data sources. This way, the objective is to relate mutated proteins, their effect on the PPIN, and the resulting organismal fitness measured under controlled laboratory conditions. Figure 3.19 shows a scheme of the data fusion. In this case, the mutations and the PPIN are the explanatory variable data blocks, and the fitness measured for each mutant takes the role of the dependent variable. Finally, a set of functional modules of the PPIN is isolated using the PLS modelling. The purpose of this approach is to gain insight into the molecular interactions that occur during the virus infection more than to construct a robust predictive model. Similar grey models have been proposed during the last few years, using exploratory (Folch-Fortuny et al. 2015; González-Martínez et al. 2014) and predictive methods (Ferreira et al. 2011), dealing with metabolic networks. To improve the predictive power of the model more genetic and proteomic information would be needed, such as the analysis of codon usage and, specially, the characterization of the protein structure. Unfortunately, this information is not available at the moment.

3.3.3 Methodology

Amino acid substitution matrix

Describing and measuring the severity of the mutations produced in the TEV genome is essential for applying the data fusion methodology in this work. The PAM2 amino acid substitution matrix is used to quantify the potential severity that a mutation produces in the virus. Although PAM2 is based on evolutionary changes over time, and it is used more often in sequence alignment methods, it is still a valuable and proved source of information regarding the likelihood of amino acid substitutions. It is assumed that if a determined change from a particular amino acid to another one is evolutionarily unlikely it is because such a change is potentially more disturbing to the protein function. Alternatively, evolutionarily common amino acid replacements are assumed to have a minor impact on the protein structure.

The scores in the matrix were used to quantify the effect of the mutations on each of the 73 mutants used in the present study (Equation 3.1). Each mutation gives a value that represents the difference between the substitution of a particular amino acid by itself (meaning no mutation at all) and the new amino acid in the sequence. For instance, mutation PC2 produces an amino acid change between F and C. The matrix establishes a score of nine for the F to F substitution (no change) and -30 for the F to C substitution. The difference (39 in this example) between these values represents how similar (chemically and structurally) both amino acids are. That value was then normalized for all mutations with the maximum possible value

for a change among the 20 amino acids (W to E replacement, with a difference value of 47). Since in the absence of epistatic interactions double mutants are potentially twice as harmful as single mutants, in order to compare all mutants (single and double) a normalizing value $2 \cdot 47 = 94$ was chosen.

Equation 3.1 then gives a value between 0 and 1 that expresses how potentially disturbing is the mutation for the protein (being 0 the most aggressive and 1 the least). This approach is a rough way to translate qualitative (mutations and amino acid changes) into quantitative data. The way this quantitative data is used later would imply that when a particular mutation is given the value of 0 the function of the protein is totally eliminated. However, this is unlikely to happen: even with very severe mutations the proteins may perform their functions with some lesser degree. This approach should be taken as an indication of the direction that the protein function may take. On the other hand, proteins are very complex and heterogeneous structures and therefore some areas of the sequence may be particularly sensible to changes (catalytic sites, docking areas, etc.). Unfortunately, the three-dimensional structure information needed to precisely quantify this severity is not yet available for any TEV protein.

Partial Least Squares regression (PLS)

Partial least squares regression (PLS) is a multivariate projection method commonly applied to model the relationships between a set of \mathbf{X} variables (descriptors or process variables) and a set of \mathbf{Y} variables (output or quality variables) reducing significantly the dimensionality of the initial data set. The PLS model finds a set of latent variables (LVs) that both describe the \mathbf{X} data and predict the \mathbf{Y} data, with the aim of maximising their covariance. In this data fusion section, since the \mathbf{Y} data comprise only a single variable y (fitness), the PLS-1 version of PLS regression is applied. When the number of \mathbf{Y} -variables increases, these variables have to be projected in the same manner as the \mathbf{X} ones. For a more detailed explanation of the PLS methodology see Section 2.3.3.

Cross-validation and Jackknife confidence intervals

Cross-validation (CV) is a resampling technique widely used in statistics and chemometrics (Wold 1978). The aim of CV is to assess the number of relevant components to be extracted in the multivariate model. This procedure groups the observations, in the present study into seven groups, and then fits as many PLS models as groups, leaving each time a single group out. Then, the sum of squares of the differences between the actual fitness values and the predicted ones is used to estimate the predictive ability of the model (Wold, Sjöström, and Eriksson 2001). CV is usually performed one component after another, until the predictive power of the model decreases.

Simultaneously with the CV, the Jackknife confidence intervals for the PLS regression coefficients are computed, at a confidence level of 95%, from all models fitted (Efron 1981). These intervals are built based on the estimated PLS regression coefficients of each round of the CV.

3.3.4 Results and discussion

The following sections explain in detail the data acquisition and the mathematical and statistical modelling. First, the PPIN of Potyvirus is built, based on an exhaustive literature review carried out in Section 3.2. Then, the effects of mutations on the fitness of TEV are measured in the natural host *N. tabacum*. Finally, the effect of the mutations on the PPIN is mathematically modelled, and the three sources of data (mutations, PPINs and fitness) are related using a multivariate statistical projection method.

Protein-protein interaction network reconstruction

All currently available PPIs of *Potyvirus* were gathered as an initial step (Section 3.2). This data was obtained from six different articles published over the last few years. In the original dataset 681 PPIs were tested, 194 PPIs were detected among the 11 viral proteins. Integrating these data from different sources in a common pool required some statistical standardization and pre-processing. To determine which interactions were relevant and an accurate representation of the *Potyvirus* PPIN topology a relevance coefficient was defined. From this analysis, a complete and biologically significant PPIN of *Potyvirus* was built. This network (Figure 3.20) is used here to relate mutations and organismal fitness and to extract information about the biological importance of proteins and their interactions. This figure is analogous to Figure 3.7 with only slight changes. The width of the connections is the same for all interactions since it was already established that these are the relevant interactions. Besides that, the layout of the network is different to facilitate posterior functional module visualization.

Mutation and fitness

The collection of single and double mutants used in this work was reported in Lalić and Elena 2012. Twenty single nucleotide substitution mutants and 53 double mutant genotypes of TEV form the dataset analysed here. The fitness of these mutants had been previously quantified by means of growth assays in the natural host *N. tabacum*. Fitness is a measure that captures the ability of a mutant virus to grow and spread through the plant during an infection cycle relative to the ability of the unmutated wild type virus (Carrasco, Iglesias, and Elena 2007).

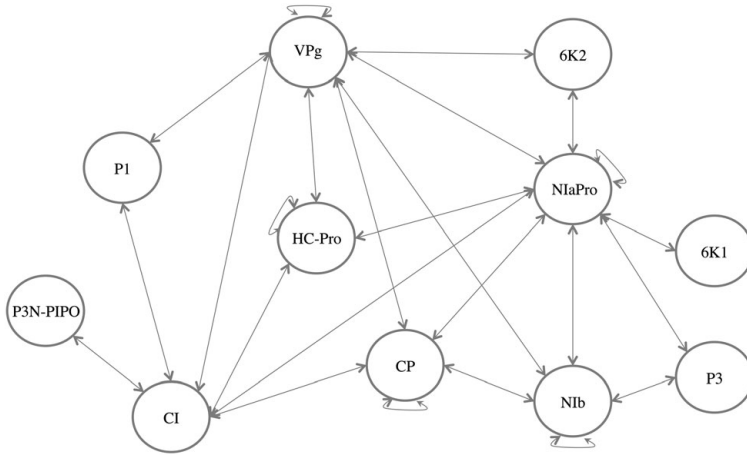


Figure 3.20. PPIN of *Potyvirus*. Eleven proteins (represented as circles) and their 25 detected interactions (represented as double-arrows).

The collection of mutants was generated at random and thus it is somehow irregular, not affecting all TEV proteins: 6K1, CP and P3N–PIPO were not mutated (see Figure 3.21). Moreover, some proteins like P1 and VPg were mutated more times than others such as 6K2, CI and NIb. Although a more complete collection of mutants would be very useful to further increase the accuracy of these findings, the collection of 73 mutants used for this study is a fair representation of the TEV genome and its 11 proteins.

The mutant collection has some features that make it an interesting and appropriate starting point for the data fusion. Six of the 20 single mutants correspond to synonymous mutations. In other words, the nucleotide substitution does not translate in an amino acid replacement in the protein sequence. In spite of being synonymous, some of these mutations had a significant effect on fitness (Carrasco, Iglesia, and Elena 2007) due to RNA stability, enhanced RNA silencing responses or improved translational efficiency, among other possibilities. Although these mutations have no effect on the protein sequence and thus no predictable effect on the PPIN either, they represent a natural source of fitness variability that is taken into account in our results. Other particularity of the data is that lethal mutations exist, meaning those that render zero fitness for the virus bearing them, i.e. these mutations do not allow the virus to survive and grow. Nine of the double mutations are lethal. These mutations are excluded from the analysis because, if included, they will mask all the variability of non-lethal mutations varying fitness in a discrete manner.

Mutation	Protein	Type	#of mutants
PC2	P1	Nonsynonymous	2
PC6	P1	Nonsynonymous	7
PC7	P1	Nonsynonymous	5
PC12	P1	Nonsynonymous	4
PC19	HC-Pro	Synonymous	10
PC22	HC-Pro	Nonsynonymous	6
PC26	HC-Pro	Synonymous	4
PC40	P3	Synonymous	5
PC41	P3	Nonsynonymous	4
PC44	P3	Synonymous	5
PC49	CI	Nonsynonymous	8
PC60	CI	Synonymous	3
PC63	6K2	Nonsynonymous	10
PC67	VPg	Nonsynonymous	4
PC69	VPg	Nonsynonymous	13
PC70	VPg	Nonsynonymous	5
PC72	VPg	Nonsynonymous	3
PC76	NIaPro	Synonymous	8
PC83	NIb	Nonsynonymous	10
PC95	NIb	Nonsynonymous	10

Figure 3.21. Mutations experimentally generated on the genome of TEV.

The effect of the mutations on the proteins can be quantified using different information. The most precise way to do it would be using structural information. Having the resolved structure of the viral proteins, it is possible to change a particular amino acid and observe how that change affects some structural variables. Some mutations would increase the stability of the protein and some would decrease it, defining a magnitude for the mutation. Unfortunately, as it was mentioned before, none of the TEV proteins has been crystallized making this approach infeasible. Another approach consists of assuming that biochemically different amino acids would induce more severe perturbations in the structure conformation. This way a mutation changing an amino acid for another very similar would produce only a slight structural disturbance and consequently only a minor protein activity variation. An extremely different amino acid would produce a much more dramatic change in the protein activity.

To represent the biochemical similarity or the distance between the original amino acid in the sequence and the new one produced by the mutation an empirical amino acid substitution matrix was used. These matrices describe the rate at which one amino acid changes to any other over time. These matrices are commonly used in the field of protein sequence alignment, calculating the probability that a particular amino acid changes over time to a new one through mutation. The underlying idea is that an amino acid substitution is more likely to survive to the filter of selection if it is similar to the original amino acid than if it is physically very different. Similar amino acids would then preserve a similar folding structure and

activity for the protein. Thus, the information contained in the entries of these matrices was used to quantify the magnitude of each mutation. Since the collection of mutants available is composed of single and double nucleotide mutations it seemed to be appropriate to use the Point Accepted Mutation (PAM) matrix (Dayhoff and Schwartz 1978) to compute the distances generated by the mutations. These matrices were developed using observed mutations in closely related proteins. Large numbers in the PAM matrix denote substitutions very likely to be removed by purifying natural selection, thus unlikely to persist in the long-term evolutionary time. Since the mutants used for this study have almost identical sequences it seemed to be more precise to use a low number PAM matrix. For this, we decided to use the PAM2 matrix (Dayhoff and Schwartz 1978) to quantify the effect of the amino acid replacement on viral proteins. It was assumed that mutations with high PAM2 values would induce a strong disruption in the protein structure and, therefore, would have a high probability to negatively affect its biological function.

Mathematical modelling

Once the distance produced by each mutation is computed from the PAM2 matrix, the effect of the mutation on the PPIN has to be modelled. However, as commented previously, some mutations result in a zero distance (synonymous mutations). Since these mutations have no effect on the network, they may directly affect fitness without crossing the PPIN. The distances generated by all mutations are provided in Annex 7.1, jointly with the fitness measured for all mutants.

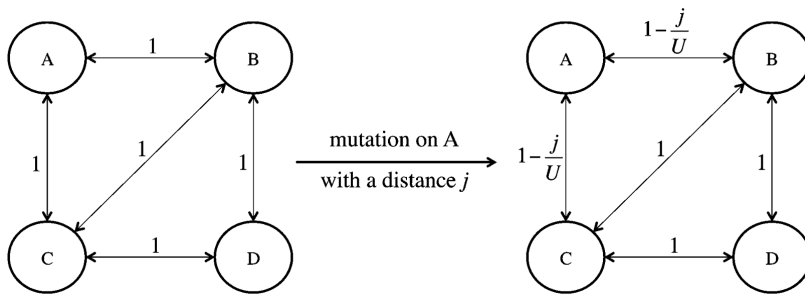


Figure 3.22. Small example of the mutation modelling. Initially, all detected interactions between proteins have a value 1. When a mutation is performed on protein A with distance j , the intensity of the PPIs $A \leftrightarrow B$ and $A \leftrightarrow C$ is lowered by j/U .

The distance registered for all non-synonymous mutations is modelled as follows. The distance generated by an amino acid replacement, which affects a particular protein, weakens the existing interactions between the influenced one and its first-step neighbours in the PPIN. Figure 3.22 shows a small example of this modelling

concept. If a mutation is produced on protein A, with a registered distance j , the interactions relating A to its neighbours, B and C, are weakened as follows:

$$\mathbf{A} \leftrightarrow \mathbf{B} = \mathbf{A} \leftrightarrow \mathbf{C} = 1 - \frac{j}{U} \quad (3.1)$$

where $\mathbf{A} \leftrightarrow \mathbf{B}$ and $\mathbf{A} \leftrightarrow \mathbf{C}$ mark the interaction between A and B, and A and C, respectively, and U is the reference value, which refers to the maximum value in the entire PAM2 matrix. In other words, the maximum distance an amino acid change defines.

It is worth noting that the distance produced in the protein is a measure of how different is the protein after the mutation. Then, this distance is translated into a strength/intensity measure in the network between the protein and its first-step neighbours. So no distance is being considered between different proteins in the PPIN.

The different data sources presented in this study must be combined properly to be analysed using a latent structure method. Since PLS, in its original form, works with two-way data matrices, the information collected on the previous subsections must be arranged in such a way that each individual (i.e. experiment) is represented by rows, and the different types of variables (i.e. mutations, interactions and fitness) by columns. So three data matrices are built: the mutation matrix \mathbf{M} has the 20 different mutations as variables, the interaction matrix \mathbf{I} has the intensity in each of the 25 interactions by columns, and the vector y has the fitness registered for each individual. All matrices have 64 rows, corresponding to the non-lethal mutants. Figure 3.23 presents an example of the matrices defined above, following the small PPIN taken as an example in Figure 3.22. In this case three individuals are considered, e.g. in Exp1 a non-synonymous mutation is performed on protein A, producing a distance j and registering a fitness y_1 . Note that on Exp3 a synonymous mutation on protein A is performed, therefore, it has no effect on I, i.e. neither $\mathbf{A} \leftrightarrow \mathbf{B}$ nor $\mathbf{A} \leftrightarrow \mathbf{C}$ are lowered in this case.

Statistical modelling

The data matrices built in the previous subsection can be analysed using different statistical techniques. Considering only mutations and fitness, a design of experiments (DOE) can be performed, but this approach presents some drawbacks here. There are 20 different mutations performed individually or two-by-two, across the original 73 individuals. A model including only mutations and fitness could be fitted using penalized regression, such as Lasso (Rasmussen and Bro 2012) or Elastic Net (Zou and Hastie 2005), to prevent rank deficiency problems. However, it is known that the PPIs affect the fitness, so in the previous approach this effect is not taken into account.

	MutA	SMutA	MutD	A~B	A~C	B~C	B~D	C~D	Fitness
Exp1	1	0	0	$1-\frac{j}{U}$	$1-\frac{j}{U}$	1	1	1	y_1
Exp2	0	0	1	1	1	1	$1-\frac{k}{U}$	$1-\frac{k}{U}$	y_2
Exp3	0	1	0	1	1	1	1	1	y_3
	M			I					y

Figure 3.23. Data matrices. Matrices **M**, **I** and vector y have the information from the mutations, interactions and fitness, respectively. Three examples are presented. On Exp1 a non-synonymous mutation is performed on A, with distance j , and fitness y_1 . A non-synonymous mutation on D is performed in Exp2, producing a distance k and fitness y_2 . On Exp3 a synonymous mutation is performed in A, producing no distance (and no effect on **I**), and a fitness y_3 . The colours correspond to the data sources described in Figure 3.19.

The other possible approach consists of relating all the interaction strengths/intensities to the fitness, using classical linear regression. The problem is that the mutations are performed on different proteins affecting different interactions, which may not be comparable in this model.

In this work, a PLS regression is applied to fuse the genomic, proteomic and phenotypic data in a single multivariate model, the first two sources being the explanatory variable blocks and the phenotypic fitness of the dependent variable. Using a PLS model, the available data are compressed into a set of latent variables that relates mutations and interactions to the observed fitness. This allows to clarify which mutations, and also which sections of the network, increase or decrease the fitness of TEV.

The different data sources, detailed in previous subsections, have to be pre-processed in order to obtain meaningful components in the PLS model. In the present case the dataset is directly autoscaled, i.e. the variables are centred and divided by their standard deviation to have mean 0 and standard deviation 1.

Regarding the statistical modelling, the PLS model can be strongly (and harmfully) affected by some of the mutants compiled for the present study. As commented above, lethal mutations decrease the fitness straight to zero, while for the non-lethal mutations it oscillates in a small range around the fitness of the wild-type virus. The inclusion of the lethal ones in the study will force the model to explain only the variation between the lethal and non-lethal, pointing simply to the mutations that have been lethal. To avoid this spurious result, and explain equally the positive and negative effect of the mutations and interactions on the fitness of TEV, these lethal genotypes have been removed from the datasets. This

relates directly to the way in which mutation severity is quantified. PAM matrices are constructed assuming non-lethal scenarios. Even the most extreme amino acid substitution is quantified as a prerequisite of biological success. Therefore it is sensible to exclude the lethal mutations from the main analysis, since the benchmark chosen to represent mutation magnitude excludes them originally.

Once the data are prepared for the analysis, a PLS model is fitted using the software *ProSensus ProMV*. To decide how many components extract from the data, the cross-validation criterion using seven groups is selected (more details in Section 3.3.3). The aim of this criterion is to choose the number of components that offer the highest predictive power.

First, a PLS model including all variables is fitted. Later on, a reduced PLS model is obtained by deleting some mutations and interactions that have a very low influence on the fitness. These mutations are PC12, PC67, PC69, and PC72. The PPIs deleted are: HC-Pro \leftrightarrow VPg, VPg \leftrightarrow VPg, VPg \leftrightarrow NIaPro and VPg \leftrightarrow CP. Basically, these variables have a non statistically significant PLS regression coefficient in the first PLS model (95% of confidence level).

Component	R^2X cumulative (%)	R^2y cumulative (%)	Q^2 cumulative (%)
1	11.8	57.6	39.5
2	23.4	70.0	46.7
3	30.1	78.3	56.7

Figure 3.24. PLS regression results (reduced model). Cumulative variances in \mathbf{X} and y explained by the model ($R^2\mathbf{M}$ and R^2y , respectively) and predictive power of the model (Q^2).

Figure 3.24 shows the results of the reduced PLS model. For the analysis, matrices \mathbf{M} and \mathbf{I} are merged in a single matrix \mathbf{X} , including all the variables collected in the study. With a 3-component model, 30.1% of the variability in \mathbf{X} explains 78.3% of variance in the fitness, y , with a predictive ability of 56.7%. It is worth noting that although network topology is definitely a major contributor to the variance of the fitness, there are some other factors that are not included in this particular approach, harming the predictive power of the PLS model. RNA structure stability and codon usage bias are two clear examples of important contributors to fitness that are not included in the analysis.

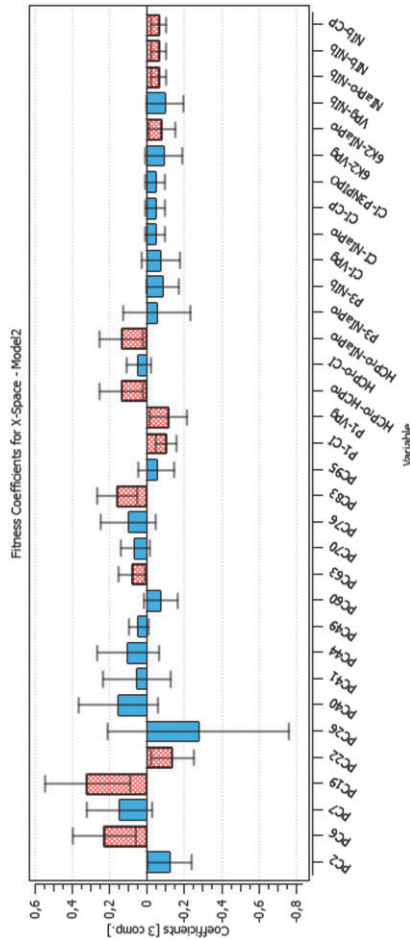


Figure 3.25. PLS regression coefficients. For each regression coefficient, the 95% Jackknife confidence interval is shown. The statistically significant variables are plotted as red bars. The mutations with a relevant effect on the fitness are PC6, PC19, PC22, PC63 and PC83. The significant PPIs are: P1 \leftrightarrow CI, P1 \leftrightarrow VPg, HC-Pro \leftrightarrow HC-Pro, HC-Pro \leftrightarrow NIaPro, 6K2 \leftrightarrow NIaPro, NIaPro \leftrightarrow NIB, NIB \leftrightarrow NIB, and NIB \leftrightarrow CP.

Figure 3.25 shows the PLS regression coefficients of the variables in the dataset. The red bars mark the statistically significant PPIs and mutations. The relevant ones are chosen based on the 95% Jackknife confidence intervals computed for their corresponding PLS regression coefficient (Efron 1981). In this way, when the interval does not include zero, the variable has a relevant effect on the fitness, either positive or negative, with a 95% confidence level.

PC22 has a statistically significant negative effect on the resulting fitness of TEV; i.e. when this mutation is generated in the genome, the fitness lowers its value (see Figure 3.25). PC6, PC19, PC63, and PC83 also affect fitness, but in a positive direction. The fitness increases when either of these mutations is present in TEV genome. It is worth noting that a PLS model using only the mutations and the fitness identifies basically the same relevant mutations as the combined mutations–interactions model, but with less explained variance and predictive power in fitness (70.1% and 47.0%, respectively).

Mutation	Protein affected	Interactions
PC6 ⁺	P1	P1 ↔ CI ⁻ , P1 ↔ VPg ⁻
PC63 ⁺	6K2	6K2 ↔ NIaPro ⁻
PC83 ⁺	NIb	NIb ↔ NIaPro ⁻ , NIb ↔ NIb ⁻ , NIb ↔ CP ⁻
PC22 ⁻	HC-Pro	HC-Pro ↔ HC-Pro ⁺ , HC-Pro ↔ NIaPro ⁺
PC19 ⁺	HC-Pro	(Synonymous mutation)

Figure 3.26. Statistically significant explanatory variables. Super-indices mark the positive/negative effect of the variable on the fitness.

The PPIs P1 ↔ CI, P1 ↔ VPg, 6K2 ↔ NIaPro, NIaPro ↔ NIb, NIb ↔ NIb, and NIb ↔ CP have a statistically significant negative effect on the fitness (see Figure 3.25). Bearing in mind the mathematical modelling, when a mutation is performed, the corresponding interactions lower their values. Therefore, the lower is the value of the interaction, the higher is the fitness computed. Alternatively, HC-Pro ↔ HC-Pro and HC-Pro ↔ NIaPro have a statistically significant positive effect on the fitness, i.e. the lower is the value of the interaction, the lower is the fitness computed. All the statistically significant variables, mutations and PPIs, are summarized in Figure 3.26. This information will be valuable to define the functional modules in the next subsection.

Functional modules

On the previous subsection, the explanatory variables, PPIs and mutations with a statistically significant effect on the organismal fitness, are identified among the rest of the variables registered. In order to finally establish the relationships among the three data sources, following the scheme proposed in Section 3.3.2 (see Figure 3.19), the genomic–proteomic–phenotypic effect must be explained using the information in Figure 3.26. If the relevant mutations and PPIs are represented on the original PPIN (see Fig. 6) some interesting conclusions can be drawn.

Mutation PC6, affecting protein P1, is positively correlated with TEV fitness. At the same time, interactions P1 ↔ VPg and P1 ↔ CI are also relevant in the PLS model, being negatively correlated with viral fitness. These mutation–fitness effects and interaction–fitness effects represent a unified mutation–interaction fitness

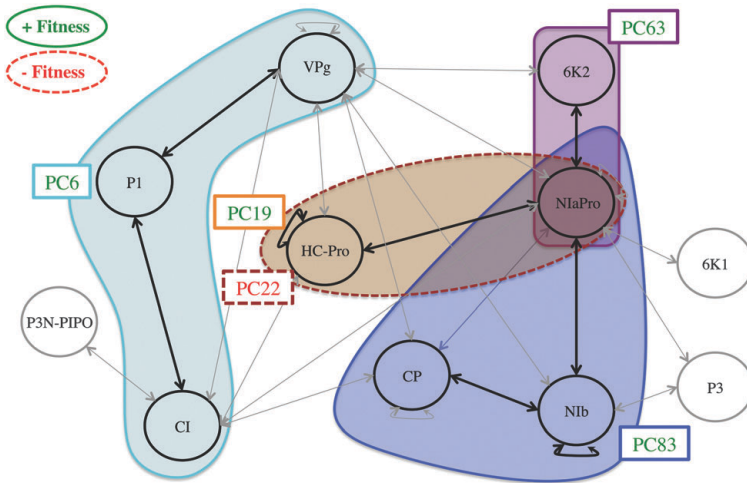


Figure 3.27. Functional modules of TEV. The cyan module is activated via mutation PC6 in protein P1 and affects proteins CI and VPg. The violet module is activated by mutation PC63 on protein 6K2 and affects protein NlaPro. The blue module is activated via mutation PC83 in protein Nlb and affects Nlb, CP and NlaPro. The brown module is activated via mutation PC22 in protein HC-Pro and affects HC-Pro and NlaPro. The synonymous mutation PC19 is performed in HC-Pro. Mutation PC22 has a negative effect on the fitness while the rest of the mutations have a positive effect.

effect. Figure 3.28 shows a scheme of this process: when PC6 is generated on P1, the interactions with its neighbours VPg and CI lower their values, and the fitness is increased as a result. A cyan ellipse in Figure 3.27 rounds this functional module.

This behaviour is also observed with the blue and violet modules (see Figure 3.27). The former one is activated via mutation PC83 on protein Nlb, and affects Nlb, NlaPro and CP. The latter starts with mutation PC63 on 6K2, affecting only its relationship with NlaPro. When these sections are activated, the fitness increases. In this way, Figure 3.28 can also represent the behaviour observed in these modules, replacing the mutation and interaction names.

Two mutations affecting HC-Pro have a statistically significant effect. When mutation PC22 is generated, the PPIs HC-Pro \leftrightarrow HC-Pro and HC-Pro \leftrightarrow NlaPro are affected (brown module in Figure 3.27) and the phenotypic fitness decreases. Alternatively, PC19 is positively correlated with the fitness: when it is introduced in HC-Pro, the fitness increases significantly. Both mutations are compatible with the mathematical modelling because PC19 is a synonymous mutation, and therefore it has no effect on the PPIN network. Figure 3.29 shows the different effects related

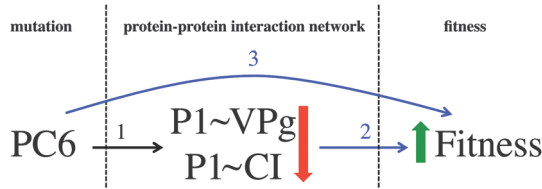


Figure 3.28. Diagram of mutation – PPIN – fitness effects. The mathematical modelling implies that, when mutation PC6 is applied, the protein–protein interactions P1↔VPg and P1↔CI lower their values (1). The data fusion results indicate that: (i) when PC6 is applied the fitness increases (2), and (ii) when the previous interactions lower their values, the fitness increases (3). The mathematical and statistical modelling are describing the effect of the mutation on the protein–protein interaction network and the effect of the network on the fitness.

to HC-Pro. This modelling would be infeasible if PC19 were a non-synonymous mutation. In this hypothetical case, since it would affect HC-Pro↔HC-Pro and HC-Pro↔NIaPro, it would be incoherent that the mutation increases the fitness and its associated interactions lower its value at the same time.

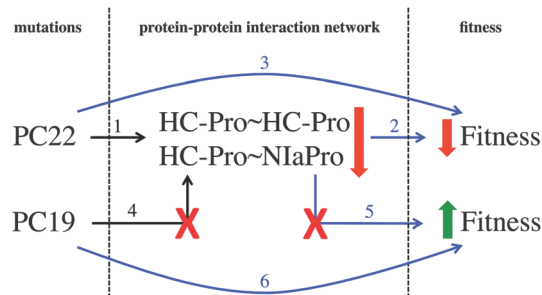


Figure 3.29. Diagram of mutations – PPIN – fitness effects in the case of multifunctional protein HC-Pro. The mathematical modelling implies that, when mutation PC22 generated in HC-Pro, the protein–protein interactions HC-Pro↔HC-Pro and HC-Pro↔NIaPro lower their values (1). The data fusion results indicate that: (i) when PC22 is introduced the fitness decreases (2), and (ii) when the previous interactions lower their values, the fitness decreases (3). When PC19 is generated, the fitness increases (6). All the effects described in this diagram are coherent among them because PC19 is a synonymous mutation; therefore it has no effect on the protein–protein interaction network (4 and 5).

Two comments are here in due regarding the functional modules (Figure 3.27). Firstly, if an interaction between two proteins is included in a module (e.g. P1↔CI) it implies that the effect of the interaction on the fitness is statistically significant, considering that it can be activated by non-synonymous mutations performed on

both proteins (i.e. P1 and CI). However, the effect is stronger when the mutation defining the module is performed (i.e. PC6 on P1), since the mutation is activating other relevant interactions (i.e. P1BVPg). Secondly, if an interaction activated by a key mutation is not included in the corresponding module (i.e. interaction 6K2 \leftrightarrow VPg, activated via mutation PC63) it implies that the effect of the interaction, considering that it can be activated by non-synonymous mutations performed on both proteins (i.e. 6K2 and VPg), is not statistically significant.

High-level and mid-level data fusion procedures obtain separate models and extract relevant features of each data matrix, respectively, to combine them in a fused model to predict the biological output (Buydens 2013). In this study, however, it was decided to apply a low-level data fusion, concatenating row-wise, matrices **M** and **I** because the mathematical modelling applied here establishes a direct relationship between the mutations and the PPIN, so the joint analysis of both matrices in a single PLS model leads to the identification of functional modules exploiting not only the mathematical modelling but also the topological interactions being affected by the different mutations.

3.4 Conclusions

The main conclusions of this chapter revolve around two central topics. First, the topology of the *Potyvirus* PPIN and the relationship between virus and plant network was analysed in detail (Section 3.2). Secondly, once the topology of the PPIN of the virus was defined, it was used as bridge between genetic (mutants) and phenotypic (organismal fitness) information. This second analysis was carried out in Section 3.3.

Topological properties of the potyvirus PPIN were studied in great detail. Data was collected from different sources and was processed and integrated in the intraviral network representation GLIN. The findings confirm the idea that intraviral network of potyvirus is highly connected and core interactions involve proteins NIaPro, VPg, CI, CP, and NIb. The four topological parameters studied seem to depend on the protein degree. Moreover, the cumulative distributions of these parameters and the degree increase in a quasi-linear way. BiFC and Y2H offer similar results and detect the most common interactions. Y2H data led to affirm that interactions with lower intensity can be as vital to virus development as the more intense ones.

In the study of host-virus interaction, VHPIN results an accurate representation of the plant-host interactome. Proteins P3 and VPg focus their effect in only one hub while HC-Pro diversifies its effect among several proteins through direct interactions. Viral proteins differ in the efficiency in which their perturbations are transmitted throughout *A. thaliana* HHPIN. Proteins P3 and VPg are the fastest

to propagate their effects while proteins CP and P1 are the slowest ones. The similarity among viral proteins in their patterns of perturbation transmission was analysed using the evolution of the Simpson index (SI) along propagation steps. This analysis highlighted common patterns of action between NIaPro, NIb, VPg, and P3.

This study opens new research avenues. This topology can be used as a base for a much more in-depth analysis of virus development with the addition of biological meaningful measures such as virus growth or fitness (see Section 3.3). On the other hand, the VHPIN analysis can be further explored using more complex metrics, graph kernels or integrating more biological information available such as sub-cellular localization or biological function. Additionally, when more studies start to use the BiFC method and the pool of reliable intravirus interactions tested and detected increases, the topology here determined can be slightly modified to meet the new data.

The PLS modelling applied to genomic, proteomic and phenotypic data sets allows for the integration of the mutations performed on viral proteins, their effects on the PPIN, and their influences on the organismal fitness experimentally quantified. In this way, three biological functional modules affecting the PPIN and influencing the fitness positively have been detected. Two additional modules are identified affecting a single protein. One influences the protein network, being negatively correlated with the organismal fitness. The other one has a positive effect on the fitness without affecting the PPIN. This implies that different mutations affecting the same protein induce different behaviours in the activity of the PPIN and the resulting fitness.

Classical clustering algorithms usually work with a standalone version of the network, detecting dense sections of the topology based solely on its interaction intensities (or basically on node degrees). In comparison to traditional clustering, the presented methodology allows working with different sources of information, combining them to squeeze the data and extract the relevant information. With this data fusion, (i) the mutations are related to topological changes in the network and their subsequent influence on the fitness, and (ii) the mutations not affecting the network can also be related to the fitness.

Data fusion reveals as a very powerful tool to analyse and relate different types of biological information. The larger the network and the collection of mutants, the more precise its findings are. The present study, analysing a relatively small PPIN (11 nodes and 25 interactions) and a small number of combinations of mutations (64 out of the 210 possible ones), results in a quite high-explained variability. However, there are intrinsic biological considerations that limit the scope of the method. These considerations, such as RNA stability, efficiency inducing the antiviral RNAi response of the plant and codon usage bias may be included in the model as additional sources of variability but much more data would be needed.

Besides this, further work of interest includes testing the proposed methodology with a larger dataset containing more mutants, and extending the analysis to larger PPINs, in order to build multivariate models with a higher predictive power, exploiting the features of the projection to latent structure methods.

Chapter 4

Constraint-based approaches for metabolic analysis of *E. coli*

- How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?

Sherlock Holmes, in Doyle, Arthur Conan. "The Sign of the Four" (Chap. 6, p. 111), 1890.

Part of the contents of this chapter appeared in the following journal articles:

- Morales, Y. et al. (2016). "PFA toolbox: a MATLAB tool for Metabolic Flux Analysis". In *BMC Systems Biology* 10, p. 46.

4.1 Introduction

This chapter revolves around three different methods used to reliably estimate metabolic fluxes and cellular growth in *E. coli* using constraint-based models. The combination of organisms studied, type of network analysed and mathematical tools used is shown in Figure 4.1.

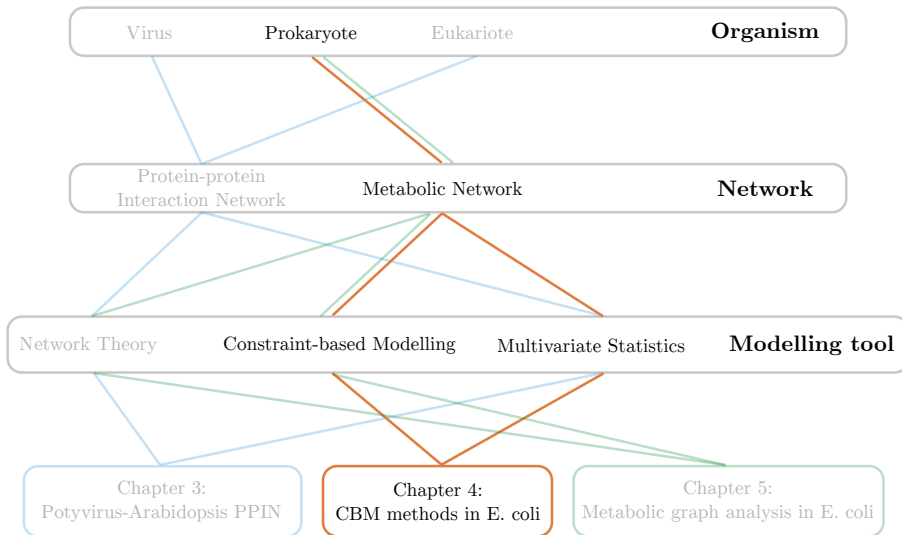


Figure 4.1. Chapter roadmap.

The first two sections (Summary and Background) address the main objectives of the work carried out in this chapter. The third and fourth sections (Materials and Methodology) explain in detail the elements necessary to achieve the proposed goals. First, in the Materials section, the dataset gathered from existing bibliography and the *E. coli* constraint-based model used are examined thoroughly. In Methodology, the methods to interrogate the model are mathematically described. Those methods are Flux Balance Analysis (FBA), Flux Variability Analysis (FVA) and Metabolic Flux Analysis (MFA). A complete mathematical description of constraint-based models is shown in Section 2.3.2.

The fourth section (Results and discussion) shows the bulk of the work done. First, the flux distributions are shown for every method. They are compared and analysed. Particular scenarios and metabolic pathways are selected to highlight the most remarkable differences and features. Later on, several simple statistics are obtained for each solution to facilitate the comparison. Finally, the results of the scenarios variability analysis are shown to address the stability and robustness

of the comparison among the different methods. The last section (Conclusions) draws the main conclusions from this chapter.

4.2 Summary

The three most popular and most widely used constraint-based methods to reliably estimate metabolic fluxes and cellular growth are used here. Those methods are Flux Balance Analysis (FBA), Flux Variability Analysis (FVA) and Metabolic Flux Analysis (MFA). Combinations of each method are tested using a large collection of scenarios (a total of 46) of *Escherichia coli* gathered from existing bibliography. In each scenario several extracellular fluxes are measured. The results for each method are compared and their respective main characteristics, features and assumptions are tested. A relatively small but popular constraint-based model of *E. coli* is used to run the different methods.

The work done and presented in this Chapter 4 achieved many different results. The level of agreement between one the most popular *E. coli* constraint-based models and a large collection of experimental scenarios was checked. FBA, FVA and MFA were applied to the *E. coli* constraint-based model to obtain the flux distributions for each scenario. The solutions obtained by the different methods were compared. Their similarity, validity and size was analysed and their capability to predict growth was measured. As a consequence of this comparison, the validity and predictive capability of the assumptions behind each method was tested. Finally, the robustness of the conclusions was tested across the complete experimental dataset.

4.3 Background

Life is an incredibly complex phenomenon. Even the smallest cell is formed by a enormous number of compounds constantly interacting with each other in many different ways. The macroscopic result of those interactions is the behaviour known as life. Mathematical models are an extremely useful tool to simplify and understand those interactions and to improve our biological knowledge. Biomolecular models help us describe, understand and predict the cell's characteristics and functions.

Biomolecular models try to organize and integrate different sources of information (usually coming from different experiments) into a coherent entity. When models grow in scale they start to show the emerging properties consequence of the whole cell and not only the particular properties of individual elements. Ideally, modelling guides experimentation by proposing hypothesis that can be tested. Better models can be built from those experiments producing new and more precise hy-

pothesis. This cycle is repeated over and over increasing our biological knowledge. In practical terms, cellular models already have a big impact in disciplines and fields such as industrial biotechnology, food industry and metabolic engineering.

When thinking about intra-cell interactions one often thinks in three large branches that, although definitely interconnected, are usually viewed and studied separately. Those are: regulation, signalling and metabolism, being the latter the main focus of this discussion. Metabolism represents the set of chemical reactions and transformations that are necessary to maintain cells alive. Very simply, it converts raw materials (a carbon source, oxygen, inorganic ions, etc.) into energy (usually in the form of phosphated substances) and a group of compounds that cell needs to produce biological structures.

Metabolism is a large cellular subsystem that is frequently understood and thought of as a network that links together thousands of chemical reactions and compounds. Nodes correspond to metabolites and the connections between them are the metabolic reactions that produce or consume them. These reactions can be either intracellular if they occur completely within the cell or exchange ones if there is transport of some substance between the cell and the exterior (i.e. uptake of substrates or secretion of by-products). The flux through the different metabolic reactions can be very different and, moreover, it can change according to external or internal conditions. This complete flux distribution of the metabolic reactions defines a metabolic state of the cell and sets a particular phenotype and behaviour that the cell shows.

However, these metabolic networks are difficult to properly model. Unfortunately, there is still not enough information about those reactions and the enzymes that catalyse them to construct a comprehensive and cell-level model of the cellular dynamics. Classical stoichiometric models, which do not take into account the dynamics of the intracellular metabolic reactions and assume steady-state concentration for every metabolite, are a viable alternative. This steady-state is a strong assumption that, nevertheless, has proven to be extremely useful for some practical applications. Constraint-based models are a particular branch of these models that have seen a spectacular success in the last two decades. Along with the steady-state mass balances for metabolites, the cell is subject of other constraints that limit significantly its behaviour. Imposing these constraints determines which metabolic states can and cannot be reached by the cell. These constraints are usually defined as a series of bounds (upper and lower bounds) for particular reaction flow rates that are known to occur under given circumstances. Constricting the cell this way, these models offer us a group of metabolic states that are possible.

A lot of methodologies use constraint-based models to tackle many different problems in many different organisms. Some methods are used to analyse properties of the studied organisms (extreme pathways, elementary modes, communities of reactions), to simulate particular genetic situations (i.e. gene knock-outs) or to

estimate the global metabolic state under certain conditions. As it was stated in Sections 4.1 and 4.3 this chapter focuses on FBA, FVA and MFA applied to a compact *E. coli* model.

4.4 Materials

In this subsection the materials used to carry out the work presented in this chapter are enumerated and described in detail. Since this is a purely computational work, the necessary experimental data required to achieve the proposed goal was gathered in an extensive search across the publicly available literature. As it was stated in Section 4.2 the main purpose of this chapter is to compare the steady-state metabolic flux distributions of *E. coli* obtained with several constraint-based methods. In order to do so, two main materials of sources of information are needed:

- A constraint-based model of *E. coli* in which the different methods will be applied.
- A collection of experimental data for *E. coli*. This collection of scenarios serve as the input data for the methods to run the model. They are usually composed by a series of extracellular flux measurement that constraint the model as explained in Section 2.3.2.

To interrogate the model and to obtain the desired flux distributions, three different methods are used: FBA, FVA and MFA. These methods are described in the second part of the present subsection.

4.4.1 *E. coli* metabolic model

The core *E. coli* model is a relatively small constraint-based model of the central carbon metabolism of *E. coli*. It was first published in Orth, Fleming, and Palsson 2010. Since then it has been extensively used because of its reduced size and yet good predicting capabilities under the most usual *E. coli* scenarios. This model includes enough reactions and pathways to enable interesting and engaging simulations, but it is also small enough to make the interpretation of the results straight-forward. It is formed by 72 metabolites and 95 reactions and it includes the main carbon metabolic pathways such as glycolysis, TCA cycle or pentose phosphate pathway.

***E. coli* constraint-based models**

The core *E. coli* model is based on the first stoichiometric reconstruction of *E. coli* fueling pathways (Varma and Palsson 1993a). Progressively, reconstructions grew in scale during the last years of the last century until reaching the first *E. coli* K-12 MG1655 genome-scale metabolic model, the *iJE660* (Edwards and Palsson 2000). This network was developed through a long process of searching the literature and the databases to make sure the stoichiometry and cofactor participation was as precise as possible. A new iteration of this network was published under the tag of *iJR904* in Reed et al. 2003. It had an extended scope, adding pathways for uptake of alternative carbon sources and a more detailed electron transport chain. Additionally, hundreds of new metabolites were included and gene-protein-reaction (GPR) associations were added. These GPR relationships are not strictly necessary to run the model are not used in this thesis. However, they give more context to the model than merely the pure reactions and make easier to simulate knock-out experiments.

In the next large update, the *iAF1260* (Feist et al. 2007) the size of the network kept growing, now adding reactions in charge of the synthesis of the cell wall building blocks. Furthermore, all reactions were assigned to cytoplasm, periplasm and extracellular space, hinting a future detailed cellular compartmentalization description. Thermodynamic information was used to set the lower bound of irreversible reactions. *iAF1260* consists of 2077 reactions, 1039 metabolites and 1260 genes. The core *E. coli* model used in this thesis is a reduced and comprised version of the *iAF1260* iteration. Since this version is already very detailed, the last *E. coli* reconstruction, the *iJO1366* (Orth, Fleming, and Palsson 2010) included just a few newly characterized genes and reactions. Many of the gaps present in *iAF1260* were filled in this last version, becoming the most complete *E. coli* reconstruction to date. It includes 1366 genes and 2251 reactions.

Main features of the core *E. coli* metabolic model

The reactions and pathways in the core *E. coli* model were chosen to represent the best studied metabolic pathways of *E. coli*. These pathways are common subjects of textbook chapters and should be very familiar to readers with a basic understanding of biochemistry. The reactions were taken from the *iAF1260* reconstruction as much as possible, although some of them were omitted or grouped together for simplifying purposes. The electron transport chain system was strongly simplified so that every reaction represents several consecutive steps in the chain while still being understandable. The present CBM contains 72 metabolites and 95 reactions. There are 20 extracellular and 52 intracellular metabolites, with a total of 54 unique metabolites (most extracellular compounds are present within the cell as well). There are 20 exchange reactions, one for each extracellular metabolite

as well as 25 transport reactions, 49 proper metabolic reactions and one biomass reaction. The complete network of the model is shown in Figure 4.2.

