SISSA/ISAS - International School for Advanced Studies

Systems biology approaches to the dynamics of gene expression and chemical reactions

Thesis submitted for the degree of Doctor Philosophiæ
in Functional and Structural Genomics

Candidate
Nicola Soranzo

Supervisor
Claudio Altafini

9th November 2009

Acknowledgements

It is really a pleasure for me to thank my supervisor, Dr. Claudio Altafini, for the help, advices and support that he has given to me over the past four years and for suggesting the challenging problems presented in this thesis. I have enjoyed very much working with him and the rest of our research group in Systems Biology. In fact, I am also particularly grateful to my colleagues and collaborators Mattia Zampieri, Giovanna De Palo, Giovanni Iacono and Daniele Bianchini.

Moreover, I would like to thank the other researchers with whom I had the opportunity and pleasure to collaborate during my PhD, Prof. Ginestra Bianconi, Prof. Lorenzo Farina, Prof. Alberto Pallavicini, and Fahimeh Ramezani.

Many thanks also to the several people with whom I have shared the student life in SISSA, the lessons, the lunch breaks and the problems: Andrea, Vanessa, Fabrizio, Rolando, my office mates Giulia, Paola, Emiliano, Roberto, Trang, Anetka, Arturo, Kamil and all the other guys of SBP sector.

Finally I wish to heartly thank my parents and my lifelong girlfriend Isabella for all their help, love and support.

Contents

Outline

Systems biology is an emergent interdisciplinary field of study whose main goal is to understand the global properties and functions of a biological system by investigating its structure and dynamics [74]. This high-level knowledge can be reached only with a coordinated approach involving researchers with different backgrounds in molecular biology, the various omics (like genomics, proteomics, metabolomics), computer science and dynamical systems theory.

The history of systems biology as a distinct discipline began in the 1960s, and saw an impressive growth since year 2000, originated by the increased accumulation of biological information, the development of high-throughput experimental techniques, the use of powerful computer systems for calculations and database hosting, and the spread of Internet as the standard medium for information diffusion [77].

In the last few years, our research group tried to tackle a set of systems biology problems which look quite diverse, but share some topics like biological networks and system dynamics, which are of our interest and clearly fundamental for this field.

In fact, the first issue we studied (covered in Part I) was the reverse engineering of large-scale gene regulatory networks. Inferring a gene network is the process of identifying interactions among genes from experimental data (tipically microarray expression profiles) using computational methods [6]. Our aim was to compare some of the most popular association network algorithms (the only ones applicable at a genome-wide level) in different conditions. In particular we verified the predictive power of similarity measures both of direct type (like correlations and mutual information) and of conditional type (partial correlations and conditional mutual information) applied on different kinds of experiments (like data taken at equilibrium or time courses) and on both synthetic and real microarray data (for *E. coli* and *S. cerevisiae*).

In our simulations we saw that all network inference algorithms obtain better performances from data produced with "structural" perturbations (like gene knockouts at steady state) than with just dynamical perturbations (like time course measurements or changes of the initial expression levels). Moreover, our analysis showed differences in the performances of the algorithms: direct methods are more robust in detecting stable relationships (like belonging to the same protein complex), while conditional methods are better at causal interactions (e.g. tran-

scription factor–binding site interactions), especially in presence of combinatorial transcriptional regulation.

Even if time course microarray experiments are not particularly useful for inferring gene networks, they can instead give a great amount of information about the dynamical evolution of a biological process, provided that the measurements have a good time resolution. Recently, such a dataset has been published [119] for the yeast metabolic cycle, a well-known process where yeast cells synchronize with respect to oxidative and reductive functions. In that paper, the long-period respiratory oscillations were shown to be reflected in genome-wide periodic patterns in gene expression.

As explained in Part II, we analyzed these time series in order to elucidate the dynamical role of post-transcriptional regulation (in particular mRNA stability) in the coordination of the cycle. We found that for periodic genes, arranged in classes according either to expression profile or to function, the pulses of mRNA abundance have phase and width which are directly proportional to the corresponding turnover rates. Moreover, the cascade of events which occurs during the yeast metabolic cycle (and their correlation with mRNA turnover) reflects to a large extent the gene expression program observable in other dynamical contexts such as the response to stresses or stimuli.

The concepts of network and of systems dynamics return also as major arguments of Part III. In fact, there we present a study of some dynamical properties of the so-called chemical reaction networks, which are sets of chemical species among which a certain number of reactions can occur. These networks can be modeled as systems of ordinary differential equations for the species concentrations, and the dynamical evolution of these systems has been theoretically studied since the 1970s [47, 65]. Over time, several independent conditions have been proved concerning the capacity of a reaction network, regardless of the (often poorly known) reaction parameters, to exhibit multiple equilibria. This is a particularly interesting characteristic for biological systems, since it is required for the switch-like behavior observed during processes like intracellular signaling and cell differentiation.

Inspired by those works, we developed a new open source software package for MATLAB, called ERNEST, which, by checking these various criteria on the structure of a chemical reaction network, can exclude the multistationarity of the corresponding reaction system. The results of this analysis can be used, for example, for model discrimination: if for a multistable biological process there are multiple candidate reaction models, it is possible to eliminate some of them by proving that they are always monostationary.

Finally, we considered the related property of monotonicity for a reaction network. Monotone dynamical systems have the tendency to converge to an equilibrium and do not present chaotic behaviors. Most biological systems have the same features, and are therefore considered to be monotone or near-monotone [85, 116]. Using the notion of fundamental cycles from graph theory, we proved some theoretical results in order to determine how distant is a given biological network from

being monotone. In particular, we showed that the distance to monotonicity of a network is equal to the minimal number of negative fundamental cycles of the corresponding J-graph, a signed multigraph which can be univocally associated to a dynamical system.

For a more thorough introduction to the different topics briefly presented here, we refer the reader to the initial chapter of each Part.

The material of this thesis has been the object of the following publications:

1. N. Soranzo, G. Bianconi and C. Altafini
   "Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data"
   Bioinformatics 23(13), pp. 1640-1647, 2007

2. M. Zampieri, N. Soranzo and C. Altafini
   "Discerning static and causal interactions in genome-wide reverse engineering problems"
   Bioinformatics 24(13), pp. 1510-1515, 2008

3. N. Soranzo, M. Zampieri, L. Farina and C. Altafini
   "mRNA stability and the unfolding of gene expression in the long-period yeast metabolic cycle"
   BMC Syst. Biol. 3:18, 2009

4. N. Soranzo and C. Altafini
   "ERNEST: a toolbox for chemical reaction network theory"
   Bioinformatics 25(21), pp. 2853-2854, 2009

5. G. Iacono, F. Ramezani, N. Soranzo and C. Altafini
   "Determining the distance to monotonicity of a biological network: a graph-theoretical approach"
   IET Syst. Biol. 4(3), pp. 223-235, 2010

Part I

Reverse engineering of gene networks

Chapter 1

Introduction

In the field of Systems Biology, the possibility of using the information provided by high throughput measurements in order to infer interactions between genes represents a first step towards a comprehensive understanding of a biological system in terms of gene functions, "partner genes", conditions for activation and dynamical behavior. The reconstrucion of gene regulatory networks [6, 54] from microarray expression profiles is certainly one of the most challenging problem for a number of reasons. First, the number of variables that come into play is very high, of the order of the thousands or tens of thousands at least, and there is normally no sufficient biological knowledge to restrict the analysis to a subset of core variables for a given biological process. Second, the number of gene expression profiles available is typically much less that the number of variables, thus making the problem underdetermined. Third, there is no standard model of the regulatory mechanisms for the genes, except for a generic cause–effect relationship between transcription factors (TFs) and corresponding binding sites (BSs). Fourth, little is known (and no high throughput measure is available) about the post-transcriptional regulation and on how it influences the regulatory pattern we see on the microarray experiments. In spite of all these difficulties, the topic of reverse engineering of gene regulatory networks is worth pursuing, as it provides the biologist with phenomenologically predicted gene–gene interactions.

Many methods have been proposed for this scope in the last few years, like Bayesian networks [50, 66], linear ordinary differential equations (ODEs) models [130], relevance networks [15, 33] and graphical models [30, 73, 87, 107].

The aim of this work is to compare a few of these methods, focusing in particular on the last two classes of algorithms, that reconstruct weighted graphs of gene–gene interactions. Relevance networks look for pairs of genes that have similar expression profiles throughout a set of different conditions, and associate them through edges in a graph. The reconstruction changes with the "similarity measure" adopted: popular choices for gene networks are covariance-based measures like the Pearson correlation [15, 33], or entropy-based like the mutual information (MI) [16, 33]. While correlation is a linear measure, MI is nonlinear. These simple

pairwise similarity methods are computationally tractable, but fail to take into account the typical patterns of interaction of multivariate datasets. The consequence is that they suffer from a high false discovery rate, i.e. genes are erroneously associated while in truth they only indirectly interact through one or more other genes.

In order to prune the reconstructed network of such false positives, one can use the notion of conditional independence from the theory of graphical modeling [36], i.e. look for residual correlation or MI after conditioning over one or more genes. These concepts are denoted as partial Pearson correlation (PPC) and conditional mutual information (CMI). First and second order PPC were used for this purposes in [30]. If n is the number of genes, the exhaustive conditioning over $n-2$ genes is instead used in [107] under the name of graphical Gaussian models (GGM). As for MI, conceptually the CMI plays the same role of the first order PPC. In our knowledge, CMI has never been used before for gene network inference, although an alternative method for pruning the MI graph proposed in [88], based on the so-called Data Processing Inequality (DPI), relies on the same idea of conditioning, namely on searching for triplets of genes forming a Markov chain.

Relevance networks and graphical models have been extensively used in recent years [84] and their results have been validated experimentally, for example in [8] where the analysis is based on a similarity index related to CMI [88], or in [40] where co-expression is used to investigate combinatorial regulation.

Since we miss a realistic large scale model of a gene regulatory network, it is not even clear how to fairly evaluate and compare these different methods for reverse engineering. A few biologically inspired (small-size) benchmark problems have been proposed, like the songbird brain model [113] or the Raf pathway [124], or completely artificial networks, typically modeled as systems of nonlinear differential equations [92, 133]. Since we are interested in large scale gene networks, we shall focus on the artificial network of [92], in which the genes represent the state variables and the mechanisms of gene–gene inhibition and activation are modeled using sigmoidal-like functions as in the reaction kinetics formalism. This network has several features that are useful for our purposes: (i) its size can be chosen arbitrarily; (ii) realistic (nonlinear) effects like state saturation or joint regulatory action of several genes are encoded in the model; (iii) perturbation experiments like gene knockout, or different initial conditions, or measurement noise are easily included.

Similar comparative studies have appeared recently in the literature [88, 124]. However, [124] evaluates Bayesian networks, GGM and correlation relevance networks on one specific, very small (11 genes) network. [88] instead compares Bayesian networks, MI relevance networks and DPI using a number of expression profiles m much larger than the number of genes n, while we are also interested in more realistic scenarios. Our investigation aims at:

- comparing conditional similarity measures (like PPCs, GGM and CMI) with "direct" measures (like correlation and MI);
- comparing linear measures (correlation and PPCs) with nonlinear ones (MI, CMI, DPI).

In particular, for the different reconstruction algorithms we are interested in the following questions:

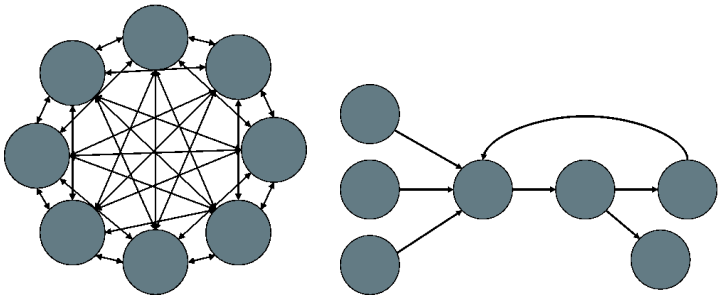
- what is the predictive power for a number of measurements $m \ll n$? How does it grow with m?
- do the algorithms scale with size?
- what is the most useful type of experiment for the purposes of network inference?

After examining these questions, inspired by several studies suggesting that co-expression is mostly related to "static" stable binding relationships, like belonging to the same protein complex (PC), rather than other types of interactions more of a "causal" and transient nature (e.g. TF–BS interactions), we tried to verify if direct or conditional network inference algorithms are indeed useful in discerning static from causal dependencies in artificial and real gene networks. Based on current literature [3, 118, 132], the interaction networks representing PCs and TF–BS are roughly characterizable by means of different recurrent regulatory motifs, that for simplicity we denote "dense modules" and "causal modules" (see Fig. 1.1). The dense modules for PC represent undirected subgraphs in which all nodes are mutually connected. The modules for the TF–BS, instead, are directed subgraphs constructed with a scale-free-like connectivity, but overall sparse graphs. In addition, in order to represent the combinatorial effect of multiple TFs on a target gene the input degrees are normally higher than the output degrees.

It is worth noticing that these two types of regulatory motifs can characterize the complexity of an organism. Going from unicellular prokaryote (*E. coli*) and eukaryote (*S. cerevisiae*) to mammals (human, rat and mouse), the distribution of annotated protein complexes shows an heavier tail towards bigger complexes (see Fig. 1.2(a)). The same happens looking at the combinatorial effect of multiple transcription factors (see Fig. 1.2(b)) [80, 81].

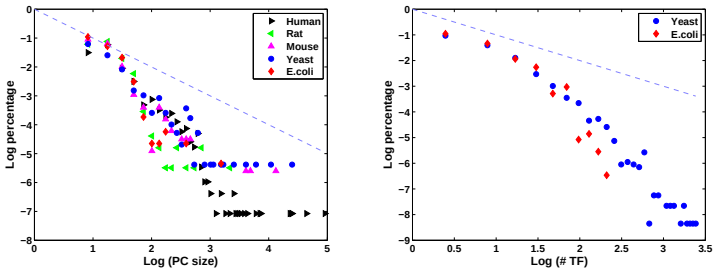
Again a comparison between the two classes of similarity metrics cited above (direct and conditional) is performed, but with the aim of analyzing their ability to infer regulatory networks characterized by the above mentioned topological structures, in a completely unsupervised manner. The five different similarity metrics are tested on an artificial and two real networks. The artificial network is meant to enable the evaluation under controlled conditions, while in the two cases of biological data the identification of true positive (TP) edges relies on real physical networks of PC and TF–BS relationships collected from the literature. We choose two simple organisms, a prokaryote (*E. coli*) and an eukaryote (*S. cerevisiae*), in order to test the consistency of the two regulatory structures for the different algorithms. For these two organisms sufficiently many PC and TF–BS have been annotated and large collections of gene expression profiles can be gathered from online repositories.