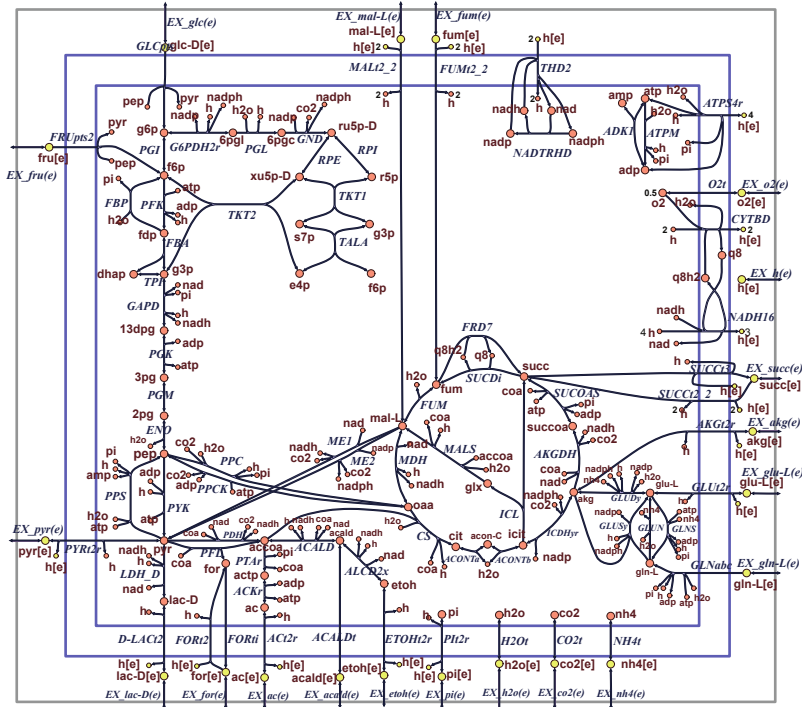


Figure 4.2. Reaction map of the core *E. coli* metabolic model (from Orth, Fleming, and Palsson 2010). The exterior grey box represents the boundary between the model and the environment, which is the ultimate source of substrates and sink for secreted waste metabolites. The blue boxes represent the cytoplasmic membrane (the periplasmic space disregarded in this model). Cytosolic metabolites are represented by orange circles and extracellular metabolites by yellow circles. The model syntax is pretty straightforward. Metabolites and reactions are given both full names and abbreviations, shown in the figure. Metabolite abbreviations are lowercase, and extracellular compounds are denoted with the suffix '[e]'. All metabolites that are not denoted as extracellular are cytosolic. Reaction abbreviations are uppercase and italicized. There are several usual suffixes used in the reaction abbreviations, i.e. 'abc' (ATP-Binding Cassette transporter), 'i' (irreversible), 'r' (reversible), and 't' (transport). Most reactions are named after the enzymes that catalyze them.

Some reactions are assumed to be irreversible when thermodynamic considerations are taken into account. In short, if the change in Gibbs energy in a biochemical reaction is highly negative, then the reaction is assumed to flow always in the

forward direction. On the other hand, the charge on each reaction is included and was determined by using pK_a of each metabolite at a pH of 7.2 (Feist et al. 2007). A total of eleven subsystems are included in the model. These subsystems are groups of reactions grouped together. Some of these subsystems are classical biochemical pathways such as glycolysis, pentose phosphate pathway or the citric acid cycle. Others are simply groups of reactions that carry out similar functions, i.e. exchange or transport reactions. All the reactions divided by subsystems are shown in Figure 4.3.

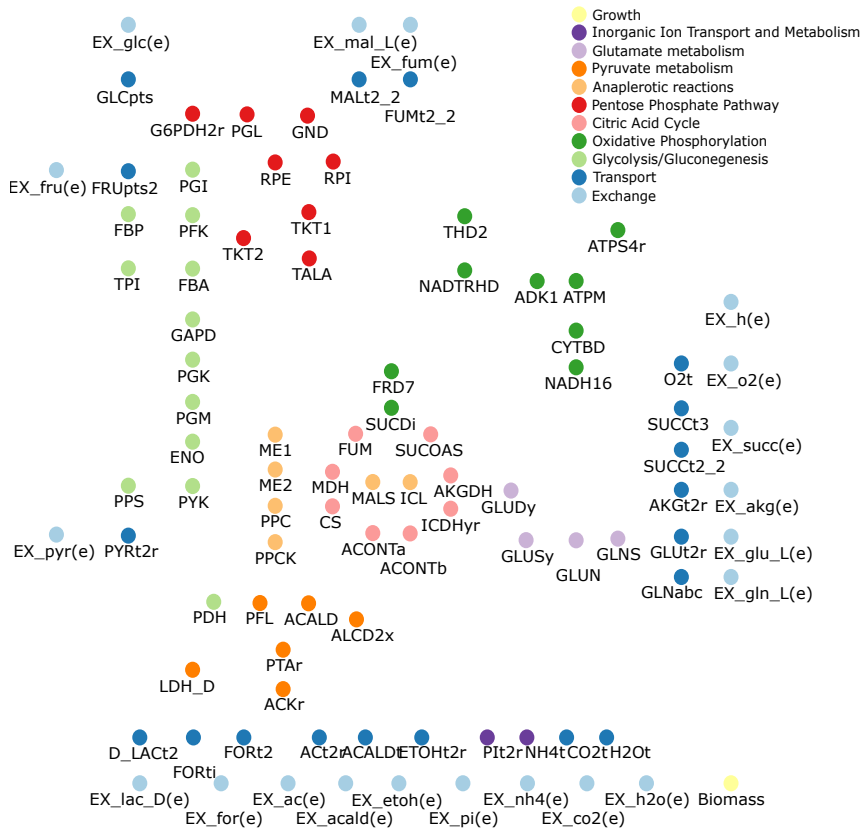


Figure 4.3. Reactions of the core *E. coli* metabolic model divided by subsystems. The layout is the same as in Figure 4.2 to facilitate comparison.

Mathematical formulation of the core E. coli metabolic model

The core *E. coli* metabolic model has stoichiometric matrix $\mathbf{S} \in \mathbb{Z}^{m,n}$, with $m = 72$ rows, each corresponding to a metabolite and $n = 95$ columns, each corresponding to a reaction. As it was stated in Section 2.3.2, the coefficients within each column of \mathbf{S} represent the stoichiometry of that particular reaction. A negative stoichiometric coefficient indicates the number of molecules of a metabolite consumed in that reaction. A positive coefficient represents production that metabolite. Since only a few metabolites are involved in each reaction, the matrix \mathbf{S} is sparse, consisting mainly of zero coefficients. Each reaction is balanced regarding chemical elements and electrical charge. The only exception to this are the exchange reactions at the boundary between the model and the environment. A metabolite leaving the model is represented by a positive flux in an exchange reaction, while the opposite occurs when the metabolite enters the model.

The net flux through all the 95 reactions in the model can be represented by a 95 dimensional vector, $\mathbf{v} \in \mathbb{R}^n$. The net flux usually has units of milli-mole per gram of dry weight per hour ($mmol \cdot gDW^{-1} \cdot hr^{-1}$). During balanced growth and when taking into account a large population of cells, it is reasonable to assume that all metabolite concentrations are constant, and therefore the system is at steady-state. Stoichiometry and the steady-state assumption are codified in Equation 2.6. Irreversibility constraints are applied setting the lower bound of some reactions to zero. Capacity constraints are codified setting the upper bound of some reactions to a maximum allowable flux. Furthermore, direct experimental information about the uptake or secretion rate of particular metabolites can be simply codified by setting both lower and upper bounds for that reaction to the experimental value. This way, the input in the model becomes fixed.

As it was stated in Section 2.3.2 the biomass reaction represents the consumption of bioprecursors and energy to ensure growth. To determine these metabolites and how much of them is needed, the dry weight composition data for an *E. coli* cell growing exponentially at 37°C under aerobic conditions on glucose minimal medium was used. The approximated doubling time was 40 min and the total dry weight was 2.81 grams. The exact composition is shown in Orth, Fleming, and Palsson 2010, page 29. The composition of the biomass reaction for *E. coli* is very well known and it has suffered only minor changes since it was first applied CBMs in Varma and Palsson 1993b (where was taken from Ingraham, Maaløe, and Neidhardt 1983). In the core *E. coli* model the biomass reaction is located in column number 13 and 23 metabolites participate in it, either as substrates or products. The main substrates are ATP, acetyl-CoA, L-glutamate, H₂O, NAD⁺, NADPH, oxalacetate and pyruvate while the most important products are ADP, 2-oxoglutarate, coenzyme A, H⁺, NADH, NADP⁺ and phosphate. On the other hand, the non-growth associated maintenance reaction (ATPM) is the 11th reaction in the model. This cost is codified by setting the lower bound of the reaction

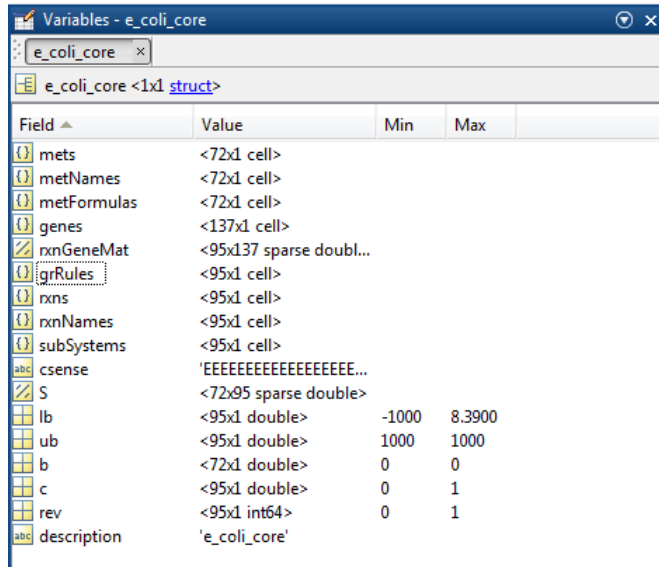
to $8.39 \text{ mmol} \cdot \text{gDW}^{-1} \cdot \text{hr}^{-1}$, simulating the processes that consume energy but do not require growth.

Computational form of the core E. coli metabolic model

The BIGG database (King et al. 2015b) is the to-go site for consulting and getting freely available constraint-based models. Most of them are available in SBML, JSON and MAT formats. In this thesis MATLAB was used to perform most of the operations so the chosen format for the model was MAT. The different categories in which the CBMs are separated are standard for the most part. Figure 4.4 shows the basic layout of a CBM in MAT format. The different fields are divided as follows:

- The three first fields correspond to metabolite's abbreviations, names and chemical composition.
- *genes* names all the genes included in the model.
- *rxnGeneMat* relates the genes with specific reactions.
- *grRules* assigns regulatory relationships between genes and reactions.
- *rxn* and *rxnNames* are the abbreviations and full names of each reaction
- *subSystems* assigns a subsystem or pathway to each reaction.
- **S** is the stoichiometric matrix
- **lb** and **ub** are the lower and upper bounds for each reaction.
- vector **c** indicates which particular reaction is being optimized (some methods explained in the following section require a specific reaction to be maximized or minimized).
- vector **rev** represents which reactions are irreversible.

The core of the CBM are the stoichiometric matrix **S**, the lower and upper bounds for each reaction **lb** and **ub** and the vector that indicates the reversibility of a reaction, vector **rev**. The first group of fields are more descriptive and all together define what is usually called a metabolic reconstruction. Therefore, CBMs are subsets of fields within metabolic reconstructions that can be used to calculate or simulate certain flux distributions.



Field	Value	Min	Max
mets	<72x1 cell>		
metNames	<72x1 cell>		
metFormulas	<72x1 cell>		
genes	<137x1 cell>		
rxnGeneMat	<95x137 sparse doubl...		
grRules	<95x1 cell>		
rxns	<95x1 cell>		
rxnNames	<95x1 cell>		
subSystems	<95x1 cell>		
csense	'EEEEEEEEEEEEEEEEEE...'		
S	<72x95 sparse double>		
lb	<95x1 double>	-1000	8.3900
ub	<95x1 double>	1000	1000
b	<72x1 double>	0	0
c	<95x1 double>	0	1
rev	<95x1 int64>	0	1
description	'e_coli_core'		

Figure 4.4. Different fields of the core *E. coli* model in MAT format.

4.4.2 Experimental *E. coli* data

In order to apply the different methods, the core *E. coli* model needs some sort of data to be used as input. This input is usually a set of experimentally measured extracellular fluxes that define the uptake and secretion rates for some of the most important metabolites such as D-glucose, oxygen and CO₂. Additionally, the growth rate is always measured too. There is abundant available data of chemostat experiments in *E. coli*. However, for the sake of removing as much of the variability as possible and to facilitate the comparison among the methods only a particular set of experiments from the same lab were chosen. All these scenarios come from works by the Systems Biology of Metabolism group of the Institute of Molecular Systems Biology of the Department of Biology at the ETH Zürich. This group, led by Professor Uwe Sauer, is at the highest level in modelling metabolism, quantitative metabolomics and intracellular flux analysis. One of their most successful branches of research is the study and analysis of the flux distributions of *E. coli*. In order to do that, they have published a series of papers in the last fifteen years. The data used in this chapter was taken from seven of those papers from 1999 to 2014 (Emmerling et al. 2002; Fischer, Zamboni, and Sauer 2004; Fong et al. 2006; Haverkorn et al. 2014; Kayser et al. 2005; Perrenoud and Sauer 2005; Sauer et al. 1999). They form a collection of 46 chemostat *E. coli* scenarios. Only the purely experimental data was used. None of the intracellular flux simulations they performed were used here. All the intracellular flux simula-

tions shown in this document were carried out as part of this chapter by me. The complete collection of scenarios is shown in Annex 7.2.

4.5 Methodology

There are several different methodologies that use the mathematical description of the cell metabolism that is the core of constraint-based modelling: the space of feasible flux distributions. Each methodology tackles a different problem, makes use of a different mathematical toolkit and is based on different biological assumptions. For an extensive review of most of these methods see Lewis, Nagarajan, and Palsson 2012 and Llaneras and Picó 2008. The scope of this thesis chapter reaches only the three most popular methods that aim to determine particular flux distributions: Metabolic Flux Analysis, Flux Balance Analysis and Flux Variability Analysis. In the three methods experimental measurements are used to obtain values for certain fluxes and therefore reduce or even completely determine the undetermined space defined in Equations 2.6, 2.7 and 2.8.

4.5.1 Metabolic Flux Analysis

Mathematical formulation

Metabolic Flux Analysis (MFA) uses a set of experimentally measured extracellular fluxes (see Section 4.4.2 and Annex 7.2) and some intrinsic knowledge of the cell's metabolism (Equation 2.6) to determine the fluxes that have not been measured. Therefore, considering a partition between the experimentally measured (\mathbf{v}_e) and the unknown fluxes (\mathbf{v}_u), the fundamental equation of MFA is obtained (Equation 4.1).

$$\mathbf{S}_u \cdot \mathbf{v}_u = -\mathbf{S}_e \cdot \mathbf{v}_e \quad (4.1)$$

Consequently, the system defined in Equation 4.1 combines the stoichiometric constraints given in Equation 2.6 and those forced by experimentally measured fluxes. If we only take into account the stoichiometric constraints defined in Equation 2.6, the system will be undetermined if n (number of reactions) is larger than m (number of metabolites). In the *E. coli* model $n = 95$ and m is equal to 72 and therefore the system is undetermined. That means there is a wide range of stoichiometric-possible flux vectors.

When constraints are added in the form of a set of fluxes experimentally measured \mathbf{v}_e , the system remains undetermined if $\text{rank}(\mathbf{S}_u) < u$ (being u the number of unknown or non-measured fluxes). In other words, if there are enough linearly inde-

pendent constraints to solely determine all the unknown fluxes \mathbf{v}_u then the system will be determined. In most cases the final resulting system is undetermined. For additional information about redundancy and determinacy in metabolic network see Klamt, Schuster, and Gilles 2002. Let's take the scenario 1 from Annex 7.2 to exemplify these considerations. In this particular example the number of measured fluxes e is equal to 6 (biomass or growth rate, D-glucose and oxygen uptake and CO_2 , acetate and pyruvate secretion), and therefore $u = n - e = 95 - 6 = 89$. Assuming that all equations in \mathbf{S}_u are independent, then the $\text{rank}(\mathbf{S}_u) = 72 < u = 89$, hence the resulting system is still undetermined. Another way to look at it is that the system becomes determined only when $n - m < e$. In this example $95 - 72 = 23 > 6$ so the system is undetermined. This situation, usually called *data scarcity*, is very common when dealing with metabolic networks and CBMs. All scenarios analysed in this chapter produce undetermined systems. Therefore a unique solution, in this case a flux distribution, is not mathematically possible. Instead, solving the Linear Programming problem we can obtain a interval for each reaction in which the flux must contained. Due to the structure of the network and the added constraints in the form of reversibility, capacity and measured fluxes those intervals, and the space of possible solutions that they define, contain very valuable biological information.

Dealing with uncertainty: Interval MFA

To represent a problem as close as possible to the biological reality of the cell's metabolic network, some authors developed a methodology for including uncertainty in the measurements and for dealing with the inherent system undetermination. This approach, which is actually a branch within MFA, is known as Flux Spectrum or Interval MFA (Llaneras and Picó 2007a; Llaneras and Picó 2007b).

To facilitate the comprehension of the method work-flow a toy example is shown in Figure 4.5. Figure 4.5A shows a very simple metabolic network with three metabolites (A, B and C) and six reactions. Reactions 1, 2 and 3 are intracellular and 4, 5 and 6 are exchange reactions. Unknown fluxes are colored in blue and experimentally measured fluxes in green. Figure 4.5B shows the mathematical description of all the constraints that form the problem. First stoichiometry, which is defined by the network structure. Then reversibility, showing which reactions are reversible (in this example only reaction 4). Third, the capacity constraints, that are defined as usual as a minimum and maximum for all the unknown fluxes (in this case reactions 1, 2, 3 and 5). Finally, the constraints are added based on experimentally measured fluxes (reactions 4 and 6). It is important to highlight the fact that the measurements are given as an interval as well, in this case 0.5 unit of flow around the measured values. In this case the measured values would be $\mathbf{v}_4 = 8$ and $\mathbf{v}_6 = 9$. This allows for a more realistic approach in the treatment of the measures, which takes into account certain uncertainty in the measures. After

solving the Linear Programming problem, Figure 4.5C shows the flux intervals for every reaction in the network. In this case due to the simplicity of the network the four unknown fluxes have the same amplitude (2 unit of flux). However, in the core *E. coli* model, being a much more complex system, some fluxes are much more constrained than others.

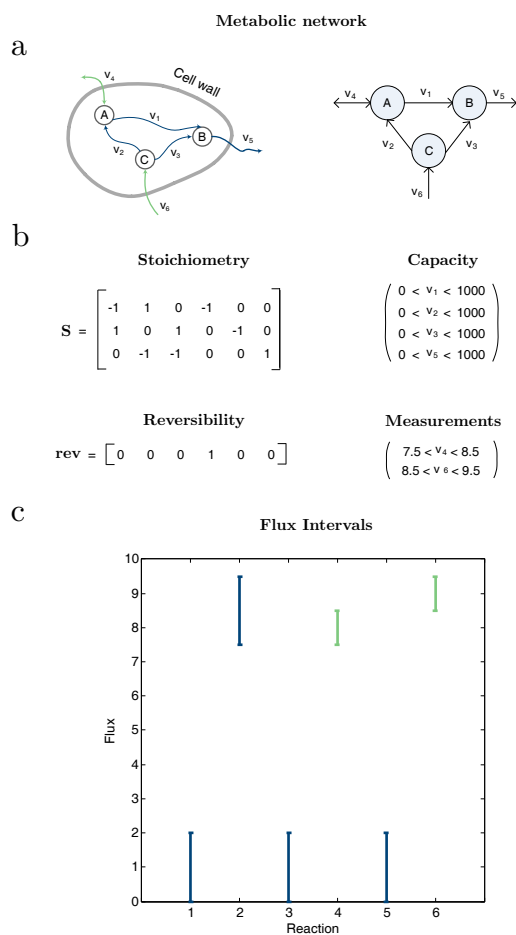


Figure 4.5. Interval MFA method applied to a toy example of metabolic network. (a) Diagram of the metabolic network. (b) Constraints defining the problem: stoichiometry, reversibility, capacity and experimental measurements. (c) Flux interval results after applying Interval MFA.

4.5.2 Flux Balance Analysis

Flux balance analysis (FBA) is a methodology that adds an optimisation principle to a constraint-based problem by assuming that cells evolved to achieve some sort of optimal behaviour under evolutionary pressure (Edwards, Covert, and Palsson 2002; Orth, Thiele, and Palsson 2010; Price et al. 2003). The most common way to formulate this assumption is to hypothesize that the metabolic objective of the cell is to maximize its growth. Mathematically, the formulation of the problem is identical to MFA but adding an additional constraint in the form of an objective function to be minimized or maximized (see Equations 4.2). The method is showcased in Figure 4.6 with the same toy example of network as before.

$$\begin{aligned} & \text{maximise: } \mathbf{c}^T \cdot \mathbf{v} \\ \text{subject to } & \left\{ \begin{array}{l} \mathbf{S} \cdot \mathbf{v} = 0 \\ \mathbf{v}_{\text{lb}} \leq \mathbf{v} \leq \mathbf{v}_{\text{ub}} \end{array} \right. \end{aligned} \quad (4.2)$$

where \mathbf{S} is the stoichiometry matrix of the model, \mathbf{v} the vector of fluxes, \mathbf{c} is an indicator vector (i.e. $c(i) = 1$ when i is the biomass reaction and zero everywhere else) so that $\mathbf{c}^T \mathbf{v}$ is the flux of the biomass reaction. The constraint $\mathbf{S} \cdot \mathbf{v} = 0$ enforces mass-conservation at stationarity, and \mathbf{v}_{lb} and \mathbf{v}_{ub} are the lower and upper bounds of each reaction's flux. Through these vectors, one can encode a variety of different scenarios (see Orth, Fleming, and Palsson 2010). The biomass reaction represents the most widely-used flux that is optimised, although there are others can be used as well (Feist and Palsson 2010; Schuetz, Kuepfer, and Sauer 2007).

Solving the resulting linear programming problem, the flux distribution that makes the best (*best* meaning *optimal*) use of the metabolic network defined and constrained to answer the objective function proposed is obtained. Therefore, predictions made with FBA are highly dependant of the selected objective function. To apply FBA three strong assumptions need to be made regarding the involved optimization:

- Cells, forced by the evolution process, evolve to reach an optimal behaviour according a concrete biological objective.
- We know which objective the cells pursue
- It is possible to mathematically describe this objective

As it was stated, biomass reaction is the most commonly used objective function. For micro-organisms, due to their relative biological simplicity, several studies proved that FBA predictions are consistent with experimental data. It makes sense that the cellular behaviour of simple organisms should be more easily understood and described than more complex organisms. Especially for *E. coli* growing

in chemostat environments the results obtained through FBA are very solid (Edwards, Ibarra, and Palsson 2001; Varma and Palsson 1994) and accepted as a viable predictions.

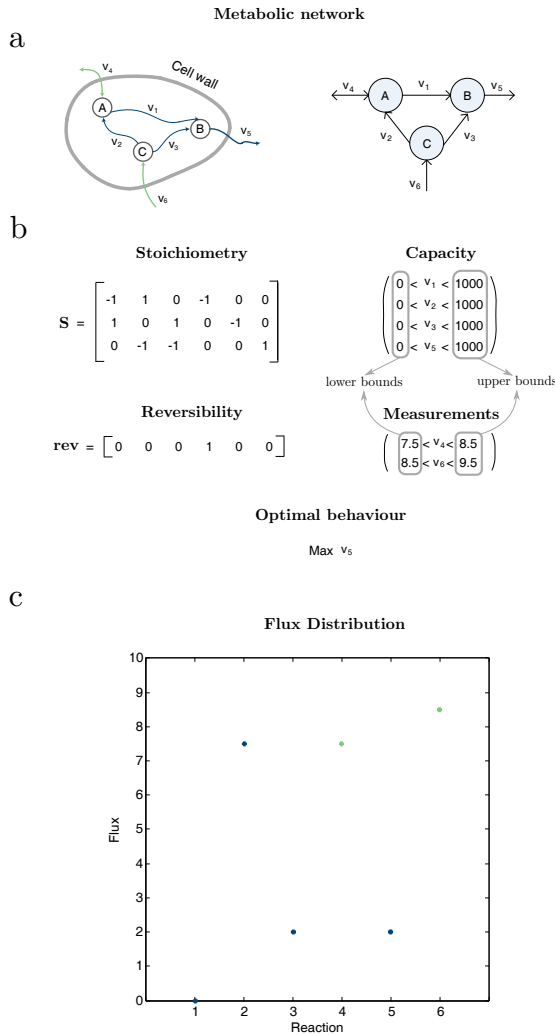


Figure 4.6. FBA method applied to a toy example of metabolic network. (a) Diagram of the metabolic network. (b) Constraints defining the problem: stoichiometry, reversibility, capacity and experimental measurements. Note the additional constraint of optimal behaviour. (c) Flux distribution results after applying FBA.

It is important to notice the relevance of the lower (\mathbf{v}_{lb}) and upper bounds (\mathbf{v}_{ub}) of each reaction's flux. These bounds contain the capacity and measurement constraints (as shown in Figure 4.6B). Unknown fluxes have bounds limited only by capacity. On the other hand, measured fluxes have bounds set very close to the measured value. As it was commented in Section 4.5.1 using an interval for representing the measured fluxes as well facilitates the inclusion of uncertainty in the measurements. The result of a typical FBA problem is a flux distribution, as it is shown in Figure 4.6C. That is an individual value for every flux in the network. However, in spite of the addition of the optimal behaviour constraint, the problem is still undetermined and therefore multiple flux distributions are equally possible. FBA usually reaches the same optimum in the objective function but different flux values are possible for all the other fluxes. Therefore, FBA outputs only one solution of many possible. In the toy example, the optimal value for the objective function \mathbf{v}_5 is 2 units of flux. Although the optimal is found and set in $\mathbf{v}_5 = 2$, many different combinations of values for every other flux reach that same optimal solution. Since the objective is not only to obtain the optimal value for the objective function, but to acquire an interval for the value of every flux, the FBA methodology alone falls short. To tackle this issue, the next and final methodology is needed.

4.5.3 Flux Variability Analysis

Metabolic networks often include flux redundancies that add to their structure robustness (Mahadevan and Schilling 2003; Price et al. 2003; Reed and Palsson 2004). FVA can be used to investigate these redundancies by calculating the complete range of numerical values for each reaction flux in the network. This is done by optimizing a particular objective flux or function, while still fulfilling set of constraints given by the system. The particular application of FVA used in this chapter is to determine the ranges or intervals of fluxes that correspond to an optimal solution obtained previously through FBA. The maximum value of the objective function (in this case the biomass reaction) is first computed and this value is used with multiple optimizations to calculate the maximum and minimum flux values through each reaction (Schellenberger et al. 2011). Figure 4.7 shows the FVA methodology applied to the same toy metabolic network as before.

FVA outputs a range of values for every flux in the network, except the flux defined previously as objective, which exact value is determined previously by FBA. Some reactions can vary more than others giving an idea of the shape of the space of possible solutions and the robustness of the network as a whole.

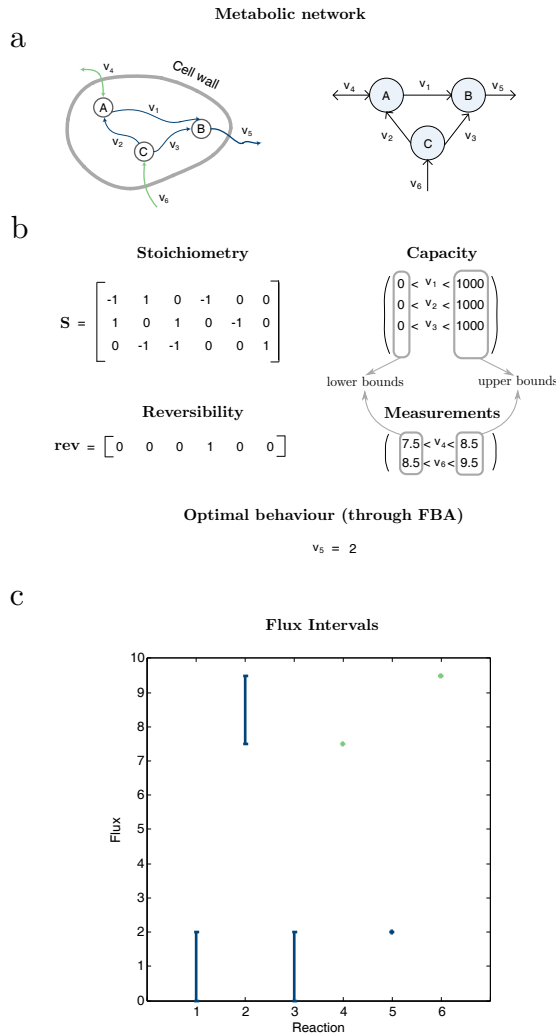


Figure 4.7. FVA method applied to a toy example of metabolic network. (a) Diagram of the metabolic network. (b) Constraints defining the problem: stoichiometry, reversibility, capacity and experimental measurements. v_5 is equal to 2 units of flux, which comes from the FBA solution. The *optimal behaviour* constraint indicates so. (c) Flux ranges or intervals results after applying FBA.

4.5.4 Comparison of methods

The three methodologies previously described produce different flux distributions (in case of FBA) or flux intervals (in case of MFA or FVA) when applied to the same network with the same constraints (see Figure 4.8). MFA (red) returns the wider intervals, including the objective function, which in the example is \mathbf{v}_5 . FBA alone (blue) produces a single solution, or flux distribution instead of a series of intervals. Finally, FVA (which previously requires FBA to locate the maximum value for \mathbf{v}_5) generates an interval solution as well. However, this FVA solution is narrower than the MFA one. Reactions \mathbf{v}_4 , \mathbf{v}_5 and \mathbf{v}_6 only allow for a concrete value of flux, while in MFA those fluxes show an interval of possible values. It is important to notice that the FVA solution is a subset of the MFA solution, being all values in the former included in the latter. Due to the simplicity of the toy network, the amplitude of each flux is the same for the MFA and the FVA flux intervals. However, more complicated networks (like the network in the core *E. coli* model) produce a many more flux differences.

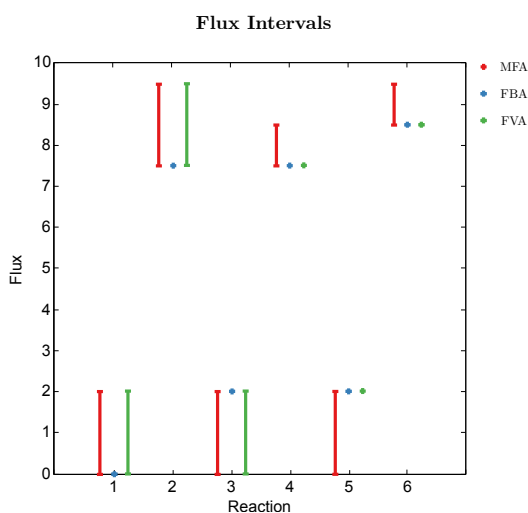


Figure 4.8. Flux intervals obtained using the three different constraint-based methods on the toy example network.

Having the particularities of MFA, FBA and FVA in mind it is finally time to decide how these methods will be used with the real data. The core of this work presented in this chapter is the comparison of such methods when applied to the core *E. coli* model in a wide range of scenarios. The three choices taken that have more biological sense are:

- Metabolic flux analysis excluding growth rate data (*MFA*). This is the most conservative approach. MFA is applied to the model with the data from the

different scenarios. However, the experimental data for growth rate, which is available in every scenario, is not used. This approach allows us to see how wide are the range of fluxes, if some of them are reversible (that is if the interval crosses the zero value) and where the real data of growth is located in the estimation that MFA provides.

- Metabolic flux analysis including growth rate data (*MFAg*). This goes a step forward using the valuable information contained in the growth rate data. The biomass reaction is highly connected and therefore having it set to a measured value will reduce the space of possible flux value for all the other reactions. This approach allows us to test how much influence has the growth rate on the rest of the network.
- Flux variability analysis (*FVA*). As a necessary step for this option, FBA must be applied first to obtain the maximum growth rate. Then with that value set, FVA calculates the range of possible values for all the other reactions. This final option allows to test two important ideas. First, we can check if the optimality principle of FBA and FVA matches the model and the experimental data (specially in the growth rate obtained). Second, it allows us to quantify the reduction of the solution space that this assumption provides when compared to the first to options.

Please note that the exact approaches taken are always written in *italics* and the original methods in normal letters. This way, *MFA* uses solely MFA, *MFAg* uses MFA with the additional data of growth rate and *FVA* uses both FBA and FVA methodologies consecutively.

4.5.5 Statistical description and analysis

Once the flux intervals for the three methods selected (*MFA*, *MFAg*, *FVA*) are determined on each of the 46 different scenarios of *E. coli*, some conclusions can be directly drawn. Inspecting the flux interval of particular reactions, or pathways, one can start to analyse the validity and usefulness of each method. Furthermore, direct comparison can be made between concrete scenarios to highlight the differences in the shape of the space of feasible solutions. However, in order to thoroughly compare the three methodologies, two basic steps must be taken. First, the space of solutions they defined must be mathematically characterized. Secondly, some sort of statistical analysis must be carried out to check if the conclusions drawn from those mathematical descriptions are robust and solid. In other words, this statistical analysis must answer the questions:

- Do the differences we see in the flux intervals in each method occur constantly throughout the complete dataset?

- Are the conclusions we draw from the methods dependant from the inputs defined in each scenario?
- Or, on the contrary, the results are robust and similar across all studied experimental conditions?

The first step is then to characterize the flux intervals of each method when applied to each scenario. Since the methods present increasing constraints from *MFA* to *MFAg* to *FVA*, in general terms each flux space is a subset of the space obtained in the previous method. Therefore, the solution space of *MFA* is the widest, the one from *MFAg* is narrower and the one from *FVA* is the smallest in terms of amplitude of flux intervals. The most notable exception is the growth rate, or biomass reaction in the model, for which sometimes there are strong discrepancies between the two *MFA* variants and the *FVA* solution. These differences are explained in detailed in Section 4.6.5.

Two ratios were defined to characterize the differences in the flux intervals between the three methods. The first one compares *MFAg* to *MFA* and it is consequently called $\mathbf{r}^{MFAg/MFA}$. The second relates *FVA* to *MFA* and it is named $\mathbf{r}^{FVA/MFA}$. First, the amplitude of the intervals is defined for each method in Equations 4.3:

$$\begin{aligned}\mathbf{int}_{i,s}^{MFA} &= \mathbf{v}_{i,s}^{MFA,max} - \mathbf{v}_{i,s}^{MFA,min}, \\ \mathbf{int}_{i,s}^{MFAg} &= \mathbf{v}_{i,s}^{MFAg,max} - \mathbf{v}_{i,s}^{MFAg,min}, \\ \mathbf{int}_{i,s}^{FVA} &= \mathbf{v}_{i,s}^{FVA,max} - \mathbf{v}_{i,s}^{FVA,min}\end{aligned}\tag{4.3}$$

where $i \in (0, n)$ being n the number of reactions and $s \in (0, e)$ where e corresponds to the number of experimental scenarios. The vectors $\mathbf{v}_{i,s}^{max}$ and $\mathbf{v}_{i,s}^{min}$ correspond to the limits of the intervals of each reaction, in each scenario and obtained with each method. Finally the $\mathbf{int}_{i,s}$ vectors represent the amplitude of each interval per reaction ($n = 95$), scenario ($e = 46$) and method (*MFA*, *MFAg*, *FVA*). The ratios $\mathbf{r}^{MFAg/MFA}$ and $\mathbf{r}^{FVA/MFA}$ capture the percentage relation between the two pairs of methods:

$$\begin{aligned}\mathbf{r}_{i,s}^{MFAg/MFA} &= \frac{\mathbf{int}_{i,s}^{MFAg}}{\mathbf{int}_{i,s}^{MFA}} \times 100, \\ \mathbf{r}_{i,s}^{FVA/MFA} &= \frac{\mathbf{int}_{i,s}^{FVA}}{\mathbf{int}_{i,s}^{MFA}} \times 100\end{aligned}\tag{4.4}$$

For example, in scenario 2, reaction 8 of the experimental data we can see the following flux results (see Figure 4.9):

$$\begin{aligned}
 \mathbf{v}_{8,2}^{MFA,max} &= 5.016, \\
 \mathbf{v}_{8,2}^{MFA,min} &= 0, \\
 \mathbf{v}_{8,2}^{MFAg,max} &= 2.572, \\
 \mathbf{v}_{8,2}^{MFAg,min} &= 0, \\
 \mathbf{v}_{8,2}^{FVA,max} &= 2, \\
 \mathbf{v}_{8,2}^{FVA,min} &= 0.819
 \end{aligned}
 \tag{4.5}$$

producing the flux intervals

$$\begin{aligned}
 \mathbf{int}_{8,2}^{MFA} &= 5.016, \\
 \mathbf{int}_{8,2}^{MFAg} &= 2.572, \\
 \mathbf{int}_{8,2}^{FVA} &= 1.180
 \end{aligned}
 \tag{4.6}$$

and finally generating the ratios

$$\begin{aligned}
 \mathbf{r}_{i,s}^{MFAg/MFA} &= 51.28\%, \\
 \mathbf{r}_{i,s}^{FVA/MFA} &= 23.52\%
 \end{aligned}
 \tag{4.7}$$

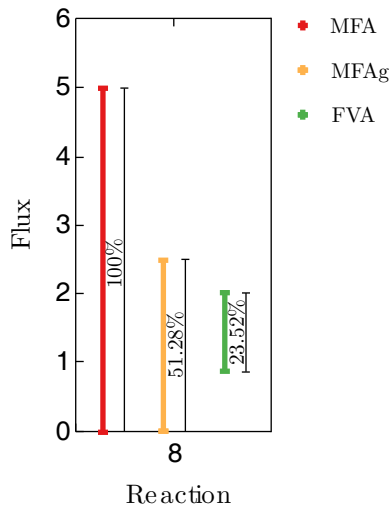


Figure 4.9. Flux intervals and ratios obtained using the three different constraint-based approaches chosen (*MFA*, *MFAg*, *FVA*) for reaction 8 of scenario 2.

That means the interval of possible flux for reaction 8 in the experimental scenario 2 for the *MFAg* solution is 48.72% smaller in size than the *MFA* solution ($100 - 51.28 = 48.72$). The *FVA* solution is 76.48% smaller than the *MFA* solution. Finally, to determine one unique ratio per scenario, the average of the ratios for all reactions in each scenario calculated. At the end, two different parameters are used to describe the intervals of each method in each of the 46 available scenarios. Those are:

- $r_s^{MFAg/MFA}$: average of the vector of ratios $\mathbf{r}_{i,s}^{MFAg/MFA}$
- $r_s^{FVA/MFA}$: average of the vector of ratios $\mathbf{r}_{i,s}^{FVA/MFA}$

Finally, to obtain a general numeric value to represent the complete dataset the averages of both ratios are calculated. These final parameters are denoted $r^{MFAg/MFA}$ and $r^{FVA/MFA}$. Once the flux solutions obtained for each scenario using each methodology are statistically described and defined, multivariate statistics (PCA and PLS) are used to check if the conclusions drawn from the comparison are robust across the complete dataset. A brief mathematical description of both methods can be found in Section 2.3.3.

All the flux calculations performed to complete this chapter were done with MATLAB (with an academic license available through UPV). *MFA* intervals were specifically obtained using the *PFA Toolbox* (Morales et al. 2016) and *FBA* and *FVA* estimations were performed using the *COBRA Toolbox* (Schellenberger et al. 2011). The multivariate statistical analysis was performed using the *ProSensus ProMV* program with an academic license available under request for undergraduate students.

4.6 Results and discussion

4.6.1 Data vs. model agreement

The first question we should ask ourselves is: does the core *E. coli* model agree with the experimental data collected? Are the *MFA*, *MFAg* and *FVA* methods able to find a solution to the model using as input the experimental dataset? Once interval solutions are found, then a comparison among them can be made but the first step is finding those solutions. The results obtained vary with the method.

MFA and *FVA* are able to find a solution in 44 of the 46 scenarios, which constitutes an impressive 95.65% of experimental scenarios solved. Only scenarios 10 and 11 remain unexplained by both methods. The cause of this inconsistency between the data and the model probably comes from the low input of both scenarios. Both define the points with the lowest intake of D-glucose, intake of O_2 and

production of CO_2 . With the D-glucose inputs defined in scenarios 10 and 11 and under optimality assumption, the required minimum O_2 input is $-2.93 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$ and $-3.42 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$, respectively. That's much higher than the reported intake values for O_2 : $-1.375 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$ for scenario 10 and $-1.781 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$ for scenario 11. Therefore, it seems that there is not enough O_2 input for the reported D-glucose intake. Besides, the reported growth is very low for both scenarios: $0.044 h^{-1}$ for scenario 10 and $0.066 h^{-1}$ for scenario 11. With such low growth the amount of energy invested in pure cell maintenance (codified in the model as reaction 11, *ATPM*) is very high compared to the energy spent for pure growth. For very low input scenarios it is hard to reach those minimum requirements of O_2 . When the inputs grow, the percentage of O_2 destined to fuel the maintenance of the cell is much lower. In general terms, the core *E. coli* model and many of these constraint-based models tend to perform poorly in very low input conditions. Since the *FVA* solutions are strictly subintervals of the *MFA* solutions, it was expected that both methods are able to solve the same scenarios.

The results obtained with *MFAg* are not as good, being able to solve 22 scenarios (1-9, 33-35 and 37-46). That establishes a 50% of solved scenarios. This shows that this method is much more fragile and measurement-dependent. The main reason behind this discrepancy with the other two methods is the fact that *MFAg* is using a fixed value for the growth rate (reaction 13 in the model, *Biomass-Ecoli-core-w-GAM*). This reaction has a massive connectivity with 16 substrates and 7 products. Therefore, a fixed value constrains the system a great deal, affecting many metabolites. Since the system defines a solution space much narrower, small inconsistencies in the measurements may prevent the method to find feasible solutions. Accordingly, the flux distributions obtained through *MFAg* are much smaller and better defined than those found through *MFA* (shown in Section 4.6.3 and Annex 7.3). However, the method has the disadvantage of working in less scenarios. This disadvantage may be useful if the objective is to make sure that the measurements are very close to those predicted by the model. It becomes a compromise problem: if we trust more the growth rate measurement than the others, we can add it to the system and compute it as a *MFAg* problem and to check if the rest of the measurements agree with it or not.

4.6.2 Growth rate estimation

One of the main applications and validations of constraint-based models is their capacity to accurately predict cellular growth. As it was already commented, this fact lays on the precise mathematical formulation of a biomass function and a series of logical consecutive constraints that reduce the space of feasible solutions and make the prediction accurate. In this dissertation one approach (*MFAg*) uses the measured cellular growth as input and therefore no estimation in this regard is possible. The other two methods, *MFA* and *FVA*, obtain estimated values for

growth rate just as they do for any other reaction in the model. *MFA* produces intervals for every reaction, including the biomass reaction that represents cellular growth. *FVA* on the other hand returns a single value for this particular reaction. This occurs because in order to apply *FVA*, a previous optimization problem must be solved. This optimization problem is known as *FBA*, which finds the maximum value for the growth rate. After that, *FVA* calculates how much each reaction can vary while keeping the biomass reaction still fixed at the maximum. Therefore, *FVA* returns a single value for growth rate for each scenario, and this value represents the maximum that is possible under such constraints.

Figure 4.10 shows the growth estimations for *MFA* and *FVA*, and the measured growth rate value for each scenario. The interval solutions from *MFA* contain the experimental growth value in 22 scenarios out of 44, which produces a 50% estimation capability. Those scenarios are 1-9, 33-35 and 37-46. It is easy to see in the graph that those points tend to be located in the upper section of the intervals, close to the maximum value. On average, the experimental values are 17.63% lower than the maximum found by *MFA*. In the other 22 scenarios (12-32 and 36) the experimental growth is higher than the maximum of the interval of the *MFA* solutions. The average underestimation (measured as the difference between the experimental value and the upper value of the *MFA* intervals) of those scenarios is 36.88%. On the other hand, the single solution obtained with *FVA* corresponds always with the maximum value allowed by *MFA*. This makes sense since both methods use the same inputs and model and the only difference is the maximization of growth rate that *FVA* imposes. In each scenario, that point can be viewed as an optimal point for growth for the organism.

The implications of both under and overestimation in both methods when compared to the experimental value for growth are significant for different reasons. When the experimental value is lower than that optimal point the organism is not growing as fast as it can accordingly with the model. This is to be expected since the model is not representing all the cell-level processes, such as transcription, translation, gene duplication and more. It only comprises the central carbon metabolism. Besides, there are numerous conditions in which the cell would decide to grow slower such as environmental stress or acclimatization to new environments. A value close to the optimum but lower makes the most sense when working with relatively simple constraint-based models.