The comparison of the inference power of the 5 algorithms shows that the gene interactions associated to co-participation in the same protein complex are better



(a) (b)

Figure 1.1: Schemes of the two regulatory motifs: in (a) a dense module, where all nodes are mutually connected. In (b) a causal module, i.e. directed graph accounting for only a few feedback loops and multi-regulated genes. The former is representing a PC, the latter (multiple) TFs acting on their BSs. PCs can be characterized as sets of proteins that interact closely with each other. As a matter of fact, searching for highly connected subgraphs is a common predictor for PCs [118, 132]. Hence in our artificial network a dense subgraph represents a PC. On the other hand, a statistical description showing a nonuniform connectivity degree on an oriented and globally sparse graph emerges from the analysis of the *E. coli* and *S. cerevisiae* known TF–BS interactions, see Fig. 1.2. It is taken here as a paradigm for the TF–BS modules in our artificial network.



(a) (b)

Figure 1.2: Log scale distribution of PC size (a), and number of TFs per gene (b) for different organisms. In yeast for example the largest complex is the cytoplasmic ribosome accounting for 81 genes, while in *E. coli* it is the flagellum complex composed of 24 genes. Both distributions hint at an increase in the size of the regulatory motifs as the complexity of the organisms increase.

detected by the direct methods, while those associated to the combinatorial effect of multiple TFs are better retrieved by the conditional metrics (in particular by the graphical Gaussian model). Apart from comparing the performances of the algorithms on the different topologies, we also aim at evaluating them on modules of different sizes (i.e. for larger PCs or with increasing numbers of TFs acting on the same BS). For this purpose it is convenient to rank the weights of each similarity matrix and look at the percentage of TPs (with respect to the total number of true edges) in the highest 1% of weights. This procedure allows us to make an unbiased and unsupervised comparison between different metrics. For the dense modules case, it is also possible to specify how well the reconstructed dense modules truly correspond to known PCs. If we do so by means of a clustering algorithm on the inferred graphs, we see that indeed the direct metrics are those allowing the most faithful PC reconstruction.

Chapter 2

Methods

2.1 Artificial datasets

The influence on the transcription of each gene due to the other genes is described by a (sparse) matrix of adjacencies $A = (a_{i,j})$. As for the topology of A, in section 3.1 we considered two classes of directed networks widely used in literature as models for regulatory networks: scale-free [7] and random [39].

Instead, for the analysis of section 3.2, A was constructed as a scale-free matrix, representing the causal module, superimposed to a matrix of densely connected subsets of nodes representing the stable modules. The procedure used to construct this artificial network is such that dense regulatory modules (of different sizes) are numerous enough to compare the inference power among the different algorithms in a statistically relevant manner.

The model we used to generate artificial gene expression datasets is the reaction kinetics-based system of coupled nonlinear continuous time ODEs introduced in [92]. The expression levels of the gene mRNAs are taken as state variables, call them x_i, $i = 1, \ldots, n$. The rate law for the mRNA synthesis of a gene is obtained by multiplying together the sigmoidal-like contributions of the genes identified as its inhibitors and activators. Consider the i-th row of A, $i = 1, \ldots, n$, and choose randomly a sign to its nonzero indexes. Denote by $j_1, \ldots, j_a$ the indexes with assigned positive values (activators of the gene x_i) and with $k_1, \ldots, k_b$ the negative ones (inhibitors of x_i). The ODE for x_i is then

$$\frac{dx_i}{dt} = V_i \prod_{j \in \{j_1,\ldots,j_a\}} \left(1 + \frac{x_j^{\nu_{i,j}}}{x_j^{\nu_{i,j}} + \theta_{i,j}^{\nu_{i,j}}}\right) \prod_{k \in \{k_1,\ldots,k_b\}} \frac{\theta_{i,k}^{\nu_{i,k}}}{x_k^{\nu_{i,k}} + \theta_{i,k}^{\nu_{i,k}}} - \lambda_i x_i, \qquad (2.1)$$

where V_i represents the basal rate of transcription, $\theta_{i,j}$ (respectively $\theta_{i,k}$) the activation (resp. inhibition) half-life, $\nu_{i,j}$ (resp. $\nu_{i,k}$) the activation (resp. inhibition) Hill coefficient (in our simulations: $\nu_{i,j}, \nu_{i,k} \in \{1, 2, 3, 4\}$), and λ_i the degradation rate constant. The ODE (2.1) always tends to a steady state, which could be 0 or a (positive) saturation value. When $x_i(0) \geq 0$, the abundance $x_i(t)$ remains positive during the entire time course, hence the solution is biologically consistent. Thus, a gene

expression profile experiment at time t corresponds to a state vector $[x_1(t) \dots x_n(t)]$ obtained by numerically integrating (2.1). For the purpose of reconstructing the network of gene–gene interactions from expression profiles, one needs to carry out multiple experiments, in different conditions, typically performed perturbing the system in many different ways. We shall consider the following cases of perturbations:

1. randomly chosen initial conditions in the integration of (2.1), plus gene knockout obtained setting to 0 the parameter V_i of the respective differential equation, as in [92];

2. only randomly chosen initial conditions in the integration of (2.1);

and the following types of measurements:

1. steady state measurements;

2. time-course experiments, in which the solution of the ODE is supposed to be measured at a certain (low) sampling rate.

The numerical integration of (2.1) is carried out in MATLAB. In all cases, a Gaussian measurement noise is added to corrupt the output.

2.2 Collected data

We downloaded the *E. coli* gene expression database M^{3D} "Many Microbe Microarrays Database" (build E_coli_v3_Build_1 from http://m3d.bu.edu, T. Gardner Lab, Boston University). This dataset consists of 445 arrays from 13 different collections corresponding to various conditions, like different media, environmental stresses (e.g. DNA damaging drugs, pH changes), genetic perturbations (upregulations and knockouts), and growth phases. The experiments were all carried out on Affymetrix GeneChip *E. coli* Antisense Genome arrays, containing 4345 gene probes. For *S. cerevisiae* we compiled a collection of microarrays containing experiments performed with cDNA chips (958 experiments for 6203 ORFs). On both datasets a global RMA normalization was performed prior to network inference.

PC network for yeast was downloaded from the MPACT subsection of the CYGD database at MIPS [58]. Only the complexes annotated from the literature and not those obtained from high throughput experiments (according to the MIPS classification scheme these last are labeled "550") were considered to limit the high rate of false positives. PC sizes for human, rat and mouse were downloaded from CORUM database [105], while for *E. coli* from the EcoCyc website [68]. We obtained TF–BS networks from the *RegulonDB* database, version 5.6, for *E. coli* [106], and from a recent collection [3] for *S. cerevisiae*. The number of edges in PC and TF–BS networks are summarized in Table 2.1.

Interaction network	N. edges	Edge type
causal modules	11716	directed
dense modules	55610	undirected

(a) Artificial network (2154 genes)

Interaction network	N. edges	Edge type
TF–BS	3071	directed
PC	2228	undirected

(b) *E. coli* (4345 genes)

Interaction network	N. edges	Edge type
TF–BS	12376	directed
PC, annotated	21616	undirected

(c) *S. cerevisiae* (6203 genes)

Table 2.1: Number of edges in the PC and TF–BS networks for: (a) the artificial network with dense modules, (b) *E. coli* and (c) *S. cerevisiae*.

2.3 Similarity measures

2.3.1 Pearson correlation (direct)

Relevance networks based on correlation were proposed already in [33]. If to each gene i we associate a random variable X_i, whose measured values we denote as $x_i(\ell)$ for $\ell = 1, \ldots, m$, the sample correlation between the random variables X_i and X_j is

$$R(X_i, X_j) = \frac{\sum_{\ell=1}^{m}(x_i(\ell) - \bar{x_i})(x_j(\ell) - \bar{x_j})}{(m-1)\sqrt{v_i v_j}},$$

where $\bar{x_i}$, v_i and $\bar{x_j}$, v_j are sample means and variances of $x_i(\ell)$ and $x_j(\ell)$ over the m measurements. When used as weight for the inferred matrix, we'll take the absolute value of R.

2.3.2 Partial Pearson correlation (conditional)

Since correlation alone is a weak concept and cannot distinguish between direct and indirect interactions (e.g. mediated by a common regulator gene), an algorithm for network inference can be improved by the use of partial correlations [30]. The minimum first order partial correlation between X_i and X_j is obtained by exhaustively conditioning the pair (X_i, X_j) over all X_k. If exists $k \neq i, j$ which explains all of the correlation between X_i and X_j, then the partial correlation between X_i and X_j becomes 0 and the pair (X_i, X_j) is conditionally independent given X_k. When this happens, following [36] we say that the triple X_i, X_j, X_k has a Markov property: on an undirected graph genes i and j are not adjacent but separated by k. This is

denoted in [36] as $X_i \perp\!\!\!\perp X_j \mid X_k$. In formulas, the minimum first order PPC is

$$R_{C_1}(X_i, X_j) = \min_{k \neq i,j} \left| R(X_i, X_j \mid X_k) \right|,$$

where

$$R(X_i, X_j \mid X_k) = \frac{R(X_i, X_j) - R(X_i, X_k)R(X_j, X_k)}{\sqrt{(1 - R^2(X_i, X_k))(1 - R^2(X_j, X_k))}}.$$

If $R_{C_1}(X_i, X_j) \simeq 0$ then there exists k such that $X_i \perp\!\!\!\perp X_j \mid X_k$. Sometimes conditioning over a single variable may not be enough, and one would like to explore higher order PPCs. The minimum second order PPC for example is given by

$$R_{C_2}(X_i, X_j) = \min_{k,\ell \neq i,j} \left| R(X_i, X_j \mid X_k, X_\ell) \right|,$$

with

$$R(X_i, X_j \mid X_k, X_\ell) = \frac{R(X_i, X_j \mid X_k) - R(X_i, X_\ell \mid X_k)R(X_j, X_\ell \mid X_k)}{\sqrt{(1 - R^2(X_i, X_\ell \mid X_k))(1 - R^2(X_j, X_\ell \mid X_k))}}$$

and so on for higher order PPCs. Since the computation is exhaustive over all n genes, the computational cost of the algorithm for the k-th order minimum PPC is of the order of $O(n^k)$, and it becomes quickly prohibitive for $k \geq 2$, if n is of the order of the thousands.

The weight matrix R can be used to rank the $(n^2 - n)/2$ possible (undirected) edges of the graph. The use of PPC allows to prune the graph of many false positives computed by correlation alone. However, the information provided by correlation and PPC is one of *independence* or *conditional independence*, i.e. a low value of correlation and PPC for a pair (X_i, X_j) guarantees that an edge between the two nodes is missing. A high value of the quantities $R(X_i, X_j)$ and $R_{C_1}(X_i, X_j)$ does not guarantee that i and j are truly connected by an edge, as $R_{C_2}(X_i, X_j)$ may be small or vanish.

In [30] it is shown how to choose a cutoff threshold for the weight matrices and how to combine together the effect of R, R_{C_1} and R_{C_2}.

2.3.3 Graphical Gaussian models (conditional)

When the $n \times n$ matrix R of elements $R(X_i, X_j)$ is invertible, and we can assume that the data are drawn from a multivariate normal distribution, then the exhaustive conditioning over $n - 2$ genes can be expressed explicitly. Denote $\Omega = R^{-1}$ the concentration matrix of elements $\Omega = (\omega_{i,j})$. Then the partial correlation between X_i and X_j is

$$R_{C_{all}}(X_i, X_j) = -\frac{\omega_{i,j}}{\sqrt{\omega_{i,i}\omega_{j,j}}}.$$

When R is not full rank, then the small-sample stable estimation procedure of [107] can be used. To compute $R_{C_{all}}$, we used the R package GeneNet version 1.0.1, available from CRAN (http://cran.r-project.org).

2.3.4 Mutual information (direct)

In a relevance network, alternatively to correlation, one can use the information-theoretic concept of MI [16, 54, 88]. Given a discrete random variable X_i, taking values in the set $\mathcal{H}_i$, its entropy [110] is defined as

$$H(X_i) = -\sum_{\phi_i \in \mathcal{H}_i} p(\phi_i) \log p(\phi_i),$$

where $p(\phi_i)$ is the probability mass function $p(\phi_i) = Pr(X_i = \phi_i)$, for $\phi_i \in \mathcal{H}_i$. The joint entropy of a pair of variables (X_i, X_j), taking values in the sets $\mathcal{H}_i$, $\mathcal{H}_j$ respectively, is

$$H(X_i, X_j) = -\sum_{\phi_i \in \mathcal{H}_i,\, \phi_j \in \mathcal{H}_j} p(\phi_i, \phi_j) \log p(\phi_i, \phi_j),$$

while the conditional entropy of X_j given X_i is defined as $H(X_j \mid X_i) = H(X_i, X_j) - H(X_i)$. The MI of (X_i, X_j) is defined as $I(X_i; X_j) = H(X_i) - H(X_i \mid X_j)$ and can be explicitly expressed as

$$I(X_i; X_j) = \sum_{\phi_i \in \mathcal{H}_i,\, \phi_j \in \mathcal{H}_j} p(\phi_i, \phi_j) \log \frac{p(\phi_i, \phi_j)}{p(\phi_i) p(\phi_j)} \geq 0.$$

When the joint probability distribution factorizes, the MI vanishes:

$$p(\phi_i, \phi_j) = p(\phi_i) p(\phi_j) \text{ for all } \phi_i \in \mathcal{H}_i, \phi_j \in \mathcal{H}_j \Longrightarrow I(X_i; X_j) = 0. \tag{2.2}$$

2.3.5 Conditional mutual information and DPI (conditional)

Similarly to PPC, also the MI can be conditioned with respect to a third variable X_k. The formula is:

$$I(X_i; X_j \mid X_k) = H(X_i \mid X_k) - H(X_i \mid X_j, X_k)$$

or, equivalently,

$$I(X_i; X_j \mid X_k) = H(X_i, X_k) + H(X_j, X_k) - H(X_k) - H(X_i, X_j, X_k).$$

All pairs of nodes can be conditioned exhaustively on each of the remaining $n - 2$ nodes and the minimum of such CMIs

$$I_C(X_i; X_j) = \min_{k \neq i,j} I(X_i; X_j \mid X_k)$$

can be taken as a measure of conditional independence. When there exists a X_k that explains the whole MI between X_i and X_j, then the triplet has the Markov property

$$I(X_i; X_j \mid X_k) = 0 \Longleftrightarrow X_i \perp\!\!\!\perp X_j \mid X_k, \tag{2.3}$$

implying $I_C(X_i; X_j) = 0$, otherwise $I_C(X_i; X_j) > 0$.