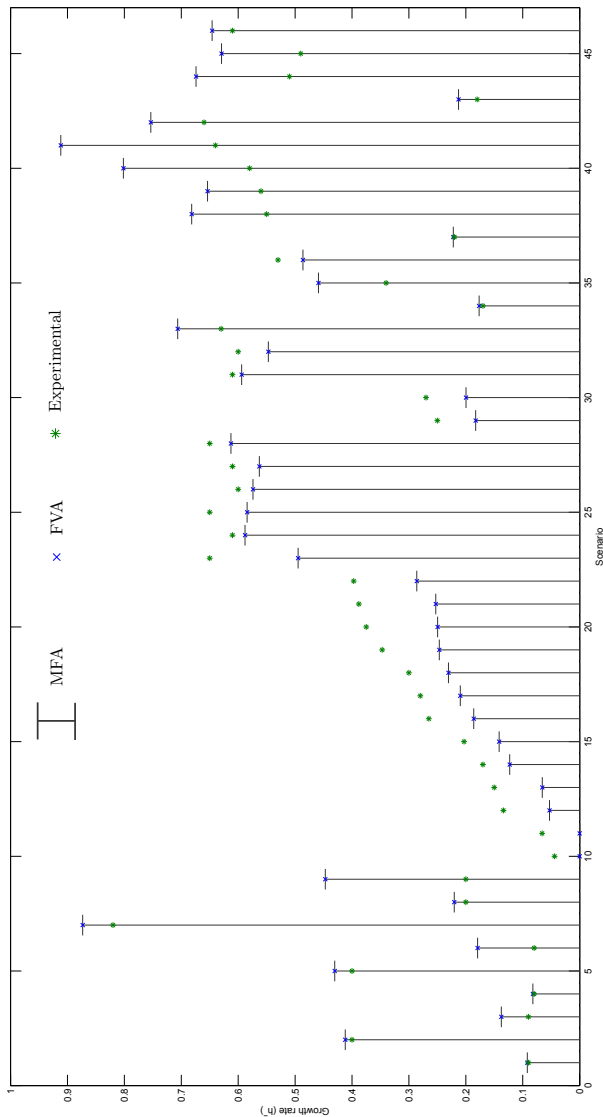


Figure 4.10. Comparison of growth rate estimations by *MFA* and *FVA* with experimentally measured growth rate.

The scenarios in which the experimental growth is higher than the maximum found by *MFA* and *FVA* project a different and more relevant issue. The data escapes the space of possible solutions built by the model. That may have two principal causes, either the model is not exhaustive enough or the data measured does not

respond to the real inputs and outputs of the cell. The model's lack of scope may be caused by alternative sources of energy or carbon for growth, in other words, there are additional inputs that the model does not cover. It could also be that the inner structure of the model is flawed or at least incomplete. The most usual cause of this is the importance of metabolic reactions that short-cut some branches in the network and are active under certain circumstances. If some of these reactions are not included in the model, then its structure and predictive capability is damaged. A typical example of this would be the Entner-Doudoroff pathway, which is involved in the gluconate metabolism, usually absent in such small models. However, in this particular case the ED pathway does not explain the under-performance of the model in those scenarios because it has been reported to be virtually absent in D-glucose fed cells (Peekhaus and Conway 1998). On the other hand, some of the data may have measurement errors that can produce the model inconsistency. The vast majority of scenarios in which the experimental growth is higher than the optimum predicted by the methods come from two papers: Perrenoud and Sauer 2005 and Kayser et al. 2005. There is a possibility that some factor was not taken into account in those particular experiments that produce a higher efficiency behaviour in *E. coli*.

MFA and *FVA* produce an average growth estimation 19.25% lower than the experimental values for the complete dataset. When taken as a whole, it is evident that the cell tends to behave in a metabolic state that is located around the principle of growth optimality. Sometimes the prediction is lower than the experimental value and sometimes is higher but the fact that there is perfect split between over and underestimation strongly suggests a metabolic state around this predicted optimal point. It is important to comment that all these scenarios correspond to chemostat cultures, where the principle of optimality makes the most sense. However, it remains to be seen if the same trend is so clearly followed in fed-batch cultures as well.

4.6.3 Flux distributions

The core of the work presented in this dissertation is the flux distributions obtained through the three different approaches (*MFA*, *MFAg* and *FVA*) for the 46 collected scenarios. The complete flux distributions, in the form of intervals, are stored and presented as graphs in Annex 7.3. The first conclusion that can be extracted from those graphs is that the distributions are very similar in shape across the entire dataset. There seems to be a robustness in the results that every method achieves. This robustness is studied in detail in Sections 4.6.5 and 4.6.6. In contrast, this section is focused on the general information and features present in all of them. In order to do that, I will use as an example the results for scenario 1, shown in Figure 4.11.

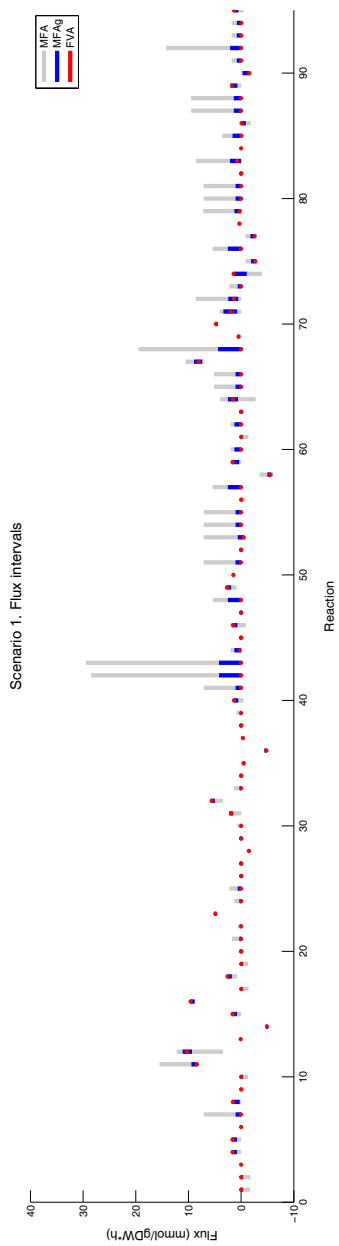


Figure 4.11. Flux distributions obtained for scenario 1.

The first consideration to be addressed when studying the flux distributions obtained is their amplitude when compared with the original space of solutions. How much reduction in the space of solutions these methods allows us? The plain simple answer would be quite a lot. The total feasible flux value for all known reactions can vary from -1000 to 1000 $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$ for reversible reactions and from 0 to 1000 $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$ for irreversible ones. These are known as capacity constraints (see Section state:modellingtools:constraint). The results of the widest solution (which is always *MFA*) show a massive reduction of the possible value for every flux. Most fluxes can take values in intervals with amplitude equal to a few $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. Even the two widest reactions can vary only from 0 to 30 $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. The reduction for *MFAg* is even greater constricting the fluxes to intervals of around 2 $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. Finally the *FVA* solution defines almost a singular solution with a concrete value for almost every flux. When the value of the cellular inputs and outputs is higher those intervals grow in size but the reduction is always very large.

A rough but informative way to quantify this reduction is to compute the sum of all values possible for all reactions in the initial space and the space defined by the methods. Assuming that every flux can take only round numbers, the size of that total flux space for scenario 1 would be $15.4 \times 10^4 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. In contrast, the total flux space possible (calculated the same way) for the *MFA* solution is $335.2 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. That defines a reduction of 99.9978% in the space of feasible solutions. Solutions for *MFAg* and *FVA* are even narrower, as Figure 4.11 highlights. Since the size of those spaces is so small compared with the original problem, the *MFA* solutions are taken as benchmarks to quantify the quality of the other two in each scenario. The *MFA* solution requires the minimum amount of information available: the structure of the network, which is constant and the measures of inputs and outputs, which vary in each scenario. That is why it is used as base in the ratios defined in Section 4.5.5. In this particular scenario the ratio $\mathbf{r}^{MFAg/MFA}$ that compares the size of the solution for *MFA* and *MFAg* has a value of 37.23%. In Figure 4.11 is clear that the blue bars that represent the *MFAg* solution are in every reaction smaller than the grey bars of the *MFA* solution. In this scenario they have a size of 37.23% of the *MFA* bars, or in other words, the *MFAg* solution achieves a reduction of 62.77% in size compared with *MFA*. The *FVA* solution has a size which is a 0.49% of the *MFA* one ($\mathbf{r}^{FVA/MFA} = 0.49\%$). That is a massive reduction (99.51%) to almost a single flux distribution in which every flux has a concrete value instead of a range of possible values.

It is important to clarify that not all combinations of fluxes are possible in the flux space obtained through the methods. For instance, picking the maximum value in each flux interval in the *MFA* solution would produce a flux distribution that would not be a solution to the problem. The intervals delimit the maximum and minimum values that any particular reaction flux can take. Therefore, most of the combination of fluxes inside those intervals are not valid solutions. However,

outside those intervals, no valid solution is possible. That's why it is appropriate to name *feasible solution space* to the set of intervals. Not because all combinations within that space are solutions but because all possible solutions are located inside that space.

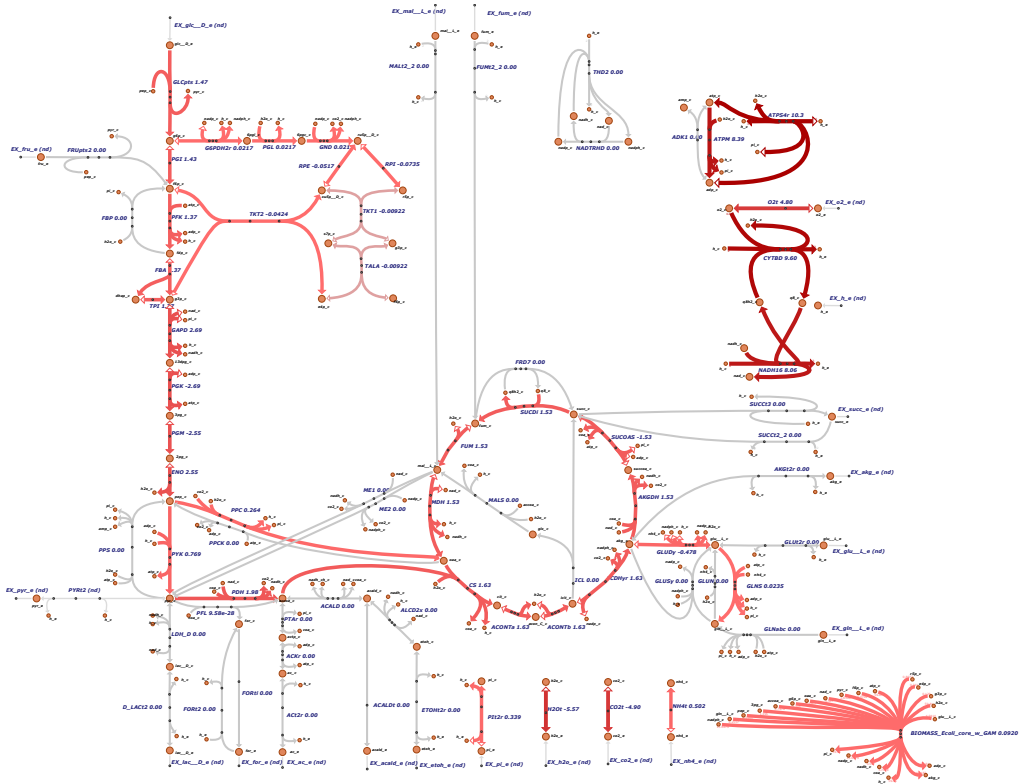


Figure 4.12. Flux map of the *FVA* solution of scenario 1.

As it has been commented before, the shape of the solution in each scenario is very similar. Reactions 20 to 39 represent the exchange with the cell's environment. They are very well defined and most of them have a value very close to zero. The information available in each scenario is encoded in the value of some of these exchange reactions. Apart from those, most fluxes oscillate between zero and $10 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. Some of the reactions have negative values because of the direction they are defined. There are two couples of reactions with larger than average interval's amplitude. The first couple are reactions 42 (FORt2) and 43 (FORti), both representing different mechanisms of formate transport. The second couple are reactions 68 (NADTRHD, NAD transdehydrogenase) and 92 (THD2, NAD(P)

transdehydrogenase) which represent the conversion and balance between intracellular NADH and NADPH. This coupled reactions have the exact same participants but in opposite direction (substrates are products and vice versa). That produces a closed loop that increases the possible values of those reactions. It is reasonable to assume that the actual value of those reactions will be similar to those around them. It does not make much sense in biological terms to have futile cycles of reactions with high fluxes. When facing this particularities the constraint-based approach lacks the precision to make accurate predictions. Nonetheless, reasonable flux constraints can be added if needed to smooth the value of those fluxes to those around them and in any case the range of the intervals in those reactions is only marginally higher than the rest.

Figure 4.12 shows the *FVA* solution for scenario 1, illustrated as a flux map on the core *E. coli* model map. Grey arrows correspond to reactions with zero flux. Reactions with flux through them are coloured from minimum value (light red) to maximum value (dark red) in absolute terms. This map highlights the metabolic state occurring in most scenarios. Slight differences occur when pyruvate or acetate are produced and secreted but the main features remain the same. Carbon enters the cell's as D-glucose and almost entirely goes through glycolysis to end up producing pyruvate. However, a small amount is rerouted through the pentose phosphate pathway to produce D-riboluse-5-phosphate (r5p-c) and D-erythrose-4-phosphate, which constitute two substrates for biomass production. Carbon enters the TCA cycle as acetyl-CoA producing citrate from oxalacetate. TCA flows as expected in aerobic conditions producing NADH and NADPH and a small amount of ATP too. A minor branch appears from 2-oxoglutarate to produce L-glutamine, also necessary for biomass production. Reactions CYTBD and NADH16 represent the electron transport chain that produce the proton unbalance that is later used by the ATP synthase (ATPS4r) to generate the core of the ATP in the cell. Inorganic ions exchange reaction represent the global consumption of oxygen (O2t), phosphate (PIt2r) and ammonium (NH4t) and the production of CO₂ (CO2t) and H₂O (H2Ot). Finally the biomass precursors produced across the entire network are spent in the biomass reaction to produce growth and a series of by-products that are re-used by the cell to keep the process going.

4.6.4 Pathway flow profiles

The amplitude of the different solutions or how the space of feasible fluxes is progressively reduced is not the only outcome that measures the validity of the approaches. Sometimes these methods produce mathematically possible solutions that are biologically infeasible. To highlight these situations Figure 4.13 shows the reactions for glycolysis and the TCA cycle for scenario 1. Reactions are displayed following the traditional direction of carbon processing. Glycolysis starts with the intake of D-glucose from the environment (GLCpts) and ends producing pyruvate at the final PYR reaction. It includes a total of ten reactions. The TCA cycle

starts with Acetyl-CoA inclusion in the stream of the cycle and ends up with the production of oxalacetate, point at which the cycle starts over again. It contains a total of nine reactions.

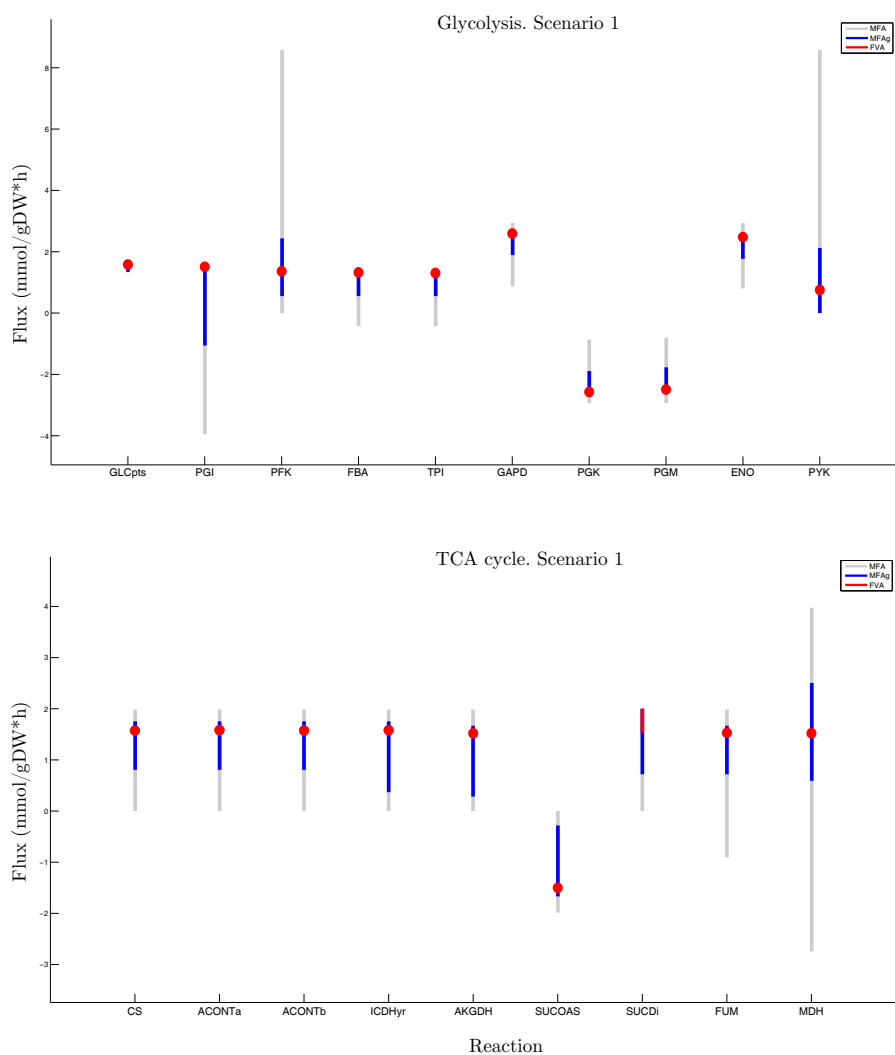


Figure 4.13. Glycolysis and TCA cycle pathways from the *MFA*, *MFAq* and *FVA* solution of scenario 1.

In glycolysis different methods produce intervals with different amplitude that, for some reactions, generate infeasible phenotypes for the conditions in which the

scenarios are based. The *MFA* solution, besides producing the widest intervals, allows for some essential reactions to have zero flux. Reactions such as FBA and TPI can have a value of exactly zero, which shuts down the first four steps of the glycolysis. Reaction PGI can even have a negative value for the flux, describing a gluconeogenesis behaviour instead of the classic glycolysis, which we know is actually happening. Some of these disturbances are produced by the pentose phosphate pathway, which is linked to glycolysis in several steps. The primary function of the pentose phosphate shunt in *E. coli* is to provide the 5-carbon and 4-carbon biosynthetic precursors, respectively α -D-ribose-5-phosphate and D-erythrose-4-phosphate. Those are necessary for amino acid and purine/pyrimidine production. That's why they are included as substrates in the biomass reaction. Therefore the flux through this pathway is much smaller than through glycolysis, which constitutes the main carbon processing route. *MFA* is not able to capture this biological feature.

The *MFAg* results for glycolysis are more accurate, because the inclusion of the growth rate as input greatly restricts the solution space. Blue bars in Figure 4.13 are much narrower and the carbon rerouted through the pentose phosphate shunt is smaller. Nevertheless, two reactions (PGI and PYK) can still have zero flux. In the PGI reaction, that would mean a complete balance between glycolysis and pentose pathway. A value of zero for PYK would stop the production of Acetyl-CoA. Both scenarios are not realistic in biological terms for the environment described for scenario 1. Finally, the results for *FVA* define a single flux value for every flux. Every reaction flows in the logical direction and the carbon spent in the pentose pathway only obeys the production of small quantities of α -D-ribose-5-phosphate and D-erythrose-4-phosphate. Note that the PGK and PGM reactions are defined opposite from the typical glycolysis direction and that's why their flux is negative. The restriction that *FVA* imposes over the system (maximization of growth rate) defines a glycolysis and pentose phosphate pathways much more precisely and in agreement with existing biological knowledge.

The situation with the TCA cycle is similar. *MFA* defines a solution that allows the complete TCA cycle to be shut down. It is only one of many solutions possible under the *MFA* approach but once again it is a clear prove that *MFA* falls short when discarding infeasible solutions. Results from *MFAg* are better: the intervals are narrower and the fluxes are positive in the standard direction. *FVA* results are also better defining almost a singular solution with reasonable flux values. Two particularities arise here that is worth mentioning. First, the *FVA* solution is not located in an extreme position of the *MFAg* solution (the red dots are not placed in an extreme of the blue bars). This does not happen in glycolysis for *FVA*. This feature highlights that fact that *FVA* solutions are located in the surface of the n -dimensional cone of solutions in some dimension but in others they are inside the volume of the cone. The exact position of the solution in the feasible solution space is function of the inputs and the inner structure of the network, which defines the

hyperplane that the method explores. Second, in reaction SUCDi *FVA* returns an interval instead of a single flux value. That's because SUCDi has a exactly opposite reaction (FRD7). It is an issue commented in Section 4.6.3 for two other couples or reactions that appear in TCA cycle as well. It basically manifests the poor performance of constraint-based methods when dealing with closed loops. In relatively small models such as the core *E. coli* model this issue is manageable but when dealing with genome-scale models, this problem becomes much more relevant.

The two pathways shown in this section for scenario 1 expose clearly the general capabilities of each approach. *MFA* produces good results starting with the least amount of information. However, big chunks of its solution are formed by biologically unrealistic solution. *MFAg* is much more precise in its predictions. Reactions usually flow in the most reasonable direction and with more accurate intervals. The addition of growth rate information appears to significantly increase the quality of the predictions. Finally *FVA* produces the best results, generating a single flux value for almost every reaction. The critical assumption in which is based, that the biomass growth is an objective that the cell tends to maximize, seems to be correct in the collection of scenarios gathered.

4.6.5 Statistical description and comparison

In this section the flux distributions of each method for each scenario are shown and analysed using the same ratios used for scenario 1 in Section 4.6.3. The main purpose is to check if the reduction of the solution space is highly dependent of each specific scenario or if the trend observed in scenario 1 is stable across the complete dataset. In order to do that the same ratios $\mathbf{r}^{MFAg/MFA}$ and $\mathbf{r}^{FVA/MFA}$ are used to relate each solution with each other. Figure 4.14 shows both ratios for each scenario studied. Note that in scenarios 12-32 and 36 *MFAg* is not able to find a valid solution and therefore the $\mathbf{r}^{MFAg/MFA}$ is not calculated. The average $\mathbf{r}^{MFAg/MFA}$ in the other 22 scenarios is 41.75% and its standard deviation is 12.57%. Therefore the reduction in the feasible solution space is similar across the complete dataset. The values for $\mathbf{r}^{FVA/MFA}$ are much smaller, as it was observed in the example of scenario 1. The average $\mathbf{r}^{FVA/MFA}$ in the complete dataset of 44 scenarios is 1.92% and its standard deviation is 4.54%.

There are only 3 out the 44 scenarios in which the size of the *FVA* solution is significantly wider. Those are scenarios 4, 6 and 8. The reasons for this discrepancy remain unclear. The only source of variability between scenarios is the magnitude of the input/output measurements. The relationship among the measurements must locate the *FVA* solution in a more wide region of the solution space. In scenario 4 for example, the amount of O₂ consumed and CO₂ produced is higher than in other scenarios with the same input of D-glucose. Scenarios 1 and 12 show a much lower measurements for O₂ and CO₂. A similar situation occurs in

scenario 8, with much higher measurements for O_2 and CO_2 than the resembling scenario 17.

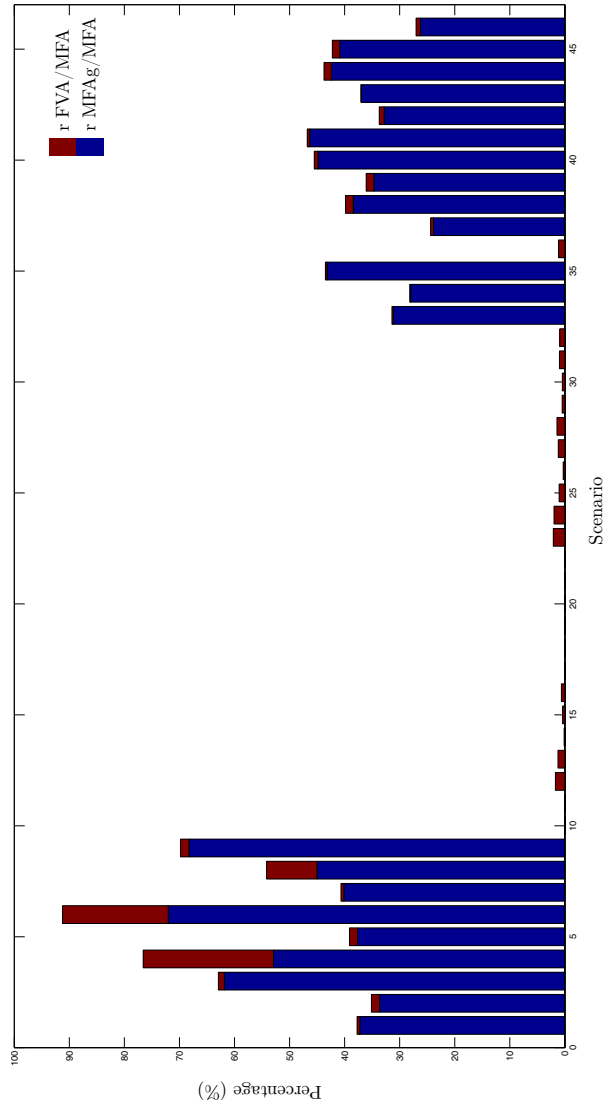


Figure 4.14. Ratios $r^{MFAg/MFA}$ and $r^{FVA/MFA}$ for the complete dataset (in stacked bars).

Scenario 6 is probably the most intriguing. It contains measurements for D-glucose, acetate, O₂ and CO₂. Acetate production seems specially relevant; this feature happens in other 27 scenarios. None of the other scenarios with acetate production has such a wide flux intervals in their solution. However, most of these scenarios contain exclusively data of D-glucose intake and acetate production. Scenarios 3, 7 and 9 are the only ones that contain the amount of O₂ consumed and CO₂ produced besides D-glucose and acetate (besides scenario 6). This limits the comparisons with other alike scenarios. In scenario 6 the proportion of O₂ intake against the carbon output (in the form of CO₂ and acetate) seems disproportionately high. However, the lack of close scenarios limits the possible explanation of the particularities of the *FVA* in scenario 6. Additional experiments around the same point should be carried out to clarify its behaviour.

4.6.6 Variability analysis

This final Results section tries to answer two questions:

- *What are the relationships among the original variables of the model?*
- *Is there any relationship between the original variables and the results of the three different approaches?*

In order to do that, a two step process is taken. First, the model itself is described in statistical terms and the relationship between the measured original variables are studied. Then, it is studied and analysed if this set of variables explains the variability of the outcome parameters obtained through the methods. These outcome parameters are $\mathbf{r}^{MFAg/MFA}$ and $\mathbf{r}^{FVA/MFA}$.

Original variables variability: Principal Component Analysis

The first step of analysing the variability and dependence among the original values is carried out by performing a Principal Component Analysis (PCA). Its main features and mathematical roots are described in Section 2.3.3. Figures 4.15 and 4.16 show the results for the PCA analysis with the complete dataset of original variables. Two components explain 95% of the variability of the original values (see Figure 4.15a).

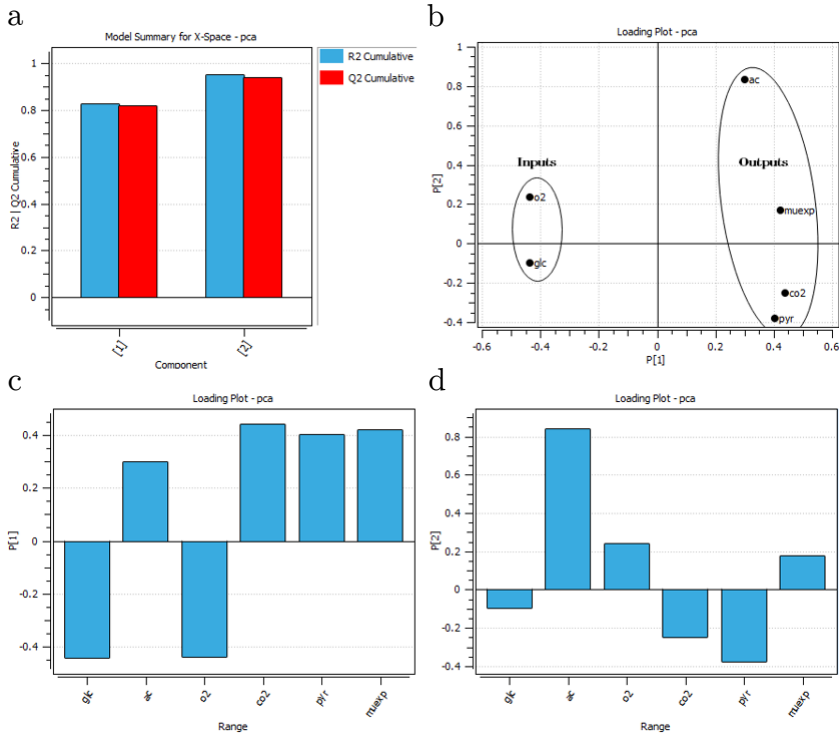


Figure 4.15. Principal Component Analysis of the dataset original variables: growth rate, intake of D-glucose and O_2 and production of CO_2 , acetate and pyruvate. **a** Principal components found. **b** Loading plot of the two components. **c** Loading plot of component 1. **d** Loading plot of component 2.

The fact that the dataset variability is explained almost totally by only a few components is to be expected. The original variables measure the inputs and outputs of *E. coli* cultures and it is only logical that there is a strong correlation between those measurements. The first component (C1) illustrates the main features of the *E. coli* metabolism very clearly. There is a strong correlation between the substrates the cell takes from the environment and the substances it produces. The more O_2 and D-glucose it takes, the more CO_2 , acetate and pyruvate produces and the higher is the growth rate. In Figure 4.15b, in the x-axis the loading of C1 (P1) illustrates this behaviour clearly. Note that the inputs have negative values because the model defines all inputs in the model as negative fluxes. In this case this facilitates the visualization. This feature is shown very clearly as well in Figure 4.15c, which is just an alternative representation of the loading plot focusing exclusively on P1. This component represents the main or core carbon metabolism of *E. coli*.

The loading of the second component (P2, in the y-axis of the same panel) paints a metabolic feature with a more subtle interpretation. The only variable strongly related to the component is the production of acetate (see Figure 4.15b and d). The other two variables significantly related with C1 are the production of CO₂ and pyruvate. These components tells us that the higher the production of acetate, the lower the production of CO₂ and pyruvate will be. This makes biological sense too, because all the carbon that enters the cell in the form of D-glucose is either consumed for growth or has to exit the cell as CO₂ (traditionally) or acetate or pyruvate (under determined circumstances). This component seems to describe clearly what is known as overflow metabolism. This refers to the apparent wasteful cellular energetic strategy of incompletely oxidize its growth substrate (D-glucose in this case) instead of using a more energetically efficient respiratory pathway, even in the presence of O₂ (which is the case here since all scenarios are aerobic). As a result of using this metabolic tactic, the cells excrete metabolites such as acetate or lactate.

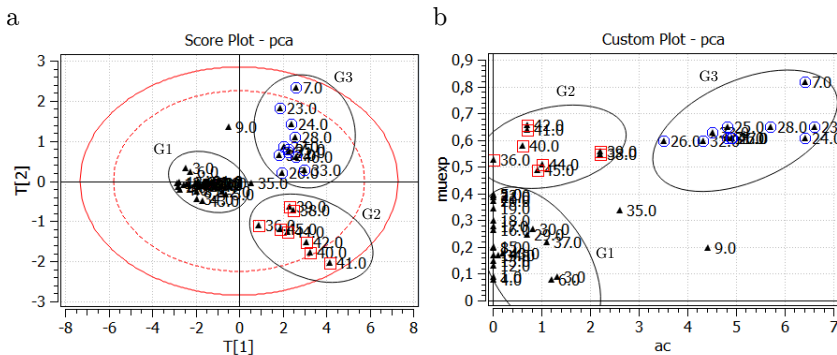


Figure 4.16. Different metabolic regimes found in the dataset through the PCA analysis. **a** Score plot of both components. **b** Plot of the experimental growth rate against the acetate secretion. Circles and colors highlight different metabolic regimes in both panels.

Overflow metabolism yields less ATP than the complete oxidation route through cellular respiration and it is a common feature across fast-growing cells. For instance, in the dataset studied in this work 28 out of 46 scenarios present some sort of overflow metabolism (usually acetate although in some cases pyruvate too). Recent research (Basan et al. 2015; Molenaar et al. 2009) has offered a general explanation for the association of overflow metabolism with fast growth. This theory postulates that enzymes required for respiration are more expensive in terms of cellular resources than those required for a partial oxidation of glucose. Given that cells have limited available energetic resources and a fixed physical volume for enzyme production, there seems to be a trade-off between a more efficient energy processing through central metabolism (respiration) and a faster growth

achieved through partial-oxidation metabolism (fermentation). The two principal components found in this dissertation support this theory: cells define their metabolic state in a dual relationship between central carbon and partial-oxidation metabolism.

Figure 4.16a shows the score plot for both principal components. It is easy to see that the 46 scenarios are mainly divided in three big clusters. Figure 4.17b displays those clusters according to their growth rate and acetate secretion. The first cluster G1 includes scenarios with a low growth rate and low or no acetate secretion. This constitutes the energy-efficient metabolic regime set by mostly central carbon metabolism. The second cluster G2 shows scenarios with higher growth rate and a slight increase in acetate production. This cluster represents the interface between both regimes. In order to grow faster the cell must start to abandon the efficient regime and start to dive into the raw maximization of growth rate. Cluster G3 shows precisely this regime of inefficient but faster growth that produces secretion of large quantities of acetate. Figure 4.16b is actually very similar to Figure 1 in Basan et al. 2015 where authors evaluate the exact point of regime shifting.

Outcome variables variability: Partial Least Squares regression

To analyse the dependence of the ratios that describe the performance of the methods with the original variables of the dataset a Partial Least Squares regression was carried out. Two different PLS were performed: PLS1 (Figure 4.17a) which studies the relationship between the original variables and the ratio $\mathbf{r}^{MFAg/MFA}$ and PLS2 (Figure 4.17b) that analyses the influence of the original variables on the ratio $\mathbf{r}^{FVA/MFA}$. Using the *MFA* solution as benchmark, PLS1 studies the performance of *MFAg* and PLS2 does it for the *FVA* approach. Both analyses can be made together but they were divided in two different PLS to facilitate interpretation.

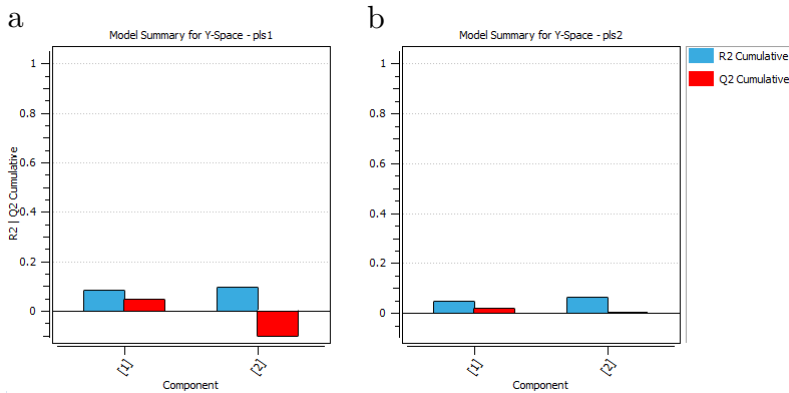


Figure 4.17. PLS regressions for $r^{MFAg/MFA}$ (a) and $r^{FVA/MFA}$ (b).

Both PLS show components that barely explain the 10% of the outcome ratios variability. Furthermore, the second component even decreases the predictive capability of the analysis (a usual stop point for PLS occurs when Q2 Cumulative decreases when a new component is added). The results are very clear: there is no correlation whatsoever between the variability of the original variables and the variability of the methods used to solve the problem. That means that no matter what value the original variables take, the methods are always going to obtain very similar results. In other words, the results obtained through the three approaches are very stable and robust across the complete dataset. Visually checking the ratios obtained across the whole dataset, it was reasonable to infer this conclusion since they seemed very similar for all scenarios but a rigorous statistical analysis was needed to ensure the validity of the conclusion. Additional work should be focused on expanding the dataset to cover other possible conditions for *E. coli* such as anaerobic conditions, alternative carbon sources or inorganic ions (phosphate an ammonium mainly) and to check if the conclusion of robustness still holds.

4.7 Conclusions

This final section aims to summarize the most relevant conclusion drawn from this chapter and propose some future work that might enrich and advance further the present work.

Constraint-based models and in this particular case the core *E. coli* model show an excellent agreement with the experimental dataset. The model is able to properly represent 95.65% of experimental scenarios (44 out of a total of 46). Constraint-based model constitute a perfectly valid approach to model metabolism in steady

state conditions. Additional work would require to extend even further the scope of the dataset to check if the model is still able to work. On the other hand, larger models (ideally genome-scale) could be used to check if they improve the level of agreement with the experimental data.

The core of the work was to obtain the flux distributions for all the scenarios with the three different approaches. The approaches, as explain in Section 4.5, use different combinations and premises of the three techniques FBA, FVA and MFA. The first approach, which is purely *MFA* is able to obtain distributions for the 44 scenarios in which the model work. The second approach, *MFAg* which is analogous to *MFA* but using the growth rate data as well is only able to solve the problem in 22 scenarios. Finally, the *FVA* approach, which uses FBA and FVA consecutively, is able to find a solution to 44 of the scenarios. Therefore, when taking into account purely the amount of scenarios that the approaches can tackle *MFA* and *FVA* achieve a 100% and *MFAg* only a 50%.

The flux distributions obtained through each approach are very similar among them. All *MFA* solutions look alike. This also happens for *MFAg* and *FVA* solutions. Just by looking at the shape of the intervals (shown in Annex 7.3) it is evident that each approach finds solutions very similar across the complete dataset. The size or width of the solutions changes significantly, though. *MFA* offer the widest solutions, being many of them biologically infeasible. *MFAg* returns narrower (the average $\mathbf{r}^{MFAg/MFA}$ is 41.75%) and more biologically sound solutions. Finally, *FVA* solutions return almost always a single value for each flux in the model (the average $\mathbf{r}^{FVA/MFA}$ in the complete dataset is 1.92%). They make the most biological sense too. From the size of the solutions they provide *FVA* would be the best, then *MFAg* and in the end *MFA*. This seems to validate the main assumption in which FBA and FVA are based: that cellular growth can be understood as a biomass production optimization process.

The *MFA* approach need the fewest assumptions and achieves reasonable results. Its solutions are very wide and it can reach every scenario in which the model works. The real utility of this method depends on the precision needed. If only a rough idea of the metabolic state of the cell is needed then it is a valid option. *MFAg* contains the additional information of the growth rate. Moreover, it assumes than growth can be described as a empirical reaction with substrates and products like any other regular metabolic reaction. Its results are much better than *MFA* in size but worse in scope. In practical terms, measuring growth rate is easy and inexpensive so it makes sense to try to add it as input in the model to get better results. If the addition of growth rate generates a system that the model is no able to successfully run, then a series of explanations arise. Some sort of measurement error may be happening or some larger more precise model may be needed. In general terms, I would try *MFAg* first if growth rate data is available. Finally *FVA* additionally assumes an optimality principle to restrain even further the solutions it achieves. According to the data used and the distributions obtained,

that assumption seems to be completely valid and a powerful tool to describe and predict the metabolic state of *E. coli*.

The Partial Least Squares regression indicates that there is no correlation between the variability of the original variables and the variability of the ratios that define the performance of the different approaches. The three approaches obtain similar results no matter the value of the original variables. Therefore, the conclusions about the performance of the different approaches are stable and robust. Additionally, the PCA analysis carried out was able to describe the system with two components. These components fit the main carbon metabolic route and the overflow metabolism. The behaviour of the cells seem to be defined by a combination of both, which agrees with previously reported metabolic regime description in *E. coli*.

Chapter 5

Metabolic graph analysis of *E. coli*

- *Yo soy yo y mi circunstancia.*

José Ortega y Gasset

Part of the contents of this chapter appeared in the following journal articles:

- M. Beguerisse-Díaz et al. (2016). “Context-dependent metabolic networks”.
In: *arXiv:1605.01639 [physics, q-bio]*

5.1 Introduction

This final chapter of the thesis is devoted to improve the network representation and analysis of metabolic networks. The two main topics around which the chapter is structured are:

- How to accurately mathematically describe metabolic networks as graphs and how to include the maximum amount of available biological information in those graphs.
- How to use those graphs for a simple *E. coli* metabolic model (the same used in Chapter 4) and how use community detection methods to find an insightful biological network structure of the bacteria.

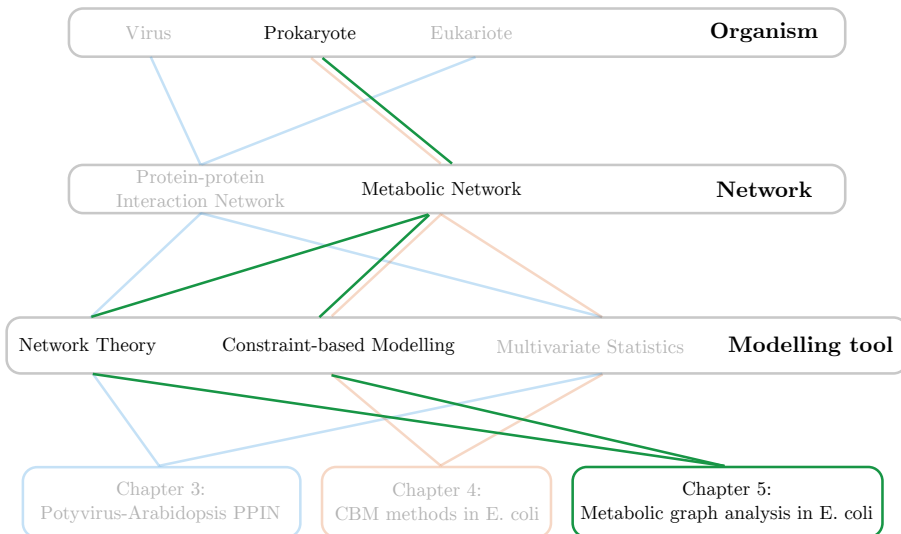


Figure 5.1. Chapter roadmap.

The first two sections (Summary and Background) address the main objectives of the work carried out in this chapter and the basics of graph representations of metabolic networks. The third section (Methodology) explains briefly the building blocks necessary to achieve the proposed goals. Those are mainly Flux Balance Analysis (see Section 4.5.2 for a detailed description) and Markov Stability community detection method (see last subsection in Section 2.3.1). However, most of the construction of the new graphs is placed in the Results and discussion section.

The fourth section (Results and discussion) shows the bulk of the work done. First, the construction of the matrices and graphs is thoroughly described. Then, the

graphs for the core *E. coli* model are built for several different biological scenarios or conditions. Finally, the community structure of these networks is analysed. The last section (Conclusions) draws the main conclusions from this chapter.

5.2 Summary

Cells adapt their metabolism to survive changes in their environment. A framework for the construction and analysis of metabolic reaction networks that can be tailored to reflect different environmental conditions is presented. Using context-dependent flux distributions from Flux Balance Analysis (FBA), directed networks with weighted links representing the amount of metabolite flowing from a source reaction to a target reaction per unit time are produced. Such networks are analyzed with tools from network theory to reveal salient features of metabolite flows in each biological context. This approach is illustrated with the directed network of the central carbon metabolism of *Escherichia coli*, and the study of its properties in four relevant biological scenarios. Results show that both flow and network structure depend drastically on the environment: networks produced from the same metabolic model in different contexts have different edges, components, and flow communities, capturing the biological re-routing of metabolic flows inside the cell. By integrating FBA-based analysis with tools from network science, results provide a framework to interrogate cellular metabolism beyond standard pathway descriptions that are blind to the environmental context.

5.3 Background

Metabolism consists of the set of reactions responsible for converting nutrients into energy and macromolecules that fuel cellular processes (Berg, Tymoczko, and Stryer 2002). These reactions are organised in metabolic pathways that are composed of metabolites such as carbon sources and intermediate precursors, and the enzymatic reactions that convert them into one another. Metabolism is naturally amenable to analysis with network science, which has also been successfully used to describe cellular systems such as protein-protein interactions (Thomas et al. 2003), transcriptional regulation (Alon 2007), and protein structure (Amor et al. 2014). In the case of metabolism, however, the insights gained from network science are somewhat more dispersed because of the lack of a widely-accepted method to construct metabolic networks. Most studies have focussed on the topological properties of metabolic pathways, such as their degree distribution (Arita 2004; Jeong et al. 2000; Wagner and Fell 2001) or their community structure (Guimera and Nunes Amaral 2005; Ravasz et al. 2002; Takemoto 2013; Zhou and Nakhleh 2012), but the conclusions are highly dependent on the way the networks are constructed (Winterbach et al. 2013).

The central object of study in network science is a *graph* that describes the entities in the network (nodes or vertices) and the connections among them (edges or arcs), which can indicate relationship or interaction (Newman 2010). Note that although in the networks literature the terms *graph* and *network* are used interchangeably; this is not the case in contexts such as the study of metabolism. To maintain consistency with the metabolic analysis literature and to avoid confusion the term *network* is reserved for the set of metabolic reactions and substrates, and the term *graph* for the mathematical object formed by a set of nodes and their connections. In the analysis of metabolism there is no unique way to construct graphs in which the nodes are chemical species or reactions and the connections. Catalytic enzymes convert multiple reactants into products with the help of other species in the network. Moreover, some enzymes catalyse several reactions and some reactions are catalysed by multiple enzymes. One can construct fundamentally different graphs for the same metabolic network. For example, one could create a graph in which the nodes are metabolites and the edges are the reactions that transform one metabolite into another (Jeong et al. 2000; Ma and Zeng 2003a; Ouzounis and Karp 2000; Wagner and Fell 2001). Alternatively, we could create a graph where nodes are reactions and edges are the metabolites shared among them (Ma et al. 2004; Samal et al. 2006; Vitkup, Kharchenko, and Wagner 2006), or a bipartite graph with two types of nodes, one type for reactions and another one for metabolites, in which the connections denote metabolites that participate in reactions (Smart, Amaral, and Ottino 2008). These graph constructions have been described extensively (Palsson 2006).