Just like for the correlation and PPC case, the two conditions (2.2) and (2.3) can be used to construct the graph of the gene network. I and I_C can also be combined together, and possibly with a cutoff threshold (computed e.g. through a bootstrapping method).

An alternative algorithm to implement the Markov property $X_i \perp\!\!\!\perp X_j \mid X_k$ is proposed in [88]. It is based on the so-called DPI and consists in dropping the edge corresponding to the minimum of the triplet $I(X_i, X_j)$, $I(X_j, X_k)$ and $I(X_i, X_k)$ for all possible triplets $i \neq j \neq k$. This method is shown in [88] to prune the graph of many false positives. Denote I_{DPI} the matrix obtained by applying the DPI. Although I_{DPI} and I_C derive from the same notion, the information they provide is not completely redundant.

In the computation of I and I_C we used the B-spline algorithm of [29]. The matrix I obtained in this way is quite similar to the MI one gets from the Gaussian Kernel method used in [88], which is known to be computationally more intense than binning into an histogram or the B-spline approach [29]. In order to evaluate how much the choice of the algorithm can influence the reconstruction, we compared two MI matrices computed using a Gaussian Kernel estimator (with the routines provided in [89]) and the B-spline approach. A typical result is shown in Fig. 2.1 for a rather conservative choice of number of bins ($q = 4$) and spline order 2. It can be seen that the two ordering of edges weights always differ for less than 10%.

While the definition of CMI can be extended to higher number of conditioning variables, from a computational point of view this becomes unfeasible for n of the order of thousands: the time complexity of our algorithm for complete data matrices is $O(n^3(mp^3 + q^3))$, where p is the spline order and q is the number of bins used.

2.4 Criteria for algorithm comparison

In order to evaluate the performances of the algorithms, we compare each (symmetric) weight matrix with the corresponding adjacency matrix A and calculate the (standard) quantities listed in Table 2.2.

The receiver operating characteristic (ROC) and the precision versus recall (PvsR) curves measure the quality of the reconstruction. To give a compact description for varying m, the area under the curve (AUC) of both quantities will be used. The ROC curve describes the trade-off between sensitivity and the false positive rate. An AUC(ROC) close to 0.5 corresponds to a random forecast, AUC(ROC) $<$ 0.7 is considered poor, $0.7 \leq$ AUC(ROC) < 0.8 fair and AUC(ROC) ≥ 0.8 good. For gene networks, as A is generally sparse, the ROC curve suffers from the high number of false positives. The PvsR curve instead is based only on comparing true edges and inferred edges, and therefore highlights the precision of the reconstruction [88]. All the quantities we consider as well as the ROC and the PvsR curves

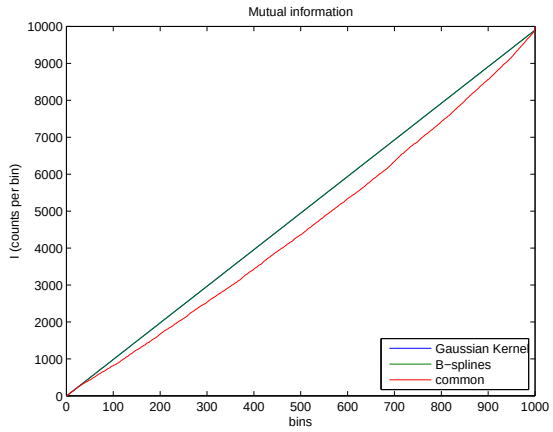


Figure 2.1: Comparison of I computed via Gaussian Kernel method from [88] and B-spline method used in this work with 4 bins and spline order 2 for a network of 100 genes and 200 experiments. The elements of the two matrices are sorted and the sorted values divided in 1000 bins. The figure shows the cumulative counts of the values of the sorted elements (y-axis) up to the i-th bin (x-axis). The counts for the two algorithms overlap (by construction), while the number of edges in common differs for less than 10% of the total.

True positives (TP)	=	correctly identified true edges
False positives (FP)	=	spurious edges
True negatives (TN)	=	correctly identified zero edges
False negatives (FN)	=	not recognized true edges
Recall (or sensitivity or TP rate)	=	$\frac{TP}{TP+FN}$
False positive rate α	=	$\frac{FP}{TN+FP}$
Precision	=	$\frac{TP}{TP+FP}$

Table 2.2: Quantities of interest in the evaluation of the algorithms

are based on sorting the edge weights (in absolute values for correlation, PPCs and GGM) and on growing the network starting from the highest weight down to the lowest one. Fixing a cutoff threshold only means altering the tail of the curves, thus we shall not make any such choice, but explore the entire range of values for the edge weights.

For the networks with dense modules of section 3.2, we used also a second criterion to test the ability of the different algorithms in retrieving the two types of regulatory modules (causal and dense) as a function of their dimension. For this task it is useful to look only at the first percentile of edge weights (i.e. top 1% of edges sorted by their weights). As inference is performed on the entire genome, this first percentile corresponds to 23188, 94373 and 192355 edges in the artificial, *E. coli* and *S. cerevisiae* networks reconstructions respectively (see Table 2.1 for the corresponding numbers of true edges). For the two types of regulatory structures (causal and dense) the true edges are binned according to the size of the module they belong to. The recall (i.e. the percentages of TP over the total number of true edges) of the reconstructed network for each bin (of module size) is used to evaluate how the reconstructions vary with size (shown in Fig. 3.4 and 3.6). More standard curves such as PvsR (Fig. 3.7) or ROC curves (Fig. 3.8) cannot show the dependence on the module size of the algorithms.

2.4.1 Clustering

In section 3.2, after selecting the most significant 1% of edges, the resulting graph is decomposed using a simple hierarchical clustering algorithm, with weighted average linkage as cost of merging, and taking a fixed number of clusters (300 in Fig. 3.5). This procedure allows to identify the most connected components, which are then matched with the dense modules/PCs. This matching is fairly robust with respect to the choice of the number of clusters (data not shown).

Chapter 3

Results

3.1 Comparison on scale-free and random artificial networks

In Fig. 3.1, the results for reconstructions of random and scale-free networks of 100 genes with the different similarity measures (R, R_{C_1}, R_{C_2}, $R_{C_{\text{all}}}$, I, I_C and I_{DPI}) are shown for different numbers m of measurements. AUC(ROC), AUC(PvsR) and the number of TP for a fixed value of acceptable FP (here 20) are displayed in the three columns. For both AUC(ROC) and AUC(PvsR), standard deviations (not shown) are around one order of magnitude smaller than the mean values, thus indicating that the repetitions are substantially faithful.

By comparing the first two rows of Fig. 3.1 it is possible to examine the influence of the network topology on the reconstruction. Under equal conditions (type and amount of experiments), all the algorithms performed better for random networks, confirming that they are easier to infer than scale-free ones [30]. Also another network parameter, the average degree, is influencing the performance of the algorithms: the predictive power is higher for sparser networks than for less sparse ones. For example, in Fig. 3.3 compare the graphs in the first row (average node degree equal to 1.5) with the ones in the second row (degree equal to 3) for artificial scale-free networks of 1000 genes.

If we now focus the attention on the scale-free topology (the most similar to known regulatory networks), it can be seen from the graphs that the performances of the reconstructions are much higher with knockout perturbations (rows 2–3) than for data produced without knockouts (row 4). This suggests that knockouts (i.e. node suppression on (2.1)) help in exploring the network structure, while perturbing only the initial conditions contributes very little predictive information.

Moreover, when perturbing the system with knockouts, steady state measurements (row 2) are able to generate good reconstructions with much less samples than time-course experiments (row 3), in agreement with the results of [6]. For steady states, the performances of the algorithms improve increasing m up to n, then stabilize (for some, like GGM, even decrease). For time-course data, instead,

the graphs tend to level off only when each gene has been knocked out once, regardless of the number of samples taken during the time series. This can be seen on the third row of Fig. 3.1, where the AUCs keep growing until 1000 samples (corresponding to 100 time series each contributing 10 samples) and only then tend to stabilize (data beyond 1000 samples are not shown in Fig. 3.1). The same trend can be observed increasing the number of samples per series (data not shown). Learning a network by means of time series alone (without any knockout) is very difficult as can be deduced from the low values of AUCs achieved in the forth row of Fig. 3.1. Notice, however, that these values get much worse (essentially random) if we consider no-knockout and steady state samples.

As for the different algorithms, the PPCs perform well in all conditions, and are significantly improving performances with respect to correlation for both AUC(PvsR) and TP for fixed FP. On the contrary, applying the DPI to MI (with a tolerance of 0.1, see [88]) only slightly improves the precision of the MI. Since the DPI simply puts to zero the weights of the edges it considers false positives, one should not forget that DPI is penalized with respect to the other measures when computing AUC(ROC). Like PPCs, GGM gives good average results, but looks promising especially for time-course experiments, where also CMI is far superior than MI and DPI.

For the random and scale-free networks reconstructed in Fig. 3.1, Fig. 3.2 reports the average runtimes (over the 10 repetitions) of the various algorithms: R_{C_2} is clearly one order of magnitude slower than the other methods. It must be remarked that for R, R_{C_1}, R_{C_2} we used MATLAB code, while for I, I_C, I_{DPI} C++ code was created (so faster than MATLAB) and $R_{C_{\text{all}}}$ was computed under R environment. Notice that I_C grows faster than the other methods with respect to the number of experiments.

Finally, it is important to remark that the results we obtained for a network of 100 genes are qualitatively and quantitatively similar to those for larger gene networks. As an example, in Fig. 3.3 (first row) a scale-free network of 1000 genes is reconstructed from knockout experiments with steady state measurements. It can be seen that all three parameters shown AUC(ROC), AUC(PvsR) and TP for fixed FP are comparable to those shown in Fig. 3.1 (second row) for an equal ratio m/n.

3.2 Discerning static and causal interactions

For the following comparisons, we used the same algorithms as before except the second order PPC (too computationally heavy for thousands of genes) and the DPI (similar to CMI).

3.2.1 Artificial dataset

In this subsection, we consider a big artificial network with a scale-free topology plus dense modules, as described in section 2.1. Our results (Fig. 3.4, left) show

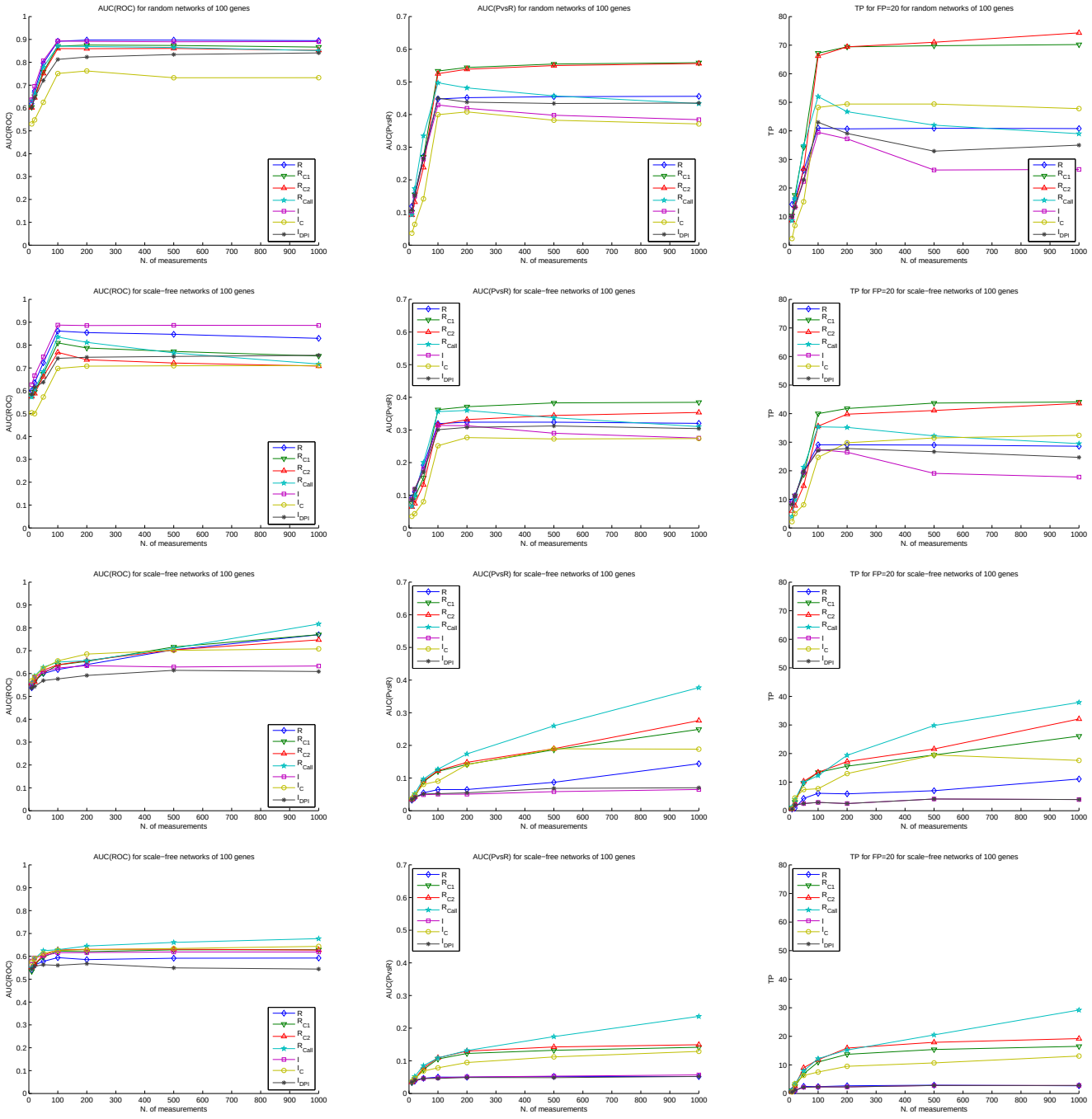


Figure 3.1: Evaluating the reconstructions via R, R_{C_1}, R_{C_2}, $R_{C_{\text{all}}}$, I, I_C and I_{DPI} algorithms on 100 gene artificial networks for increasing numbers of measurements. Top row: random topology, knockout perturbations and steady state measurements. Second row: scale-free topology, knockout perturbations and steady state measurements. Third row: scale-free, knockout and time-course experiments. Fourth row: scale-free, only initial conditions perturbations and time-course experiments. On the two time courses 10 (equispaced) samples are taken on each time course. The x axis label "N. of measurements" refers to the total number of samples taken (for example 200 means 200 experiments of steady state type, but only 20 experiments on the two time courses). Left column: AUC(ROC). Central column: AUC(PvsR). Right column: number of TP for a number of FP equal to 20. Values shown are means over 10 repetitions.

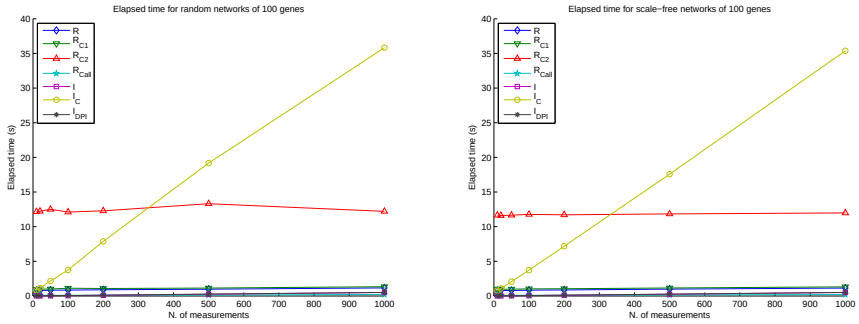


Figure 3.2: Runtime of the algorithms for the random (left) and scale-free (right) networks of 100 genes shown in Fig. 3.1.

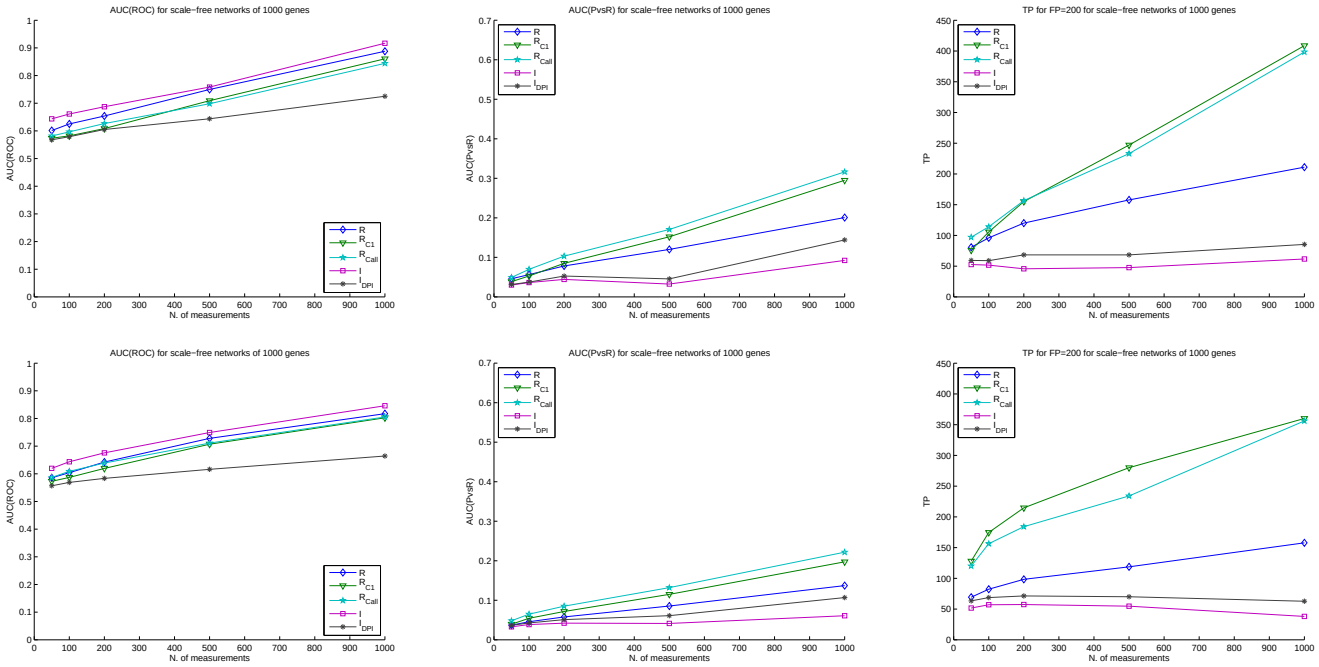


Figure 3.3: Evaluating the reconstructions via R, R_{C_1}, $R_{C_{\text{all}}}$, I and I_{DPI} on 1000 gene artificial networks of scale-free type, for increasing numbers of measurements, all for knockout experiments and all at steady state. First row: average node degree 1.5. Second row: average node degree 3. Left column: AUC(ROC). Central column: AUC(PvsR). Right column: number of TP for a number of FP equal to 200. Values shown are means over 3 repetitions.

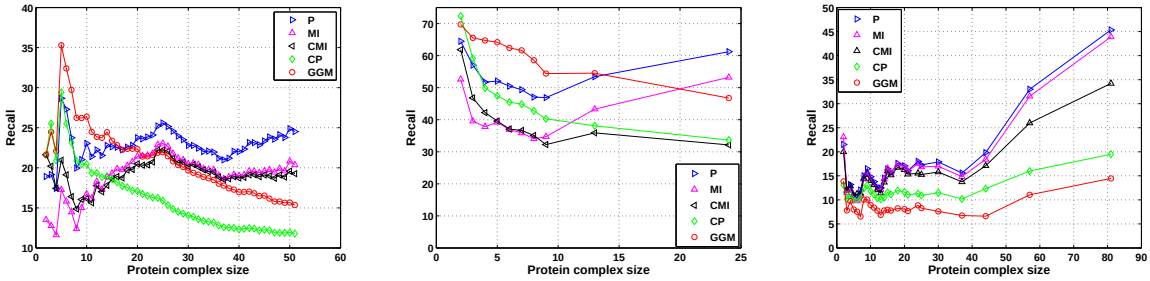


Figure 3.4: Recall (i.e. TP rate) for the network reconstructions at the first percentile with the five different similarity metrics for increasing size of the PCs, in: (left) artificial dataset, (middle) *E. coli* and (right) *S. cerevisiae*. In all three cases, considering the percentage of TPs for the whole PC network, the two direct metrics can be ranked in the same order (correlation followed by MI) and are performing better than the corresponding conditional metrics.

that increasing the size of the dense modules, conditional metrics perform worse than direct metrics.

Also the clustering of the reconstructed network shows the same qualitative difference and in fact the best results are obtained for the direct measures (correlation, MI). In Fig. 3.5 the percentage of complexes completely contained in: one cluster, two clusters, three clusters and more than three are shown.

On the other hand for the causal modules (Fig. 3.6, left), the performances of the conditional metrics are higher than the direct ones in correspondence of the largest modules. Notice how for all 5 algorithms the absolute performances drop dramatically when the number of transcription factors increases, as we expect due to the more complicated combinatorial effect.

As for the PvsR curves of Fig. 3.7, the qualitative difference between direct and conditional metrics in the two regulatory structures are substantially confirmed, although in the TF–BS curve (bottom row) the differences are minimal (precision is much lower than in PC).

3.2.2 *E. coli* dataset

Owing to the different genome organization and architecture, in prokaryotes regulatory mechanisms are much simpler than in eukaryotes. Genes are organized in transcriptional units, with one promoter for many consecutive genes, a feature absent in monocistronic eukaryotic DNA. *E. coli* has only a few large PCs and also the combinatorial regulation of transcription is lower, so we expect the different algorithms to have more similar performances.

We calculate the PvsR and ROC curves of the five different metrics (Fig. 3.7 and 3.8, middle) and plot the percentage of TPs in the most significant percentile of edges, for increasing sizes of the PCs (Fig. 3.4, middle) and combinatoriality of

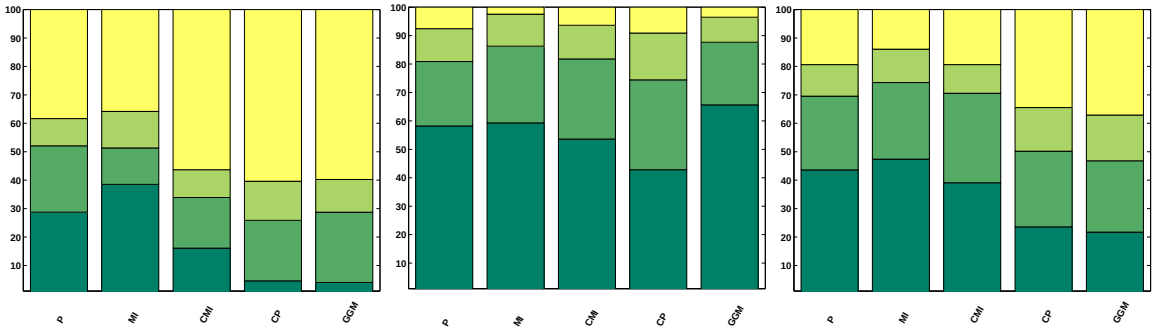


Figure 3.5: Clustering of the inferred graphs. For the artificial (left), *E. coli* (middle) and *S. cerevisiae* (right) networks, the color scale represents the percentage of PCs belonging to a single cluster (darkest), two clusters, three clusters and more than three (lightest). Correlation and MI are almost unanimously outperforming the three conditional metrics (the only exception being GGM for *E. coli*).

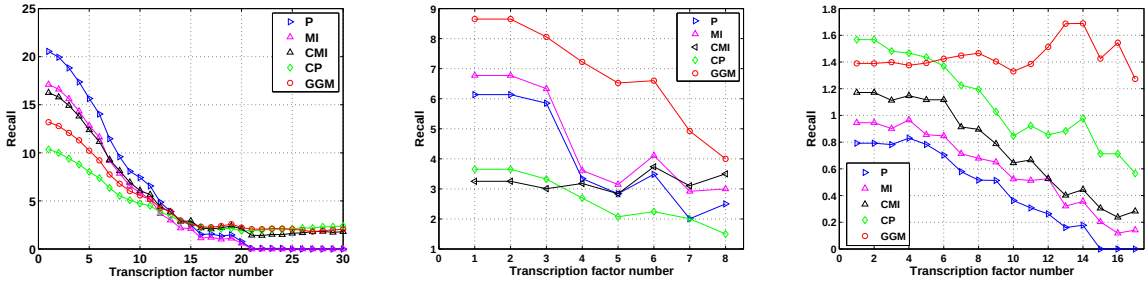


Figure 3.6: Combinatorial transcription regulation. Recall for TF–BS modules with increasing number of TFs on the same BS, in: (left) artificial, (middle) *E. coli*, and (right) *S. cerevisiae* datasets. In all three plots the downward trends in the ability to recover causal modules is visible and in all three the conditional measures seem to outperform the direct measures when combinatorial complexity increases.

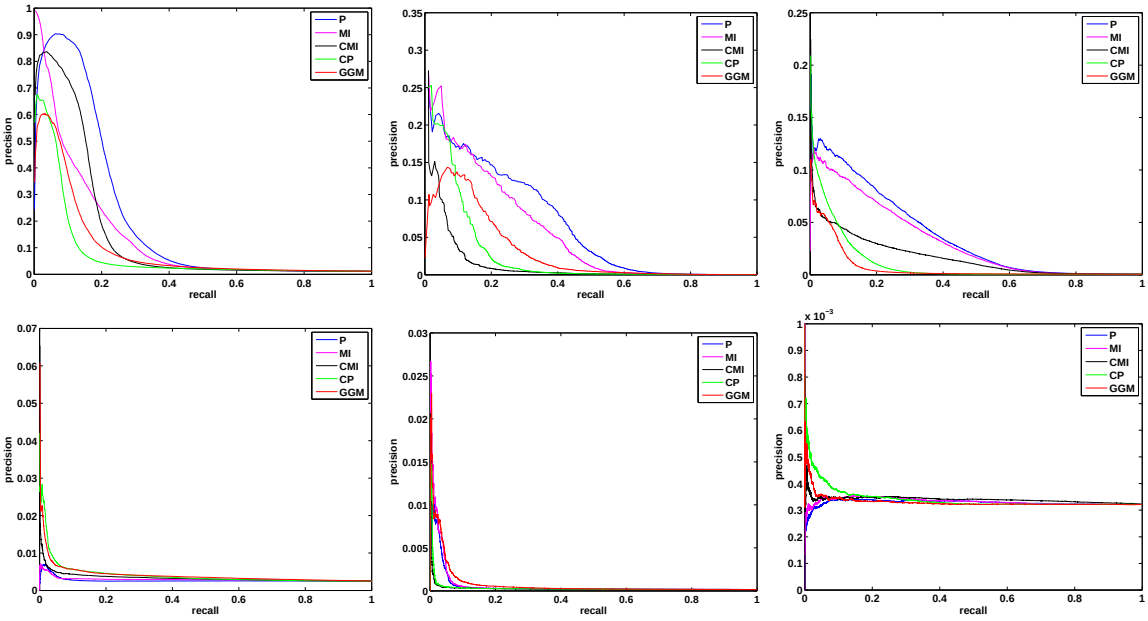


Figure 3.7: PvsR curves for the reconstruction of PC and TF–BS networks. Top row: PvsR curves of the five different similarity metrics using the PCs as the true network for the artificial (left), *E. coli* (middle) and *S. cerevisiae* (right) datasets. In all three cases the two direct metrics (correlation and MI) seem to be performing better than the corresponding conditional metrics. The curves are very high in the artificial network case because the density of true PC edges is higher than in the two organisms, see Table 2.1. Bottom row: PvsR curves of the five different similarity metrics using TF–BS interactions as the true network for the artificial (left), *E. coli* (middle) and *S. cerevisiae* (right) datasets. In absolute terms, the inference power is much lower than for PCs. Notice, however, how the conditional metrics still give the best results (in *E. coli*, correlation and MI are performing slightly better than PPC and CMI, but GGM is still outperforming all the others; compare also Fig. 3.6, middle panel).