Metabolic reactions operate in a preferred direction depending on the physiological state of the cell and environmental conditions (Berg, Tymoczko, and Stryer 2002). For example, in glucose-rich environments some glycolytic enzymes operate preferably in their forward mode, while during gluconeogenesis their catalytic activity reverses direction. Other reactions can only operate in one direction (i.e., they are irreversible). Although this notion of directionality is a key feature of metabolism, many of the existing graph constructions do not incorporate the direction of the edges (Palsson 2006; Wagner and Fell 2001). Undirected graphs neglect important information about the connectivity of the metabolic network, such as the distinction between reactions that compete for the same metabolite, the ones who produce the same metabolites, and those that have a supplier-consumer relationship.

Furthermore, current graph constructions are based on metabolic networks that simultaneously include all interactions from all known pathways in an organism. Although the resulting networks are the blueprint for the whole metabolic activity of a cell, in reality cells switch specific pathways on/off to sustain their energetic budget in different environments. Such *blueprint graphs* therefore contain many node connections that are relevant only in specific growth conditions, which distorts the network topology and the biological insights drawn from it.

In this chapter two new graph constructions are proposed to study metabolic networks and their adaptation to the environment in which the connections among the reactions and their intensity have a clear interpretation. The first of these graphs is the *Probabilistic Flux Reaction Graph* (PRG or \mathbf{D}_p), a weighted, directed graph in which the nodes are reactions, and connections occur between reactions that have a supplier-consumer relationship. The weight of the connections corresponds to the probability that a metabolite chosen uniformly at random is produced by the source reaction and consumed by the target in the absence of more information about the environmental context. The second graph is the *Flux-Balance Graph* (FBG or \mathbf{M}_v), a directed, weighted graph where the nodes are again the reactions, but the weight of the connections is the total flow of metabolites per unit time from the source reaction to the target in a specific environmental context. To obtain the edge-weights flux distributions from Flux Balance Analysis (Orth, Thiele, and Palsson 2010; Rabinowitz and Vastag 2012) are used. In both graphs, an edge between nodes indicates that metabolites are produced by the source reaction and consumed by the target reaction. This definition accounts for metabolic directionality and thus captures the natural flow of chemical mass from carbon sources to metabolic products.

One advantage of the Probabilistic Flux Reaction Graph is that the weight of the connections created by metabolites that appear in many reactions (e.g., *pool* such as ATP, NADH and other co-factors) is very small, but not zero; this is a result of the probabilistic formulation of the graph connectivity. Pool metabolites typically lead to graphs dominated by the connections created by them, obscuring other more informative features of the network. A common workaround in the literature is to prune pool metabolites from the network, but this is done with *ad-hoc* heuristics (Kreimer et al. 2008; Ma and Zeng 2003b; Samal and Martin 2011; Silva et al. 2008), and arbitrarily destroys information that may affect the interpretation of the results. The approach presented here, motivated by results in Croes et al. 2006, circumvents this issue by appraising the information contained in the interactions generated by all metabolites in a consistent manner without the need to remove any species from the analysis.

The Flux Balance Graph incorporates the environmental context into the graph connectivity by using Flux Balance Analysis (FBA), a widespread method to predict metabolic fluxes in genome-scale metabolic networks (Orth, Thiele, and Palsson 2010; Rabinowitz and Vastag 2012). Given upper and lower bounds for the flux of each reaction, FBA predicts flux distributions that maximise an objective function, such as the growth rate of the organism, although others are possible (Schuetz, Kuepfer, and Sauer 2007). The flux of the reaction is subject constraints that describe the state of the environment. Optimal environment-dependent fluxes from FBA are incorporated into the FBG by defining the weight of the connections to be the total flux of metabolites produced by the source reaction and consumed by the target reaction. As a result, the edge weights in the FBG can be directly

interpreted as fluxes in units of mass per time ($\frac{\text{mmol}}{\text{g}_{\text{DW}} \cdot \text{h}}$). The FBG thus allows for a systematic study of genome-scale metabolic adaptations in response to changing carbon sources and other environmental perturbations.

We showcase the utility of the PRG and FBG in the core model for the metabolism of *Escherichia coli* metabolism (Orth, Fleming, and Palsson 2010). Our results show that the structure of the FBG can vary dramatically depending on the environmental context under consideration. The FBG for different environmental conditions was constructed and Markov stability method (Delvenne, Yaliraki, and Barahona 2010) was used to detect node communities in different timescales. The results suggest that the structure of metabolic networks varies drastically across different environmental conditions, casting doubt on the utility of a single metabolic blueprint to describe cellular adaptations. The proposed graphs can be readily applied to study the topology of genome-scale metabolic networks (Ravasz et al. 2002), find network clusters and their link to environmental changes (Ma et al. 2004; Samal et al. 2006; Takemoto 2013), and to assess the robustness of metabolic connectivity (Smart, Amaral, and Ottino 2008). The objective is that these new constructions for metabolic graphs stimulate new applications of network science to cellular metabolism.

5.4 Methodology

The two methods used in this chapter are the flux estimating method Flux Balance Analysis (see Section 4.5.2) and the community detection method Markov Stability (see Section 2.3.1). A very brief summary of both methodologies is shown in this section.

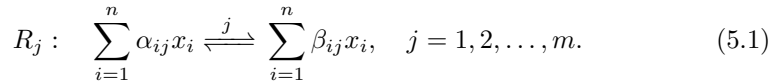
Flux Balance Analysis (FBA) (Orth, Thiele, and Palsson 2010; Rabinowitz and Vastag 2012) is a widely-adopted approach to analyse metabolism and cellular growth. FBA calculates the reaction fluxes that optimise growth in specific biological contexts. The main hypothesis behind FBA is that cells adapt their metabolism to maximise growth in different biological conditions. The conditions are encoded as constraints on the fluxes of certain reactions; for example, exchange reactions that import nutrients and other necessary compounds from the exterior. On the other hand, the communities in each network were extracted using the Markov Stability (MS) community detection framework (Delvenne, Yaliraki, and Barahona 2010; Delvenne et al. 2013). This framework uses diffusion processes on the network to find groups of nodes (i.e., communities) that retain flows for longer than one would expect on a comparable random network; in addition, MS incorporates directed flows seamlessly into the analysis (Beguerisse-Díaz et al. 2014; Lambiotte, Delvenne, and Barahona 2014).

5.5 Results and discussion

5.5.1 Graphs of metabolic networks: constructing graphs that incorporate directionality and context

The objective is to create graph descriptions of metabolic models in which the connections among the reactions have a clear interpretation, and can thus be interrogated with a wide range of tools from network science. To illustrate the construction of the different graphs described in this chapter, a toy example (Figure 5.2a) of a metabolic network that describes nutrient uptake, biosynthesis of metabolic intermediates, secretion of waste products, and biomass production (Rabinowitz and Vastag 2012) is used, all of which are key components of genome-scale metabolic models.

Let's consider metabolic networks composed of n metabolites and m reactions:



The numbers α_{ij} and β_{ij} are the depletion and production coefficients of the species in each reaction. The evolution of the metabolite concentrations follows the differential equation

$$\dot{\mathbf{x}} = \mathbf{S}\mathbf{v}, \quad (5.2)$$

where $\mathbf{x}(t)$ is an n -dimensional vector of metabolite concentrations $x_i(t)$ and \mathbf{v} is an m -dimensional vector of reaction rates. The $n \times m$ matrix \mathbf{S} is the stoichiometric matrix with entries $s_{ij} = \beta_{ij} - \alpha_{ij}$, that is, the net number of x_i molecules produced ($s_{ij} > 0$) or consumed ($s_{ij} < 0$) by the j -th reaction. Due to chemical or thermodynamic constraints, some reactions in \mathbf{v} are known to be irreversible (i.e., $v_j \geq 0$ for certain reactions R_j). This information can be summarised in an m -dimensional reversibility vector \mathbf{r} with entries

$$r_i = \begin{cases} 1 & \text{if } R_i \text{ is reversible,} \\ 0 & \text{otherwise.} \end{cases} \quad (5.3)$$

A unipartite graph with m reaction nodes can be described by its $m \times m$ adjacency matrix with entries that represent the connection between nodes i and j . In contrast, a bipartite graph has two types of nodes and connections only occur between nodes of different type, resulting in an $n \times m$ adjacency matrix. One of the simplest ways (Palsson 2006) to construct a graph from a metabolic model is

to create a bipartite graph with adjacency matrix $\widehat{\mathbf{S}}$, a boolean version of \mathbf{S} such that

$$\hat{s}_{ij} = \begin{cases} 1 & \text{if } s_{ij} \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

This graph connects metabolites to reactions based on whether metabolites participate on a given reaction, either as reactants or products. Figure 5.2B shows the bipartite graph associated to the toy model. The most common approach to construct a unipartite graph of reactions from $\widehat{\mathbf{S}}$ is via the $m \times m$ adjacency matrix

$$\mathbf{A} = \widehat{\mathbf{S}}^T \widehat{\mathbf{S}}. \quad (5.4)$$

In the graph \mathbf{A} , two reaction nodes are connected if they share metabolites as reactants or products (Figure 5.2b), while self loops represent the total number of metabolites that participate in a reaction. Though widely studied (Palsson 2006; Wagner and Fell 2001), the graph \mathbf{A} suffers from three key limitations: (i) It does not distinguish between forward and backward flow of metabolites between reactions (its adjacency matrix is symmetric), and hence it cannot incorporate information on the reversibility of reactions. (ii) It does not distinguish between connections caused by important metabolites, such as carbon sources or metabolic intermediates, and connections caused by pool metabolites that participate in numerous reactions, such as water, ions or enzymatic cofactors. As a consequence, \mathbf{A} has a large number of connections that obscure its structure and the connectivity between its reaction nodes. (iii) It is a rigid description that does not account for varying cellular contexts such as changes in carbon sources, growth conditions or environmental shocks.

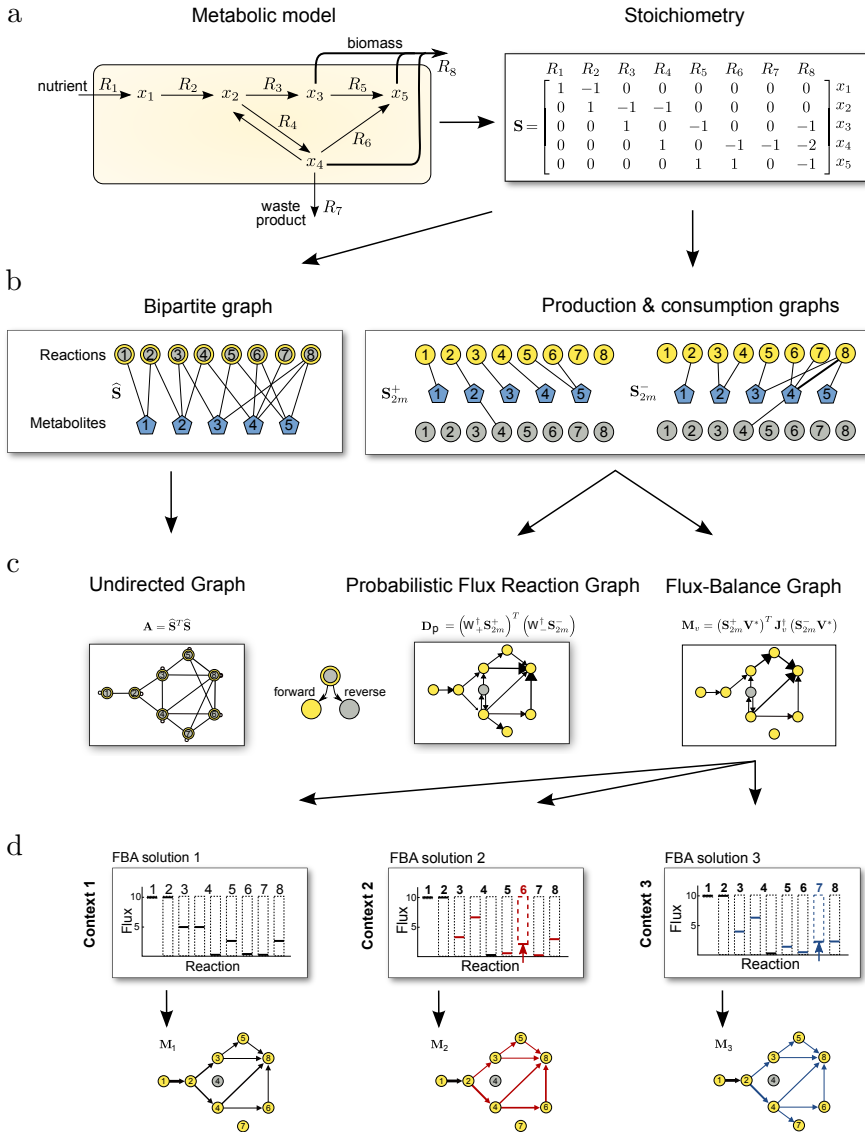


Figure 5.2. Different graph representations for metabolic networks (caption continues in the next page).

Figure 5.2. (a) Toy model of a metabolic network describing nutrient uptake, biosynthesis of metabolic intermediates, secretion of waste products, and biomass production (Rabinowitz and Vastag 2012). Reaction R_4 is reversible and reaction R_8 represents biomass production defined as $R_8 : x_3 + 2x_4 + x_5$. Associated stoichiometric matrix to the model (\mathbf{S}). (b) Direct construction of graphs from the stoichiometric matrix of the toy model. The bipartite graph is defined by the boolean matrix $\widehat{\mathbf{S}}$. By splitting reactions into their forward and backward rates, the production (\mathbf{S}_{2m}^+) and consumption (\mathbf{S}_{2m}^-) bipartite graphs are built. (c) The traditional undirected reaction network (Pals-son 2006) can be constructed from $\widehat{\mathbf{S}}$. The proposed Probabilistic Flux Reaction Graph (\mathbf{D}_p) is an edge re-weighted version of \mathbf{D} (see text) developed to equalize the importance of each metabolite. Finally, the Flux-Balance Graphs (\mathbf{M}_v) balance the weight of each metabolite according to the FBA reaction flux distributions. Annex 7.4 includes all the matrices to create these networks. (d) Three examples of Flux-Balance Graphs (\mathbf{M}_v) for the toy example. The graphs were constructed with equation (5.15) applied to the stoichiometric matrix shown in Figure 5.2a and with three solutions from FBA under different constraints. In contexts 2 and 3, the lower flux bounds for reactions 6 and 7 were perturbed, causing changes in the flux distribution that translate into different weighting of the graph edges. The dashed boxes represent the flux bounds employed in each case, and solid lines describe the optimal fluxes predicted by FBA in units of $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. In all cases, the optimal flux for reaction R_4 is positive and thus the corresponding backward reaction node R_4 is disconnected.

To address the limitations of the graph \mathbf{A} , the construction of two types of graphs that can be applied to any stoichiometric model is proposed:

- A *directed graph in the absence of context*, yet capturing flows, that penalises the connections caused by pool metabolites and that has a concrete probabilistic interpretation. The proposed graph that meets those conditions is the *Probabilistic Flux Reaction Graph* (PRG or \mathbf{D}_p).
- A family of *metabolic graphs* that are context-dependent and that re-balance the edge weights according to flux distributions predicted by Flux Balance Analysis (FBA). The proposed family of graphs are the *Flux-Balance Graphs* (FBG or \mathbf{M}_v), which depend on a vector v of fluxes obtained through FBA.

In both graphs, directionality of reactions is included by redefining the connections between reaction nodes, namely: *two reactions are connected if one produces a metabolite that is consumed by the other*. This definition leads to graphs that naturally account for the reversibility of reactions and allow for a seamless integration of different biological contexts modelled through FBA.

Inspired by techniques for the analysis of kinetic models for biochemical reaction networks (Chellaboina et al. 2009), one first decomposes the reaction vector into its forward and backward rates as $\mathbf{v} = \mathbf{v}^+ - \mathbf{v}^-$, where \mathbf{v}^+ and \mathbf{v}^- are both non-negative. For irreversible reactions, the backward rates are $\mathbf{v}^- = 0$, and thus the following relation holds:

$$\begin{aligned}\mathbf{v} &= \mathbf{v}^+ - \mathbf{v}^-, \\ &= \mathbf{v}^+ - \text{diag}(\mathbf{r}) \mathbf{v}^-, \end{aligned}$$

where \mathbf{r} is the reversibility vector defined in (5.3). The metabolic model is then rewritten as

$$\dot{\mathbf{x}} = \mathbf{S}\mathbf{v} = \underbrace{[\mathbf{S} \quad -\mathbf{S}]}_{\mathbf{S}_{2m}} \underbrace{\begin{bmatrix} \mathbf{I}_m & 0 \\ 0 & \text{diag}(\mathbf{r}) \end{bmatrix}}_{\mathbf{v}_{2m}} \begin{bmatrix} \mathbf{v}^+ \\ \mathbf{v}^- \end{bmatrix}, \quad (5.5)$$

where \mathbf{I}_m is the $m \times m$ identity matrix. The matrix \mathbf{S}_{2m} and vector \mathbf{v}_{2m} are augmented versions of original stoichiometric matrix \mathbf{S} and reaction vector \mathbf{v} . As a result, the augmented model in (5.5) has n metabolites and $2m$ reactions. In the following sections we detail how to construct the proposed graphs over the augmented set of reaction nodes.

Directed Probabilistic Flux Reaction Graphs of metabolism in the absence of context

To construct directed graphs first the auxiliary *production* and *consumption* stoichiometric matrices are extracted from \mathbf{S}_{2m} :

$$\begin{aligned}\mathbf{S}_{2m}^+ &= \frac{1}{2} (\text{abs}(\mathbf{S}_{2m}) + \mathbf{S}_{2m}), \\ \mathbf{S}_{2m}^- &= \frac{1}{2} (\text{abs}(\mathbf{S}_{2m}) - \mathbf{S}_{2m}), \end{aligned} \quad (5.6)$$

where $\text{abs}(\mathbf{S}_{2m})$ is a matrix whose entries are the absolute values of \mathbf{S}_{2m} . Note that the matrices satisfy $\mathbf{S}_{2m} = \mathbf{S}_{2m}^+ - \mathbf{S}_{2m}^-$. The entries s_{ij}^+ of the matrix \mathbf{S}_{2m}^+ are the number of molecules of metabolite x_i produced by reaction R_j , whereas the entries s_{ij}^- of \mathbf{S}_{2m}^- are the number of molecules of metabolite x_i consumed by reaction R_j .

Similarly, as in the bipartite graph $\widehat{\mathbf{S}}$, the boolean versions of the matrices \mathbf{S}_{2m}^+ and \mathbf{S}_{2m}^- define two bipartite graphs, a production and a consumption graph, shown in Figure 5.2b for the toy metabolic model. From these two matrices the adjacency matrix of a *directed* graph is constructed:

$$\mathbf{D} = \widehat{\mathbf{S}}_{2m}^{+T} \widehat{\mathbf{S}}_{2m}^-, \quad (5.7)$$

where the tilde denotes the boolean versions of the respective matrices. The entries d_{ij} of \mathbf{D} represent the total number of metabolites *produced* by reaction R_i that are *consumed* by reaction R_j . The adjacency matrix of \mathbf{D} has size $2m \times 2m$ and defines a directed graph on the space of reactions split into their forward and backward components. In contrast to the undirected graph \mathbf{A} , the adjacency matrix of \mathbf{D} is not symmetric and thus captures the existence of irreversible reactions in the original metabolic model. Note that \mathbf{D} and \mathbf{A} are not directly comparable as they are defined on different sets of nodes. We can obtain a directed graph on the same set of nodes as \mathbf{A} with the $m \times m$ adjacency matrix

$$\mathbf{A}_{\text{dir}} = \mathbf{C}^T \mathbf{D} \mathbf{C}, \quad (5.8)$$

where $\mathbf{C} = [\mathbf{I}_m \quad \mathbf{I}_m]^T$. The directed graph defined by \mathbf{A}_{dir} contains the connections in \mathbf{A} , but excludes spurious edges caused by the non-existent backward rates in irreversible reactions. If the metabolic model contains only reversible reactions, i.e., when the reversibility vector \mathbf{r} contains only ones, from equations (5.5) and (5.8) $\mathbf{A}_{\text{dir}} = \mathbf{A}$ is recovered.

Although an improvement on \mathbf{A} , the graph \mathbf{D} still is limited by the fact that uninformative connections created by pool metabolites have the same weight as connections from other more informative metabolites. To ameliorate their effect on the structure of a metabolic network, a common approach is to remove such highly-connected metabolites from the network description (Ma and Zeng 2003b; Samal and Martin 2011; Silva et al. 2008). However, removing metabolites can change the network structure drastically, and without a clear definition of which pool metabolites should be pruned, it is typically done with heuristics in an *ad-hoc* manner that depends on the particular network and scientific question at hand. As an alternative, instead of advocating the ad hoc removal of pool metabolites, their connections are weighted according to their relative importance (Croes et al. 2006).

More specifically, a probabilistic formulation is proposed to quantify the contribution of metabolite x_k to the weight of the connection from reaction R_i to R_j . The probability that a molecule of x_k chosen at random is produced by R_i and consumed by R_j is

$$P(\text{a molecule of } x_k \text{ is produced by } R_i \text{ and consumed by } R_j) = \frac{s_{ki}^+ s_{kj}^-}{w_k^+ w_k^-},$$

where $w_k^+ = \sum_{h=1}^{2m} s_{kh}^+$ and $w_k^- = \sum_{h=1}^{2m} s_{kh}^-$ are the total number of molecules of x_k produced and consumed by *all* reactions. We collect this information in the vectors

$$\mathbf{w}^+ = \mathbf{S}_{2m}^+ \mathbf{1}_{2m} \quad \text{and} \quad \mathbf{w}^- = \mathbf{S}_{2m}^- \mathbf{1}_{2m},$$

where $\mathbb{1}_{2m}$ is an $2m$ -dimensional vector of ones. For example, if x_k is only produced by R_i and only consumed by R_j then $s_{ki}^+ = w_k^+$, $s_{kj}^- = w_k^-$, and the probability that any molecule of x_k flows from R_i to R_j is 1.

We propose constructing a weighted, directed Probabilistic Flux Reaction Graph (PRG) with adjacency matrix

$$\mathbf{D}_p = \frac{1}{n} \left(\mathbf{W}_+^\dagger \mathbf{S}_{2m}^+ \right)^T \left(\mathbf{W}_-^\dagger \mathbf{S}_{2m}^- \right), \quad (5.9)$$

where \mathbf{W}_+^\dagger and \mathbf{W}_-^\dagger are the pseudo-inverses of $\text{diag}(\mathbf{w}^+)$ and $\text{diag}(\mathbf{w}^-)$, respectively. The weight of the connection from R_i to R_j is

$$d_{p_{ij}} = \frac{1}{n} \sum_{k=1}^n \frac{s_{ki}^+}{w_k^+} \cdot \frac{s_{kj}^-}{w_k^-}, \quad (5.10)$$

which is the probability that *any metabolite molecule* chosen at random is produced by R_i and consumed by R_j . The construction of \mathbf{D}_p is analogous to \mathbf{D} in (5.7), but it has the important difference that the connections have a clear probabilistic interpretation and accurately quantify the importance of the relation between any pair of reactions. It is clear from equations (5.9) and (5.10) that entry-wise 1-norm of the graph's adjacency matrix is

$$\|\mathbf{D}_p\|_1 = \sum_{i,j} d_{p_{ij}} = 1,$$

which is a consequence of the probabilistic formulation of this graph. The graph encoded in \mathbf{D}_p describes the probabilistic producer-consumer relationships of the reactions; the graphs that describe the two other types of relationship between reactions, *competition* and *synergy* can be constructed as well:

$$\mathbf{D}_c = \frac{1}{n} \left(\mathbf{W}_-^\dagger \mathbf{S}_{2m}^- \right)^T \left(\mathbf{W}_+^\dagger \mathbf{S}_{2m}^+ \right), \quad (5.11)$$

$$\mathbf{D}_s = \frac{1}{n} \left(\mathbf{W}_+^\dagger \mathbf{S}_{2m}^+ \right)^T \left(\mathbf{W}_-^\dagger \mathbf{S}_{2m}^- \right). \quad (5.12)$$

These two graphs are undirected, their edge-weights describe the probability that any two reactions consume (\mathbf{D}_c) or produce (\mathbf{D}_s) a metabolite picked uniformly at random.

Flux-Balance Graphs

To incorporate the effect of different cellular contexts on the graph of a metabolic network, the edges of the graph are re-weighted using the metabolic flux distributions predicted by Flux Balance Analysis (FBA). FBA computes a vector of fluxes \mathbf{v}^* that optimize a cellular objective, usually maximisation of biomass or growth, assuming that cellular metabolism is in quasi steady state with respect to the remaining cellular processes, i.e., $\dot{\mathbf{x}} = \mathbf{S}\mathbf{v} = 0$. A key characteristic of FBA is that specific environmental conditions can be included as upper and lower bounds on the reaction fluxes, which in turn act as constraints for the optimisation problem. These constraints describe, for example, the availability of nutrients, oxygen, and toxins. The key elements of FBA were summarized in Section 4.5.2.

The weight of an edge between reactions nodes R_i and R_j is defined as the *total flow of metabolites produced by R_i that are consumed by R_j* . Under such definition, the entries of the adjacency matrix of the Flux-Balance Graph (FBG), which is called \mathbf{M}_v , are

$$m_{ij} = \sum_{k=1}^n (\text{flow of } x_k \text{ produced by } v_i) \times \left(\frac{\text{flow of } x_k \text{ consumed by } v_j}{\text{total flow of } x_k} \right). \quad (5.13)$$

The main assumption behind this definition is that the amount of metabolite produced by one reaction is distributed among the reactions that consume it in proportion to their flux. For example, if in the FBA solution the total flux of metabolite x_k is $10 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$, with reaction R_i producing x_k at a rate $1.5 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$ and reaction R_j consuming x_k at a rate $3.0 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$, then the mass flow of x_k from R_i to R_j is $1.5 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}} \times (3.0/10) = 0.45 \frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. The definition in equation (5.13) adds together the mass flows of *all* metabolites produced by R_i and consumed by R_j , and thus m_{ij} represents the *total flow* between the two reactions. Self loops describe the metabolic flux of autocatalytic reactions, i.e., those in which some products are also reactants.

The edge weights m_{ij} can be computed directly from the stoichiometric matrices \mathbf{S}_{2m}^+ and \mathbf{S}_{2m}^- defined in equation (5.6), and the FBA solution \mathbf{v}^* . The augmented flux vector is built

$$\mathbf{v}_{2m}^* = \begin{bmatrix} \mathbf{v}^+ \\ \mathbf{v}^- \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \text{abs}(\mathbf{v}^*) + \mathbf{v}^* \\ \text{abs}(\mathbf{v}^*) - \mathbf{v}^* \end{bmatrix},$$

which splits the forward and backward fluxes from the FBA solution in a similar way as \mathbf{v}_{2m} in equation (5.5). If the i -th and j -th entries of \mathbf{v}_{2m}^* are denoted as $v_{2m\ i}^*$ and $v_{2m\ j}^*$, respectively, we have that reaction R_i produces the metabolite x_k at a

rate $s_{ki}^+ v_{2m,i}^*$, while reaction R_j and consumes x_k at a rate $s_{kj}^- v_{2m,j}^*$. Substituting in equation (5.13) we get

$$m_{ij} = \sum_{k=1}^n s_{ki}^+ v_{2m,i}^* \times \left(\frac{s_{kj}^- v_{2m,j}^*}{\sum_{j=1}^{2m} s_{kj}^- v_{2m,j}^*} \right). \quad (5.14)$$

The adjacency matrix of the graph with entries m_{ij} is then

$$\mathbf{M}_v = (\mathbf{S}_{2m}^+ \mathbf{V}^*)^T \mathbf{J}_v^\dagger (\mathbf{S}_{2m}^- \mathbf{V}^*), \quad (5.15)$$

where $\mathbf{V}^* = \text{diag}(\mathbf{v}_{2m}^*)$, \mathbf{J}_v^\dagger is the pseudo-inverse of $\text{diag}(J_v)$, and J_v is the vector of production and consumption fluxes

$$J_v = \mathbf{S}_{2m}^+ \mathbf{v}_{2m}^* = \mathbf{S}_{2m}^- \mathbf{v}_{2m}^*.$$

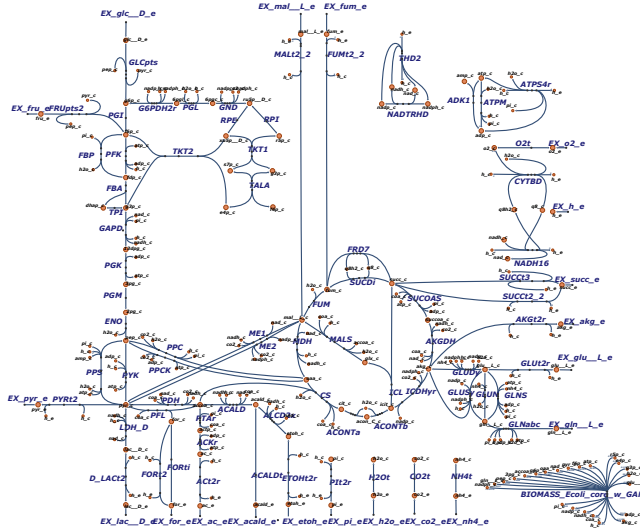
The equality is a consequence of the steady state condition $\dot{\mathbf{x}} = \mathbf{S}_{2m} \mathbf{v}_{2m}^* = (\mathbf{S}_{2m}^+ - \mathbf{S}_{2m}^-) \mathbf{v}_{2m}^* = 0$.

A fundamental feature of the FBGs \mathbf{M}_v is that the connections have a precise physical interpretation. The weights of the connections correspond to the total flow of metabolites between reactions in units of $\frac{\text{mmol}}{\text{gDW}\cdot\text{h}}$. This feature allows to directly link the connectivity of the graph to the mass flow of metabolites through the network. Since there is a graph \mathbf{M}_v specific to each FBA solution \mathbf{v}^* , this approach is a versatile framework to produce metabolic reaction graphs for different environmental conditions.

Figure 5.2d shows three metabolic graphs for the example in Figure 5.2a. In each case FBA solutions are computed under a fixed uptake flux and constrained the remaining fluxes to account for different cellular contexts. In context 1 the fluxes are constrained to be strictly positive and no larger than the nutrient uptake flux. In contexts 2 and 3, additional lower bounds are imposed for the fluxes through reactions R_6 and R_7 , respectively. The results illustrate how changes in the optimal flux distributions translate into different graph connectivities and edge weights. Context 2 leads to a graph with a similar connectivity to context 1, but with a noticeable redistribution of edge weights in the graph, while context 3 displays an extra connection between reactions R_4 and R_7 , which is absent in the other two cases.

5.5.2 Graph analysis of the core *E. coli* metabolic model

a



b

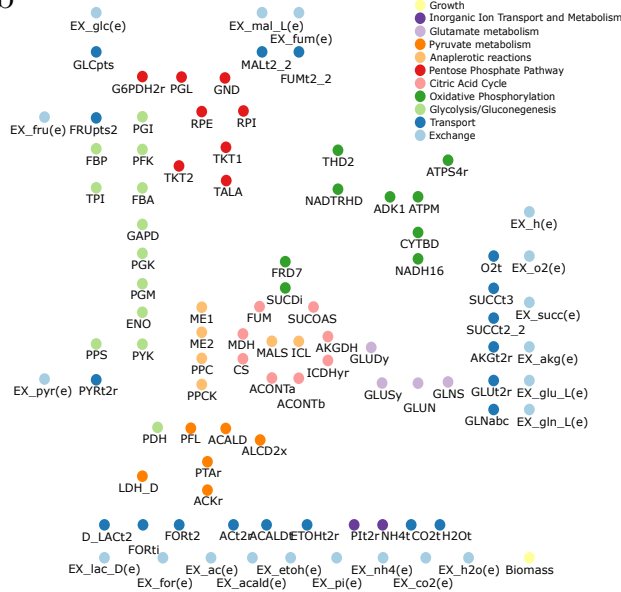


Figure 5.3. The *E. coli* core metabolic model. (a) Map of the *E. coli* core metabolic model (Orth, Fleming, and Palsson 2010) created with the online tool Escher (King et al. 2015b). (b) Reactions of the model coloured by pathways.

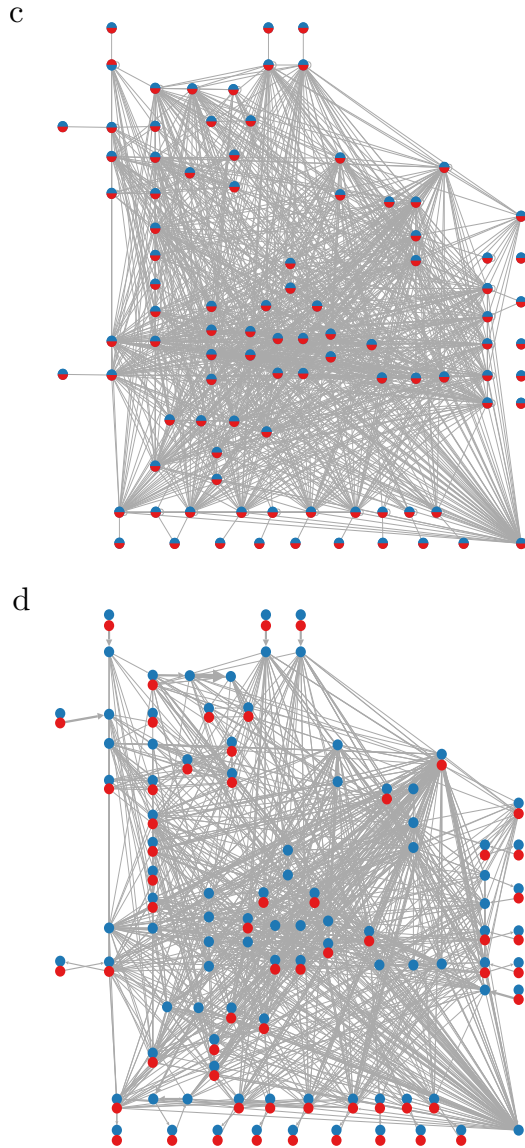


Figure 5.3. The *E. coli* core metabolic model. (c) Undirected reaction graph \mathbf{A} from equation (5.4). Since the graph is constructed ignoring directionality, each node represents forwards and backwards reactions. (d) Probabilistic Reaction Flux Graph \mathbf{D}_p of the metabolic model constructed from equation (5.9). Reversible reactions are represented by two overlapping nodes (one blue node for the forward reaction and another red for the backward when possible). The layout of the reactions in the map and the graphs is the same.

The graphs described in the previous section are constructed and analysed here for the *E. coli* core metabolic model (Orth, Fleming, and Palsson 2010). This model is relatively small, with 72 metabolites (20 extracellular and 52 intracellular) and 95 reactions (20 exchange reactions, 25 transport reactions, 49 metabolic reactions and one biomass reaction). The canonical way to describe metabolic reactions is in terms of subsystems or pathways that consist of reactions that serve a specific function (Folch-Fortuny et al. 2015; Schilling, Letscher, and Palsson 2000; Schuster, Fell, and Dandekar 2000). For example, the reactions that form glycolysis convert D-glucose into pyruvate and produce adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide (NADH). In this model, the reactions are grouped into 11 metabolic pathways that represent the main biochemical routes in the central carbon metabolism (Figure 5.3a).

The notion of directed flow is at the core of the construction of the graphs proposed in this chapter. Therefore the interest resides in studying the directed pattern of flow between reactions in these networks and observing how they change in different contexts. Such patterns can be understood as a generalisation of the notion of pathway, which are tailored to specific contexts. This analysis will help us understand which reactions form groups of interaction, under what circumstances, and what are their defining features. The Flux-Balance Graphs \mathbf{M}_v proposed can help to transition from analyses that do not incorporate networks nor context, to a framework of flexible, context-dependent, graph-based analyses of the cell's metabolism.

Markov Stability community detection framework (Delvenne, Yaliraki, and Barahona 2010; Delvenne et al. 2013) is used to extract groups of reactions in the different graphs. This framework employs diffusion processes (flows) of different duration on graphs, which is ideally suited to study the proposed reaction graphs. Markov Stability defines a *community* in precisely the way that interests us: a group of nodes in which flows are retained at specific scales. The duration of the diffusion process acts as a resolution parameter on the size of the communities (Schaub et al. 2012). A consequence of defining communities in terms of flow-retention is that Markov Stability can naturally incorporate the directionality of the connections (Beguerisse-Díaz et al. 2014; Lambiotte, Delvenne, and Barahona 2014) (see last subsection in Section 2.3.1), which is crucial to analyse metabolism in a realistic way. Therefore, communities are studied in reaction graphs: groups of reactions that are tightly linked by the flow of metabolites they produce and consume. Each community is formed by reactions that retain metabolites as much as possible.

The community structure of the graphs obtained from the core *E. coli* metabolic model (\mathbf{A} , \mathbf{D}_p , and the metabolic graphs \mathbf{M}_v for a selection of \mathbf{v}) is analysed in the next subsections. This analysis will enable to answer questions such as: *How does context affect the community structure of the network?, is it useful to describe the networks in terms of the same pathways in very different scenarios?, or what*

is the multiscale organisation of metabolism and how does it relate to the standard pathways?

Probabilistic flux reaction graph of the *E. coli* core metabolic model

As previously mentioned, the graph **A**, obtained from equation (5.4), contains connections between reactions that share metabolites in any capacity and does not distinguish between reversible and irreversible reactions (Fig. 5.3c). This graph has 95 reactions and 1158 connections. In contrast, the graph **D_p** (Fig. 5.3d) contains 154 reactions (i.e., all forward reactions and all legitimate reverse reactions) and 1,604 connections. Due to its construction in equation (5.9), the connections created by pool metabolites are weighted correctly, so that more weight is placed on connections that describe the flow of less-abundant, yet more informative, metabolites. Graphs **A** and **D_p** reveal the underlying complexity of the connectivity of the reactions which is typically absent from pathway representations. Annex 7.5 highlights additional differences between **A** and **D_p** in pathway composition (Figure 7.27a-b), pagerank (Figure 7.27c-d) and community structure (Figure 7.27e-f).

The community structures of the graphs **A** and **D_p** are extracted. The undirected graph **A** has a robust partition into seven communities (see Figure 7.27e in Annex 7.5 for a detailed description of each community). These communities are largely determined by the connections created by pool metabolites. For example, community C1_A is mainly formed by reactions that consume/produce ATP and water. The biomass reaction (the largest consumer of ATP) is not a member of this community. This reaction uses the majority of the ATP produced in the cell for cellular growth; however, the construction of the graphs considers *any* connection that involves ATP equally. Other communities in this graph are also determined by pool metabolites such as NAD⁺ and NADP⁺ (C3_A). The community structure in the network **A** highlights its limitations due to the absence of biological context and the overwhelming amount of uninformative connections.

Figure 7.27f in Annex 7.5 shows the PRG **D_p**. This graph has a robust partition into five communities. The communities in this graph emphasise flows of metabolites that are important for specific functions, and can be described in terms of pathways. For example community C1_{D_p} contains the pentose phosphate pathway and the first steps of glycolysis. Communities contain reactions that tend to belong to the same pathways. Although an improvement on **A**, the graph **D_p** is still context-independent and thus not suitable to study metabolism in specific scenarios.

Flux-Balance Graphs of the E. coli core metabolic model

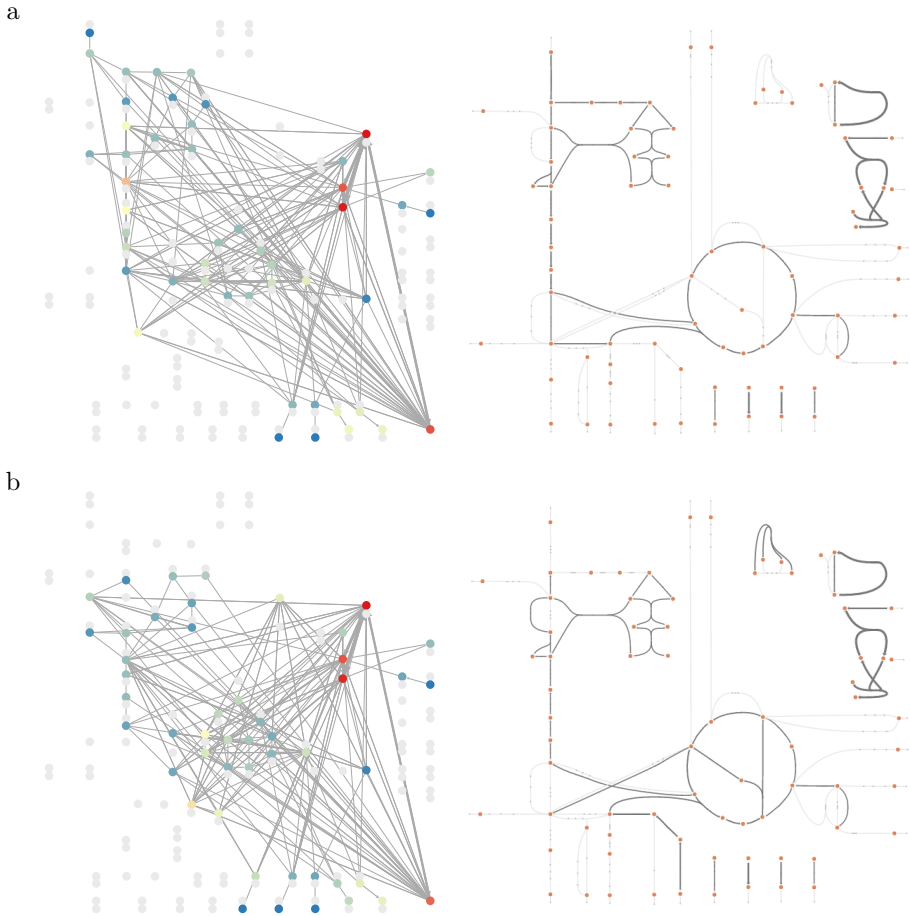


Figure 5.4. Metabolic graphs and their corresponding flux maps of *E. coli* in different contexts. Each of these graphs is obtained from equation (5.15) and the solution of an FBA problem in four different contexts: (a) aerobic growth with D-glucose as a source of carbon; (b) aerobic growth with ethanol; the inactive reactions in each context (i.e., with zero flux) are shown in grey. The thickness of the connections is proportional to the edge weights within each graph.

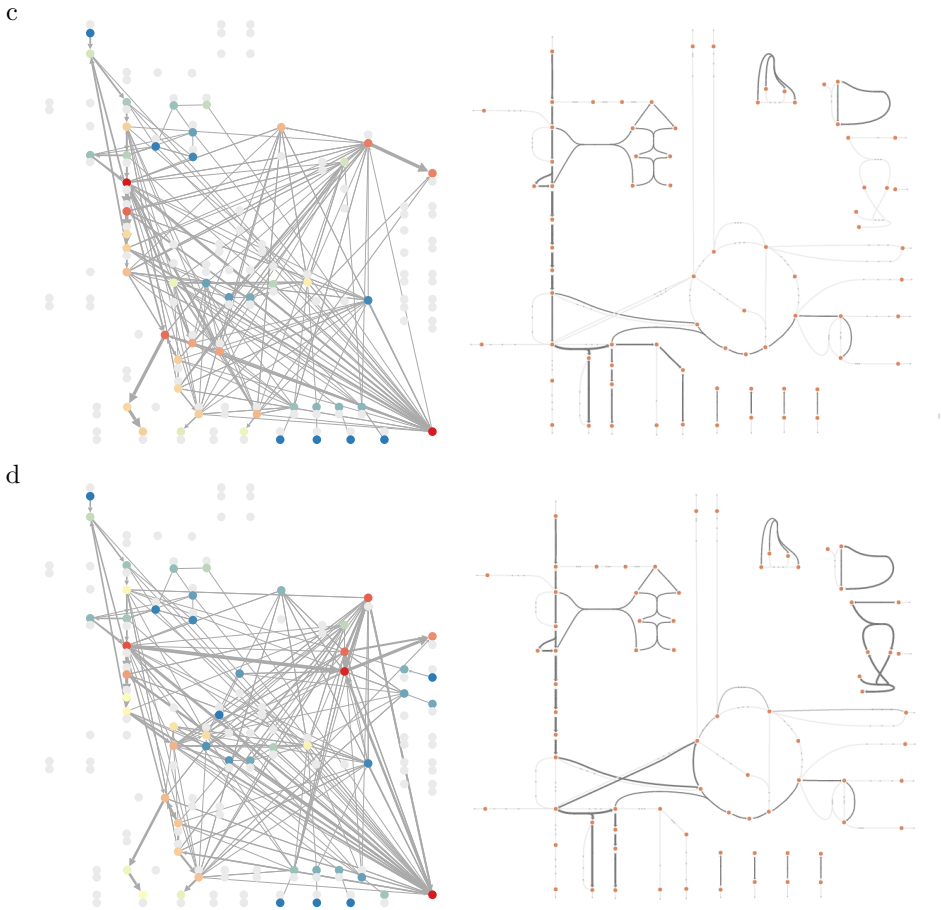


Figure 5.4. Metabolic graphs and their corresponding flux maps of *E. coli* in different contexts. Each of these graphs is obtained from equation (5.15) and the solution of an FBA problem in four different contexts: (c) anaerobic with D-glucose; (d) aerobic growth with D-glucose but limited ammonium and phosphate; the inactive reactions in each context (i.e., with zero flux) are shown in grey. The thickness of the connections is proportional to the edge weights within each graph.