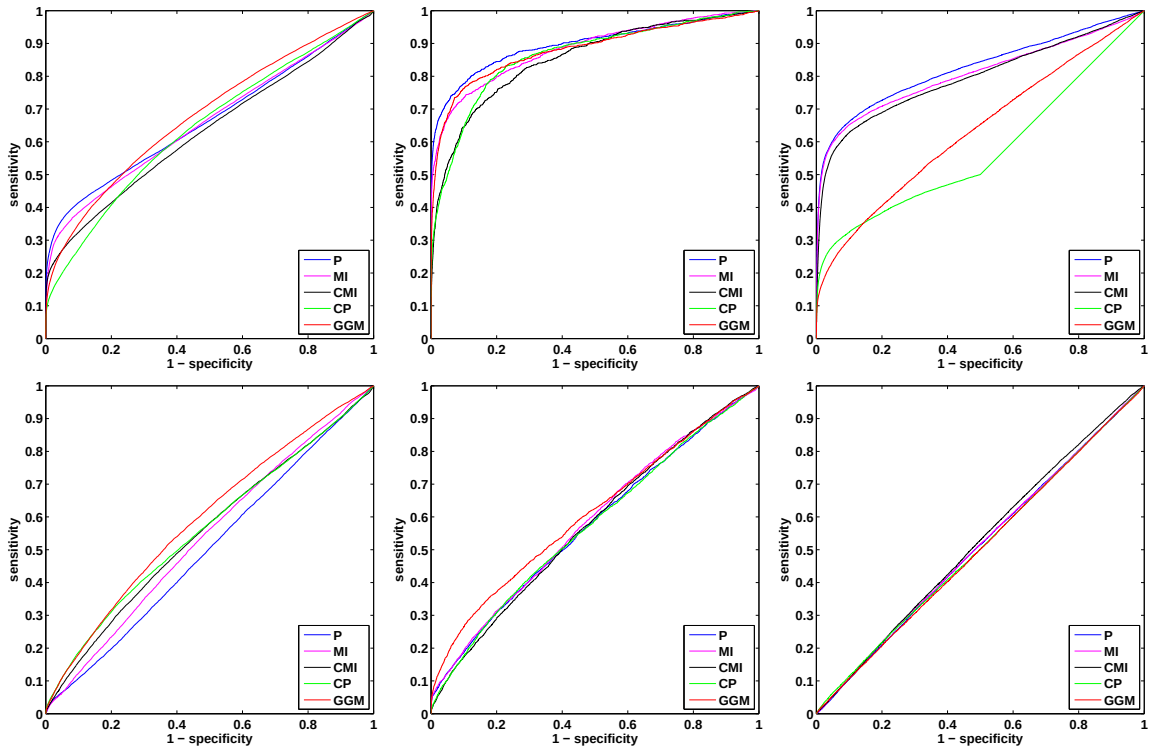


Figure 3.8: ROC curves for the reconstruction of PC and TF–BS networks. Top row: ROC curves of the five different similarity metrics for the PC network, in artificial (left), *E. coli* (middle) and *S. cerevisiae* (right) dataset. In all three cases the correlation is performing better than all the other metrics. Bottom row: ROC curves of the five different similarity metrics for the TF–BS network, in artificial (left), *E. coli* (middle) and *S. cerevisiae* (right) dataset. These curves highlight the general inability of inferring TF–BS relationships from co-expression indexes. Notice how the conditional metrics (in particular the GGM) in the artificial and *E. coli* networks give (slightly) better results.

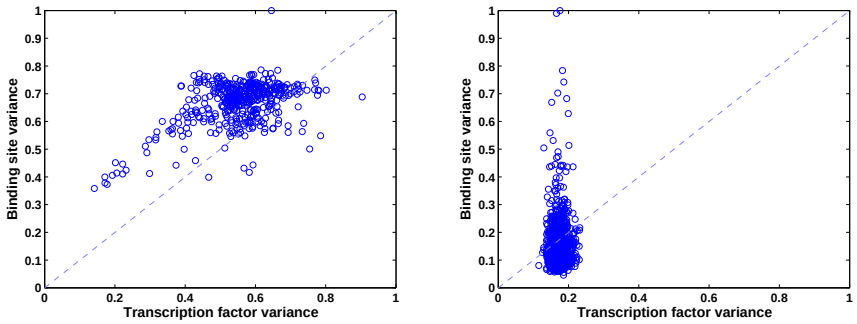
TFs (Fig. 3.6, middle). PCs are identified slightly better by the two direct metrics, although the number of relatively large complexes is too low to have statistical significance. The different performances emerging from the clustering (Fig. 3.5, middle) indicate that the highest correspondence between PCs and clusters are provided by GGM followed by correlation and MI. Considering as an example the flagellum complex (accounting for 24 genes), if the clustering procedure is performed by means of correlation and MI, the complex belongs entirely to a single cluster, which contains also other genes functionally related to the flagellum, like chemotactic genes and other genes involved in flagellar biogenesis and motility. Instead for CMI, PPC and GGM this complex is split respectively into 6, 8 and 2 clusters.

Regarding TF–BS relationships, we expect the ability in recovering true interactions to be inversely proportional to the multiplicity of TFs. This is particularly true for the algorithms performing well on low multiplicity TF (correlation, MI and GGM), while CMI has a counterintuitive slightly positive trend for multiregulated targets.

3.2.3 *S. cerevisiae* dataset

In *S. cerevisiae*, Fig. 3.4 (right panel) shows clearly that for small complexes the performances of conditional metrics are comparable with those of correlation and MI, up to a critical size above which the inference power of CMI and GGM remains almost constant while the direct metrics increase their percentage of TPs. The results are consistent with the ones obtained for the artificial data. Qualitatively, the results on the two organisms are the same, although the percentages of TPs are higher in the simpler one (see also Fig. 3.7, top row). In addition, the critical size of the dense modules for which conditional similarities start to fail is almost similar to the one obtained in the artificial network and *E. coli*, suggesting an intrinsic peculiarity of such similarity metrics. The clustering performances (Fig. 3.5, right) for the five algorithms are coherent with those of the *E. coli* and artificial networks and once again better results are obtained for the correlation and MI metrics.

If we move to the network of TF–BS (Fig. 3.6, right), we immediately notice that all the three conditional metrics perform better than the direct ones, although in absolute terms results are much worse than for *E. coli*. One reason for the low inference power regarding TF–BS could be that regulation is not just combinatorial but also combinatorially different in different environmental conditions. Another could be that TFs do not show the large variations in expression that can be seen for the corresponding regulated genes, but instead keep their expressions at low basal levels (see Fig. 3.9(b)).



(a) (b)

Figure 3.9: TF vs BS variance. Scatter plots representing the expression variance of the TFs against that of the corresponding BSs. In both organisms ((a) *E. coli* and (b) *S. cerevisiae*) most of the times the BS variance is broader than for the corresponding TF. Notice how especially in *S. cerevisiae* all TFs have low variance and how most node pairs in TF–BS edges live in the low variance corner.

Chapter 4

Conclusion

For the networks generated with the model (2.1), we find that steady state systematic gene knockout experiments are the most informative for the purpose of reconstructing this type of networks, yielding an AUC(ROC) > 0.7 even with $m \ll n$. In particular for this class of perturbations the linear similarity measures are enough. The nonlinear measures MI and CMI instead are less precise. For time series, the situation is different: relevance networks perform poorly even when $m \gg n$. In this context, conditional measures are relatively good. The marked difference between inference on steady state + knockouts and the more "classical" dynamical inference from time series alone without knockouts, is probably due to the highly nonlinear content of the transient evolution of (2.1). Reverse engineering nonlinear dynamical systems is notoriously a very difficult problem, and not even the use of nonlinear similarity measures is enough to attain a decent predictive power. At steady state, such nonlinear behavior has collapsed into a set of algebraic relations (corresponding to $dx_i/dt = 0$), which become sufficiently informative if "structurally" perturbed, e.g. by means of node suppressions. In short, structural perturbations are more efficient than dynamical perturbations for the purposes of (nonlinear) network inference.

Other interesting observations are the following:

- After a certain threshold $m_0 \geq n$ the inference ratio of all algorithms tends to stabilize. To improve the predictive capabilities, other types of perturbations should probably be used (like simultaneous multiple knockouts, external stimuli, etc.).
- AUC(ROC) around 0.9 are reached only by MI, correlation and GGM in the steady state knockout simulations.
- Conditioning is useful to improve the false discovery rate, and the TP it identifies are to a large extent different from those detected without conditioning.
- Of all algorithms tested only second order PPC and CMI are too computationally intensive to be used in a truly large network (tens to hundreds of thousands of genes).

- MI, CMI and DPI depend heavily on the implementation algorithm, and, at least in our B-spline implementation, on the underlying model of probability distribution (for time-course experiments the quality of the reconstruction improves considerably with the pre-application of a rank transform to the data). Correlations instead, are much less sensitive. For example replacing Pearson correlation with Spearman correlation yields no substantial difference.

- The best performances versus runtime are achieved by the GGM algorithm.

- Sparse networks are easier to identify than dense (or less sparse) ones, regardless of the algorithms used.

- Even with $m \ll n$ (realistic situation), using steady state knockout experiments all algorithms have a decent predictive power.

If unsupervised graph learning problems are notoriously difficult [36, 96], the conditions under which these problems must be studied for large scale gene regulatory network inference (less data than nodes) are even more challenging. Nevertheless, we can see through simulation and through reasonable biological assumptions on real data that the predictive power of current methods is indeed non-zero, and that a certain amount of structural information can be extracted even in this regime by means of computationally tractable algorithms, although the precision is very low and the number of false positives unavoidably very high.

Moreover, the results reported show that indeed different reverse engineering algorithms have performances which are tailored to different "characteristic" regulatory modules. PCs are characterized by a very stable binding and give rise to a sort of post-transcriptional regulation, where gene products have to be expressed in a constant stoichiometric ratio and are mutually dependent one from the other, features absent in cause–effect relationships such as transcriptional activation. For the network generated with the model and the two real ones, we tested the ability to recover dense modules/PCs and causal transcriptional modules of different sizes. Several important observations emerge from the results. The critical size of a dense module for which direct similarity measures begin to perform better than the corresponding conditional ones is between 10 and 20 on both artificial and real datasets. The dense modules that characterize PCs are better captured by direct similarity measures, especially for large dense modules. This is almost the same in both organisms, in spite of the different complexity and the low experiments/genes ratio. On the contrary, the conditional similarity measures are more suited to deal with causal dependencies such as TF–BS interactions, especially when the combinatorial complexity of the regulation increases. It is evident and predictable that the ability to recover TF–BS interactions is roughly inversely proportional to the number of TF regulating a gene. At the same time, it is worth pointing out that conditional metrics are more robust in taming this multiplicity effect of TFs. Needless to say the inference power of all the algorithms is higher in the simpler organism,

for both PC and TF–BS networks. This reflects the more complex eukaryote regulation, as deducible also from Fig. 1.2. Finally it is worth remarking that although direct metrics are better at detecting "static" interactions and conditional metrics at detecting causal ones, in absolute terms all algorithms are far more powerful at discovering the static than the causal gene–gene dependencies (as can be deduced comparing the first and second rows in both Fig. 3.7 and 3.8, or comparing the recall percentages of Fig. 3.4 and Fig. 3.6).

The predictive power of a reverse engineering algorithm is clearly a function of several aspects. First of all system complexity, data quality and numerosity. In addition, inference power depends on the type of interaction and the associate topology. Showing that indeed the algorithms yield different performances coherently with the features they are meant to extrapolate from the data (direct for static and stable interactions, conditional for causal interactions) is already a significant and encouraging observation.

Part II

The role of mRNA stability in the coordination of the yeast metabolic cycle

Chapter 5

Introduction

Ultradian self-sustaining energy-metabolic oscillations arising spontaneously in high density *Saccharomyces cerevisiae* continuous cultures exposed to glucose-limited growth have been known and studied for decades [94, 98], and have more recently been observed to induce genome-wide periodic patterns in different series of microarray experiments [76, 119], although with widely different periodicities, $\sim$ 40 min for [76] and $\sim$ 300 min for [119].

Many studies aim at understanding the mechanisms inducing these sustained oscillations and the rigorous temporal compartmentalization they induce, see [95, 100] for surveys. Suggested causes range from a single critical pathway (like the feedback effect of cysteine on the sulfur assimilation pathway [126]) to the alternation of aerobic and anaerobic respiratory modes (as deduced by the fluctuations in the concentration of dissolved O_2 and of other observed metabolites [119]), from the interaction with cell cycle [20, 51] to the mutual incompatibility of different redox biochemical processes [83, 127].

The scope of this work is to emphasize a different aspect, intrinsically dynamical and post-transcriptional, which is likely to play an important role in the coordination of the "slower" yeast metabolic cycle (YMC) of [119], namely mRNA stability. We will show that there is a roughly linear relationship between the average half life (HL) of the transcripts, clustered according to expression or function, and the phase at which their concentration peaks in the cycle. More generally, there seems to be a strong correlation between HL and the shape of the pulses of gene expression: genes with short HL have short and sharp (almost impulsive in the time scale considered) pulses, while genes with long HL have pulses that are not only delayed but also broader and with more gentle slopes.

In recent years, post-transcriptional control is being recognized as an important aspect of gene regulation, especially in eukaryotic DNA, which lacks operonal structure [11, 13, 53, 91]. It can occur in many guises, through mRNA turnover [18, 57, 78, 123, 125], or through "RNA regulons" [70], i.e. groups of genes coordinately guided in the RNA processing, localization and protein synthesis by RNA-binding proteins (RBPs) [56, 109], or even through the mediation of a metabolic

substrate (typically a nutrient [17, 71, 122] or an enzyme [59]). Our result confirms the importance of post-transcriptional control, and points at mRNA turnover as a regulatory mechanism at a genome-wide level. Its peculiarity consists in putting the time axis into the picture in an intrinsically dynamical way. Consequently, in order to be observed, it requires times series sampled at a sufficiently high frequency and dynamics in the right time window, a combination seldom occurring in current expression profiling datasets. So for example the correlation between HL and phase/shape of the oscillations cannot be observed in the much faster YMC of [76], where HL and the period are of comparable duration, hence the system has no time to decay before the arrival of the next wavefront.

In order to emphasize the dynamical aspects, we shall treat the YMC as the time response of a genome-wide dynamical system to a sequence of impulsive "inputs" of transcription activation. We will show that grouping genes in terms of progressively delayed and broadened responses to a sequence of "input pulses" of transcriptional activation allows to see in a remarkably fine detail the causal chain of events constituting the transcriptional program of the cell. The few ambiguities resulting from this classification can be interpreted in terms of some other annotation, typically compartmental localization.

In the following we shall proceed in two complementary ways: first the YMC time series are clustered in a completely unsupervised manner, only according to gene expression. The linear relationship between pulse phase (also pulse width) and HL then emerges in a straightforward way. Next, we consider families of genes whose products share some common annotation, for example genes on the same pathway or genes that are subunits of the same protein complex, and look at the type of time series they produce and at their "position" along the YMC. Both approaches confirm that the YMC represents an organized cascade of events, in response to precisely equispaced bursts of transcriptional activation, with the temporal order reflecting the transcript turnover rate.

Extrapolating from the specific YMC context, this cascade of events is observable to a good extent also in other gene expression time series (such as the response to a pulse of nutrient of [104], or the stress responses of [55]), suggesting it might reflect a prototypical dynamical mode of action of transcriptional response.

Chapter 6

Methods

6.1 Data sources

The YMC time series of [119], the compendium of 790 gene profile experiments (all performed with the Affymetrix GeneChip Yeast Genome S98 platform) and the data series from [104] were downloaded from Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/). The time series of [104] are performed with cDNA, hence values of the area under the profiles are intended as relative (to the basal mRNA abundance). For each gene, the values obtained for the two different glucose stimuli are averaged. Five stress responses from [55] (two heat shocks of different amplitude, hydrogen peroxide, diamide, and sorbitol responses) are considered. The amplitudes are averaged over the five data series (the signs of these responses are known to be highly similar, see [55]).

The metabolic pathways used are those of the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/). Also the assembling into the 15 macrocategories follows the KEGG hierarchy.

The protein complexes were downloaded from the MPACT subsection (http://mips.helmholtz-muenchen.de/genre/proj/mpact/) of the CYGD database at MIPS. Only complexes manually annotated from the literature are considered; those obtained from high throughput experiments are disregarded (according to the MIPS classification scheme these last are labeled “550”).

The HLs are computed averaging the values of the three experimental datasets [57, 78, 123]. While the magnitudes of the HLs in the three collections show some differences, in “normalized” terms (looking e.g. at rank-ordered values), the agreement between the three sets is sufficiently good, see [57] for a comparison. No turnover data specific for long-term continuous cultures are currently available. However, it is not unlikely that even in this setting the relative differences of HL rates (and also their ordering) remains more or less unchanged. In any case, we expect the correlation phase/HL to improve in presence of more tailored mRNA turnover data.

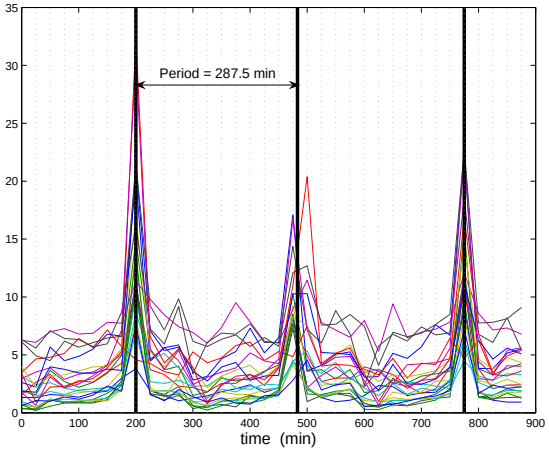


Figure 6.1: When the period of the bursts in the YMC is computed via Fourier analysis, as is done in [119], the answer is 300 min. However, a closer look at the genes having impulse-like behavior (in this Figure the three RNA polymerases) reveals that the sampling is not perfectly synchronized with the period observed: in a time window twice the period (200÷775 min) there are 23 samples instead of the expected 24, and the "11.5" samples per period ratio seems to yield a more accurate matching of the peaks. The resulting period is therefore $11.5 \cdot 25 = (775-200)/2 = 287.5$ min. Notice how this explains why the second peak is less resolved that the first and third one in basically all Figures shown in this work.

6.2 Time series analysis

The period of the YMC, computed in the time domain looking at the most impulsive-like categories (in Fig. 6.1 the 3 RNA polymerases), is estimated as 287.5 min (see Fig. 6.1 for a detailed description). As in [119], genes are filtered using a periodogram test. In order to retain only genes with a well-defined periodicity, we fixed a more selective p.value than [119] thereby reducing the number of genes to 1951. To each of the genes labeled as periodic, we associated a phase, computed maximizing the correlation with respect to a train of 360 shifted sinusoids (resolution of 1°). Thus, a phase delay ϕ can be transformed into a time delay τ by means of the relation $\tau = \phi \frac{287.5}{360}$ (min).

Means and standard deviations of the phases of periodic signals must be computed "mod 360°", and are normally subject to large numerical errors and ill-conditioning. A typical example is the following: assume two periodic genes are assigned the phases $\phi_1 = 350°$ and $\phi_2 = 6°$. Owing to the 360° periodicity, the peaks of the two genes are very close, but the average phase is $(\phi_1 + \phi_2)/2 = 178°$, which is obviously wrong. The correct answer requires a shift from the principal values of the periodic signal: $(\phi_1 - 360° + \phi_2)/2 = -2°$. To avoid problems with biased mean values and/or the appearance of inelegant negative phases around the "crucial" transcription bursts, the 0 phase was chosen so as to anticipate of $\sim 30°$

these events. Under this convention, each period "begins" approximately 24 min before the transcription bursts.

For each gene, the pulse width is computed estimating on each period the interval in which the expression level stays above the median value across consecutive samples.

6.3 Least squares regressions

In Fig. 7.1, the least square fitting in the HL/phase plot is given by the equation

$$\phi = 9.25\ \text{HL} - 104.83^\circ, \quad R^2 = 0.86, \quad \text{p.value} \sim 10^{-7}.$$

The corresponding p.value is computed via a Fisher test statistics. Since we have determined the period as 287.5 min and the zero phase $\sim 30^\circ$ before the impulsive bursts shown for example in Fig. 6.1, the equation in terms of time delay with respect to the bursts, $\tau_t \simeq \frac{287.5}{360}(\phi - 30^\circ)$, is

$$\tau_t \simeq 7.39\ \text{HL} - 107.68\ \text{(min)}.$$

Within most clusters, the standard deviation in terms of ϕ is minimal; it is higher in terms of HL, see Table 7.1. Hence if we use weighted least squares regression, while the fitted curves we get are still very similar in the range of values of interest, the differences are in the coefficient of determination R^2:

Method	Regression	R^2	p.value
l. s. weighted w.r.t. ϕ	$\phi = 8.95\ \text{HL} - 101.24$	0.92	$\sim 10^{-9}$
l. s. weighted w.r.t. HL	$\phi = 9.03\ \text{HL} - 99.76$	0.54	$\sim 10^{-4}$

Let ω be the width of the pulses, then the corresponding least squares fits are

Method	Regression	R^2	p.value
l. s.	$\omega = 0.24\ \text{HL} - 1.05$	0.72	$\sim 10^{-5}$
l. s. weighted w.r.t. ω	$\omega = 0.21\ \text{HL} - 0.99$	0.32	$\sim 10^{-5}$
l. s. weighted w.r.t. HL	$\omega = 0.23\ \text{HL} - 0.71$	0.31	$\sim 10^{-2}$

Repeating the linear regression for the three plots in Fig. 7.3(b), we get:

- phase/HL (top plot)

Method	Regression	R^2	p.value
l. s.	$\phi = 6.84\ \text{HL} - 33.05$	0.64	$\sim 10^{-4}$
l. s. weighted w.r.t. ϕ	$\phi = 4.39\ \text{HL} + 27$	0.82	$\sim 10^{-6}$
l. s. weighted w.r.t. HL	$\phi = 5.26\ \text{HL} + 6$	0.68	$\sim 10^{-4}$

- width/HL (middle plot)

Method	Regression	R^2	p.value
l. s.	$\omega = 0.15$ HL + 1.41	0.44	$\sim 10^{-3}$
l. s. weighted w.r.t. ω	$\omega = 0.19$ HL + 0.5	0.53	$\sim 10^{-3}$
l. s. weighted w.r.t. HL	$\omega = 0.11$ HL + 2.58	0.8	$\sim 10^{-5}$

- width/phase (bottom plot)

Method	Regression	R^2	p.value
l. s.	$\omega = 0.02\phi + 1.9$	0.8	$\sim 10^{-6}$
l. s. weighted w.r.t. ϕ	$\omega = 0.03\phi + 1.66$	0.96	$\sim 10^{-10}$
l. s. weighted w.r.t. ω	$\omega = 0.03\phi + 1.12$	0.8	$\sim 10^{-6}$

Finally, for the dynamical model of Fig. 7.12, if ψ is the order of the transfer function model used, $\psi \in [1, 4]$, we have

Method	Regression	R^2	p.value
l. s.	$\psi = 0.09$ HL + 0.32	0.52	$\sim 10^{-3}$
l. s. weighted w.r.t. HL	$\psi = 0.06$ HL + 1.1	0.73	$\sim 10^{-5}$

6.4 Clusterization

The clusterization of the time profiles in Fig. 7.1 is performed via a k-means algorithm using as distance a nonnormalized correlation function. Varying the number of clusters and/or the (randomly chosen) initial cluster assignments, the results (in terms of the regressions) are basically unchanged.

6.5 A minimal dynamical model: low-pass transfer functions and their dynamical system realizations

The aim of this Section is to set up a minimal dynamical model describing the response to the periodic bursts of transcriptional activation represented as "impulsive inputs" to the system. Such a model has to be able to reproduce the following features observable in the dataset:

- impulse responses get delayed and broadened in a way which is roughly proportional to HL;
- profile changes get progressively less steep with HL;
- the system "discharges" completely (i.e. the mRNA concentrations return to a basal level) in absence of further pulses.

At the same time, to be internally consistent a dynamical model has to:

- respect causality (i.e. be non-anticipating);

- preserve positivity of the mRNA concentrations.

In the Engineering practice of Systems Theory, one of the most elementary formalism that can be used to build dynamical models is the input-output design based on Laplace transform and elementary transfer functions [35], see e.g. [9] for an application to a transcriptional time series.

The concentration of mRNA of a gene y can be described as the response to the pulse of transcriptional activation u by the linear integral

$$y(t) = \int_0^t g(t-\tau)u(\tau)d\tau. \tag{6.1}$$

In the Laplace domain, a convolution integral such as (6.1) corresponds to

$$Y(s) = \mathcal{L}\left[\int_0^t g(t-\tau)u(\tau)d\tau\right] = G(s)U(s) \tag{6.2}$$

where s is the Laplace variable and $G(s)$ is called a transfer function. If $u(t)$ is a perfect impulse δ_0 (Dirac delta) then $U(s) = \mathcal{L}[\delta_0(t)] = 1$. When the transfer function $G(s)$ represents a linear differential equation (i.e. it derives from a linear convolution such as (6.1)), it can be expressed as a rational polynomial in the Laplace variable s. A simple such polynomial is

$$G_1(s) = \frac{s+n_1}{s+d_1} \tag{6.3}$$

where $s = -d_1$ is called the pole of G_1 and $s = -n_1$ its zero. Choosing $d_1 > 0$ the transfer function is stable (the pole is in the left half of the complex plane), i.e. a bounded input will always result in a bounded output. When $n_1 > 0$ the system is said to be minimum phase. In this context this is an important condition in order to guarantee positivity of the output signal for all times.

The requirements above can be translated into easy-to-handle design specifications on the values of the poles and zeros of the transfer function. For example, the first requirement (at least for what concerns pulse broadening) is met by the class of so-called low-pass filters, the most basic of which has the form given in (6.3), provided we choose $0 < d_1 < n_1$. The term "low-pass" literally means that low frequencies in the input signal pass unchanged through the transfer function $G_1(s)$, while high frequencies get damped, hence the impulsive input exits from $G_1(s)$ smoothed and with more gentle slopes. Such a transfer function is proper and therefore respects causality; it discharges completely as required (since it has no integrator, i.e. no factors of the form $1/s$ in $G_1(s)$). Strictly speaking, it is not a positive filter [1], however as long as $u(t) > 0$ and $0 < d_1 < n_1$ it is also $y(t) > 0$. In the Laplace domain, a time delay T_1 has Laplace transform equal to $e^{-T_1 s}$. This operator does not add poles or zeros to (6.3), but yields the irrational transfer function

$$y = G_1(s)e^{-T_1 s}u. \tag{6.4}$$

In the time domain, each convolution integral (6.1) can be expressed as a linear input-output systems (of ordinary differential equations). For the transfer function in (6.3) and the delay operator in (6.4) this corresponds to

$$\begin{aligned}\frac{dx(t)}{dt} &= -d_1 x(t) + (n_1 - d_1)u(t - T_1)\\ y(t) &= x(t) + u(t - T_1),\end{aligned}$$

i.e. the pole d_1 plays the role of "degradation rate" while the activation amplitude is proportional to $n_1 - d_1$ (> 0). The typical impulse response of a low-pass filter transfer function such as (6.3) is shown in the top plot of Fig. 7.12(b). Given a pulse shape, the capabilities of a single low-pass filter in terms of broadening and smoothing of the responses are limited, hence, in order to obtain a progressive effect of delayed and broadened impulse responses, several delayed low-pass filters should be put in cascade. For example the order-2 transfer function obtained concatenating 2 filters is

$$G_2(s) = \frac{(s + n_1)(s + n_2)}{(s + d_1)(s + d_2)},$$

or, in the time domain,

$$\begin{aligned}\frac{dx_1(t)}{dt} &= -d_1 x_1(t) + (n_1 - d_1)u(t - T_2)\\ \frac{dx_2(t)}{dt} &= -d_2 x_2(t) + (n_2 - d_2)\left(x_1(t) + u(t - T_2)\right)\\ y_2(t) &= x_1(t) + x_2(t) + u(t - T_2).\end{aligned}$$

In this case both d_1 and d_2 contribute to form the degradation profile of the mRNA concentration $y_2(t)$. Likewise, both dynamical variables x_1 and x_2 contribute to shape the pulse of a gene. Typically this model induces a steeper upregulation and a slower degradation front, coherently with what we observe on the YMC time series. The intermediate variables x_i are only meant to describe the complexity of the input-output relationship. Qualitatively, they might reflect intermediate steps in the gene expression program. For example, the transcription of the genes of the central metabolism is activated downstream of the genes for translation and amino acid synthesis, which in their turn follow the principal bursts of transcription machinery (polymerases and other RNA processing components). Downstream activation of the genes of a category translates in this modeling framework into delayed and broadened pulses. Typical output responses for 1, 2, 3, and 4 such concatenated blocks are shown in Fig. 7.12(b).

A simple parameter search can be set up to identify values of n_i, d_i and T_i, $i = 1, \ldots, 4$, that guarantee for each gene a sufficiently well-reproduced time course. The best transfer function order for each gene is identified as that maximizing the correlation between true and model-based time series.

Chapter 7

Results

The $\sim$ 2000 genes labeled as periodic by a periodogram test are subdivided into 16 clusters, see Fig. 7.1. In Fig. 7.1(a) the clusters are sorted in increasing order of HL (computed as the average of the HLs of the cluster elements). It is immediately evident that the typical profiles, both in terms of the phase of the peaks (for each gene the phase is computed maximizing the correlation with respect to a train of shifted sinusoids) and of their width (although in a less regular way) is modified in an almost continuous manner as we move along the clusters figures. Notice in particular how the peaks of the first clusters match the "valleys" of the last ones. For the average phase on each cluster, the phase/HL relationship is almost linear (Fig. 7.1(b)).

The scatter plot in Fig. 7.1(d) confirms this linear proportionality, but also shows a growing variance along the HL axis (see Table 7.1 for details). One of the reasons may be that the HL measures are imprecise (see comparison between HL datasets in [57]), and should probably be considered as context-specific, parametrically dependent on a set of physiological conditions (see also HL sources description in section 6.1).