Four different Flux-Balance Graphs from the *E. coli* core metabolic model are now examined. These graphs illustrate how different circumstances force important changes in the pattern of metabolite flows. The scenarios analysed are:

- M_{glc} : Aerobic growth with D-glucose as a carbon source.
- M_{etoh} : Aerobic growth with ethanol.

- $\mathbf{M}_{\text{anaero}}$: Anaerobic growth on D-glucose.
- \mathbf{M}_{lim} : Aerobic growth with D-glucose, and limited phosphate and ammonium.

In each case an FBA problem in which the constraints encode a different scenario is solved. The optimal reaction fluxes represent the state of the cell's metabolism in each context. The graphs are constructed using each optimal flux \mathbf{v}^* and equation (5.15). In all cases, the connected components of the graphs have fewer nodes and connections than \mathbf{D}_p because the solution of the FBA contains numerous reactions with zero net flux.

Figure 5.4 shows that the four networks are remarkably different from each other: the nodes in the giant component, the weights of the connections are different, as well as the pagerank of the reactions. The flux map of each FBA solution is located next to each network to facilitate interpretation and analysis.

The community structure of each graph \mathbf{M}_v is shown in Figure 5.5. Additionally, a Sankey diagram (Rosvall and Bergstrom 2010) between the traditional pathways and the communities found in each partition is provided as well. The main features of the communities of each graph are explained below. Annex 7.6 describes in detailed all communities and contains the output figures for the Markov Stability process (Figure 7.28 in Annex 7.6).

The graph \mathbf{M}_{glc} has a robust partition into three communities (see Figure 5.5a and Annex 7.6.3) with a concrete interpretation: community $\mathbf{C1}_{\text{glc}}$ contains reactions in charge of processing carbon from D-glucose to pyruvate. Community $\mathbf{C2}_{\text{glc}}$ harbours the bulk of the cell's ATP production. Community $\mathbf{C3}_{\text{glc}}$ contains the reactions in charge producing NADH and NADPH (cell's reductive power). The graph from growth in ethanol (\mathbf{M}_{etoh} Figure 5.5b) also has a partition into three communities resembling those in \mathbf{M}_{glc} with subtle, yet important differences. For example, in $\mathbf{C1}_{\text{etoh}}$ the change of source of carbon from D-glucose to ethanol has transformed glycolysis into gluconeogenesis by a reversal of the flux in the reactions in $\mathbf{C1}_{\text{glc}}$. Moreover, this community contains reactions in charge of the production of growth precursors (e.g., ME2, PPCK, GLUDy, and ICDHyr). The biomass reaction is now contained in $\mathbf{C1}_{\text{etoh}}$ due the increased flow of precursors relative to ATP production.

The absence of oxygen has, predictably, a profound impact in the cell. The graph $\mathbf{M}_{\text{anaero}}$ (Figure 5.5c) reflects this new metabolic regime. The connectivity and the communities in this graph are different from the other aerobic scenarios. For example, the reactions CYTBD and NADH16 which are the first two steps of the electron transport chain are absent. This has a profound effect (due to their high connectivity in normal circumstances) in the flow of metabolites and, consequently, on the community structure of the graph.

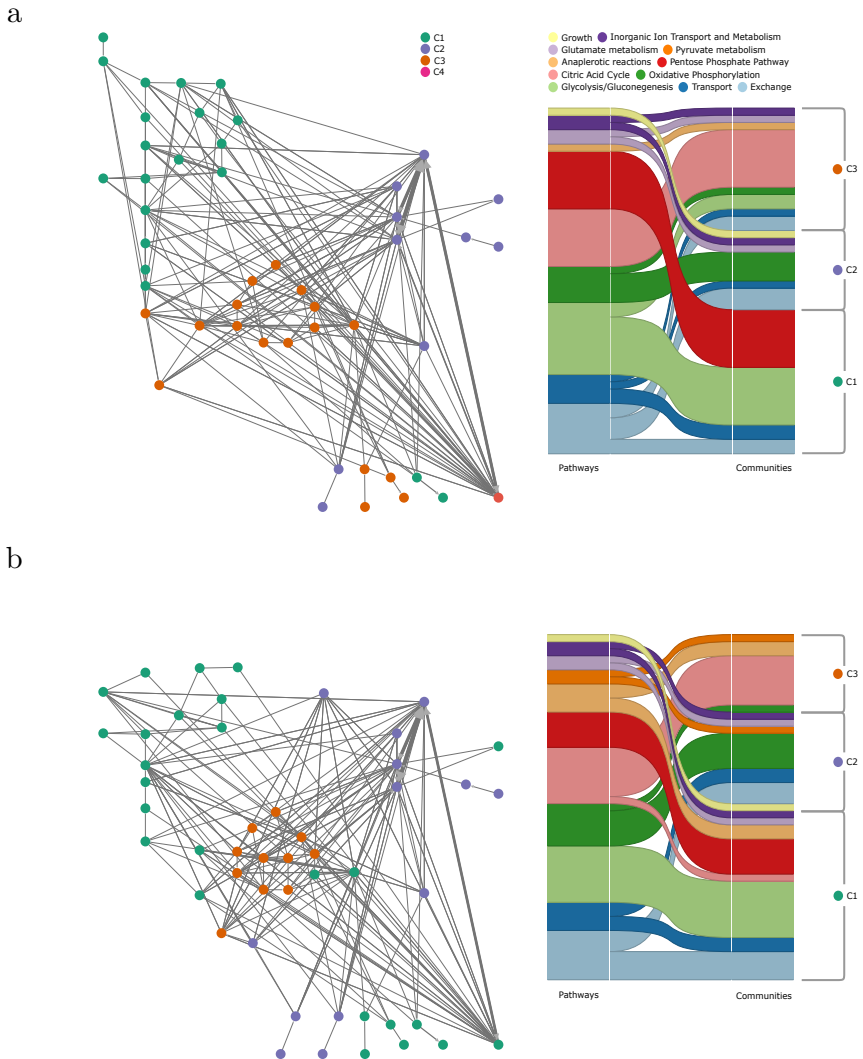


Figure 5.5. Community structure and pathway composition of metabolic graphs of *E. coli* in different contexts. (a) Aerobic growth with D-glucose. (b) aerobic growth with ethanol. The reactions are coloured according to their community (see Annex 7.6, subsections 7.6.3 and 7.6.4 for a detailed analysis of each partition). Next to each graph a Sankey diagram (Rosvall and Bergstrom 2010) shows the pathway composition of each community found.

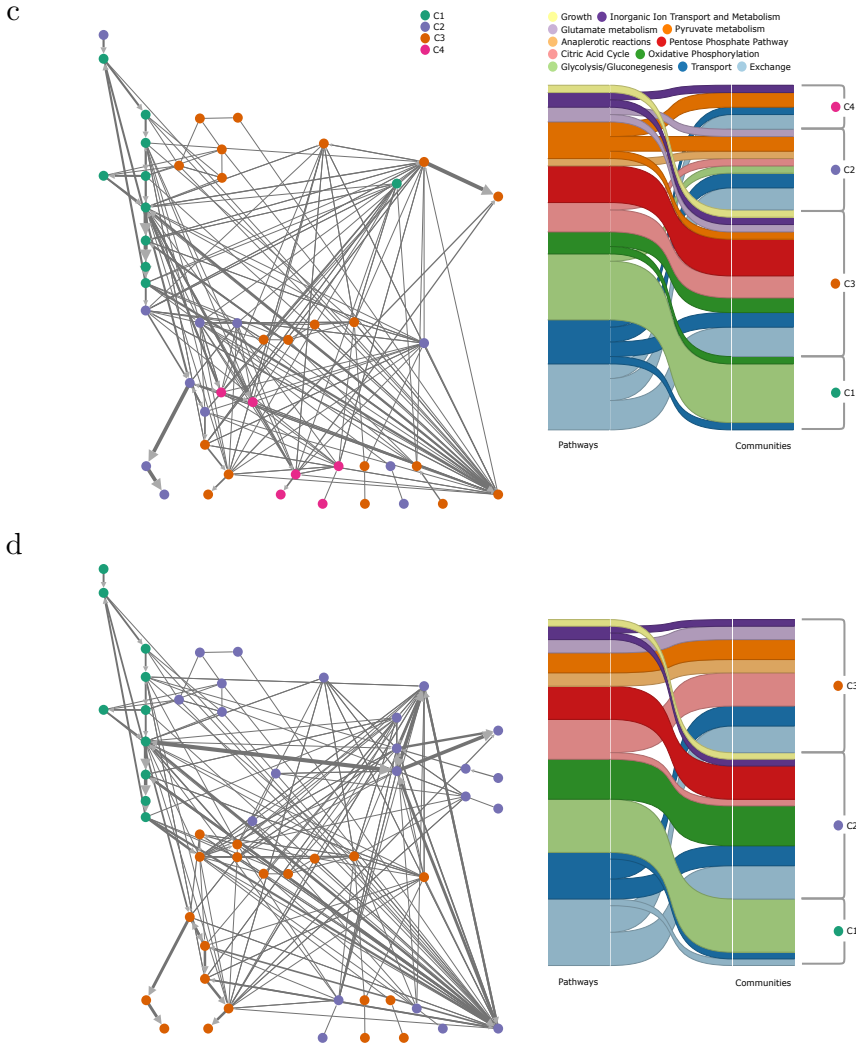


Figure 5.5. Community structure and pathway composition of metabolic graphs of *E. coli* in different contexts. (c) anaerobic with D-glucose. (d) aerobic growth with D-glucose and limited ammonium and phosphate. The reactions are coloured according to their community (see Annex 7.6, subsections 7.6.5 and 7.6.6 for a detailed analysis of each partition). Next to each graph a Sankey diagram (Rosvall and Bergstrom 2010) shows the pathway composition of each community found.

The graph M_{lim} (Figure 5.5d) also depicts the metabolic network under severe conditions. In this scenario, the community structure of the FBG reflects a phenomenon known as *overflow metabolism* (Basan et al. 2015; Vemuri et al. 2007),

which occurs when the cell takes in more carbon than it can process. In this case overflow metabolism is due to limited phosphate and NH_4 : the over-abundance of carbon is secreted from the cell, there is a strong decrease in the growth rate, and a partial shut-down of the TCA cycle. Community C3_{lim} is similar to C3_{glc} and C3_{etoh} , with the addition of the secretion routes of acetate and formate.

5.5.3 Structure of FBGs at multiple resolutions

One partition in each of the four Flux-Balance Graphs has been already analysed. However, complex graphs such as these often have important partitions into communities at different levels of granularity (Bacik et al. 2015; Delvenne, Yaliraki, and Barahona 2010). In the Markov Stability framework, we can explore these scales by scanning through different values of Markov time (see last subsection in Section 2.3.1).

For instance, Figure 5.6a shows the number of communities found in the graph \mathbf{M}_{glc} as one scans through a range of Markov times. Short Markov times result in fine-grained partitions (e.g., near $t = 0.3$ there are over 20 communities); Markov time is increased, we will find coarser partitions until (as Markov time tends to infinity) a single community containing all the nodes is found. Five Markov times at which robust partitions into 11, 7, 5, 3, and 2 communities have been selected. To analyse the partitions (and the partition into pathways) simultaneously, a Sankey diagram is constructed (Figure 5.6b). These diagrams allow to visualise the composition of the different partitions and how they are related in terms of their members.

In the example from Figure 5.6b, we start on the left side with the reactions of \mathbf{M}_{glc} divided in metabolic pathways. As we move to the right, we can see how the reactions in each of the pathways assemble into the partitions obtained with Markov Stability. This figure highlights different features and properties of the *E. coli* metabolic network in aerobic conditions and with glucose as the sole source of carbon. In some cases, the reactions of pathways such as the oxidative phosphorylation or glycolysis are grouped mostly together in the same community. Interestingly, the TCA cycle, although it appears as a cohesive unit for most Markov times is split in two at $t = 19.72$. The pathways with reactions in charge of exchanging substances with the exterior of the cell (exchange and transport) are spread among the partitions in all Markov times; these are pathways in which the reactions do not interact amongst themselves. These pathways act more like ‘roles’ (i.e., importing and moving substrates) than like cohesive metabolic sub-units. Other pathways such as the pentose phosphate pathway is divided into different communities except at $t = 6.01$ when, as the TCA cycle, comes together before splitting up again. This phenomenon illustrates that some biological features may only become relevant at specific resolutions.

The question of how to deal with this resolution dependency results of vital importance. In general there is no set rule to pick a concrete Markov Time in which to analyse the community structure of each network. It is a case by case decision. There are however some general guidelines that increase the usefulness or interest of particular resolution. Times in which the Variation of Information decreases abruptly point in mayor network re-structuring moments which may be of interest. Furthermore, long time intervals in which the VI stays very low or zero show very stable partitions which are potentially more interesting. Finally, there is compromise that has to be reached: enough Markov Time to reach stable partitions with communities of significant size. This depends on the size of the network too, since larger networks will need longer times.

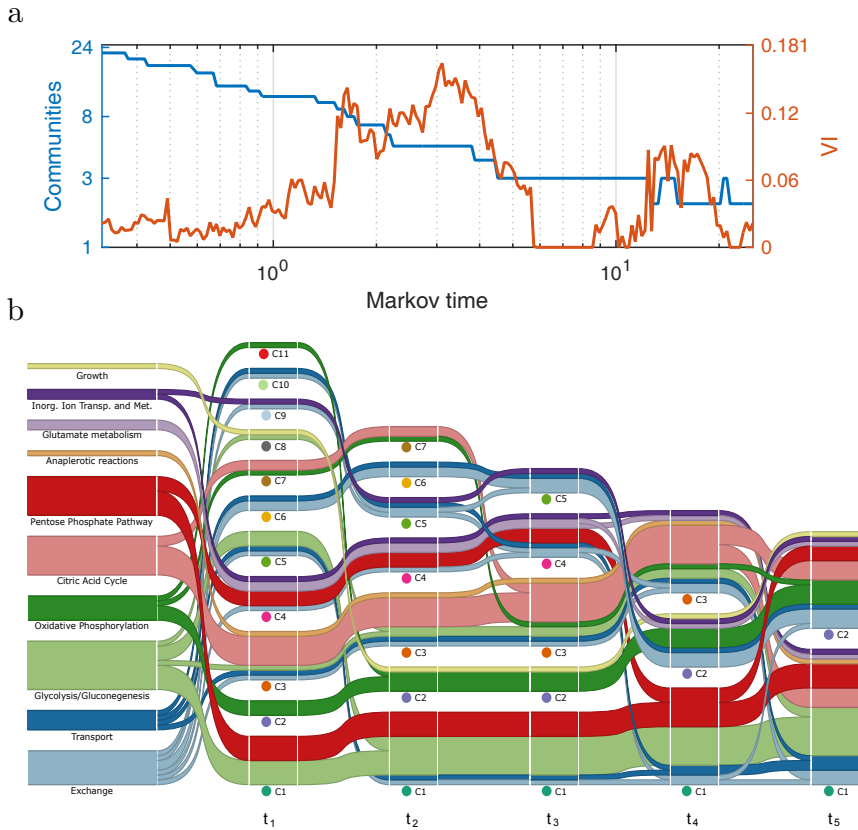


Figure 5.6. Multiscale community structure of the graph M_{glc1} . (a) Number of communities (blue line) and Variation of Information (red line) in the M_{glc1} graph as we scan Markov times (see text and Methods section) for communities. Five Markov times in which the network can be split into 11, 7, 5, 3, and 2 communities and the VI is low or has a pronounced dip were selected. (b) Alluvial diagram (Rosvall and Bergstrom 2010) showing how the reactions that form each of the pathways (left) assemble in communities of different size as Markov time is scanned. Note that in this graph there are no reactions (with positive flux) in the pyruvate metabolism subsystem.

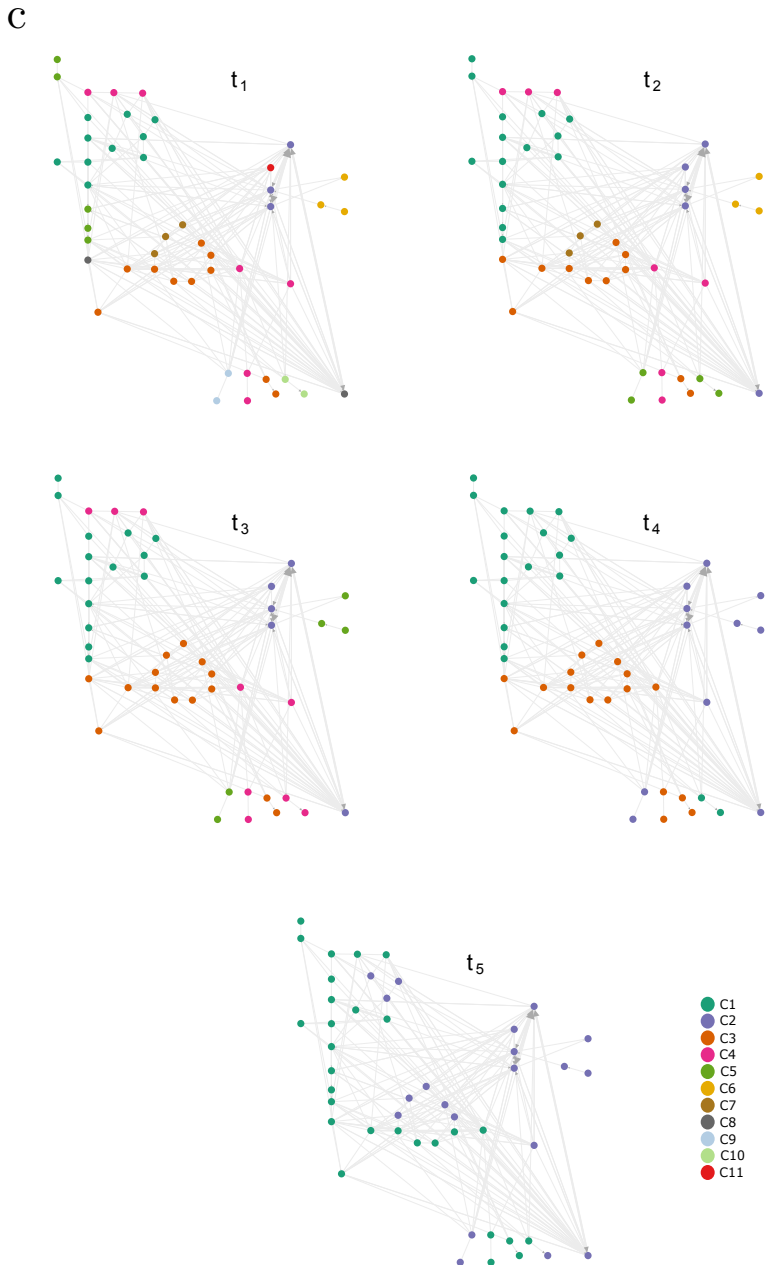


Figure 5.6. Multiscale community structure of the graph M_{glc} . (c) Community structure of the graph M_{glc} in the selected Markov times.

5.6 Conclusions

Metabolic networks have been described by multiple and fundamentally different graphs, unlike other cellular networks that have a natural correspondence between species-interactions and nodes-edges of a graph (Alon 2007). This has been long acknowledged, and in essence there are three types of metabolic graphs (Palsson 2006): a graph with reactions as nodes, a graph with metabolites as nodes, and a graph with both reaction and metabolites as nodes. Moreover, each one of these graphs can be directed or undirected, and the edge weights may be computed according to different principles. Another source of confusion is that metabolic pathways are typically drawn as maps of metabolites with directed arrows to represent enzymatic reactions among them (King et al. 2015b). Although these diagrams resemble a graph, they are not amenable to network-theoretic analyses unless we ascribe a specific meaning to the arrows between nodes. These subtleties in the construction of metabolic graphs are important because they can affect the interpretation of results (Arita 2004; Ouzounis and Karp 2000). Yet there is still a lack of consensus in the community and, as a result, comparisons across studies are difficult and cast doubt on the generality of the observed topological properties.

The objective in this chapter was to address some of the missing elements in the construction of graphs for metabolic networks. The construction of two types of graphs were presented where nodes represent reactions and directed edges represent metabolites produced by one reaction and consumed by another. In the Probabilistic Flux Reaction Graph (PRG), the edge-weights describe the probability that a any two reactions produce-consume a molecule of any metabolite. This probabilistic formulation can tame the overwhelming number of connections generated by pool metabolites without the need to remove them from the network description, as commonly done in the literature (Kreimer et al. 2008; Ma and Zeng 2003b; Samal and Martin 2011; Silva et al. 2008). To incorporate the effect of the environment, the Flux Balance Graph (FBG) is proposed, in which edge weights are the total flux of metabolites between reactions predicted by Flux Balance Analysis (FBA). By computing FBA solutions for different exchange fluxes between the cell and its environment, one can systematically build metabolic graphs for different compositions of the growth media. When applied to the core *E. coli* metabolic model the topology of the FBG effectively captures known metabolic adaptations such as the glycolytic-gluconeogenic switch, overflow metabolism, and the effects of anoxia.

The proposed FBG draws a novel connection between modern network theory (which studies graphs) and constraint-based methods widely employed in metabolic modelling (Orth, Fleming, and Palsson 2010; Rabinowitz and Vastag 2012). Previous attempts to incorporate constraint-based models into metabolic graphs include, for example, the use of FBA solutions to search for node clusters (Samal et al. 2006), and the study of network robustness upon removal of FBA-constrained

reaction nodes (Smart, Amaral, and Ottino 2008). Using a graph built from a genome-scale metabolic model *S. cerevisiae*, it has been shown that the connectivity of reaction nodes does not correlate with FBA fluxes that maximise growth (Vitkup, Kharchenko, and Wagner 2006). This brings into question the amount of physiologically-relevant information contained in graphs that are built from the whole metabolic blueprint of a cell but that are blind to the environmental and biological context. In contrast, the FBG exploits the physiological predictions from FBA to construct metabolic graphs that are more informative of cell physiology and are directly grounded on specific environmental conditions. The resulting graphs are smaller and less connected than those built from the complete metabolic blueprint, but they shed further light on the organisation of metabolic activity in realistic physiological conditions.

A number of promising applications of these results open up. First, the most immediate application of the FBG is to study how environmental inputs shape the community structure and evolution of metabolic networks (Takemoto 2013; Zhou and Nakhleh 2012). Such analysis has potential in biomedicine, for example, by finding metabolic conditions that maximise the efficacy of drugs treatments (Chang et al. 2010; Csermely, Ágoston, and Pongor 2005). Second, the FBG can be readily used to quantify metabolic robustness via graph statistics upon node (e.g., reaction) removal (Smart, Amaral, and Ottino 2008). Third, the proposed approach can be extended to include dynamic adaptations of metabolic activity, for example, by using dynamic extensions of FBA (Mahadevan, Edwards, and Doyle 2002; Rügen, Bockmayr, and Steuer 2015; Waldherr, Oyarzún, and Bockmayr 2015), or by incorporating static (Colijn et al. 2009) and time-varying (Oyarzún 2011) enzyme concentrations. Fourth, the FBG could provide a novel route for robustness analysis of FBA solutions (Gudmundsson and Thiele 2010). A common challenge is that FBA solutions are not unique, but instead they belong to a solution space containing an infinite number of flux distributions that are equally optimal (Orth, Thiele, and Palsson 2010). The connectivity of the FBG depends on the particular FBA solution used to construct it, and thus one can exploit non-uniqueness to quantify the robustness of solutions in a space of graphs.

Chapter 6

Conclusions

- *Good night, and good luck.*

Edward R. Murrow

The contributions of this work were listed in Chapter 1, and particular conclusions can be found at the end of each main chapter. Here some general conclusions are drawn and discussed together besides some proposed lines for future work.

- **(Chapter 3)** Intraviral PPIN is highly connected. The four topological parameters studies seem to depend on the protein degree. Moreover, the cumulative distributions of these parameters and the degree increase in a quasi-linear way. The results were very similar across the detection methods used. Interactions with lower intensity can be as vital to virus development as the more intense ones.
- **(Chapter 3)** Some viral proteins focus their effect in only one host hub, while other diversify their effects among several proteins through direct interactions. There are significant differences in the propagation speed across different viral proteins. Some proteins spread its effect in similar patterns, hinting a common functionality.
- **(Chapter 3)** The PLS modelling applied to genomic, proteomic and phenotypic data sets allows for the integration of the mutations performed on viral proteins, their effects on the PPIN, and their influences on the organismal fitness experimentally quantified. Three biological functional modules affect-

ing the PPIN and influencing the fitness positively have been detected. Two additional modules are identified affecting a single protein. Different mutations affecting the same protein induce different behaviours in the activity of the PPIN and the resulting fitness.

- **(Chapter 3)** Data fusion allows unveiling two significant features: (i) the mutations are related to topological changes in the network and their subsequent influence on the fitness, and (ii) the mutations not affecting the network can also be related to the fitness.
- **(Chapter 4)** The *E. coli* constraint-based model show an excellent agreement with the experimental dataset. Constraint-based model constitute a perfectly valid approach to model metabolism in steady state conditions. Taking into account purely the amount of scenarios that the approaches can tackle *MFA* and *FVA* achieve a 100% and *MFAg* only a 50% success rate.
- **(Chapter 4)** The flux distributions obtained through each approach are very similar among them. All *MFA* solutions look alike. This also happens for *MFAg* and *FVA* solutions. Just by looking at the shape of the intervals it is evident that each approach finds solutions very similar across the complete dataset.
- **(Chapter 4)** The size or width of the solutions changes significantly, though. *MFA* offer the widest solutions, being many of them biologically infeasible. *MFAg* returns narrower and more biologically sound solutions. Finally, *FVA* solutions return almost always a single value for each flux in the model. They make the most biological sense too. From the size of the solutions they provide *FVA* is the best, then *MFAg* and in the end *MFA*. This seems to validate the main assumption in which FBA and FVA are based: that cellular growth can be understood as a biomass production optimization process.
- **(Chapter 4)** The Partial Least Squares regression indicates that there is no correlation between the variability of the original variables and the variability of the ratios that define the performance of the different approaches. The three approaches obtain similar results no matter the value of the original variables. Therefore, the conclusions about the performance of the different approaches are stable and robust.
- **(Chapter 4)** The PCA analysis carried out was able to describe the system with two components. These components fit the main carbon metabolic route and the overflow metabolism. The behaviour of the cells seem to be defined by a combination of both, which agrees with previously reported metabolic regime description in *E. coli*.
- **(Chapter 5)** The construction of two types of graphs were presented where nodes represent reactions and directed edges represent metabolites produced

by one reaction and consumed by another. In the Probabilistic Flux Reaction Graph (PRG), the edge-weights describe the probability that a any two reactions produce-consume a molecule of any metabolite. To incorporate the effect of the environment, the Flux Balance Graph (FBG) is proposed, in which edge weights are the total flux of metabolites between reactions predicted by Flux Balance Analysis (FBA).

- **(Chapter 5)** Computing FBA solutions for different exchange fluxes between the cell and its environment, one can systematically build metabolic graphs for different compositions of the growth media. When applied to the core *E. coli* metabolic model the topology of the FBG effectively captures known metabolic adaptations such as the glycolytic-gluconeogenic switch, overflow metabolism, and the effects of anoxia.
- **(Chapter 5)** The proposed FBG draws a novel connection between modern network theory (which studies graphs) and constraint-based methods widely employed in metabolic modelling. The FBG exploits the physiological predictions from FBA to construct metabolic graphs that are more informative of cell physiology and are directly grounded on specific environmental conditions. The resulting graphs are smaller and less connected than those built from the complete metabolic blueprint, but they shed further light on the organisation of metabolic activity in realistic physiological conditions.
- **(Chapter 5)** Metabolic network and community structure of *E. coli* under different environmental scenarios produces significantly different graphs. The most stable partition is formed by three communities found for the aerobic scenario. This robust partition has a concrete interpretation: community C1_{glc} contains reactions in charge of processing carbon from D-glucose to pyruvate. Community C2_{glc} harbours the bulk of the cell's ATP production. Community C3_{glc} contains the reactions in charge producing NADH and NADPH (cell's reductive power).

This thesis addressed problems related to the analysis and modelling of molecular biological networks. The objective was to develop and use a series of techniques and methodologies that shed light and increased our understanding of complex biological systems. The main conclusion of all the work done and the line that joins all the dots together is the power and capability of network analysis in molecular biology. This system approach has resulted in a series of valuable studies where the focus was always on the whole system and its general and emergent properties. Ideally, this thesis will contribute to establish systems biology or the systemic or integral approach to biology as an absolutely fundamental part of almost any biological research done in the future.

Future work

The study carried out in this thesis opens new research avenues. From the analysis of Chapter 3, the VHPIN analysis can be further explored using more complex metrics, graph kernels or integrating more biological information available such as sub-cellular localization or biological function. Specially relevant would be to integrate in this topology the regulatory network of the host. The interactions between viral proteins and host transcription factors are very sensible points to study further. Besides this, further work of interest includes testing the proposed methodology with a larger dataset containing more mutants, and extending the analysis to larger PPINs, in order to build multivariate models with a higher predictive power, exploiting the features of the projection to latent structure methods.

Regarding Chapter 4 additional work would require to extend even further the scope of the dataset to check if the constraint-model is still able to work. On the other hand, larger models (ideally genome-scale) could be used to check if they improve the level of agreement with the experimental data. Moreover, constraint-based model of higher organisms should be addressed as well. In these organisms, it is much harder to define a objective function and therefore the methods that do not assume any could improve their performance.

Finally, the proposed graph description of metabolic networks proposed in Chapter 5 could be used in many different applications. Community detection through Markov Stability is one particular example of many possible. Using real fluxomics data is the most straight forward application. Calculate the synergy and competition version of the PRG would out an even more accurate relationship between reactions. The reverse matrix (using metabolites as nodes instead of reactions) would give us even more insight into the metabolic system. Furthermore, it would be interesting as well to analyse the transition between to extreme states (such as aerobic and anaerobic conditions) an the study if the structure of the networks varies uniformly or if there are jumps between stable intermediate metabolic states. In addition, the applications already discussed in Section 5.6 are also worth pursuing.

Chapter 7

Annexes

- *Success is the sum of details.*

Harvey S. Firestone

7.1 Annex I: Mutations performed on TEV, distances registered, and fitness measured.

Mutant nº	Mutation 1	Mutation 2	Distance 1	Distance 2	Fitness	Relative Fitness
<i>wild-type</i>	-	-	-	-	1.3461	1
1	PC6	PC7	14	17	1.3889	1.0318
2	PC6	PC19	14	0	1.3842	1.0283
3	PC7	PC19	17	0	1.3802	1.0253
4	PC12	PC19	19	0	1.374	1.0207
5	PC6	PC49	14	14	1.3813	1.0261
6	PC7	PC63	17	14	1.3843	1.0284
7	PC6	PC63	14	14	1.3764	1.0225
8	PC6	PC69	14	17	1.3904	1.0329
9	PC7	PC69	17	17	1.3806	1.0256
10	PC2	PC69	39	17	1.3445	0.9988
11	PC12	PC83	19	14	1.3698	1.0176
12	PC12	PC95	19	18	1.3553	1.0068
13	PC6	-	14	-	1.3477	1.0012
14	PC12	-	19	-	1.3371	0.9933
15	PC2	-	39	-	1.331	0.9888
16	PC7	-	17	-	1.3198	0.9805
17	PC19	PC40	0	0	1.3817	1.0264
18	PC19	PC41	0	19	1.3781	1.0238
19	PC26	PC63	0	14	1.331	0.9888
20	PC19	PC69	0	17	1.3856	1.0293
21	PC19	PC70	0	24	1.381	1.0259
22	PC26	PC69	0	17	1.3122	0.9748
23	PC22	PC72	16	23	1.2831	0.9532
24	PC22	PC69	16	17	1.2288	0.9129
25	PC19	PC83	0	14	1.3836	1.0279
26	PC19	PC95	0	18	1.3786	1.0241
27	PC19	-	0	-	1.3308	0.9886
28	PC22	-	16	-	1.2795	0.9505
29	PC26	-	0	-	1.2586	0.9350
30	PC41	PC49	19	14	1.3875	1.0308
31	PC44	PC49	0	14	1.3835	1.0278
32	PC44	PC63	0	14	1.3454	0.9995

Figure 7.1. Mutations performed on TEV, distances registered, and fitness measured (part 1).

Mutant n°	Mutation 1	Mutation 2	Distance 1	Distance 2	Fitness	Relative Fitness
<i>wild-type</i>	-	-	-	-	1.3461	1
33	PC40	PC69	0	17	1.3538	1.0057
34	PC44	PC69	0	17	1.3497	1.0027
35	PC44	PC76	0	0	1.3762	1.0224
36	PC40	PC83	0	14	1.4027	1.0420
37	PC41	PC83	19	14	1.3811	1.0260
38	PC40	-	0	-	1.3291	0.9874
39	PC44	-	0	-	1.3237	0.9834
40	PC41	-	19	-	1.31	0.9732
41	PC49	PC70	14	24	1.3851	1.0290
42	PC49	PC67	14	17	1.384	1.0282
43	PC49	PC83	14	14	1.3682	1.0164
44	PC49	PC95	14	18	1.3547	1.0064
45	PC60	PC83	0	14	1.3534	1.0054
46	PC60	PC95	0	18	1.3392	0.9949
47	PC60	-	0	-	1.32	0.9806
48	PC49	-	14	-	1.3164	0.9779
49	PC63	PC70	14	24	1.3814	1.0262
50	PC63	PC69	14	17	1.3603	1.0105
51	PC63	PC95	14	18	1.3577	1.0086
52	PC63	-	14	-	1.3205	0.9810
53	PC67	PC76	17	0	1.3564	1.0077
54	PC70	PC83	24	14	1.3887	1.0316
55	PC69	PC95	17	18	1.3886	1.0316
56	PC72	PC83	23	14	1.3881	1.0312
57	PC72	-	23	-	1.3359	0.9924
58	PC67	-	17	-	1.3327	0.9900
59	PC70	-	24	-	1.3268	0.9857
60	PC69	-	17	-	1.3156	0.9773
61	PC76	PC95	0	18	1.3522	1.0045
62	PC76	-	0	-	1.3392	0.9949
63	PC83	-	14	-	1.3371	0.9933
64	PC95	-	18	-	1.3306	0.9885

Figure 7.2. Mutations performed on TEV, distances registered, and fitness measured (part 2).

7.2 Annex II: collected *E. coli* experimental data

The complete collection of *E. coli* scenarios used in this dissertation is shown here. The data of each scenario contains information about the original paper in which was published, the year, the strain of *E. coli* used, the recipient used for growth and its volume and experimental measures of growth rate and a series of extracellular uptake and secretion rates. The complete list of papers (shown also in the bibliography section) used is the following:

- U. Sauer et al. (1999). “Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism”. In: *Journal of Bacteriology* 181.21, pp. 6679–6688
- Marcel Emmerling et al. (2002). “Metabolic Flux Responses to Pyruvate Kinase Knockout in *Escherichia coli*”. In: *Journal of Bacteriology* 184.1, pp. 152–164
- Eliane Fischer, Nicola Zamboni, and Uwe Sauer (2004). “High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived ^{13}C constraints”. In: *Analytical Biochemistry* 325.2, pp. 308–316
- A. Perrenoud and U. Sauer (2005). “Impact of Global Transcriptional Regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on Glucose Catabolism in *Escherichia coli*”. In: *Journal of Bacteriology* 187.9, pp. 3171–3179
- Anke Kayser et al. (2005). “Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state”. In: *Microbiology (Reading, England)* 151.3, pp. 693–706
- Stephen S. Fong et al. (2006). “Latent pathway activation and increased pathway capacity enable *Escherichia coli* adaptation to loss of key metabolic enzymes”. In: *The Journal of Biological Chemistry* 281.12, pp. 8024–8033
- B.R.B. Haverkorn et al. (2014). “Large-scale ^{13}C -flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*”. In: *Molecular Systems Biology* 7.1, pp. 477–477

Scenario	Year	Paper	Strain	Scale	Volume (ml)	Growth Rate (h ⁻¹)	Specific Rates (mmol l ⁻¹ h ⁻¹)					Biomass Yield g ^{dw} (g l ⁻¹) ⁻¹	Carbon Balance (%)							
							Glucose uptake	O ₂ uptake	CO ₂ secretion	Acetate secretion	Pyruvate secretion									
7	2004	High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C	MG1655	Reactor	800	0.82	11	18.2	18.6	6.4	0.42	99								
23							Shake Flask	30	0.85	7.1	6.6	0.41								
24													96-Deep-well plate	2	0.61	8	6.4	0.41		
25	2005	Impact of global transcription factor AcvA, CcpA, Fru, and Mlc on Glucose Catabolism in <i>Escherichia coli</i> (2005)	MG1655	BW25113	500	0.65+/-0.01	7.6+/-0.2	4.8+/-0.2	0.48+/-0.01	0.41										
26				AcvA			0.60+/-0.01					7.2+/-0.2	3.5+/-0.0	0.46+/-0.01						
27				AcvB			0.61+/-0.01					7.4+/-0.2	4.9+/-0.1	0.45+/-0.01						
28				CcpA			0.65+/-0.01					8.1+/-0.3	5.7+/-0.4	0.44+/-0.02						
29				Ccp			0.25+/-0.01					2.5+/-0.2	0.7+/-0.1	0.56+/-0.02						
30				CcpA			0.27+/-0.01					2.7+/-0.2	0.8+/-0.1	0.56+/-0.01						
31				Fru			0.61+/-0.01					7.7+/-0.5	4.8+/-0.0	0.44+/-0.02						
32				Mlc			0.60+/-0.01					7.1+/-0.0	4.3+/-0.2	0.47+/-0.00						
33				2006			Latent pathway activation and increased pathway capacity enable <i>Escherichia coli</i> adaptation to loss of key enzymes					MG1655	Erlenmeyer Flask	500	0.63+/-0.03	8.8+/-0.5	2.2+/-0.2	0.39+/-0.01	0.40+/-0.02	
34																Pgl				
35	PglE1	0.34+/-0.06	5.8+/-0.3		2.6+/-0.5	0.32+/-0.07														
36	PglE2	0.53+/-0.03	5.6+/-0.5		0.0+/-0	0.53+/-0.02														
37	Ppc	0.22+/-0.01	3.0+/-0		1.1+/-0	0.4+/-0.01														
38	PpcE1	0.55+/-0.04	8.1+/-0.1		2.2+/-0.3	0.37+/-0.02														
39	PpcE2	0.56+/-0.01	7.8+/-0.3		2.2+/-0.2	0.39+/-0.01														
40	Ppa	0.58+/-0.02	9.1+/-0.9		0.6+/-0.2	0.36+/-0.02														
41	PpaE1	0.64+/-0.04	10.3+/-0.6		0.7+/-0.3	0.34+/-0.01														
42	PpaE2	0.66+/-0	8.6+/-0.5		0.7+/-0.2	0.43+/-0.02														
43	Tpi	0.18+/-0.02	2.7+/-0	0.2+/-0.1	0.33+/-0.02															
44	TpiE1	0.51+/-0.02	7.8+/-0.8	1.0+/-1.0	0.36+/-0.01															
45	TpiE2	0.49+/-0.02	7.3+/-0.8	0.9+/-0.9	0.37+/-0.02															

Figure 7.3. Collection of *E. coli* experimental scenarios (part 1).

Scenario	Year	Paper	Strain	Scale	Volume (ml)	Growth Rate		Specific Rates (mmol g ⁻¹ h ⁻¹)				Biomass Yield (g g ⁻¹)	Carbon Balance (%)		
						μ (h ⁻¹)	Glucose uptake	CO ₂ uptake	CO ₂ secretion	Acetate secretion	Pyruvate secretion			0.1 (pyruvate)	0.1 (succinate)
1-6	2001	Metabolic Flux Response to Pyruvate Kinase Knockout in <i>Escherichia coli</i>	JM101	Chemostat	1000	0.09	1.4±0.2	4.6±0.7	4.9±0.5	0	0	0.37±0.03	99±9		
						0.4	4.8±0.4	11.8±1.4	12.4±1.1	0	0	0.46±0.03	94±7		
						0.09	2.2±0.3	5.1±0.9	5.2±0.4	1.3±0.1	0.1 (pyruvate)	0.22±0.02	93±7		
						0.08	1.4±0.1	5.5±0.9	5.6±0.5	0	0	0.33±0.02	104±8		
						0.4	5.0±0.4	12.8±1.6	13.7±1.3	0	0	0.44±0.03	95±7		
						0.08	2.7±0.3	7.0±1.5	6.3±0.7	1.2±0.1	0.2 (oxoglutarate), 0.1 (fumarate)	0.17±0.01	96±8		
46	2011	Transcriptional control of respiratory and fermentative metabolism in <i>Escherichia coli</i>	BW25113	Shake Flask	500	0.61	8.26±0.5		4.89±1.52						
						0.2	2.8	8.3	8.7	0	0				
8	1999	Metabolic Flux Ratio Analysis of Genetic and Environmental Modulation of <i>Escherichia coli</i> Central Carbon Metabolism	ME1655	Shake Flask	1000	0.2	6.4	10.6	9.6	4.4	0.8				
						0.044	0.48	1.38	1.27	0	0				
10						0.066	0.73	1.78	1.98	0					
11						0.134	1.43	3.41	3.98	0					
12						0.15	1.59	3.66	3.98	0					
13						0.17	1.81	5.24	4.84	0					
14						0.203	2.08	5.50	5.36	0					
15						0.265	2.79	6.19	6.61	0					
16						0.28	2.81	7.41	6.89	0					
17						0.3	3.01	7.41	7.39	0					
18						0.347	3.36	7.56	7.73	0					
19						0.375	3.56	7.55	7.55	0					
20						0.388	3.57	7.63	7.61	0					
21						0.397	4.02	8.22	8.68	0					
22															

Figure 7.4. Collection of *E. coli* experimental scenarios (part 2).

7.3 Annex III: MFA , MFAg and FVA interval estimates for the experimental scenarios

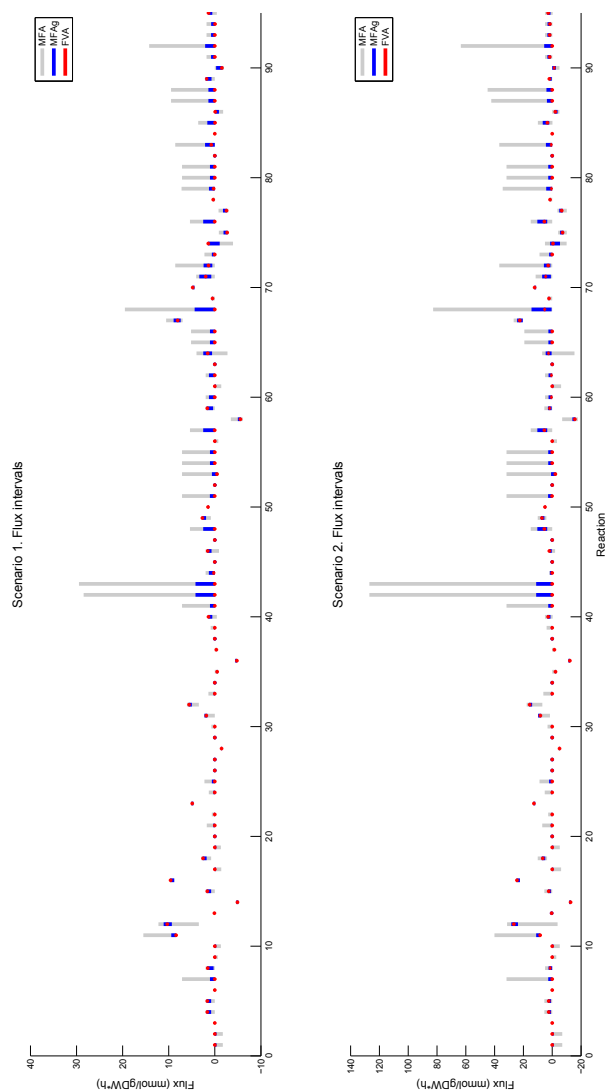


Figure 7.5. MFA , MFAg and FVA interval estimates for scenarios 1 and 2.

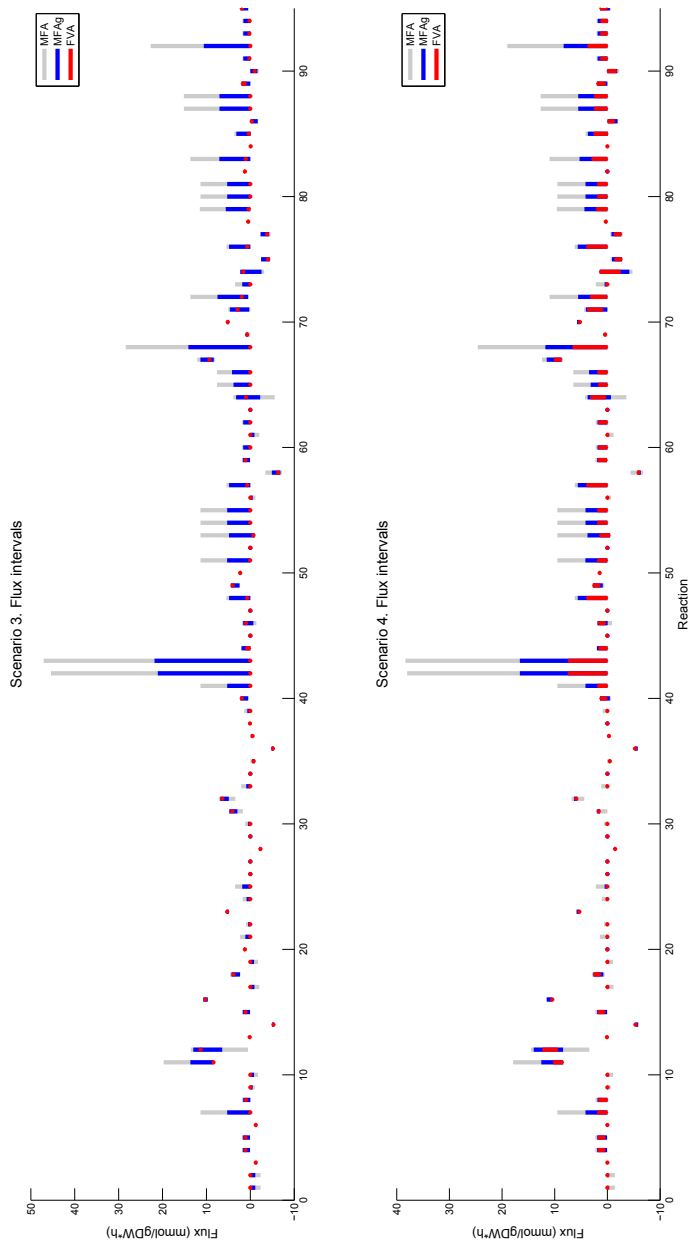


Figure 7.6. *MFA* , *MFAg* and *FVA* interval estimates for scenarios 3 and 4.

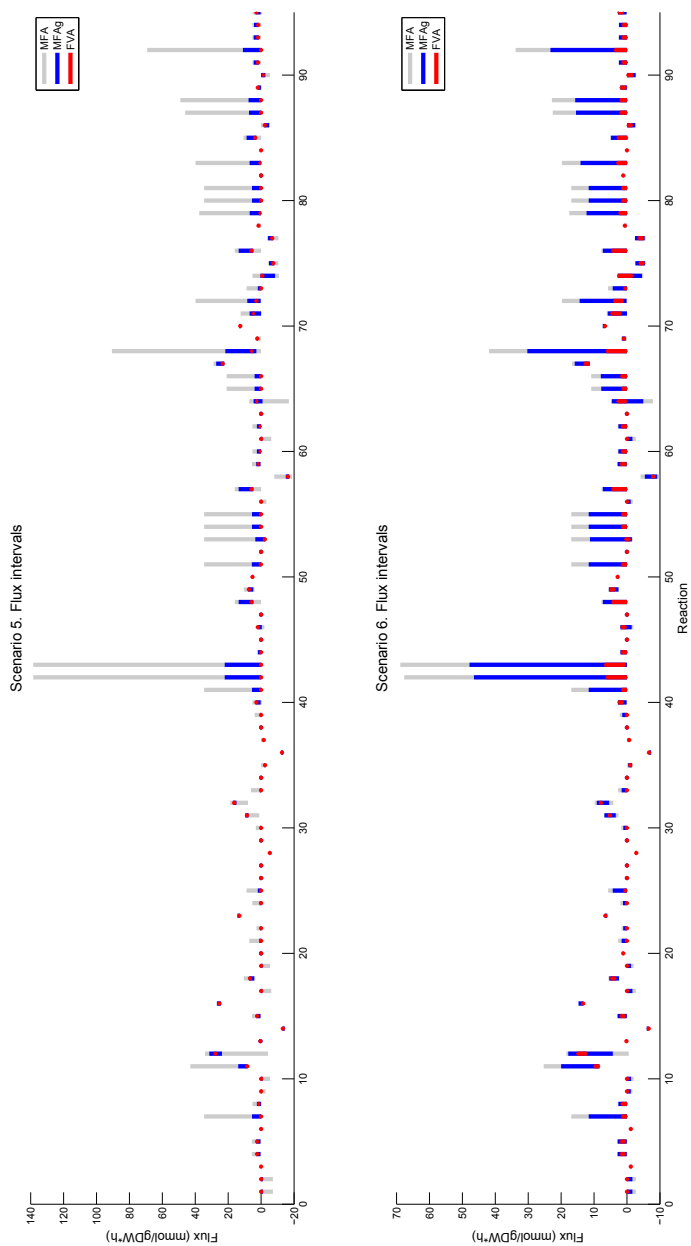


Figure 7.7. MFA , MFAg and FVA interval estimates for scenarios 5 and 6.

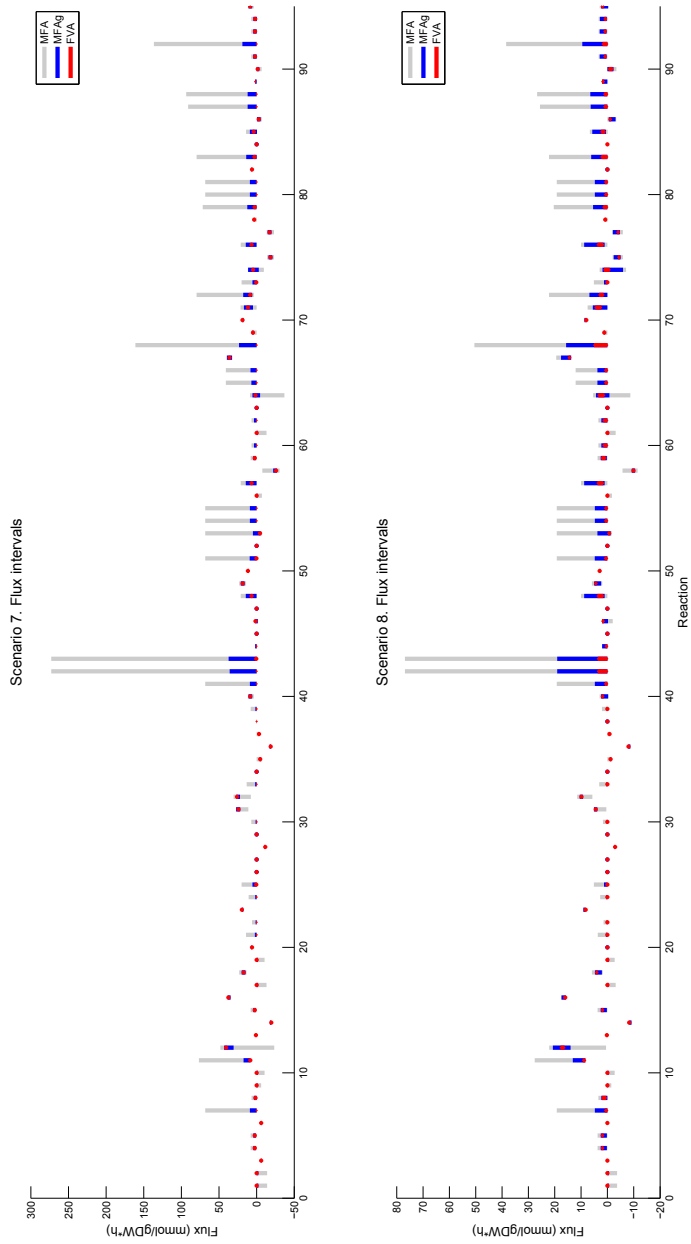


Figure 7.8. *MFA* , *MFAg* and *FVA* interval estimates for scenarios 7 and 8.

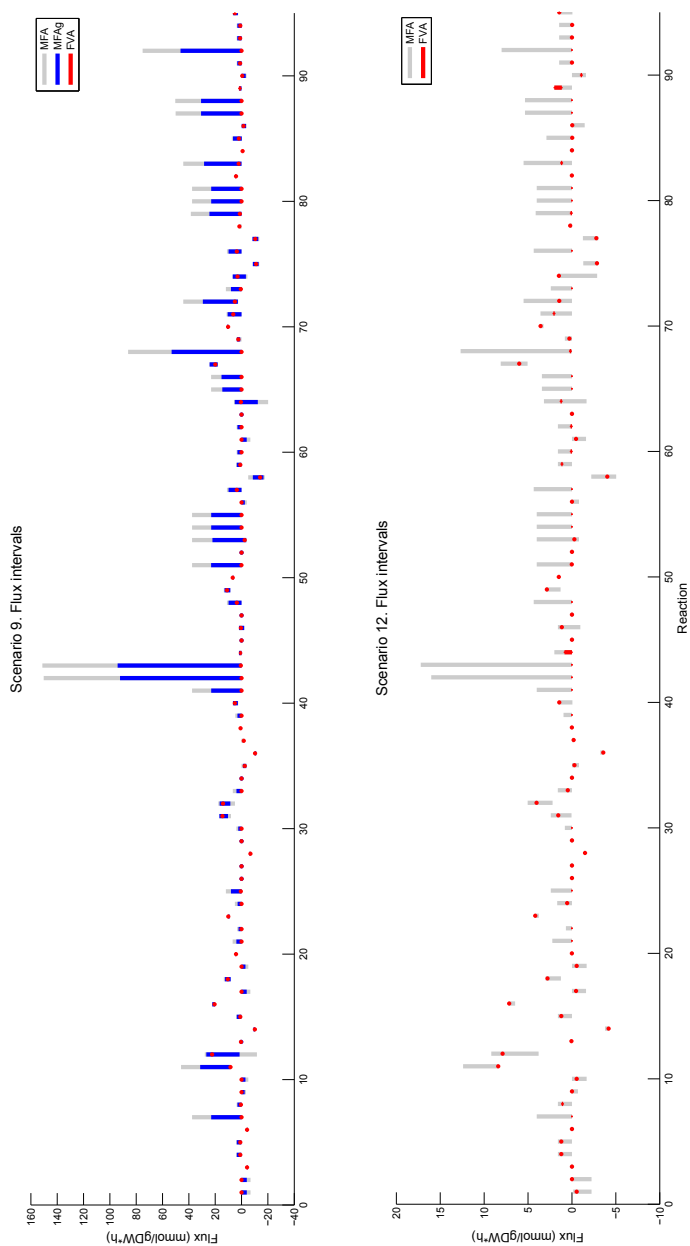


Figure 7.9. MFA , MFAg and FVA interval estimates for scenarios 9 and 12.

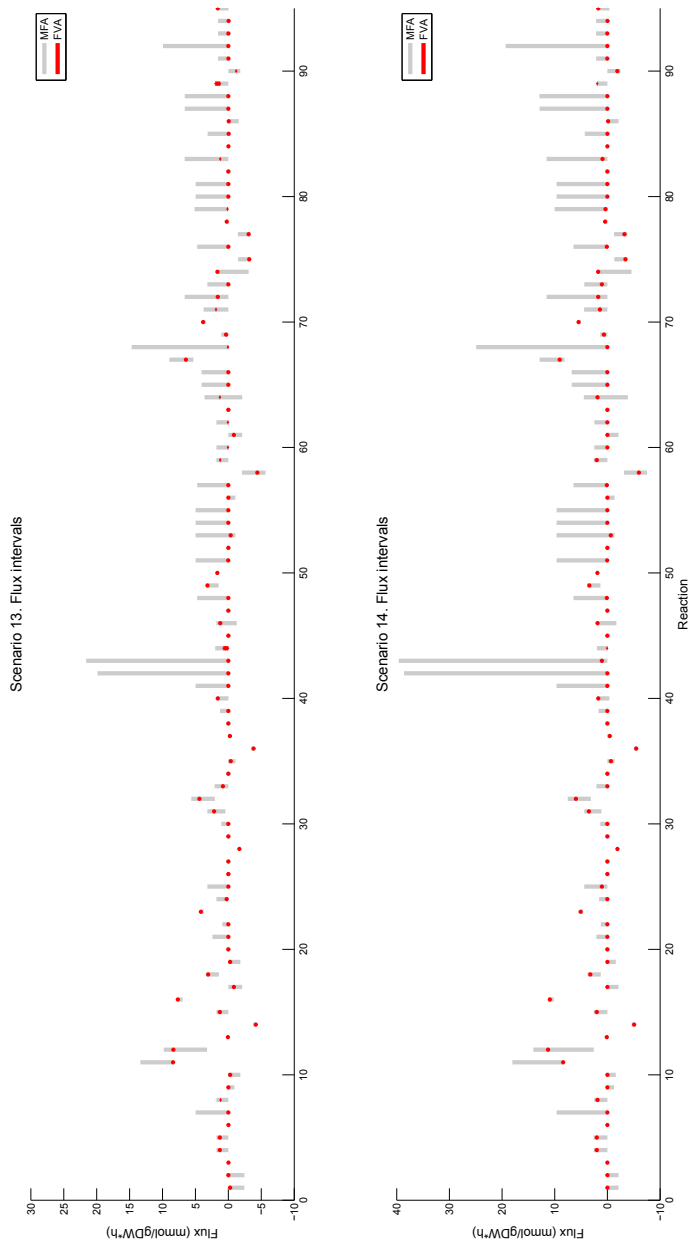


Figure 7.10. MFA , MFA_g and FVA interval estimates for scenarios 13 and 14.

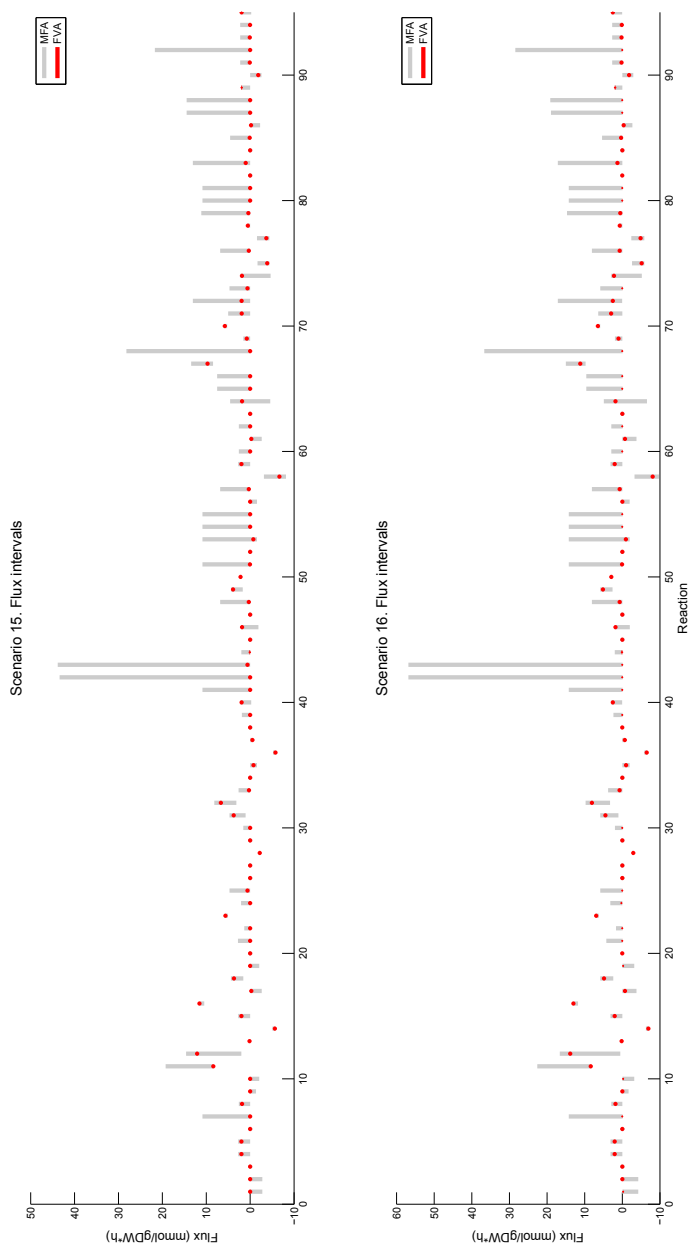


Figure 7.11. MFA , MFAg and FVA interval estimates for scenarios 15 and 16.

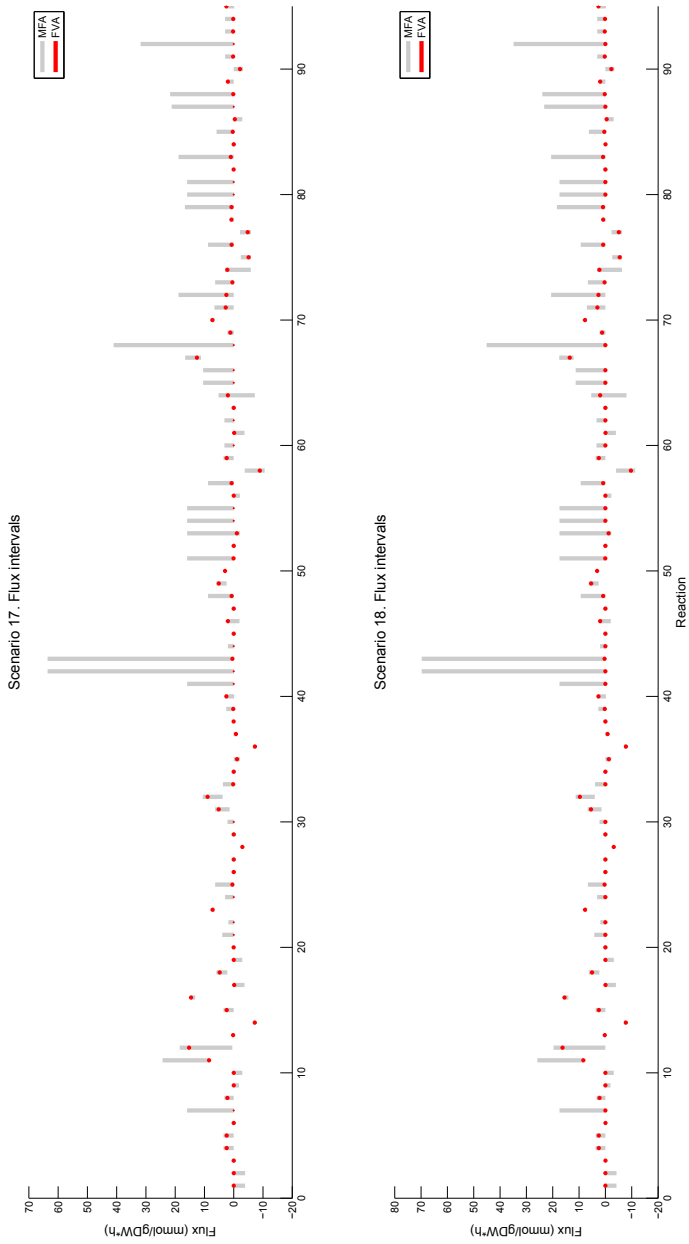


Figure 7.12. *MFA* , *MFA_g* and *FVA* interval estimates for scenarios 17 and 18.

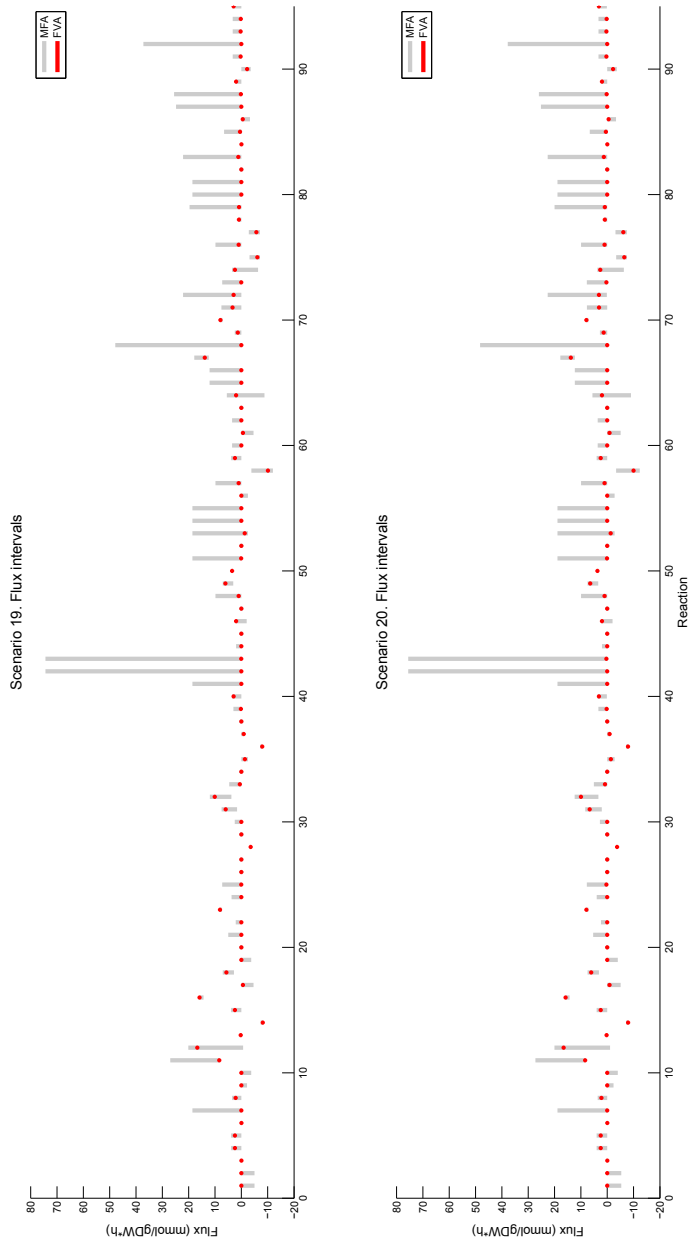


Figure 7.13. MFA , MFAg and FVA interval estimates for scenarios 19 and 20.

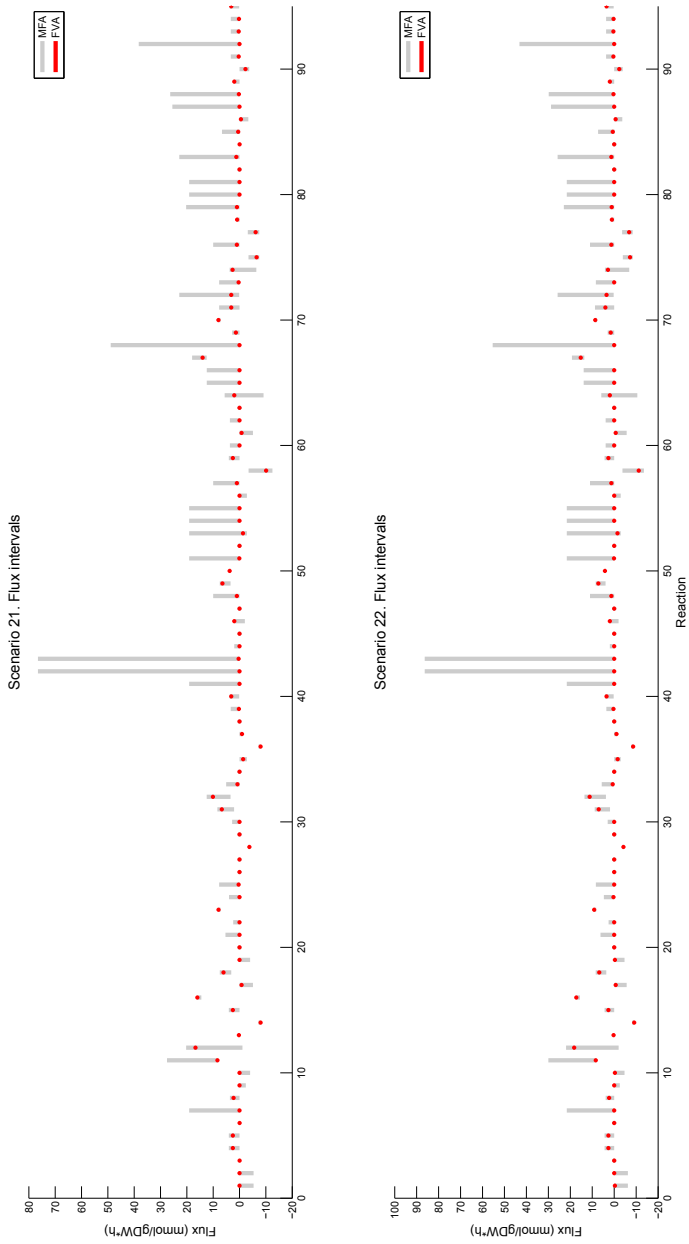


Figure 7.14. MFA , MFA_g and FVA interval estimates for scenarios 21 and 22.

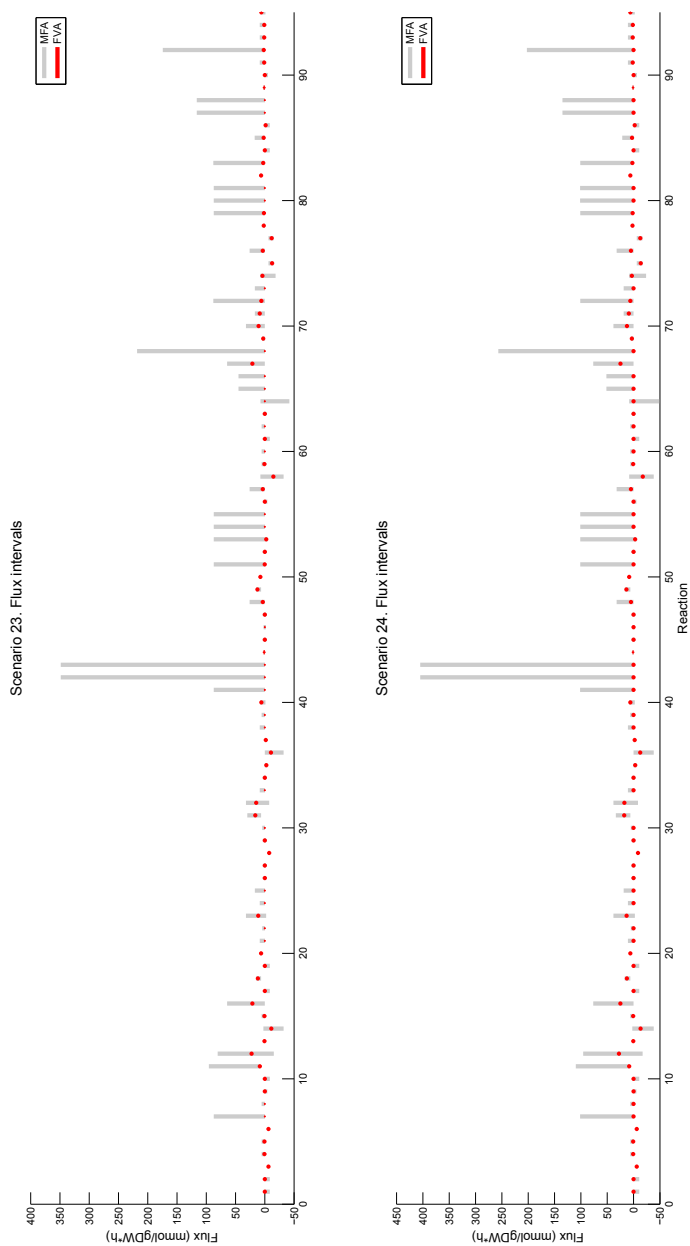


Figure 7.15. MFA , MFAg and FVA interval estimates for scenarios 23 and 24.

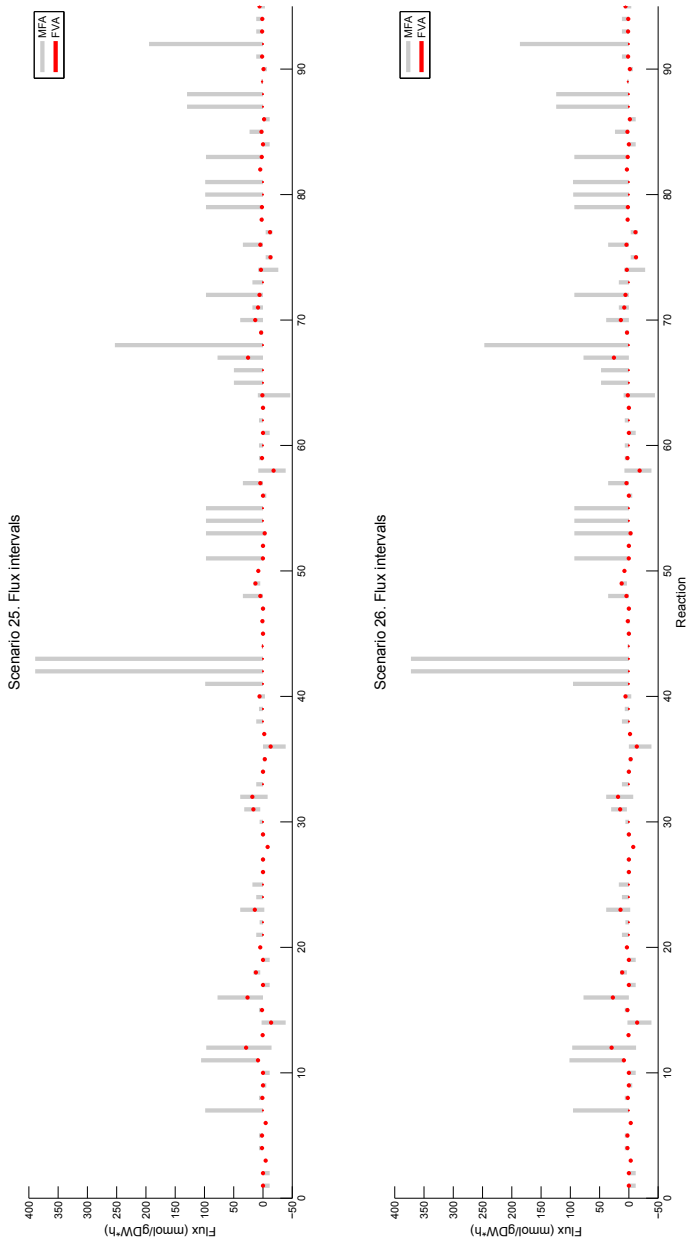


Figure 7.16. *MFA* , *MFA_q* and *FVA* interval estimates for scenarios 25 and 26.

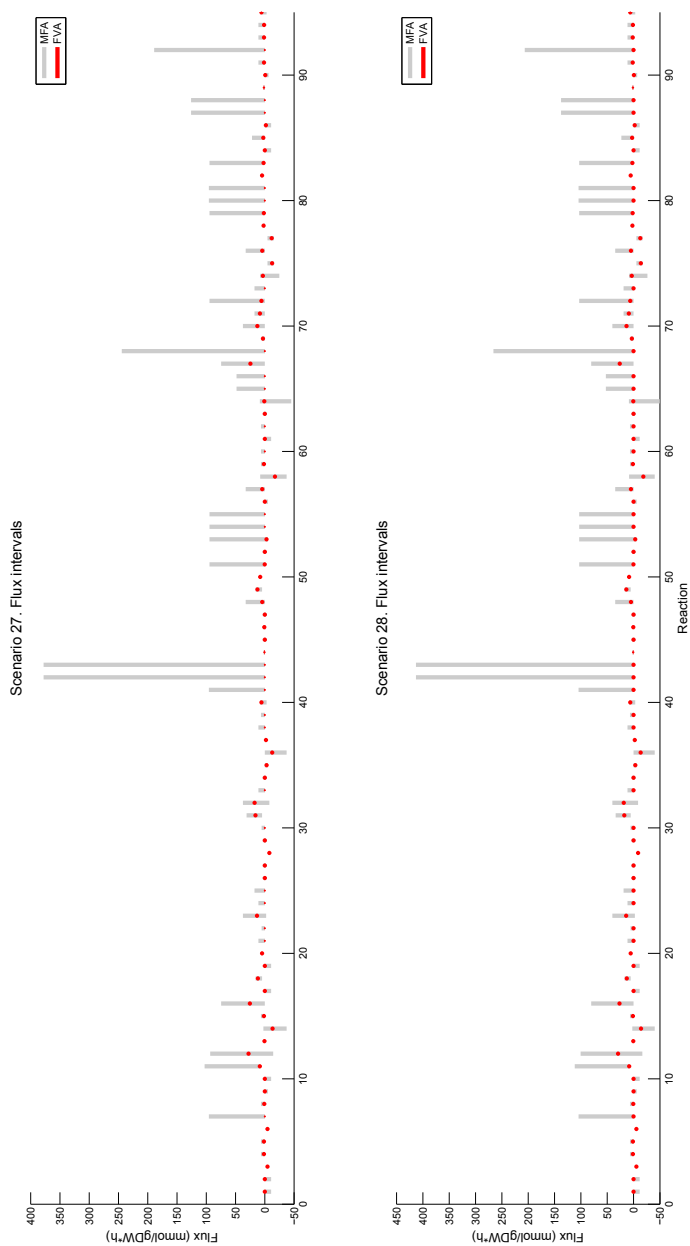


Figure 7.17. MFA , MFAg and FVA interval estimates for scenarios 27 and 28.

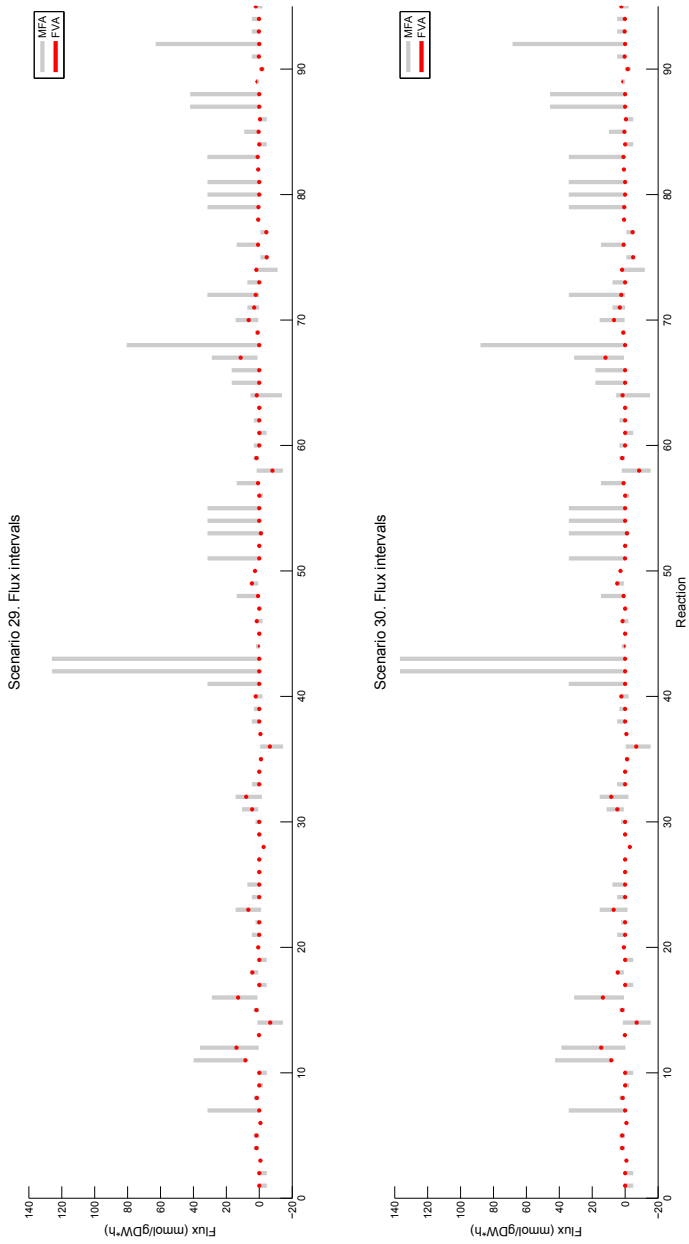


Figure 7.18. MFA , MFA_g and FVA interval estimates for scenarios 29 and 30.

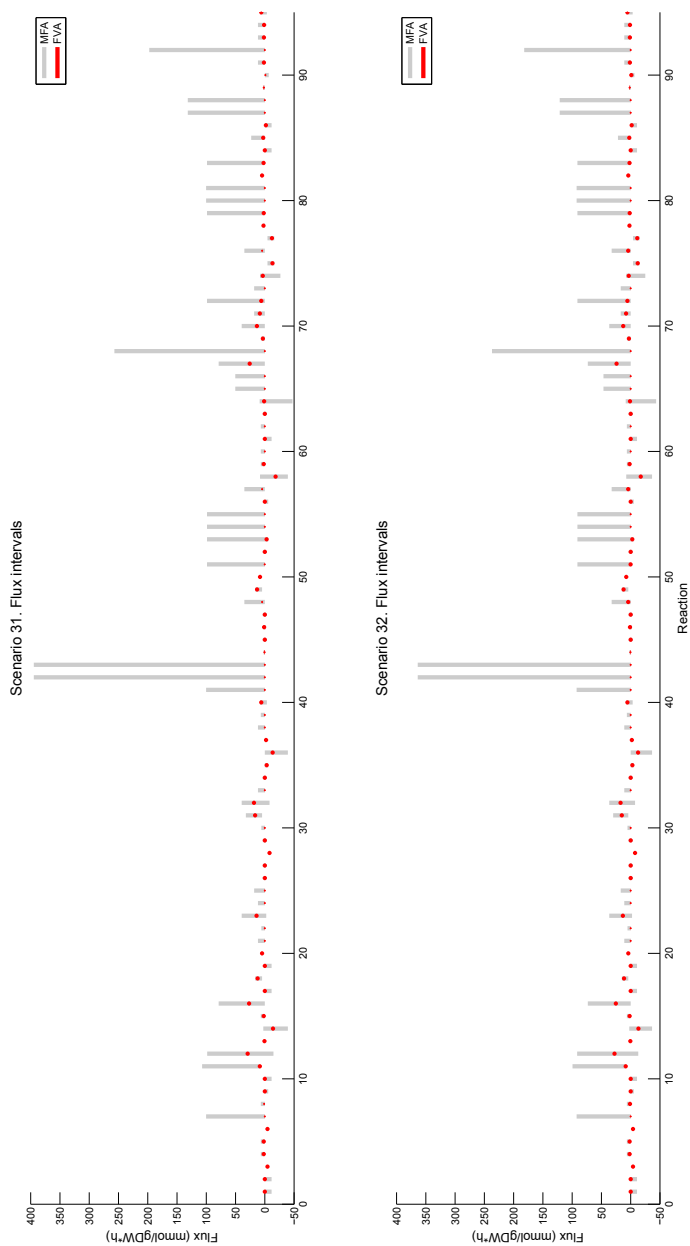


Figure 7.19. MFA , MFAg and FVA interval estimates for scenarios 31 and 32.

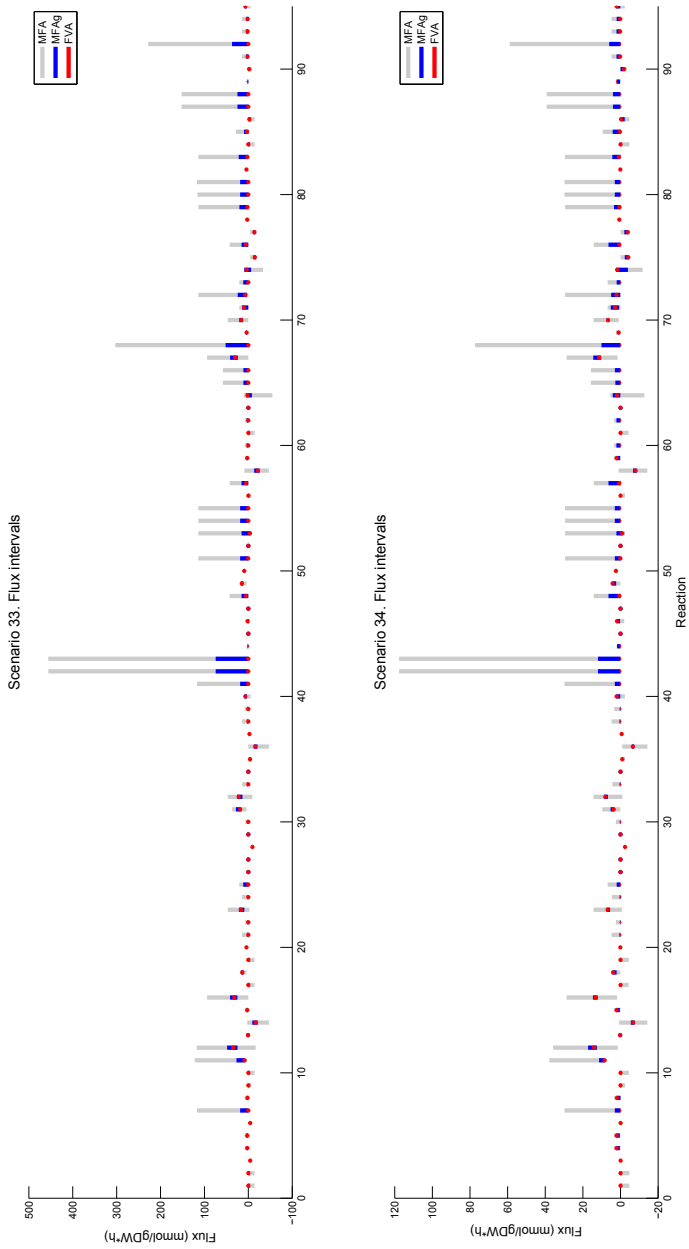


Figure 7.20. MFA , MFAg and FVA interval estimates for scenarios 33 and 34.

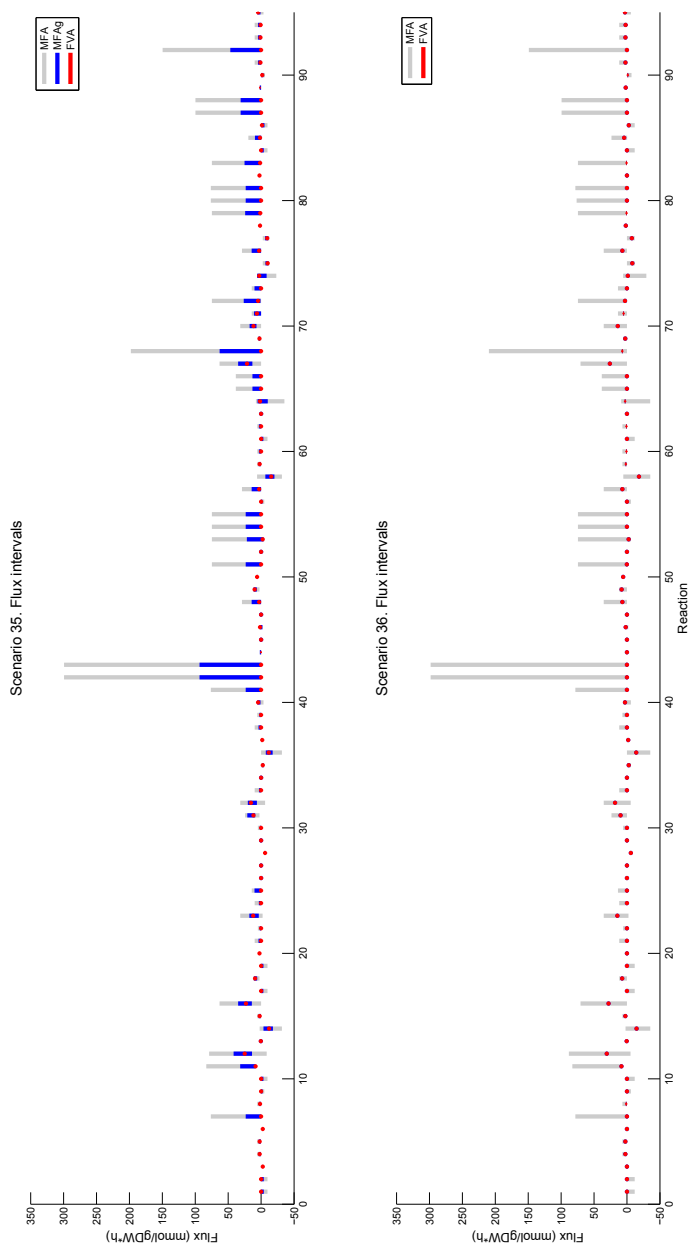


Figure 7.21. MFA , MFAg and FVA interval estimates for scenarios 35 and 36.

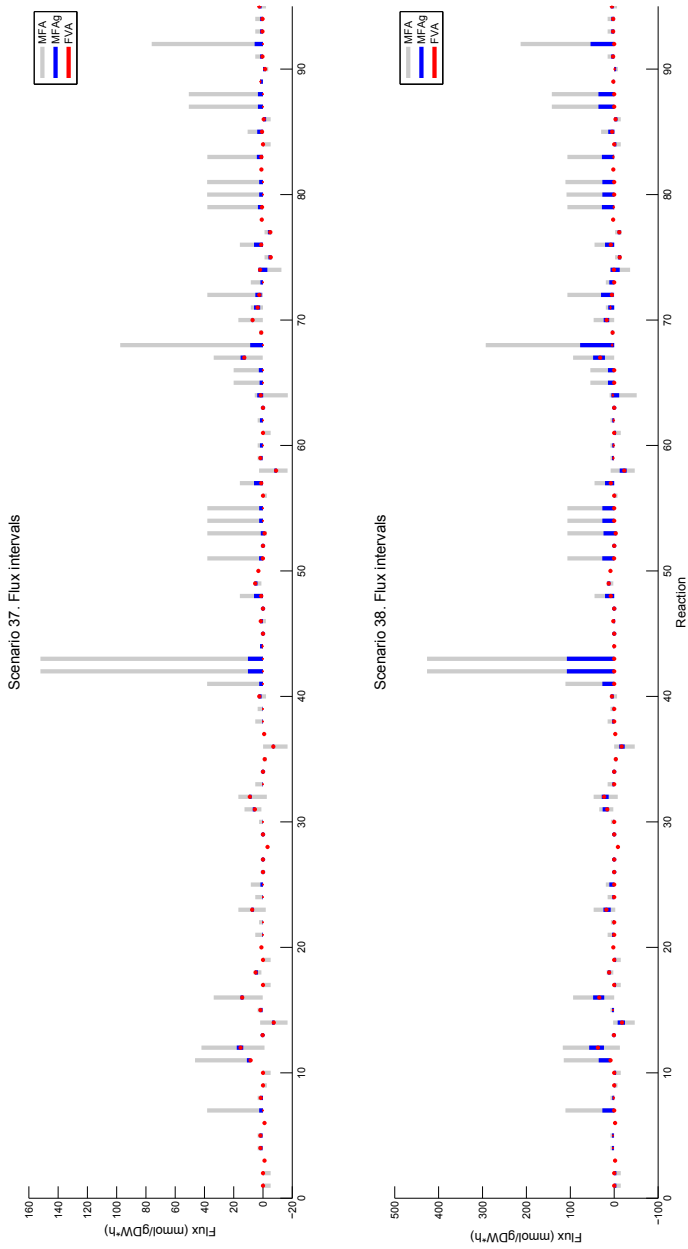


Figure 7.22. MFA , MFAg and FVA interval estimates for scenarios 37 and 38.