The deviations from linearity of clusters 6 and 9 admit a reasonable explanation, mostly in terms of compartmental localization. Cluster 6 is essentially composed of retrotransposons (all Ty1 and Ty2) and long term repeat mRNAs (mostly of δ type) for a total of 73 out of 102 genes. For most of these genes (59) an HL measure is missing. Hence the average HL for this cluster (and this cluster alone) may be biased or unreliable. Cluster 9 instead is almost entirely composed of cytoplasmic ribosomal subunits (109 out of 151 genes). In between, Clusters 7 and 8 contain to a large extent genes with mitochondrial localization and/or function (mitochondria organization and biogenesis, protein import into mitochondrial matrix, oxidoreductase activity for Cluster 7, mitochondrial ribosomes, envelope and membranes for Cluster 8). As is explained in detail in the next paragraph, the large deviation from linearity seen in Cluster 9 can be due to an extremely fast and short lived response of the mRNAs deputed to the biosynthesis of the cytoplasmic ribosomal complexes, not deducible from the available HL data, neither from the

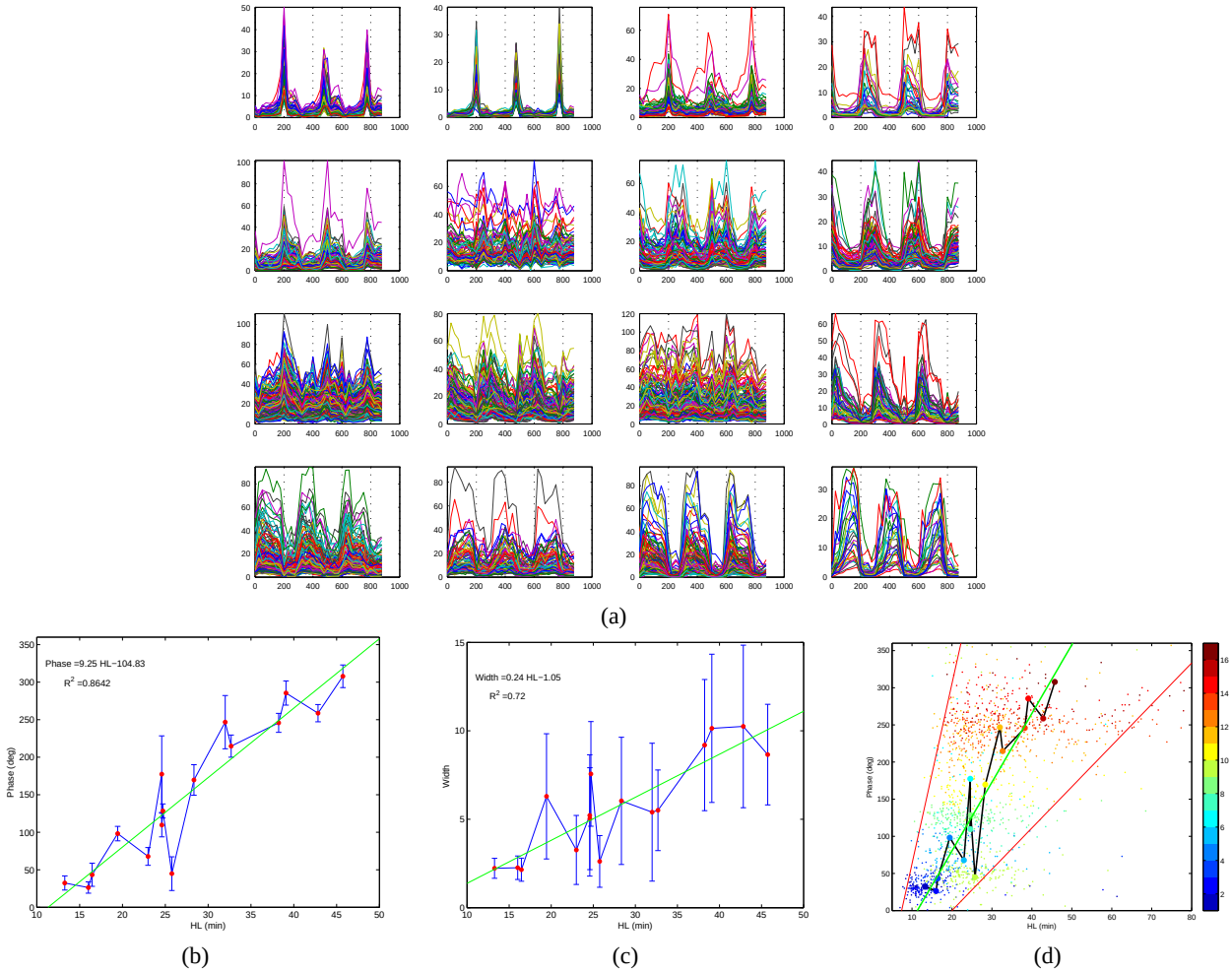


Figure 7.1: In (a) the time series (on the x axis: time in min) of the periodic genes is clustered according to a nonnormalized correlation distance function (see Table 7.1 for details on the clusters). The clusters are then sorted (from left to right from top to bottom) according to the average HL. In (b) the average HL is plotted against the average phase for each cluster, while in (c) the average HL is shown against the average pulse width. In the scatter plot of HL versus phase (d), the color indicates the cluster number (see colorbar on the right). As can be noticed, along the HL axis the standard deviation of a cluster grows with the mean, see Table 7.1 for exact values, and the cloud of points looks like a cone (the cone delimited by the two red lines contains 95% of the periodic genes). Still the increase of the phase with the HL is clearly visible. In the least-squares linear fit in (b) (green) half of the L_2 norm of the residues is due to Cluster 9 (cytoplasmic ribosomes, see text). The p.value for both linear regressions is $< 10^{-5}$. Further details on these regressions are provided in section 6.3.

Cl.	Genes	HL		Phase		Width		Ontology
		Mean	SD	Mean	SD	Mean	SD	
1	101	13.26	(9.54)	32.4	(9.4)	2.2	(0.56)	RNA, rRNA, and tRNA processing and metabolism, ribosome biogenesis and assembly
2	58	16.02	(19.07)	26.3	(7.6)	2.3	(0.66)	RNA, rRNA and tRNA processing and metabolism, RNA helicase, ribosome assembly
3	101	16.46	(8.65)	43.3	(15.4)	2.1	(0.65)	RNA polymerase, translation initiation, regulation, and termination, nucleotide biosynthesis
4	34	19.44	(10.19)	98.2	(9.7)	6.3	(3.54)	transferase activity, DNA replication, cell cycle
5	102	22.99	(10.27)	67.7	(11.8)	3.3	(1.95)	glycine metabolism, nitrogen and sulfur metabolism, amino acid biosynthesis
6	102	24.59	(11.67)	177.4	(51.0)	5.2	(3.43)	retrotransposons, long term repeats
7	124	24.59	(13.45)	109.6	(15.8)	5.0	(2.88)	mitochondrial membrane organization and biogenesis, mitochondrial transport
8	151	24.72	(11.80)	128.3	(9.4)	7.6	(2.96)	mitochondrial ribosome, envelope, and membranes
9	232	25.76	(13.78)	44.8	(22.5)	2.6	(1.46)	cytoplasmic ribosomes, translation processes
10	154	28.34	(16.36)	169.7	(20.3)	6.0	(3.59)	ion/cation transmembrane transport, electron transport, oxidative phosphorylation
11	230	31.99	(19.05)	246.7	(35.5)	5.4	(3.90)	endopeptidase activity, protein catabolic process, proteasome, actin filament organization, glycolysis, gluconeogenesis
12	65	32.69	(18.68)	214.8	(14.8)	5.5	(2.28)	lipid and alcohol metabolic process, peroxisome
13	223	38.24	(28.35)	245.8	(12.6)	9.2	(3.71)	kinase activity, vacuolar transport, membrane organization and biogenesis
14	128	39.10	(29.27)	285.5	(16.1)	10.1	(4.19)	arginine biosynthesis, protein folding
15	117	42.83	(28.02)	258.7	(11.5)	10.2	(4.59)	hydrolase activity, fatty acid oxidation, cytokinesis
16	29	45.74	(26.30)	307.8	(15.1)	8.7	(2.84)	catalytic activity

Table 7.1: Statistics for the 16 clusters for Fig. 7.1

current literature (in [131] it is affirmed that cytoplasmic ribosomal genes tend to be stabilized by nutrient uptake).

Although less precise, also the relation between HL and pulse width on each cluster (Fig. 7.1(c)) is approximately linear. Unlike the phase/HL proportionality, this last result is expected from simple dynamical considerations, as longer HL means longer "kernel width", see also the dynamical model explanation below.

The emergence of a linear relation between HL and phase once the genes are arranged in classes according to profile similarity suggests that a corresponding cascade of causally organized events may be taking place during the YMC. To some extent this is already visible through an ontological analysis of the clusters of Fig. 7.1 (see Table 7.1), but in order to investigate more in detail the biological meaning and significance of such a genomic "assembly line" we computed HLs, phases and pulse widths along the main yeast pathways and for some of the annotated yeast protein complexes. The data for the pathways (see Fig. 7.2) are then lumped together into the 15 functional macrocategories shown in Fig. 7.3. In terms of these macrocategories (sorted by phase), the result is that the mRNAs activation reflects tightly the gene expression program expected to take place in the cell, especially for the "fast" categories, i.e. transcription, nucleotide metabolism and translation starting essentially synchronously in the time scale of the YMC, followed by DNA replication and repair and amino acid metabolism. Progressing further toward the slow processes, one encounters the metabolism of energy, carbohydrates and lipids. Also for this classification, the progression in terms of phase along the cycle is substantially faithful to the increase in HL (in the top plot of Fig. 7.3(b) the most significant outlier is still the category "translation" already mentioned, see also Fig. 7.5), and the progression in phase is paralleled by an increase in pulse width (see bottom plot of Fig. 7.3(b)).

7.1 HL and the short-period YMC

The HL of a gene is defined as the time needed to halve the concentration of mRNA in absence of new transcription. Hence in order for a "full" degradation of mRNA to be observed, the interval between two consecutive waves of transcription has to be at least twice or three times the HL. For yeast, the mean HL extrapolated from [57, 78, 123] is $\sim 26 \pm 17$ min. Hence for the long-period YMC the response to bursts of transcription has the time to exhaust completely before the arrival of the next wavefront. On the contrary, for the short-period YMC described in [76] the period is approximately 40 min, meaning that excitation and degradation fronts are substantially overlapping.

7.2 A detailed functional analysis

Using the ordering by phase of pathways and protein complexes (see Fig. 7.2 and 7.4), we can zoom on these categories in much more detail. The first phase of

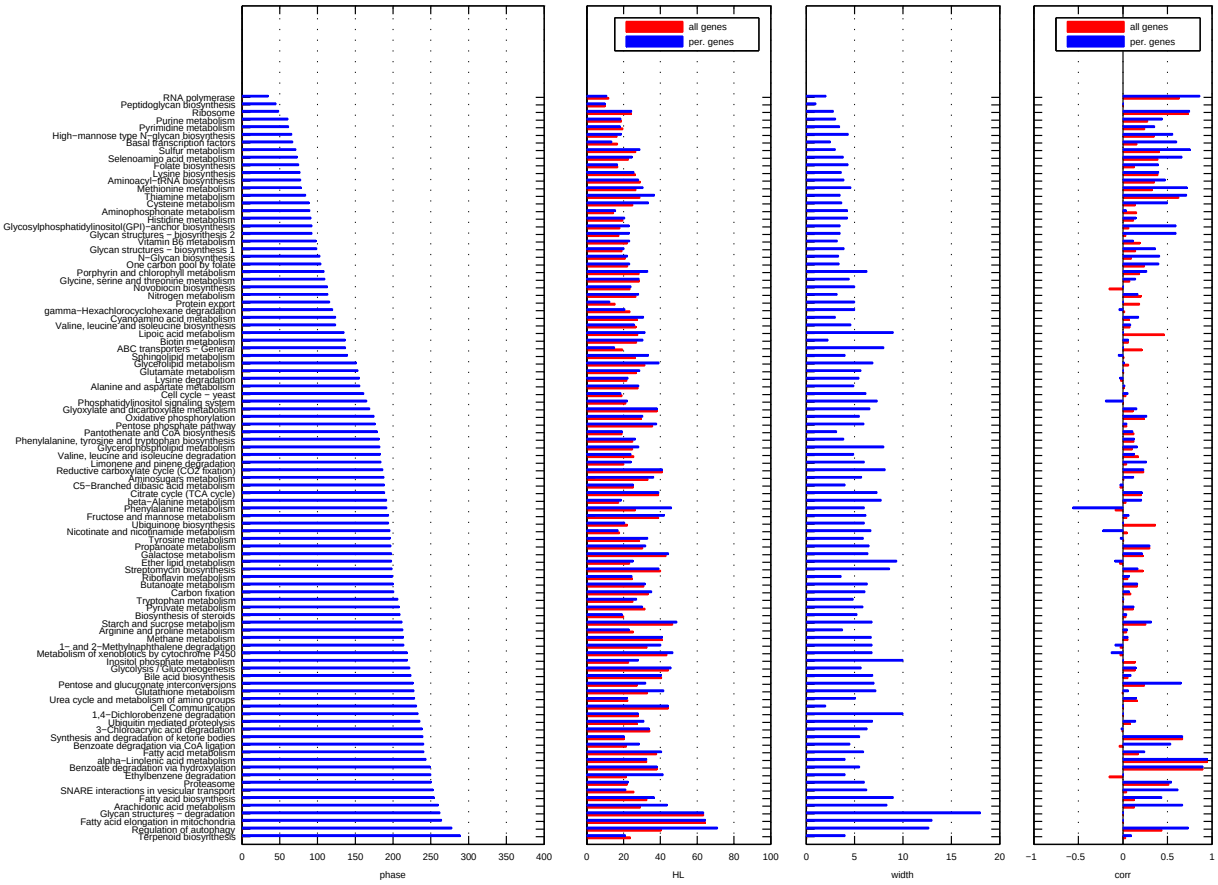


Figure 7.2: Average of phase, HL, area and Pearson correlation along KEGG pathways, for all genes (red) and for periodic genes (blue), sorted by phase of the periodic genes.