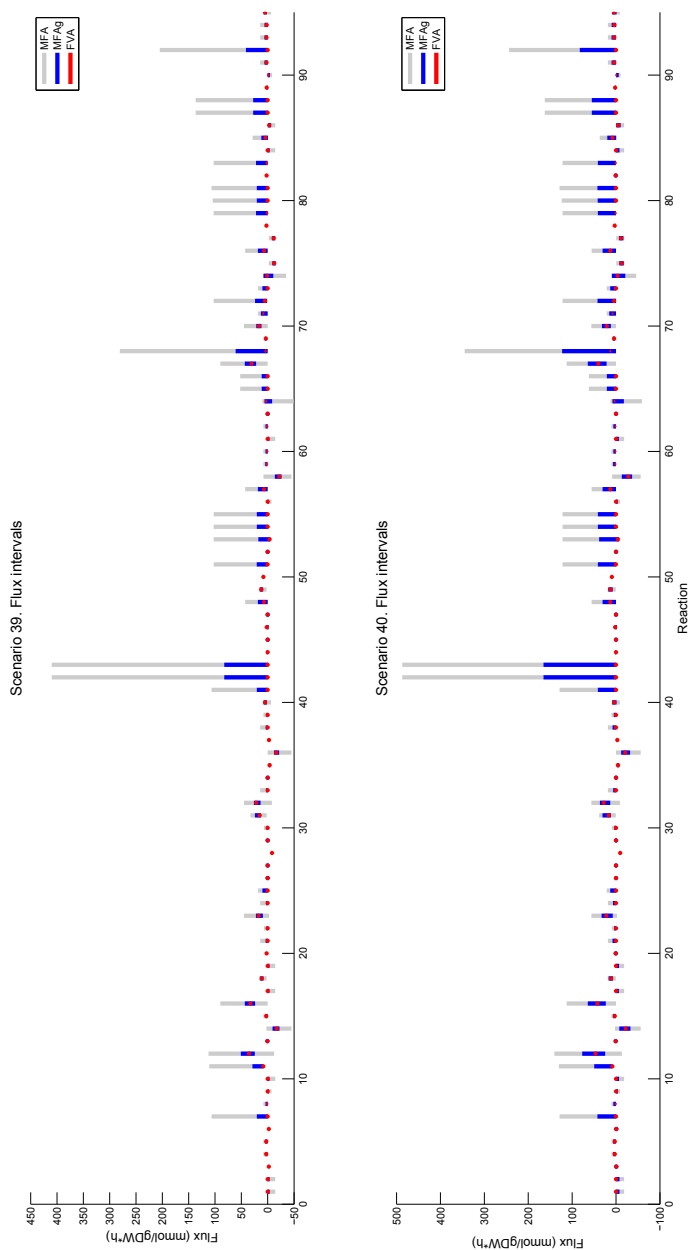


Figure 7.23. MFA , MFAg and FVA interval estimates for scenarios 39 and 40.

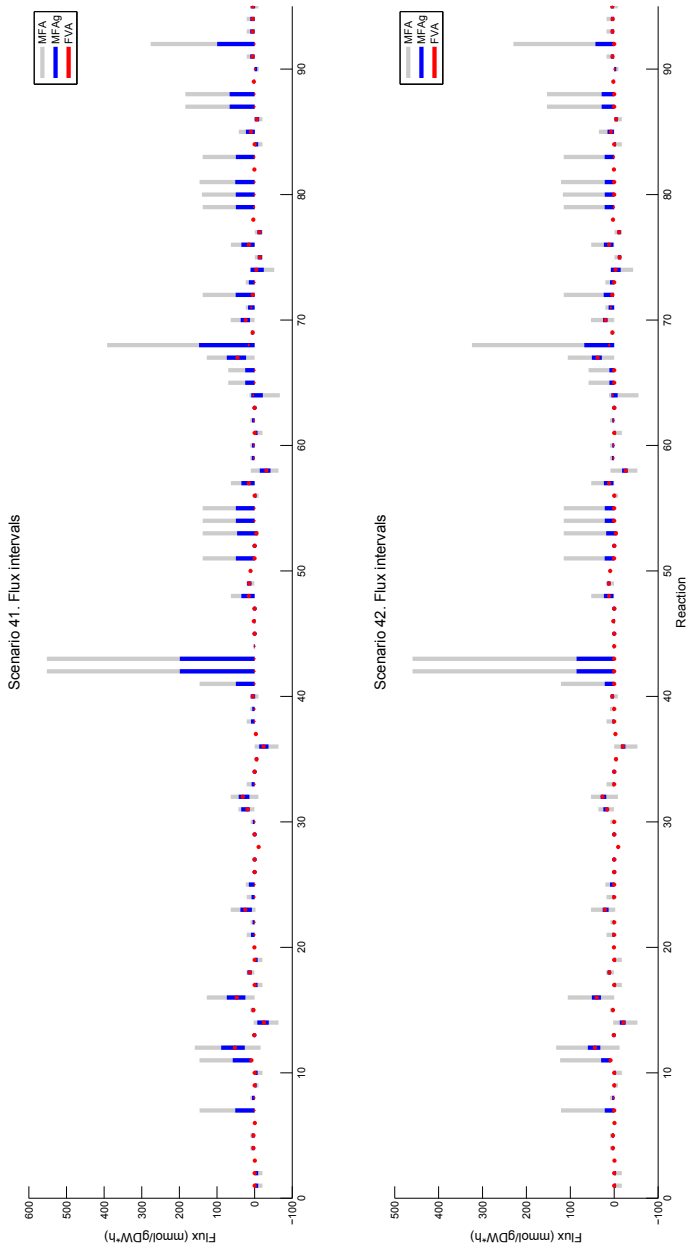


Figure 7.24. MFA , MFAg and FVA interval estimates for scenarios 41 and 42.

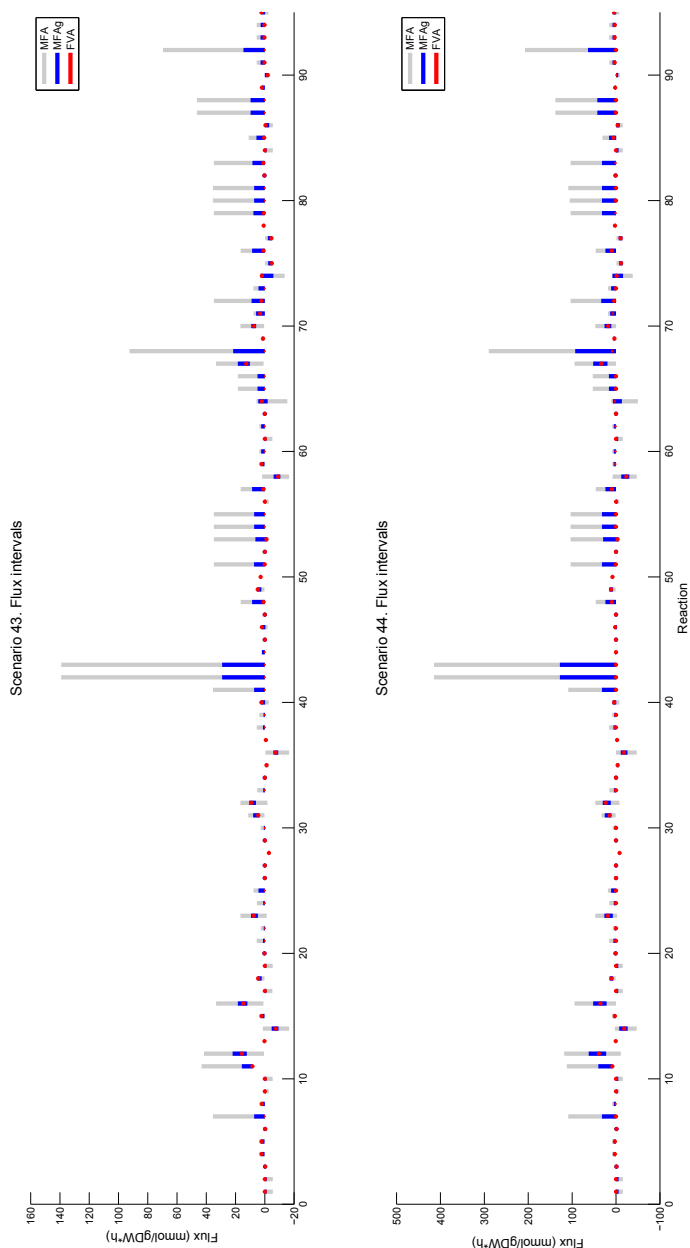


Figure 7.25. MFA , MFAg and FVA interval estimates for scenarios 43 and 44.

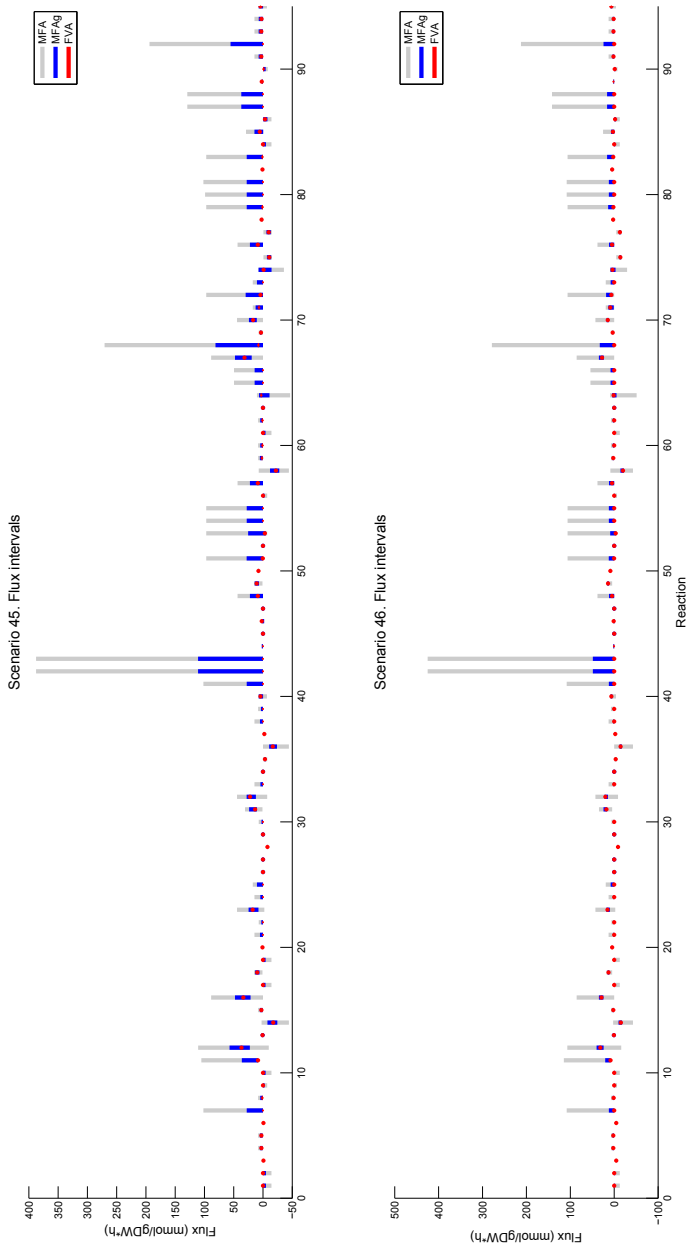


Figure 7.26. MFA , MFAg and FVA interval estimates for scenarios 45 and 46.

7.4 Annex IV: Toy model example matrices.

Matrices for the examples in Figure 5.2 from Rabinowitz and Vastag 2012. Stoichiometric matrix and unipartite projection using equation (5.4):

$$\mathbf{S} = \begin{bmatrix} & R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 & R_8 \\ A & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ B & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ C & 0 & 0 & 1 & 0 & -1 & 0 & 0 & -1 \\ D & 0 & 0 & 0 & 1 & 0 & -1 & -1 & -2 \\ E & 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \end{bmatrix},$$

$$\mathbf{A} = \widehat{\mathbf{S}}^T \widehat{\mathbf{S}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 2 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 & 0 & 2 & 2 \\ 0 & 0 & 0 & 1 & 1 & 2 & 1 & 2 & 2 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 & 1 & 3 & 3 \end{bmatrix}$$

FBA scenario 1 setting and context-dependent network:

$$\begin{array}{rcccc} & \mathbf{v}_{lb_1} & \mathbf{v}_{ub_1} & \mathbf{v}_1^* & \\ R_1 : & 10 & 10 & 10 & \\ R_2 : & 0 & 10 & 10 & \\ R_3 : & 0 & 10 & 4.992 & \\ R_4 : & -10 & 10 & 5.008 & \\ R_5 : & 0 & 10 & 2.492 & \\ R_6 : & 0 & 10 & 0.008 & \\ R_7 : & 0 & 10 & 0 & \\ R_8 : & 0 & 10 & 2.5 & \\ R_{4r} : & -10 & 10 & 0 & \end{array},$$

$$\mathbf{M}_{v_1} = \begin{bmatrix} & R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_8 \\ R_1 & 0 & 10 & 0 & 0 & 0 & 0 & 0 \\ R_2 & 0 & 0 & 4.992 & 5.008 & 0 & 0 & 0 \\ R_3 & 0 & 0 & 0 & 0 & 2.492 & 0 & 2.5 \\ R_4 & 0 & 0 & 0 & 0 & 0 & 0.008 & 5 \\ R_5 & 0 & 0 & 0 & 0 & 0 & 0 & 2.492 \\ R_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0.008 \\ R_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

FBA scenario 2 setting and context-dependent network:

	\mathbf{v}_{lb_2}	\mathbf{v}_{ub_2}	\mathbf{v}_2^*
R_1	10	10	10
R_2	0	10	10
R_3	0	10	3.877
R_4	-10	10	6.123
R_5	0	10	1.877
R_6	0	10	0.123
R_7	2	10	2
R_8	0	10	2
R_{4r}	-10	10	0

$$\mathbf{M}_{v_2} = \begin{bmatrix} & R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 & R_8 \\ R_1 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 0 \\ R_2 & 0 & 0 & 3.877 & 6.123 & 0 & 0 & 0 & 0 \\ R_3 & 0 & 0 & 0 & 0 & 1.877 & 0 & 0 & 2 \\ R_4 & 0 & 0 & 0 & 0 & 0 & 0.123 & 2 & 4 \\ R_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.877 \\ R_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.123 \\ R_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ R_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

FBA scenario 3 setting and context-dependent network:

	\mathbf{v}_{lb_3}	\mathbf{v}_{ub_3}	\mathbf{v}_3^*
R_1	10	10	10
R_2	0	10	10
R_3	0	10	3
R_4	-10	10	7.001
R_5	0	10	0.5
R_6	2	10	2.001
R_7	0	10	0
R_8	0	10	2.5
R_{4r}	-10	10	0

$$\mathbf{M}_{v_3} = \begin{bmatrix} & R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_8 \\ R_1 & 0 & 10 & 0 & 0 & 0 & 0 & 0 \\ R_2 & 0 & 0 & 2.999 & 7.001 & 0 & 0 & 0 \\ R_3 & 0 & 0 & 0 & 0 & 0.499 & 0 & 2.5 \\ R_4 & 0 & 0 & 0 & 0 & 0 & 2.001 & 5 \\ R_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0.499 \\ R_6 & 0 & 0 & 0 & 0 & 0 & 0 & 2.001 \\ R_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

7.5 Annex V: Additional comparison between graphs \mathbf{A} and \mathbf{D}_p

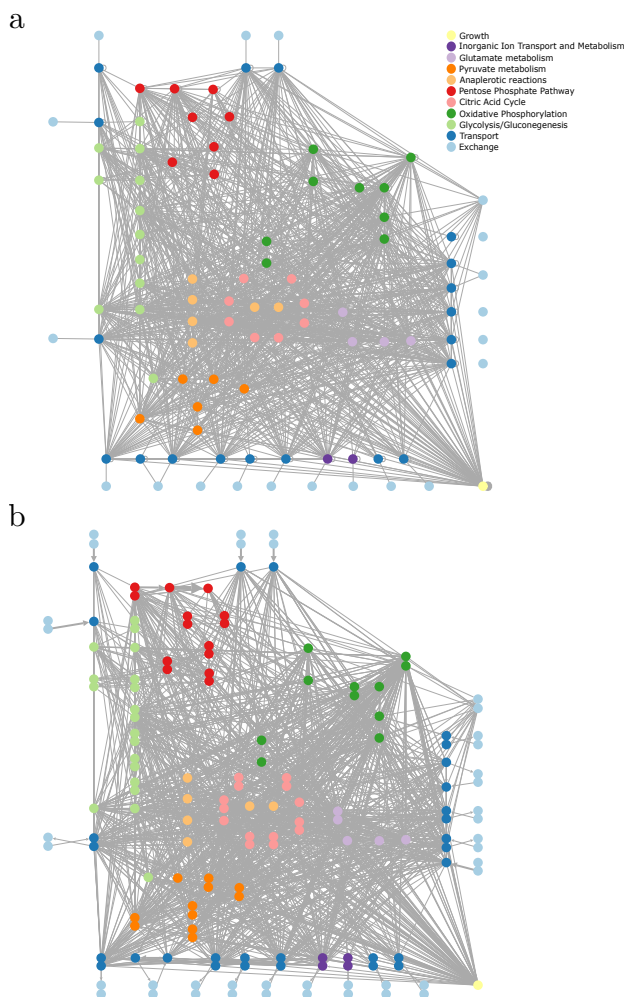


Figure 7.27. Additional comparison of the graphs in the absence of context. (a) The graph \mathbf{A} divided by pathways. (b) The graph \mathbf{D}_p divided by pathways.

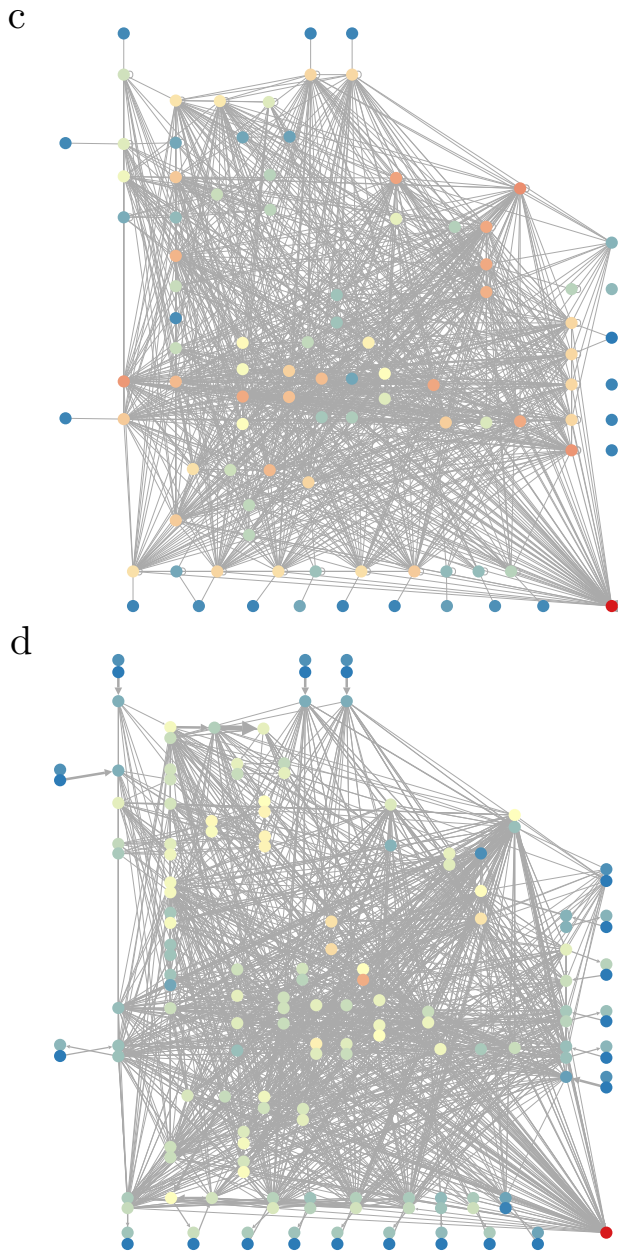


Figure 7.27. Additional comparison of the graphs in the absence of context. (c) The graph A with nodes coloured by their pagerank value. (d) The graph D_p with nodes coloured by their pagerank value.

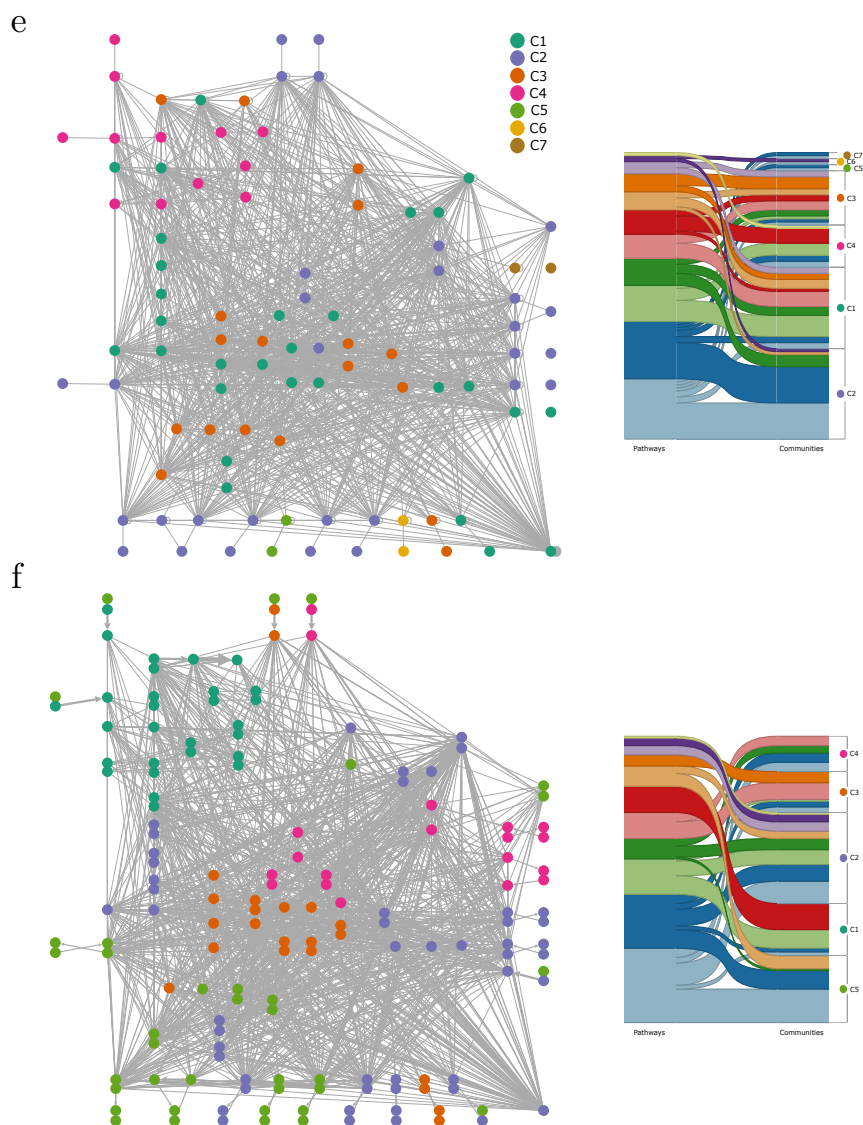


Figure 7.27. Additional comparison of the graphs in the absence of context. (e) The graph \mathbf{A} divided into seven communities and its corresponding alluvial diagram linking traditional pathways with those communities. (f) The graph \mathbf{D}_p divided into five communities and its corresponding alluvial diagram linking traditional pathways with those communities.

7.6 Annex VI: communities from Markov Stability

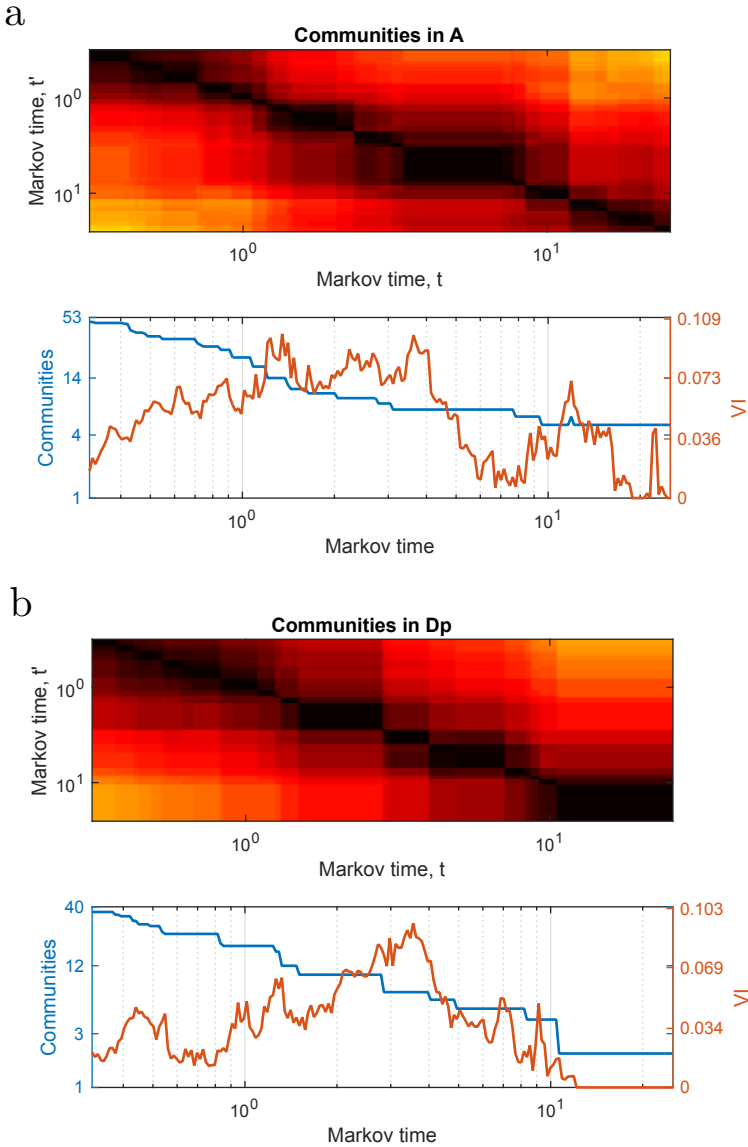


Figure 7.28. Number of communities (blue line) and Variation of Information (red line) in Markov time for **A** and **D_p**.

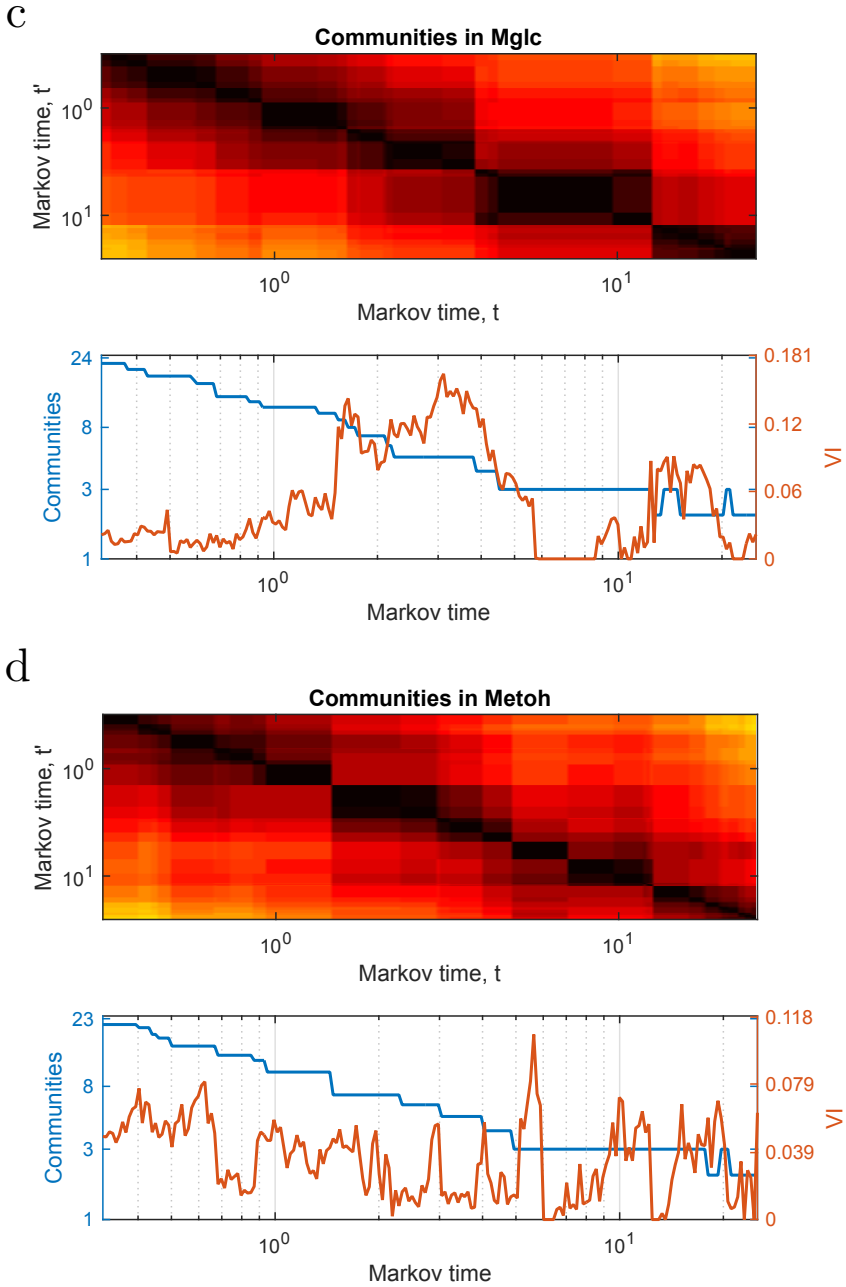


Figure 7.28. Number of communities (blue line) and Variation of Information (red line) in Markov time for M_{glc} and M_{etoh} .

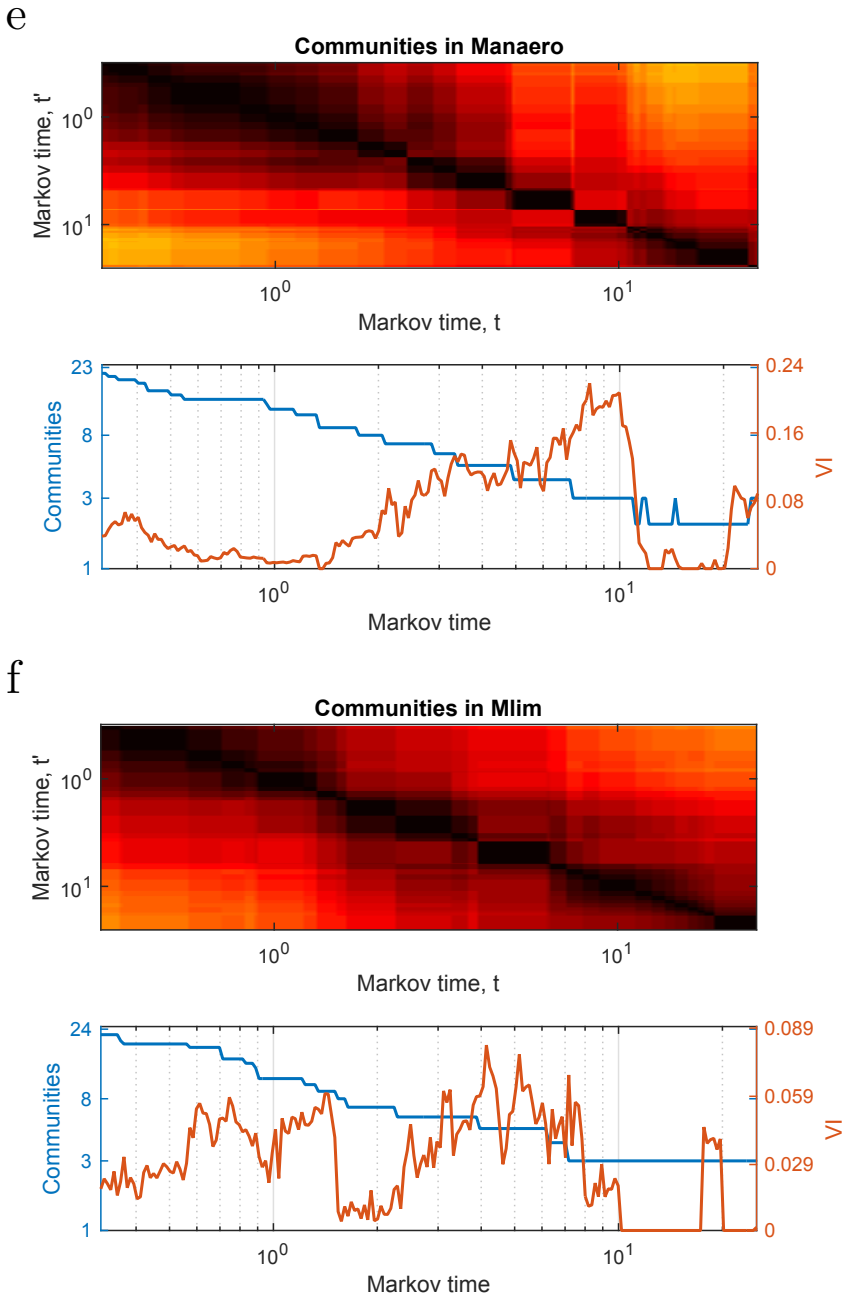


Figure 7.28. Number of communities (blue line) and Variation of Information (red line) in Markov time for M_{anaero} and M_{lim} .

7.6.1 Graph A

The communities in this graph were found at Markov time $t = 6.01$ (Figure 7.28a). The communities at this resolution (Figure 7.27e) are:

- Community C1_A (turquoise) contains all the reactions that consume or produce ATP and water, another two pool metabolites. Production of ATP comes mostly from oxidative phosphorylation (ATPS4r) and substrate level phosphorylation reactions such as phosphofructokinase (PFK), phosphoglycerate kinase (PGK) and succinyl-CoA synthase (SUCOAS). Reactions that consume ATP include glutamine synthetase (GLNS) and ATP maintenance equivalent reaction (ATPM). The reactions L-glutamine transport via ABC system (GLNabc), acetate transport in the form of phosphotransacetilase (PTAr), and acetate kinase (ACKr) are also part of this community. Additionally, C1_A (green) contains also reactions that involve H₂O. Under normal conditions water is assumed to be abundant in the cell, thus the biological link that groups these reactions together is tenuous.
- Community C2_A (purple) includes the reactions NADH dehydrogenase (NADH16) and cytochrome oxidase (CYTBD). These two reactions involve pool metabolites (such as H₂O, H⁺, NADH) which create a large number of connection. Other members include fumarate reductase (FR7) and succinate dehydrogenase (SUCDi) which couple the TCA cycle with the electron transport chain (through ubiquinone-8 reduction and ubiquinol-8 oxidation). Reactions that include export and transport of most secondary carbon sources (such as pyruvate, ethanol, lactate, acetate, malate, fumarate, succinate or glutamate) are included in the community as well. These reactions are included in the community because of their influence in the proton balance of the cell. Most of these reactions do not occur under normal circumstances. This community highlights the fact that in the absence of biological context, many reactions that do not normally interact can be grouped together.
- Community C3_A (orange) contains reactions that produce or consume nicotinamide adenine dinucleotide (NAD⁺), nicotinamide adenine dinucleotide phosphate (NADP⁺), or their reduced variants NADH and NADPH. The main two reactions of the community are NAD(P) transhydrogenase (THD2) and NAD⁺ transhydrogenase (NADTRHD). There are also reactions related to the production of NADH or NADPH in the TCA cycle such as isocitrate dehydrogenase (ICDHyr), 2-oxoglutarate dehydrogenase (AKGDH) and malate dehydrogenase (MDH). The community also includes reactions that are not frequently active such as malic enzyme NAD (ME1) and malic enzyme NADH (ME2) or acetate dehydrogenase (ACALD) and ethanol dehydrogenase (ALCD2x).

- Community C4_A (fuchsia) contains the main carbon intake of the cell (glucose), the initial steps of glycolysis, and most of the pentose phosphate shunt. These reactions are found in this community because the metabolites involved in these reactions (e.g., alpha-D-ribose-5-phosphate (r5p) or D-erythrose-4-phosphate (e4p)) are only found in these reactions. This community includes the biomass reaction due to the number of connections created by growth precursors.
- Communities C5_A (green), C6_A (magenta) and C7_A (brown) are small communities that contain oxygen intake, ammonium intake and acetaldehyde secretion reactions respectively.

7.6.2 Graph D_p

The communities in this graph were found at Markov time $t = 6.28$ (Figure 7.28b). The communities at this resolution (Figure 7.27f) are:

- Community C1_{D_p} (turquoise) includes the first half of the glycolysis and the complete pentose phosphate pathway. The metabolites that create the connections among these reactions such as D-fructose, D-glucose, or D-ribulose.
- Community C2_{D_p} (purple) contains the main reaction that produces ATP through substrate level (PGK, PYK, ACKr) and oxidative phosphorylation (ATPS4r). The flow of metabolites among the reactions in this community includes some pool metabolites such as ATP, ADP, H₂O, and phosphate; however, there are connections created by metabolites that only appear in a handful of reactions such as adenosine monophosphate (AMP) whose sole producer is phosphoenolpyruvate synthase (PPS) and its sole consumer is ATPS4r, being this connection is clearly visible. This community also contains the biomass reaction.
- Community C3_{D_p} (orange) includes the core of the citric acid (TCA) cycle such as citrate synthase (CS), aconitase A/B (ACONTa/b), and anaplerotic reactions such as malate synthase (MALS), malic enzyme NAD (ME1), and malic enzyme NADP (ME2). This community also includes the intake of cofactors such as CO₂.
- Community C4_{D_p} (fuchsia) contains reactions that are secondary sources of carbon such as malate and succinate, as well as oxidative phosphorylation reactions.
- Community C5_{D_p} (green) contains some reactions part of the pyruvate metabolism subsystem such as D-lactate dehydrogenase (LDH-D), pyruvate formate lyase (PFL) or acetaldehyde dehydrogenase (ACALD). In addition, it also includes

the transport reaction for the most common secondary carbon metabolites such as lactate, formate, acetaldehyde and ethanol.

7.6.3 Graph M_{glc}

This graph has 48 reactions with nonzero flux and 227 edges. At Markov time $t = 7.66$ (Figure 7.28c) this graph has a partition into three communities (Figure 5.5a):

- Community $C1_{\text{glc}}$ (turquoise) comprises the intake of glucose and most of the glycolysis and pentose phosphate pathway. The function of the reactions in this community consists of carbon intake and processing glucose into phosphoenolpyruvate (PEP). This community produces essential biocomponents for the cell such as alpha-D-Ribose 5-phosphate (rp5), D-Erythrose 4-phosphate (e4p), D-fructose-6-phosphate (f6p), glyceraldehyde-3-phosphate (g3p) or 3-phospho-D-glycerate (3pg). Other reactions produce energy ATP and have reductive capabilities for catabolism.
- Community $C2_{\text{glc}}$ (purple) contains the electron transport chain which produces the majority of the energy of the cell. In the core *E coli* metabolic model the chain is represented by the reactions NADH dehydrogenase (NADH16), cytochrome oxidase BD (CYTBD) and ATP synthase (ATPS4r). This community also contains associated reactions to the electron transport such as phosphate intake (EXpi(e), PIt2), oxygen intake (EXo2(e), O2t) and proton balance (EXh(e)). This community also includes the two reactions that represent energy maintenance costs (ATPM), and growth (biomass); this is consistent with the biological scenario because ATP is the main substrate for both ATPM, and the biomass reaction.
- Community $C3_{\text{glc}}$ (orange) contains the TCA cycle at its core. The reactions in this community convert PEP into ATP, NADH and NADPH. In contrast with $C0_{\text{glc}}$, there is no precursor formation here. Beyond the TCA cycle, pyruvate kinase (PYK), phosphoenolpyruvate carboxylase (PPC) and pyruvate dehydrogenase (PDH) appear in this community. These reactions highlight the two main carbon intake routes in the cycle: oxalacetate from PEP through phosphoenol pyruvate carboxylase (PPC), and citrate from acetyl coenzyme A (acetyl-CoA) via citrate synthase (CS). Furthermore, both routes begin with PEP, so it is natural for them to belong to the same community along with the rest of the TCA cycle. Likewise, the production of L-glutamate from 2-oxoglutarate (AKG) by glutamate dehydrogenase (GLUDy) is strongly coupled to the TCA cycle.

7.6.4 Graph M_{etoh}

This graph contains 49 reactions and 226 edges. At Markov time $t = 6.28$ (Figure 7.28d) this graph has a partition into three communities (Figure 5.5b):

- Community $C1_{\text{etoh}}$ (turquoise) in this graph is similar to its counterpart in M_{glc} , but with important differences. For example, the reactions in charge of the glucose intake (EXglc(e) and GLCpts) are no longer part of the network (i.e., they have zero flux), and reactions such as malic enzyme NADP (ME2) and phosphoenolpyruvate carboxykinase (PPCK), which now appear in the network, belong to this community. This change in the network reflects the cell's response to a new biological situation. The carbon intake through ethanol has changed the direction of glycolysis into gluconeogenesis (Berg, Tymoczko, and Stryer 2002) (the reactions in $C0_{\text{glc}}$ in Figure 5.5a are now operating in the reverse direction in Figure 5.4b). The main role of the reactions in this community is the production of bioprecursors such as PEP, pyruvate, 3-phospho-D-glycerate (3PG) glyceraldehyde-3-phosphate (G3P), D-fructose-6-phosphate (F6P), and D-glucose-6-phosphate, all of which are substrates for growth. Reactions ME2 and PPCK also belong to this community due to their production of PYR and PEP. Reactions that were in a different community in M_{glc} , such as GLUDy and ICDHy which produce precursors L-glutamate and NADPH respectively, are now part of $C0_{\text{etoh}}$. This community also includes the reactions that produce inorganic substrates of growth such as NH_4 , CO_2 and H_2O .
- Community $C2_{\text{etoh}}$ (purple) contains the electron transport chain and the bulk of ATP production, which is similar to $C2_{\text{glc}}$. However, there are subtle differences that reflect changes in this new scenario. Ethanol intake and transport reactions (EXetoh(e) and ETOHt2r) appear in this community due to their influence in the proton balance of the cell. In addition, $C2_{\text{etoh}}$ contains NADP transhydrogenase (THD2) which is in charge of NADH/NADPH balance. This reaction is present here due to the NAD consumption involved in the reactions ACALD and ethanol dehydrogenase (ALCD2x), which belong to this community as well.
- Community $C3_{\text{etoh}}$ (orange) contains most of the TCA cycle. The main difference between this community and $C1_{\text{glc}}$ is that here acetyl-CoA is extracted from acetaldehyde (which comes from ethanol) by the reaction acetaldehyde dehydrogenase reaction (ACALD), instead of the classical pyruvate from glycolysis. The glyoxylate cycle reactions isocitrate lyase (ICL) and malate synthase (MALS) which now appear in the network, also belong to this community. These reactions are tightly linked to the TCA cycle and appear when the carbon intake is acetate or ethanol to prevent the loss of carbon as CO_2 .

7.6.5 Graph M_{anaero}

This graph contains 47 reactions and 212 edges. At Markov time $t = 6.01$ (Figure 7.28e) this graph has a partition into four communities (Figure 5.5c):

- Community $C1_{\text{anaero}}$ (turquoise) contains the reactions responsible D-glucose intake (EX_{glc}) and most of the glycolysis. The reaction that represents the cellular maintenance energy cost, ATP maintenance requirement (ATPM), is included in this community because of the increased strength of its connection to the substrate-level phosphorylation reaction phosphoglycerate kinase (PGK). Also note that reactions in the pentose phosphate pathway do not belong to the same community as the glycolysis reactions (unlike in M_{glc} and M_{etoh}).
- Community $C2_{\text{anaero}}$ (purple) contains the conversion of PEP into formate through the sequence of reactions PYK, PFL, FORT_i and EX_{for(e)}. More than half of the carbon secreted by the cell becomes formate.
- Community $C3_{\text{anaero}}$ (orange) includes the biomass reaction and the reactions in charge of supplying it with substrates. These reactions include the pentose phosphate pathway (now detached from $C0_{\text{glc}}$), which produce essential growth precursors such as alpha-D-ribose-5-phosphate (r5p) or D-erythrose-4-phosphate (e4p). The TCA cycle is present as well because its production of two growth precursors: 2-oxalacetate and NADPH. Finally, the reactions in charge of acetate production (ACK_r, ACT_{2r} and EX_{ac(e)}) are also members of this community through the ability of ACK_r to produce ATP. Glutamate metabolism reaction GLUD_y is also included in this community. It is worth mentioning that the reverse of ATP synthase (ATPS_{4r}) is present in this community because here, unlike in M_{glc} , ATPS_{4r} consumes ATP instead of producing it. When this flux is reversed, then ATPS_{4r} is in part responsible for pH homeostasis.
- Community $C4_{\text{anaero}}$ (fuchsia) includes the main reactions involved in NADH production and consumption, which occurs via glyceraldehyde-3-phosphate dehydrogenase (GAPD). NADH consumption occurs in two consecutive steps in ethanol production: in ACALD and ALCD_{2x}. The phosphate intake and transport reactions EX_{pi(e)} and PIT_{2r} belong to this community because most of the phosphate consumption takes place at GAPD. Interestingly, the core reaction around which the community forms (GAPD) is not present in the community. It is included in earlier Markov times but when communities start to get larger the role of GAPD becomes more relevant as a part of the glycolysis than its role as a NADH hub. This is a good example of how the graph structure and the clustering method are able to capture two different roles in the same metabolite.

7.6.6 Graph M_{lim}

This graph has 52 nodes and 228 edges. At Markov time $t = 13$ this graph (Figure 7.28f) has a partition into three communities (Figure 5.5d):

- Community $C0_{\text{lim}}$ (turquoise) contains the glycolysis pathway (detached from the pentose phosphate pathway). This community is involved in precursor formation, ATP production, substrate-level phosphorylation and processing of D-glucose into PEP.
- Community $C2_{\text{lim}}$ (purple) contains the bioenergetic machinery of the cell; the main difference to the previous scenarios is that the electron transport chain has a smaller role in ATP production (ATPS4r), and substrate-level phosphorylation (PGK, PYK, SUCOAS, ACKr) becomes more important. In M_{lim} the electron transport chain is responsible for the 21.8% of the total ATP produced in the cell while in M_{glc} it produces 66.5%. The reactions in charge of intake and transport of inorganic ions such as phosphate (EXpi(e) and Pit2r), O_2 (EXO₂(e) and O₂t) and H_2O (EXH₂O and H₂Ot) belong to this community as well. This community includes the reactions in the pentose phosphate pathway that produce precursors for growth: transketolase (TKT2) produces e4p, and ribose-5-phosphate isomerase (RPI) produces r5p.
- Community $C3_{\text{lim}}$ (orange) is the community that differs the most from those in the other aerobic growth networks (M_{glc} and M_{etoh}). This community gathers reactions that under normal circumstances would not be so strongly related but that the limited availability of ammonium and phosphate have forced together; its members include reactions from the TCA cycle, the pentose phosphate pathway, nitrogen metabolism and by-product secretion. The core feature of the community is carbon secretion as formate and acetate. Reactions PPC, malate dehydrogenase (MDH) reverse and ME2 channel most of the carbon to the secretion routes in the form of formate and acetate. The production of L-glutamine seems to be attached to this subsystem through the production of NADPH in ME2 and its consumption in the glutamate dehydrogenase NAPD (GLUDy).

Bibliography

- Abdi, Hervé (2010). “Partial least squares regression and projection on latent structure regression (PLS Regression)”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.1, pp. 97–106.
- Abdi, Hervé and Lynne J. Williams (2010). “Principal component analysis”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, pp. 433–459.
- Adams, Michael J., John F. Antoniw, and Frederic Beaudoin (2005). “Overview and analysis of the polyprotein cleavage sites in the family Potyviridae”. In: *Molecular Plant Pathology* 6.4, pp. 471–487.
- Aerts, Stein et al. (2006). “Gene prioritization through genomic data fusion”. In: *Nature Biotechnology* 24.5, pp. 537–544.
- Albert, Réka and Albert-László Barabási (2002). “Statistical mechanics of complex networks”. In: *Reviews of Modern Physics* 74.1, pp. 47–97.
- Albert, Réka et al. (2011). “Computationally efficient measure of topological redundancy of biological and social networks”. In: *Physical Review E* 84.3, p. 036117.
- Alberts, B. et al. (2014). *Molecular Biology of the Cell*. Garland Science.
- Allison, R., R. E. Johnston, and W. G. Dougherty (1986). “The nucleotide sequence of the coding region of tobacco etch virus genomic RNA: evidence for the synthesis of a single polyprotein”. In: *Virology* 154.1, pp. 9–20.
- Alon, Uri (2007). “Network motifs: theory and experimental approaches.” In: *Nature reviews. Genetics* 8.6, pp. 450–61.

- Amor, B et al. (2014). “Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection.” In: *Molecular bioSystems* 10.8, pp. 2247–58.
- Amos, Martyn et al. (2015). “Bacterial computing with engineered populations”. In: *Phil. Trans. R. Soc. A* 373.2046, p. 20140218.
- Arabidopsis Interactome Mapping Consortium (2011). “Evidence for network evolution in an Arabidopsis interactome map”. In: *Science (New York, N.Y.)* 333.6042, pp. 601–607.
- Arita, Masanori (2004). “The metabolic world of Escherichia coli is not small.” In: *Proceedings of the National Academy of Sciences of the United States of America* 101.6, pp. 1543–7.
- Bacik, Karol A et al. (2015). “Flow-based network analysis of the Caenorhabditis elegans connectome”. In: *arXiv:1511.00673*.
- Barabási, Albert-László and Réka Albert (1999). “Emergence of Scaling in Random Networks”. In: *Science* 286.5439, pp. 509–512.
- Barabási, Albert-László and Zoltán N. Oltvai (2004). “Network biology: understanding the cell’s functional organization”. In: *Nature Reviews Genetics* 5.2, pp. 101–113.
- Barrett, Christian L., Markus J. Herrgard, and Bernhard Palsson (2009). “Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation”. In: *BMC Systems Biology* 3, p. 30.
- Basan, Markus et al. (2015). “Overflow metabolism in Escherichia coli results from efficient proteome allocation”. In: *Nature* 528.7580, pp. 99–104.
- Baumbach, Jan, Sven Rahmann, and Andreas Tauch (2009). “Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms”. In: *BMC systems biology* 3, p. 8.
- Beguerisse-Díaz, M. et al. (2016). “Context-dependent metabolic networks”. In: *arXiv:1605.01639 [physics, q-bio]*.
- Beguerisse-Díaz, Mariano et al. (2014). “Interest communities and flow roles in directed networks: the Twitter network of the UK riots.” In: *J R Soc Interface* 11.101.

- Berg, J.M., J.L. Tymoczko, and L. Stryer (2002). *Biochemistry, Fifth Edition*. W. H. Freeman. ISBN: 9780716730514.
- Berggård, Tord, Sara Linse, and Peter James (2007). “Methods for the detection and analysis of protein-protein interactions”. In: *Proteomics* 7.16, pp. 2833–2842.
- Bertalanffy, L.v. (1968). *General system theory: foundations, development, applications*. G. Braziller.
- Blondel, Vincent D et al. (2008). “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008.
- Boccaletti, S. et al. (2006). “Complex networks: Structure and dynamics”. In: *Physics Reports* 424, pp. 175–308.
- Börnke, Frederik (2008). “Protein Interaction Networks”. In: *Analysis of Biological Networks*. Ed. by Björn H. Junker and Falk Schreiber. John Wiley & Sons, Inc., pp. 207–232.
- Bosque, G. et al. (2014). “Topology analysis and visualization of Potyvirus protein-protein interaction network”. In: *BMC Systems Biology* 8.1, p. 129.
- Brückner, Anna et al. (2009). “Yeast two-hybrid, a powerful tool for systems biology”. In: *International Journal of Molecular Sciences* 10.6, pp. 2763–2788.
- Bruggeman, F. J. and H. V. Westerhoff (2007). “The nature of systems biology”. In: *Trends in Microbiology* 15.1, pp. 45–50.
- Buydens, Lutgarde (2013). “Towards Tsunami-Resistant Chemometrics”. In: *The Analytical Scientist* 813, pp. 24–30.
- Carrasco, Purificación, Francisca de la Iglesia, and Santiago F. Elena (2007). “Distribution of Fitness and Virulence Effects Caused by Single Nucleotide Substitutions in Tobacco Etch Virus”. In: *Journal of Virology* 81.23, pp. 12979–12984.
- Carrera, Javier, Santiago F. Elena, and Alfonso Jaramillo (2012). “Computational design of genomic transcriptional networks with adaptation to varying environments”. In: *Proceedings of the National Academy of Sciences* 109.38, pp. 15277–15282.

- Carrera, Javier et al. (2009). “Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions”. In: *Genome Biology* 10, R96.
- Chang, Roger L. et al. (2010). “Drug off-target effects predicted using structural analysis in the context of a metabolic network model.” In: *PLoS Comput Biol* 6.9, e1000938.
- Charles Averre, North Carolina State University (2009). *Sweet Potato Feathery Mottle Virus (Potyvirus SPFMV)*. [Online; accessed July 29, 2016]. URL: <http://www.forestryimages.org/browse/detail.cfm?imgnum=1563214>.
- Chassagnole, Christophe et al. (2002). “Dynamic modeling of the central carbon metabolism of Escherichia coli”. In: *Biotechnology and Bioengineering* 79.1, pp. 53–73.
- Chellaboina, V. et al. (2009). “Modeling and analysis of mass-action kinetics”. In: *IEEE Control Systems* 29.4, pp. 60–78.
- Cho, Dong-Yeon, Yoo-Ah Kim, and Teresa M. Przytycka (2012). “Network Biology Approach to Complex Diseases”. In: *PLoS Comput Biol* 8.12, e1002820.
- Chung, Betty Y.W. et al. (2008). “An overlapping essential gene in the Potyviridae”. In: *Proceedings of the National Academy of Sciences* 105.15, pp. 5897–5902.
- Colijn, Caroline et al. (2009). “Interpreting Expression Data with Metabolic Flux Models: Predicting *Mycobacterium tuberculosis* Mycolic Acid Production”. In: *PLoS Comput Biol* 5.8, e1000489.
- Conesa, Ana et al. (2010). “A multiway approach to data integration in systems biology based on Tucker and N-PLS”. In: *Chemometrics and Intelligent Laboratory Systems* 104.1, pp. 101–111.
- Costa, Rafael S. et al. (2014). “An extended dynamic model of Lactococcus lactis metabolism for mannitol and 2,3-butanediol production”. In: *Molecular BioSystems* 10.3, pp. 628–639.
- Covert, M W, C H Schilling, and B Palsson (2001). “Regulation of gene expression in flux balance models of metabolism”. In: *Journal of theoretical biology* 213.1, pp. 73–88.

- Covert, Markus W. et al. (2004). “Integrating high-throughput and computational data elucidates bacterial networks”. In: *Nature* 429.6987, pp. 92–96.
- Crick, F. (1970). “Central Dogma of Molecular Biology”. In: *Nature* 227, pp. 561–563.
- Croes, Didier et al. (2006). “Inferring meaningful pathways in weighted metabolic networks”. In: *Journal of Molecular Biology* 356.1, pp. 222–236.
- Csermely, Péter, Vilmos Ágoston, and Sándor Pongor (2005). “The efficiency of multi-target drugs: the network approach might help drug design”. In: *Trends in Pharmacological Sciences* 26.4, pp. 178–182.
- Culver, James N. and Meenu S. Padmanabhan (2007). “Virus-induced disease: altering host physiology one interaction at a time”. In: *Annual Review of Phytopathology* 45.1, pp. 221–243.
- Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. J. Murray.
- Dayhoff, M. O. and R. M. Schwartz (1978). “Chapter 22: A model of evolutionary change in proteins”. In: *Atlas of Protein Sequence and Structure*. Vol. 5, pp. 345–358.
- De Las Rivas, Javier and Celia Fontanillo (2010). “Protein-protein interactions essentials: key concepts to building and analyzing interactome networks”. In: *PLoS computational biology* 6.6, e1000807.
- Delvenne, J.-C., S. N. Yaliraki, and M. Barahona (2010). “Stability of graph communities across time scales”. In: *Proceedings of the National Academy of Sciences* 107.29, pp. 12755–12760.
- Delvenne, Jean-Charles et al. (2013). “The Stability of a Graph Partition: A Dynamics-Based Framework for Community Detection”. In: *Dynamics On and Of Complex Networks, Volume 2*. Ed. by Animesh Mukherjee et al. Modeling and Simulation in Science, Engineering and Technology. Springer New York, pp. 221–242.
- Dittrich, Marcus T. et al. (2008). “Identifying functional modules in protein-protein interaction networks: an integrated exact approach”. In: *Bioinformatics* 24.13, pp. 223–231.

- Domier, L. L. et al. (1986). “The nucleotide sequence of tobacco vein mottling virus RNA”. In: *Nucleic acids research* 14.13, pp. 5417–5430.
- Duarte, Natalie C. et al. (2007). “Global reconstruction of the human metabolic network based on genomic and bibliomic data”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.6, pp. 1777–1782.
- Dunn, Irving J. (2003). *Biological Reaction Engineering: Dynamic Modelling Fundamentals with Simulation Examples*. Wiley-VCH.
- Edwards, J. S., R. U. Ibarra, and B. O. Palsson (2001). “In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data”. In: *Nature Biotechnology* 19.2, pp. 125–130.
- Edwards, J. S. and B. Ø. Palsson (1999). “Systems Properties of the *Haemophilus influenzae* Rd Metabolic Genotype”. In: *Journal of Biological Chemistry* 274.25, pp. 17410–17416.
- Edwards, J. S. and B. O. Palsson (2000). “The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities”. In: *Proceedings of the National Academy of Sciences* 97.10, pp. 5528–5533.
- Edwards, Jeremy S., Markus Covert, and Bernhard Palsson (2002). “Metabolic modelling of microbes: the flux-balance approach”. In: *Environmental Microbiology* 4.3, pp. 133–140.
- Efron, Bradley (1981). “Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods”. In: *Biometrika* 68.3, pp. 589–599.
- Elena, Santiago F., Javier Carrera, and Guillermo Rodrigo (2011). “A systems biology approach to the evolution of plant-virus interactions”. In: *Current Opinion in Plant Biology* 14.4, pp. 372–377.
- Elena, Santiago F. and Guillermo Rodrigo (2012). “Towards an integrated molecular model of plant-virus interactions”. In: *Current Opinion in Virology* 2.6, pp. 719–724.
- Emmerling, Marcel et al. (2002). “Metabolic Flux Responses to Pyruvate Kinase Knockout in *Escherichia coli*”. In: *Journal of Bacteriology* 184.1, pp. 152–164.
- Erdős, Paul (1959). “On Random Graphs I.” In: *Publicationes Mathematicae (Debrecen)* 6, pp. 290–297.

- Even, Sergine, Nic D. Lindley, and Muriel Coccagn-Bousquet (2003). “Transcriptional, translational and metabolic regulation of glycolysis in *Lactococcus lactis* subsp. *cremoris* MG 1363 grown in continuous acidic cultures”. In: *Microbiology* 149.7, pp. 1935–1944.
- Feist, Adam M. and Bernhard O Palsson (2008). “The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*”. In: *Nature Biotechnology* 26.6, pp. 659–667.
- Feist, Adam M and Bernhard O Palsson (2010). “The biomass objective function”. In: *Current Opinion in Microbiology* 13.3, pp. 344–349.
- Feist, Adam M. et al. (2007). “A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information”. In: *Molecular Systems Biology* 3, p. 121.
- Fell, D. A. (1992). “Metabolic control analysis: a survey of its theoretical and experimental development.” In: *Biochemical Journal* 286.2, pp. 313–330.
- Ferreira, Ana R. et al. (2011). “Projection to latent pathways (PLP): a constrained projection to latent variables (PLS) method for elementary flux modes discrimination”. In: *BMC Systems Biology* 5, p. 181.
- Fields, S and O Song (1989). “A novel genetic system to detect protein-protein interactions”. In: *Nature* 340.6230, pp. 245–246.
- Fischer, Eliane, Nicola Zamboni, and Uwe Sauer (2004). “High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived ^{13}C constraints”. In: *Analytical Biochemistry* 325.2, pp. 308–316.
- Florida Division of Plant Industry (2007). *Tobacco Etch Virus (Potyvirus TEV)*). [Online; accessed July 29, 2016]. URL: <http://www.insectimages.org/browse/detail.cfm?imgnum=5266033>.
- Folch-Fortuny, A. et al. (2015). “MCR-ALS on metabolic networks: Obtaining more meaningful pathways”. In: *Chemometrics and Intelligent Laboratory Systems* 142, pp. 293–303.
- Folch-Fortuny, A. et al. (2016). “Fusion of genomic, proteomic and phenotypic data: the case of potyviruses”. In: *Molecular BioSystems* 12.1, pp. 253–261.

- Fong, Stephen S. et al. (2006). “Latent pathway activation and increased pathway capacity enable *Escherichia coli* adaptation to loss of key metabolic enzymes”. In: *The Journal of Biological Chemistry* 281.12, pp. 8024–8033.
- Forshed, Jenny et al. (2008). “Proteomic Data Analysis Workflow for Discovery of Candidate Biomarker Peaks Predictive of Clinical Outcome for Patients with Acute Myeloid Leukemia”. In: *Journal of Proteome Research* 7.6, pp. 2332–2341.
- Fortunato, Santo and Marc Barthélemy (2007). “Resolution limit in community detection”. In: *Proceedings of the National Academy of Sciences* 104.1, pp. 36–41.
- Fossum, Even et al. (2009). “Evolutionarily conserved herpesviral protein interaction networks”. In: *PLoS pathogens* 5.9, e1000570.
- Fouss, Francois et al. (2012). “An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification”. In: *Neural networks: the official journal of the International Neural Network Society* 31, pp. 53–72.
- Franceschini, Andrea et al. (2013). “STRING v9.1: protein–protein interaction networks, with increased coverage and integration”. In: *Nucleic Acids Research* 41.1, pp. 808–815.
- Fuxman Bass, J. I. et al. (2013). “Using networks to measure similarity between genes: association index selection”. In: *Nature Methods* 10.12, pp. 1169–1176.
- Gibbs, Adrian and Kazusato Ohshima (2010). “Potyviruses and the digital revolution”. In: *Annual review of phytopathology* 48, pp. 205–223.
- Girvan, M. and M. E. J. Newman (2002). “Community structure in social and biological networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12, pp. 7821–7826.
- González-Martínez, J. M. et al. (2014). “Metabolic flux understanding of *Pichia pastoris* grown on heterogenous culture media”. In: *Chemometrics and Intelligent Laboratory Systems* 134, pp. 89–99.
- Gudmundsson, Steinn and Ines Thiele (2010). “Computationally efficient flux variability analysis.” In: *BMC bioinformatics* 11.1, p. 489.

- Guimera, Roger and Luís A. Nunes Amaral (2005). “Functional cartography of complex metabolic networks”. In: *Nature* 433.7028, pp. 895–900.
- Guo, Deyin et al. (2001). “Towards a protein interaction map of potyviruses: protein interaction matrixes of two potyviruses based on the yeast two-hybrid system”. In: *Journal Of General Virology* 82, pp. 935–939.
- Guzmán, G. I. et al. (2015). “Model-driven discovery of underground metabolic functions in *Escherichia coli*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.3, pp. 929–934.
- Hartwell, Leland H. et al. (1999). “From molecular to modular cell biology”. In: *Nature* 402, pp. C47–C52.
- Haverkorn, B.R.B. et al. (2014). “Large-scale ^{13}C -flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*”. In: *Molecular Systems Biology* 7.1, pp. 477–477.
- Hegele, Anna et al. (2012). “Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome”. In: *Molecular Cell* 45.4, pp. 567–580.
- Herrgard, Markus J. et al. (2006). “Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*”. In: *Genome Research* 16.5, pp. 627–635.
- Ho, Yuen et al. (2002). “Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry”. In: *Nature* 6868, pp. 180–183.
- Holme, Petter, Mikael Huss, and Hawoong Jeong (2003). “Subnetwork hierarchies of biochemical pathways”. In: *Bioinformatics (Oxford, England)* 19.4, pp. 532–538.
- Hotelling, H. (1933). “Analysis of a complex of statistical variables into principal components”. In: *Journal of Educational Psychology* 24.6, pp. 417–441.
- Hu, Chang-Deng, Yurii Chinenov, and Tom K Kerppola (2002). “Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation”. In: *Molecular cell* 9.4, pp. 789–798.
- Hucka, M. et al. (2003). “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models”. In: *Bioinformatics* 19.4, pp. 524–531.

- Hwang, D. et al. (2005). “A data integration methodology for systems biology”. In: *Proceedings of the National Academy of Sciences* 102.48, pp. 17296–17301.
- Ideker, Trey et al. (2002). “Discovering regulatory and signalling circuits in molecular interaction networks”. In: *Bioinformatics* 18, pp. 233–240.
- Ingraham, John L., Ole Maaløe, and Frederick Carl Neidhardt (1983). *Growth of the bacterial cell*. Sinauer Associates.
- Ito, Takashi et al. (2001). “A comprehensive two-hybrid analysis to explore the yeast protein interactome”. In: *Proceedings of the National Academy of Sciences* 98.8, pp. 4569–4574.
- Jeong, H et al. (2000). “The large-scale organization of metabolic networks.” In: *Nature* 407.6804, pp. 651–4.
- Junker, B. H. (2008). “Networks in Biology”. In: *Analysis of Biological Networks*. Ed. by B. H. Junker and F. Schreiber. John Wiley & Sons, Inc., pp. 1–14.
- Kanehisa, Minoru and Susumu Goto (2000). “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 28.1, pp. 27–30.
- Kang, SH, WS Lim, and KH Kim (2004). “A protein interaction map of soybean mosaic virus strain G7H based on the yeast two-hybrid system.” In: *Molecules and cells* 18.1, pp. 122–126.
- Karr, J. R. et al. (2012). “A Whole-Cell Computational Model Predicts Phenotype from Genotype”. In: *Cell* 150.2, pp. 389–401.
- Kayser, Anke et al. (2005). “Metabolic flux analysis of *Escherichia coli* in glucose-limited continuous culture. I. Growth-rate-dependent metabolic efficiency at steady state”. In: *Microbiology (Reading, England)* 151.3, pp. 693–706.
- Kerppola, Tom K. (2006). “Design and Implementation of Bimolecular Fluorescence Complementation (BiFC) Assays for the Visualization of Protein Interactions in Living Cells”. In: *Nature protocols* 1.3, pp. 1278–1286.
- Kerppola, Tom K. (2008). “Bimolecular fluorescence complementation (BiFC) analysis as a probe of protein interactions in living cells”. In: *Annual Review of Biophysics* 37, pp. 465–487.
- Kerrien, Samuel et al. (2012). “The IntAct molecular interaction database in 2012”. In: *Nucleic Acids Research* 40.1, pp. 841–846.

- Keseler, Ingrid M. et al. (2013). “EcoCyc: fusing model organism databases with systems biology”. In: *Nucleic Acids Research* 41.D1, pp. D605–D612.
- King, Zachary A. et al. (2015a). “BiGG Models: A platform for integrating, standardizing and sharing genome-scale models”. In: *Nucleic Acids Research*, gkv1049.
- King, Zachary A. et al. (2015b). “Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways”. In: *PLoS Comput Biol* 11.8, e1004321.
- Kitano, H. (2002). “Computational systems biology”. In: *Nature* 420.6912, pp. 206–210.
- Klamt, Steffen, Utz-Uwe Haus, and Fabian Theis (2009). “Hypergraphs and Cellular Networks”. In: *PLoS Comput Biol* 5.5, e1000385.
- Klamt, Steffen, Stefan Schuster, and Ernst Dieter Gilles (2002). “Calculability analysis in underdetermined metabolic networks illustrated by a model of the central metabolism in purple nonsulfur bacteria”. In: *Biotechnology and Bioengineering* 77.7, pp. 734–751.
- Klipp, E. et al. (2005). “Basic Principles”. In: *Systems Biology in Practice*. Wiley-VCH Verlag GmbH & Co. KGaA, pp. 1–17.
- Klipp, E. et al. (2011). *Systems Biology*. John Wiley & Sons.
- Kodama, Yutaka and Chang-Deng Hu (2010). “An improved bimolecular fluorescence complementation assay with a high signal-to-noise ratio”. In: *BioTechniques* 49.5, pp. 793–805.
- Kreimer, Anat et al. (2008). “The evolution of modularity in bacterial metabolic networks.” In: *Proc Natl Acad Sci U S A* 105.19, pp. 6976–6981.
- Lalić, J and S F Elena (2012). “Magnitude and sign epistasis among deleterious mutations in a positive-sense plant RNA virus”. In: *Heredity* 109.2, pp. 71–77.
- Lambiotte, R., J. Delvenne, and M. Barahona (2014). “Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks”. In: *Network Science and Engineering, IEEE Transactions on* 1.2, pp. 76–90.
- Lancichinetti, Andrea, Santo Fortunato, and Janos Kertesz (2009). “Detecting the overlapping and hierarchical community structure of complex networks”. In: *New Journal of Physics* 11.3, p. 033015.

- Lee, Hae Woo et al. (2012). “Data fusion-based assessment of raw materials in mammalian cell culture”. In: *Biotechnology and Bioengineering* 109.11, pp. 2819–2828.
- Lee, Jong Min et al. (2008). “Dynamic Analysis of Integrated Signaling, Metabolic, and Regulatory Networks”. In: *PLOS Comput Biol* 4.5, e1000086.
- Lerman, Joshua A. et al. (2012). “In silico method for modelling metabolism and gene product expression at genome scale”. In: *Nature Communications* 3.3, p. 929.
- Lewis, Nathan E., Harish Nagarajan, and Bernhard O. Palsson (2012). “Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods”. In: *Nature Reviews. Microbiology* 10.4, pp. 291–305.
- Lin, Lin et al. (2009). “Protein-protein interactions in two potyviruses using the yeast two-hybrid system”. In: *Virus Research* 142, pp. 36–40.
- Link, Hannes, Dimitris Christodoulou, and Uwe Sauer (2014). “Advancing metabolic models with kinetic information”. In: *Current Opinion in Biotechnology* 29, pp. 8–14.
- Llaneras, F. and J. Picó (2007a). “An interval approach for dealing with flux distributions and elementary modes activity patterns”. In: *Journal of Theoretical Biology* 246.2, pp. 290–308.
- Llaneras, Francisco and Jesús Picó (2007b). “A procedure for the estimation over time of metabolic fluxes in scenarios where measurements are uncertain and/or insufficient”. In: *BMC Bioinformatics* 8, p. 421.
- Llaneras, Francisco and Jesús Picó (2008). “Stoichiometric modelling of cell metabolism”. In: *Journal of Bioscience and Bioengineering* 105.1, pp. 1–11.
- Llaneras, Francisco and Jesús Picó (2010). “Which Metabolic Pathways Generate and Characterize the Flux Space? A Comparison among Elementary Modes, Extreme Pathways and Minimal Generators”. In: *Journal of Biomedicine and Biotechnology* 2010, pp. 1–14.
- Ma, Hong-Wu and An-Ping Zeng (2003a). “The connectivity structure, giant strong component and centrality of metabolic networks.” In: *Bioinformatics (Oxford, England)* 19.11, pp. 1423–30.

- Ma, Hong-Wu et al. (2004). “Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph.” In: *Bioinformatics (Oxford, England)* 20.12, pp. 1870–6.
- Ma, Hongwu and An-Ping Zeng (2003b). “Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms”. In: *Bioinformatics* 19.2, pp. 270–277.
- Mahadevan, R. and C. H. Schilling (2003). “The effects of alternate optimal solutions in constraint-based genome-scale metabolic models”. In: *Metabolic Engineering* 5.4, pp. 264–276.
- Mahadevan, Radhakrishnan, Jeremy S. Edwards, and Francis J Doyle 3rd (2002). “Dynamic flux balance analysis of diauxic growth in *Escherichia coli*.” In: *Biophys J* 83.3, pp. 1331–1340.
- Mayr, E. (1998). *This is Biology: The Science of the Living World*. Harvard University Press.
- Meila, Marina (2007). “Comparing clusterings: an information based distance”. In: *Journal of Multivariate Analysis* 98.5, pp. 873–895.
- Mering, C. von et al. (2003). “Genome evolution reveals biochemical networks and functional modules”. In: *Proceedings of the National Academy of Sciences* 100.26, pp. 15428–15433.
- Merits, Andres et al. (2002). “Proteolytic processing of potyviral proteins and polyprotein processing intermediates in insect and plant cells”. In: *The Journal of general virology* 83, pp. 1211–1221.
- Mitchell, P. (1961). “Coupling of Phosphorylation to Electron and Hydrogen Transfer by a Chemi–Osmotic type of Mechanism”. In: *Nature* 191.4784, pp. 144–148.
- Mitchell, P. (1966). “Chemiosmotic Coupling in Oxidative and Photosynthetic Phosphorylation”. In: *Biological Reviews* 41.3, pp. 445–501.
- Molenaar, Douwe et al. (2009). “Shifts in growth strategies reflect tradeoffs in cellular economics”. In: *Molecular Systems Biology* 5, p. 323.
- Morales, Y. et al. (2016). “PFA toolbox: a MATLAB tool for Metabolic Flux Analysis”. In: *BMC Systems Biology* 10, p. 46.

- Newman, M. E. J. (2002). “Assortative Mixing in Networks”. In: *Physical Review Letters* 89.20, p. 208701.
- Newman, M. E. J. (2003). “The Structure and Function of Complex Networks”. In: *SIAM Review* 45.2, pp. 167–256.
- Newman, M. E. J. (2006). “Finding community structure in networks using the eigenvectors of matrices”. In: *Physical Review E* 74.3.
- Newman, M. E. J. and M. Girvan (2004). “Finding and evaluating community structure in networks”. In: *Physical Review E* 69.2.
- Newman, Mark (2010). *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc. ISBN: 0199206651, 9780199206650.
- Nie, Yaling and Jingkai Yu (2013). “Mining breast cancer genes with a network based noise-tolerant approach”. In: *BMC systems biology* 7, p. 49.
- Nielsen, Jens and John Villadsen (1992). “Modelling of microbial kinetics”. In: *Chemical Engineering Science* 47.17, pp. 4225–4270.
- Niu, Hongxing et al. (2013). “Dynamic modeling of methylotrophic *Pichia pastoris* culture with exhaust gas analysis: From cellular metabolism to process simulation”. In: *Chemical Engineering Science* 87, pp. 381–392.
- O’Brien, Edward J. et al. (2013). “Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction”. In: *Molecular Systems Biology* 9.1, p. 693.
- Orth, Jeffrey D., R. M. T. Fleming, and Bernhard O Palsson (2010). “Reconstruction and use of microbial metabolic networks: the Core *Escherichia coli* metabolic model as an educational guide”. In: *EcoSal Plus* 4.1.
- Orth, Jeffrey D., Ines Thiele, and Bernhard O Palsson (2010). “What is flux balance analysis?” In: *Nature Biotechnology* 28.3, pp. 245–248.
- Orth, Jeffrey D. et al. (2011). “A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011”. In: *Molecular Systems Biology* 7, p. 535.
- Ouzounis, C. A. and P. Karp (2000). “Global Properties of the Metabolic Map of *Escherichia coli*”. In: *Genome Research* 10.4, pp. 568–576.

- Oyarzún, D. A. (2011). “Optimal control of metabolic networks with saturable enzyme kinetics”. In: *IET systems biology* 5.2, pp. 110–9.
- Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Stanford InfoLab.
- Palsson, B. Ø. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press.
- Palsson, B. Ø. (2015). *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press.
- Papin, Jason A., Jennifer L. Reed, and Bernhard O. Palsson (2004). “Hierarchical thinking in network biology: the unbiased modularization of biochemical networks”. In: *Trends in Biochemical Sciences* 29.12, pp. 641–647.
- Papin, Jason A. et al. (2003). “Metabolic pathways in the post-genome era”. In: *Trends in Biochemical Sciences* 28.5, pp. 250–258.
- Peekhaus, N. and T. Conway (1998). “What’s for Dinner?: Entner-Doudoroff Metabolism in *Escherichia coli*”. In: *Journal of Bacteriology* 180.14, pp. 3495–3502.
- Perrenoud, A. and U. Sauer (2005). “Impact of Global Transcriptional Regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on Glucose Catabolism in *Escherichia coli*”. In: *Journal of Bacteriology* 187.9, pp. 3171–3179.
- Phizicky, E M and S Fields (1995). “Protein-protein interactions: methods for detection and analysis”. In: *Microbiological reviews* 59.1, pp. 94–123.
- Price, Nathan D. et al. (2003). “Genome-scale microbial in silico models: the constraints-based approach”. In: *Trends in Biotechnology* 21.4, pp. 162–169.
- Pržulj, N., D. A. Wigle, and I. Jurisica (2004). “Functional topology in a network of protein interactions”. In: *Bioinformatics* 20.3, pp. 340–348.
- Pržulj, Nataša (2011). “Protein-protein interactions: Making sense of networks via graph-theoretic modeling”. In: *BioEssays* 33.2, pp. 115–123.
- Rabinowitz, Joshua D. and Livia Vastag (2012). “Teaching the design principles of metabolism.” In: *Nat Chem Biol* 8.6, pp. 497–501.

- Rajagopala, Seesandra V. et al. (2014). “The binary protein-protein interaction landscape of *Escherichia coli*”. In: *Nature Biotechnology* 32.3, pp. 285–290.
- Rasmussen, Morten Arendt and Rasmus Bro (2012). “A tutorial on the Lasso approach to sparse modeling”. In: *Chemometrics and Intelligent Laboratory Systems* 119, pp. 21–31.
- Ravasz, E. et al. (2002). “Hierarchical Organization of Modularity in Metabolic Networks”. In: *Science* 297.5586, pp. 1551–1555.
- Reed, Jennifer L. and Bernhard O Palsson (2004). “Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states”. In: *Genome Research* 14.9, pp. 1797–1805.
- Reed, Jennifer L. et al. (2003). “An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)”. In: *Genome Biology* 4.9, R54.
- Reichardt, Joerg and Stefan Bornholdt (2006). “Statistical Mechanics of Community Detection”. In: *Physical Review E* 74.1.
- Resat, H., L. Petzold, and M. F. Pettigrew (2009). “Kinetic Modeling of Biological Systems”. In: *Methods in molecular biology* 541, pp. 311–335.
- Revers, Frédéric et al. (1999). “New advances in understanding the molecular biology of plant/potyvirus interactions”. In: *Molecular Plant-Microbe Interactions* 12.5, pp. 367–376.
- Riechmann, J. L., S Laín, and J. A. García (1992). “Highlights and prospects of potyvirus molecular biology”. In: *The Journal of general virology* 73, pp. 1–16.
- R.J. Reynolds Tobacco Company (1990). *Potato Virus Y (Potyvirus PVY)*. [Online; accessed July 29, 2016]. URL: <http://www.insectimages.org/browse/detail.cfm?imgnum=1440040>.
- Rodrigo, Guillermo et al. (2012). “A Meta-Analysis Reveals the Commonalities and Differences in *Arabidopsis thaliana* Response to Different Viral Pathogens”. In: *PLoS ONE* 7.7, e40526.
- Rojas, J. et al. (2004). “Chemometric analysis of screen-printed biosensor chronoamperometric responses”. In: *Sensors and Actuators B: Chemical* 102.2, pp. 284–290.

-
- Rosvall, Martin and Carl T. Bergstrom (2010). “Mapping change in large networks”. In: *PLoS ONE* 5.1, e8694.
- Rual, Jean-François et al. (2005). “Towards a proteome-scale map of the human protein-protein interaction network”. In: *Nature* 437.7062, pp. 1173–1178.
- Rügen, Marco, Alexander Bockmayr, and Ralf Steuer (2015). “Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional FBA.” In: *Scientific reports* 5, p. 15247.
- Russell, Robert B. and Patrick Aloy (2008). “Targeting and tinkering with interaction networks”. In: *Nature Chemical Biology* 4.11, pp. 666–673.
- Samal, Areejit and Olivier C. Martin (2011). “Randomizing Genome-Scale Metabolic Networks”. In: *PLoS ONE* 6.7, e22295.
- Samal, Areejit et al. (2006). “Low degree metabolites explain essential reactions and enhance modularity in biological networks.” In: *BMC bioinformatics* 7, p. 118.
- Satuluri, Venu, Srinivasan Parthasarathy, and Duygu Ucar (2010). “Markov Clustering of Protein Interaction Networks with Improved Balance and Scalability”. In: *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*. BCB '10. New York, NY, USA: ACM, pp. 247–256.
- Sauer, U. et al. (1999). “Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism”. In: *Journal of Bacteriology* 181.21, pp. 6679–6688.
- Savageau, M. A. (1969). “Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions”. In: *Journal of Theoretical Biology* 25.3, pp. 365–369.
- Schaub, Michael T. et al. (2012). “Markov dynamics as a zooming lens for multi-scale community detection: non clique-like communities and the field-of-view limit”. In: *PLoS ONE* 7.2, e32210.
- Schaub, Michael T. et al. (2014). “Structure of complex networks: Quantifying edge-to-edge relations by failure-induced flow redistribution”. In: *Network Science* 2.01, pp. 66–89.

- Schellenberger, Jan et al. (2011). “Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0”. In: *Nature Protocols* 6.9, pp. 1290–1307.
- Schilling, C. H., D. Letscher, and B. O. Palsson (2000). “Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective.” In: *J Theor Biol* 203.3, pp. 229–248.
- Schuetz, R. et al. (2012). “Multidimensional optimality of microbial metabolism”. In: *Science* 336.6081, pp. 601–604.
- Schuetz, Robert, Lars Kuepfer, and Uwe Sauer (2007). “Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*”. In: *Molecular Systems Biology* 3.1.
- Schuster, S., D. A. Fell, and T. Dandekar (2000). “A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.” In: *Nat Biotechnol* 18.3, pp. 326–332.
- Shannon, Paul et al. (2003). “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: *Genome research* 13.11, pp. 2498–2504.
- Shen, W T et al. (2010). “Protein interaction matrix of Papaya ringspot virus type P based on a yeast two-hybrid system”. In: *Acta virologica* 54.1, pp. 49–54.
- Silva, Marcio Rosa da et al. (2008). “Metabolic Networks”. In: *Analysis of Biological Networks*. Ed. by Bjorn H. Junker and Falk Schreiber. John Wiley & Sons, Inc., pp. 233–253.
- Simeonidis, Evangelos, Sriram Chandrasekaran, and Nathan D Price (2013). “A guide to integrating transcriptional regulatory and metabolic networks using PROM (probabilistic regulation of metabolism)”. In: *Methods in molecular biology (Clifton, N.J.)* 985, pp. 103–112.
- Smart, Ashley G, Luis A N Amaral, and Julio M Ottino (2008). “Cascading failure and robustness in metabolic networks.” In: *Proceedings of the National Academy of Sciences of the United States of America* 105.36, pp. 13223–8.
- Spence, N. J. et al. (2007). “Economic impact of Turnip mosaic virus, Cauliflower mosaic virus and Beet mosaic virus in three Kenyan vegetables”. In: *Plant Pathology* 56.2, pp. 317–323.

- Spirin, Victor and Leonid A. Mirny (2003). “Protein complexes and functional modules in molecular networks”. In: *Proceedings of the National Academy of Sciences* 100.21, pp. 12123–12128.
- Stark, Chris et al. (2006). “BioGRID: a general repository for interaction datasets”. In: *Nucleic Acids Research* 34.1, pp. 535–539.
- Stephanopoulos, George, Aristos A. Aristidou, and Jens Nielsen (1998). *Metabolic Engineering: Principles and Methodologies*. Academic Press.
- Suter, Bernhard, Saranya Kittanakom, and Igor Stagljar (2008). “Two-hybrid technologies in proteomics research”. In: *Current Opinion in Biotechnology* 19.4, pp. 316–323.
- Takemoto, Kazuhiro (2013). “Does habitat variability really promote metabolic network modularity?” In: *PLoS one* 8.4, e61348.
- Thiele, Ines et al. (2009). “Genome-scale reconstruction of Escherichia coli’s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization”. In: *PLoS computational biology* 5.3, e1000312.
- Thomas, A. et al. (2003). “On the structure of protein–protein interaction networks”. In: *Biochemical Society Transactions* 31.6, pp. 1491–1496.
- Uetz, P. et al. (2000). “A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*”. In: *Nature* 403.6770, pp. 623–627.
- Uetz, Peter et al. (2006). “Herpesviral protein networks and their interaction with the human proteome”. In: *Science (New York, N. Y.)* 311.5758, pp. 239–242.
- Urcuqui-Inchima, S, A L Haenni, and F Bernardi (2001). “Potyvirus proteins: a wealth of functions”. In: *Virus research* 74, pp. 157–175.
- Van Mechelen, Iven and Age K. Smilde (2010). “A generic linked–mode decomposition model for data fusion”. In: *Chemometrics and Intelligent Laboratory Systems* 104.1, pp. 83–94.
- Varma, A and B. O. Palsson (1994). “Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110”. In: *Applied and environmental microbiology* 60.10, pp. 3724–3731.

- Varma, Amit and Bernhard O. Palsson (1993a). “Metabolic capabilities of *Escherichia coli*: I. synthesis of biosynthetic precursors and cofactors”. In: *Journal of Theoretical Biology* 165.4, pp. 477–502.
- Varma, Amit and Bernhard O. Palsson (1993b). “Metabolic Capabilities of *Escherichia coli* II. Optimal Growth Patterns”. In: *Journal of Theoretical Biology* 165.4, pp. 503–522.
- Vemuri, G. N. et al. (2007). “Increasing NADH oxidation reduces overflow metabolism in *Saccharomyces cerevisiae*.” In: *Proc Natl Acad Sci U S A* 104.7, pp. 2402–2407.
- Venkatesan, Kavitha et al. (2008). “An empirical framework for binary interactome mapping”. In: *Nature Methods* 6.1, pp. 83–90.
- Vercammen, D., F. Logist, and J. V. Impe (2014). “Dynamic estimation of specific fluxes in metabolic networks using non-linear dynamic optimization”. In: *BMC Systems Biology* 8, p. 132.
- Vitkup, Dennis, Peter Kharchenko, and Andreas Wagner (2006). “Influence of metabolic network structure and function on enzyme evolution.” In: *Genome biology* 7.5, R39.
- Voevodski, Konstantin, Shang-Hua Teng, and Yu Xia (2009). “Finding local communities in protein networks”. In: *BMC Bioinformatics* 10, p. 297.
- Wagner, A. (2001). “The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes”. In: *Molecular Biology and Evolution* 18.7, pp. 1283–1292.
- Wagner, A. and D. A. Fell (2001). “The small world inside large metabolic networks”. In: *Proceedings. Biological Sciences / The Royal Society* 268.1478, pp. 1803–1810.
- Waldherr, Steffen, Diego A. Oyarzún, and Alexander Bockmayr (2015). “Dynamic optimization of metabolic networks coupled with gene expression”. In: *Journal of Theoretical Biology* 365, pp. 469–485.
- Ward, C. W. and D. D. Shukla (1991). “Taxonomy of potyviruses: current problems and some solutions”. In: *Intervirology* 32.5, pp. 269–296.
- Watts, Duncan J. and Steven H. Strogatz (1998). “Collective dynamics of small-world networks”. In: *Nature* 393.6684, pp. 440–442.

- Wei, Taiyun et al. (2010). “Formation of complexes at plasmodesmata for potyvirus intercellular movement is mediated by the viral protein P3N-PIPO”. In: *PLoS pathogens* 6.6, e1000962.
- Westerhoff, H. V. and B. Ø. Palsson (2004). “The evolution of molecular biology into systems biology”. In: *Nature Biotechnology* 22.10, pp. 1249–1252.
- Winterbach, Wynand et al. (2013). “Topology of molecular interaction networks”. In: *BMC Systems Biology* 7.1, p. 90.
- Wold, Svante (1978). “Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models”. In: *Technometrics* 20.4, pp. 397–405.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (2001). “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and Intelligent Laboratory Systems* 58.2, pp. 109–130.
- Wuchty, S. (2001). “Scale-free behavior in protein domain networks”. In: *Molecular Biology and Evolution* 18.9, pp. 1694–1702.
- Xu, Yun, Elon Correa, and Royston Goodacre (2013). “Integrating multiple analytical platforms and chemometrics for comprehensive metabolic profiling: application to meat spoilage detection”. In: *Analytical and Bioanalytical Chemistry* 405.15, pp. 5063–5074.
- Yambao, Ma et al. (2003). “The central and C-terminal domains of VPg of Clover yellow vein virus are important for VPg-HCPro and VPg-VPg interactions”. In: *Journal of General Virology* 84.10, pp. 2861–2869.
- Yook, Soon-Hyung, Zoltán N Oltvai, and Albert-László Barabási (2004). “Functional and topological characterization of protein interaction networks”. In: *Proteomics* 4.4, pp. 928–942.
- Yu, Haiyuan et al. (2008). “High Quality Binary Protein Interaction Map of the Yeast Interactome Network”. In: *Science* 322.5898, pp. 104–110.
- Zheng, Hongying et al. (2011). “Mapping the self-interacting domains of TuMV HC-Pro and the subcellular localization of the protein”. In: *Virus Genes* 42.1, pp. 110–116.

- Zhou, Wanding and Luay Nakhleh (2011). “Properties of metabolic graphs: biological organization or representation artifacts?” In: *BMC Bioinformatics* 12, p. 132.
- Zhou, Wanding and Luay Nakhleh (2012). “Convergent evolution of modularity in metabolic networks through different community structures.” In: *BMC evolutionary biology* 12.1, p. 181.
- Zilian, E. and E. Maiss (2011). “Detection of plum pox potyviral protein-protein interactions in planta using an optimized mRFP-based bimolecular fluorescence complementation system”. In: *Journal of General Virology* 92.12, pp. 2711–2723.
- Zou, Hui and Trevor Hastie (2005). “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.