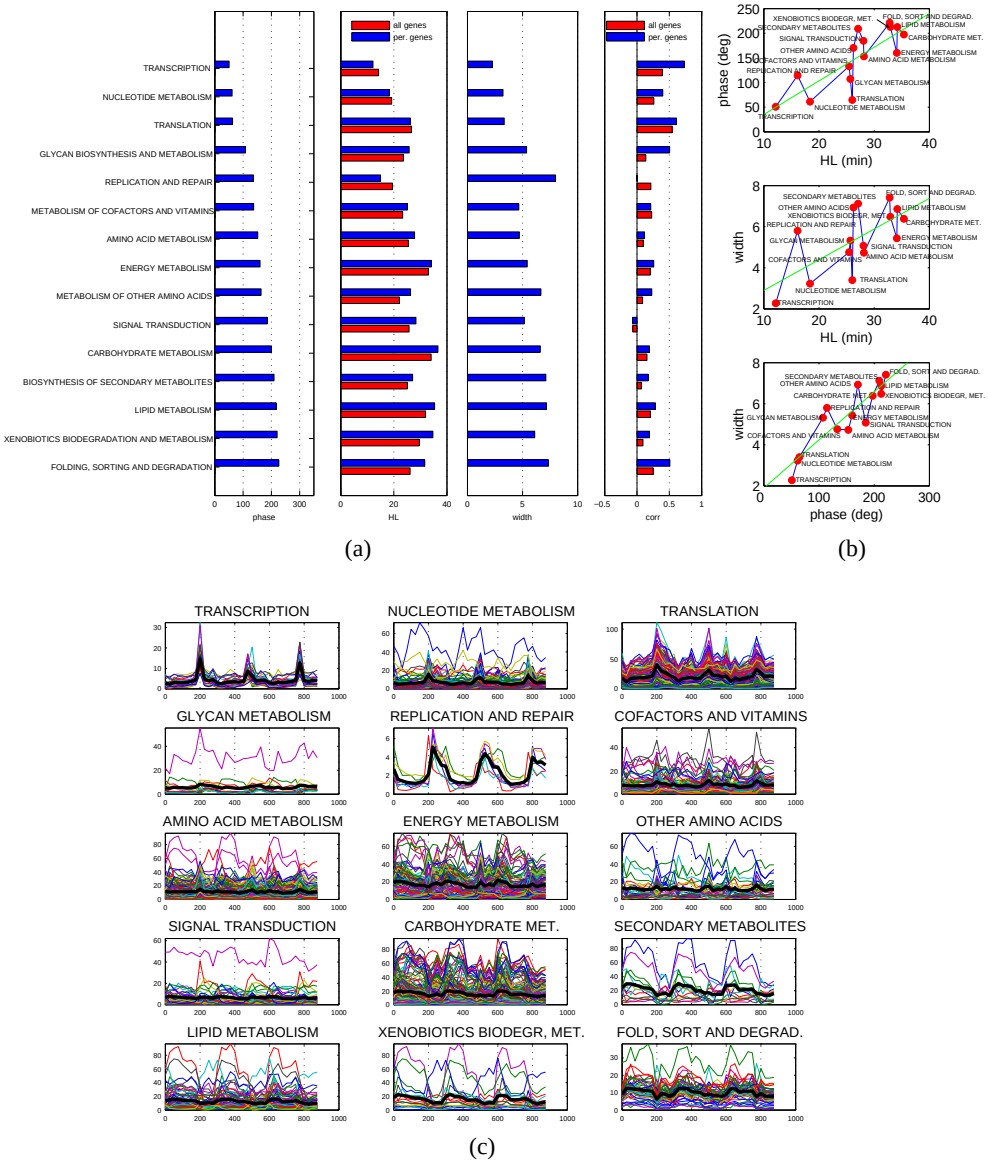


Figure 7.3: (a): The periodic genes of the YMC are grouped according to KEGG pathways (see Fig. 7.2) and then in the 15 macrocategories shown. For each macrocategory we calculate the average phase, HL, pulse width, and correlation of the periodic genes (in blue), and also the average HL and correlation of all genes (in red). Sorting by phase reveals the expected concatenation of events of the yeast gene expression program, especially in the first part with transcription preceding protein synthesis and DNA replication, followed by the slower categories of central metabolism. (b): Comparing HL and phase (or pulse width) roughly the same type of direct proportionality still appear. The trend in the average profiles of each category (black thick lines in (c)) reflects to a large extent that of Fig. 7.1. The third plot in (b) shows that also phase and pulse width are directly correlated: pulses that are delayed are also broadened. Linear regressions for these plots are discussed in section 6.3.

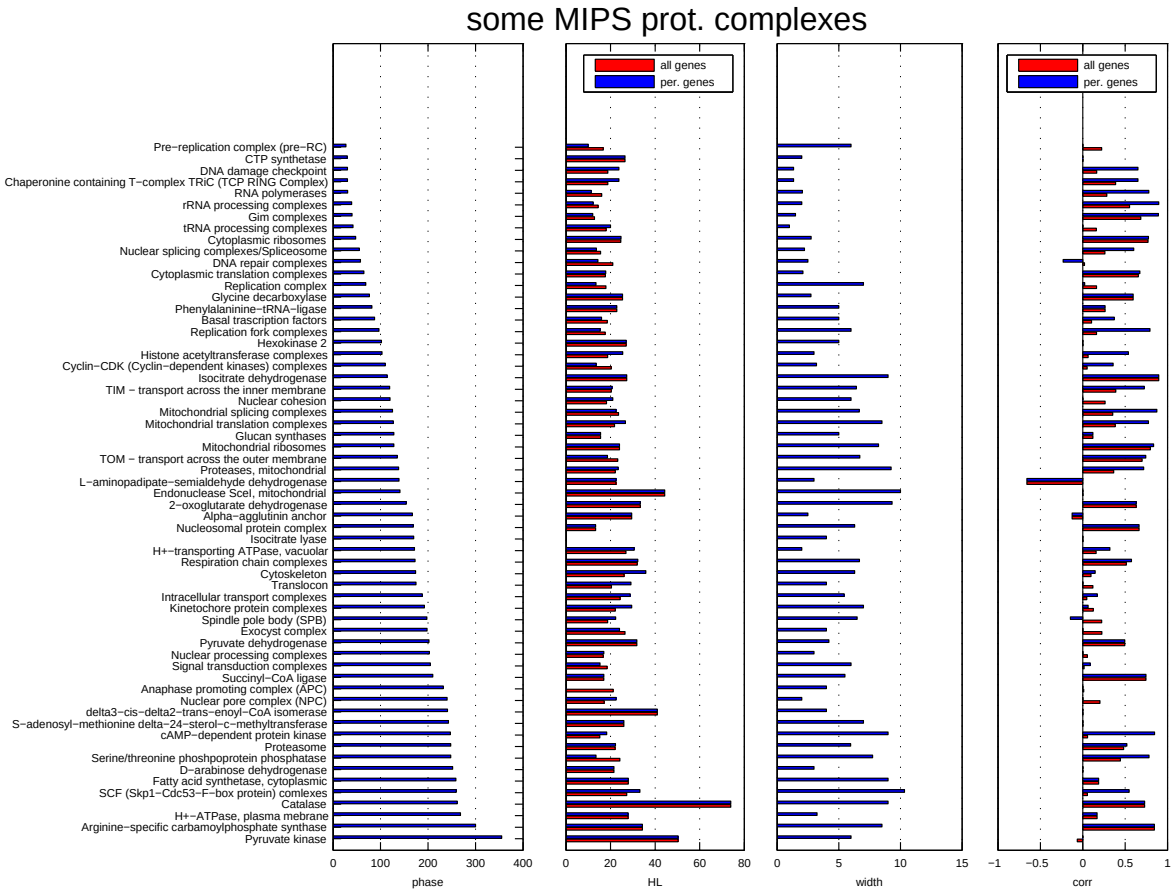


Figure 7.4: Average phase, HL, pulse width and intracomplex Pearson correlation for a few MIPS protein complexes, sorted by phase (on the periodic genes).

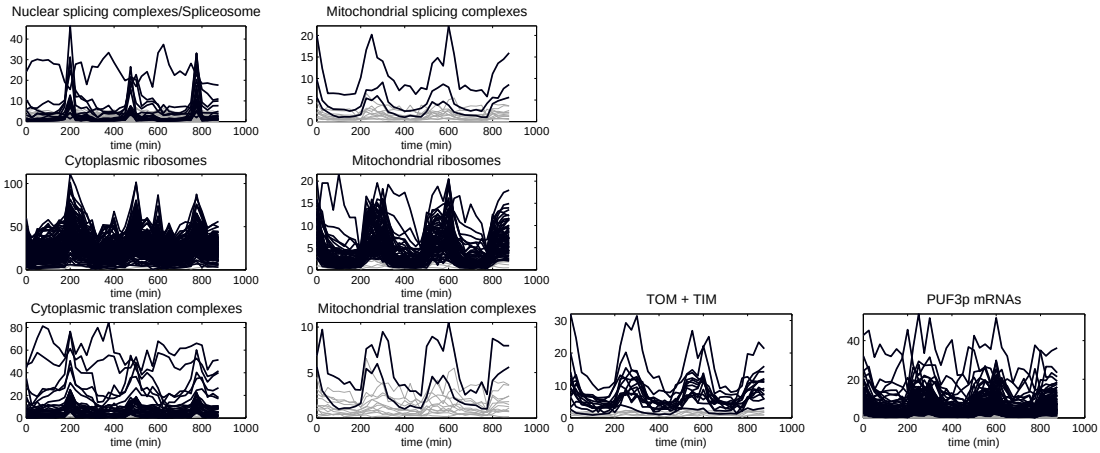
this cascade consists of the activation of the transcription machinery with the synchronous bursts of transcription of the three RNA polymerases (see Fig. 6.1) and of most of the RNA processing components, like the tRNA processing complexes (RNase P) and rRNA processing complexes (exosome, RNase MRP, *SIK1*, *NOP1*), with the nuclear splicing complexes following closely. While the mRNAs for the polymerases are highly coordinated, the same cannot be said for the basal transcription factors (TFs) required for their initiation. Overall only a few of these genes follow the bursting trend of the RNA polymerases, notably, among them, *SPT15*, which forms the TATA-binding protein and is also a component of the polymerase I core factor and of TFIIIB. Most other genes involved with these general TFs do not show any periodic pattern, and their mRNA concentrations never surpass very low levels.

From Fig. 7.3, the peak of mRNA concentrations associated with the category “translation” seems to be synchronous with the RNA processing burst. However,

a more careful analysis reveals that this phase is an average of two "compartmentalized" activations of the translation machinery, having fairly different phases: while cytoplasmic translation follows almost simultaneously the RNA machinery, the mitochondrial translation activation has a phase lag of more than one sixth of the period. In terms of time delay, this amounts to approximately 50 min, see Fig. 7.5. More in detail, most of the mRNAs of ribosomal small and large subunits for both cytoplasmic and mitochondrial localizations are highly correlated within their complex (average Pearson correlation for both is around 0.8) and correlated with the translation complexes at the corresponding location. In particular, among the cytoplasmic translation complexes, the initiation factors eIF and the termination factors eRF are very coordinated and respond very fast, while of the three elongation factors only eEF2 and eEF3 are well-coordinated, whereas the larger complex eEF1 shows a less-defined response pattern, with only the subunit eEF1$-\beta$ clearly expressed. Overall for the class of translation complexes the pattern of activation of the response reflects closely the corresponding HL distributions [123] (eIF and eRF have short HL, eEF has not). Notice that a simple comparison of the HLs of the cytoplasmic and mitochondrial ribosomal and translation machineries (both approximately 24 min) does not show the significant difference which can be seen on the time series profiles and which is instead revealed by the phase delay analysis. For cytoplasmic ribosomal biogenesis, a similar anomaly is encountered also in the stress/stimuli responses analyzed below. For mitochondria, the same type of pattern is verified also by other complexes, for example by both the translocases located in the outer and inner mitochondrial membranes (*TOM* and *TIM*) which are known to mediate the protein import into the mitochondria, see Fig. 7.5.

A neat organization can be seen also in the phase of the nucleotide and amino acid metabolism: while pyrimidine and purine synthesis, as well as e.g. the CTP synthase enzyme involved in pyrimidine biosynthesis, are synchronous with the burst of transcription, the peaks for most of the enzymes involved in amino acid pathways tend to be in phase with the activation of the translational machinery. Also the synthesis of aminoacyl-tRNAs, necessary for the delivery of the amino acids to the ribosomes during translation has a similar phase. As expected, the "synthesis" pathway of an amino acid always anticipates its "degradation" pathway (see Fig. 7.2). In order to start translation, the initiator tRNA carrying methionine is required, and in fact, among the amino acid metabolic pathways, methionine is one of the fastest. As a matter of fact, the pathways of sulfur metabolism and of the sulfur-related amino acids (methionine, cysteine, as well as the closely related selenoamino acid metabolic pathway) present very similar and very compact time series, with an early (synchronous with the main burst) but long lasting activation (duration of the pulse is more than 100 min). This tight coordination may hint at a special role played by the sulfur pathways in the yeast population synchronization [114, 121].

To conclude the protein synthesis, the nascent polypeptide chains must fold into 3D structures. The molecular chaperonin-containing T-complex and the Gim complex, which help in the folding, behave synchronously with the main burst.



(a) (b)

Complex (periodic genes only)	Phase	HL	Width
Nuclear splicing complexes	55.11	13.7	2.22
Cytoplasmic ribosomes	47.5	24.7	2.77
Cytoplasmic translation complexes	64.9	17.8	2.1
Mitochondrial splicing complexes	125.3	22.4	6.67
Mitochondrial ribosomes	128.2	24	8.23
Mitochondrial translation complexes	127	26.7	8.5
TOM + TIM translocases	127.1	19.6	6.57
PUF3p mRNAs	129.7	25.4	7.38

(c)

Figure 7.5: (a): Cytoplasmic vs. mitochondrial splicing, ribosomal (small and large subunits are lumped together) and ribosomal translational complexes. All genes are nuclear-encoded. Black profiles represent mRNAs classified as periodic. Within each of the two cellular compartments, the time courses of gene expression are similar and fairly coordinated. Even the amount of correlation among the complexes subunits is similar, with e.g. ribosomal mRNAs in both compartments being more tightly coordinated than the corresponding translational machineries. The bursts for the cytoplasmic localizations are much sharper, higher and shorter than in the mitochondria. These last accumulate an average phase lag of $\sim 90°$, or around 50 minutes of delay (recall that the phase is computed by autocorrelation with a train of sinusoids, hence the value for the phase represents the "center" of the pulse). The cytoplasmic ribosomal complex substantially overlaps with cluster 9 of Fig. 7.1(a), while the mitochondrial ribosomal complex is contained in cluster 8 of the same Figure. (b): Mitochondrial translocases across outer and inner membranes, and mRNAs having Puf3p as a RBP (220 genes, 134 periodic). Of the 236 mRNAs belonging to at least one of the mitochondrial categories shown in the Figure, 62 have Puf3p as RBP. This tells us that in this case the "localization" constraint is stronger than co-sharing a single RBP, but that the two conditions are coupled and induce a similar pattern of dynamical regulation.

On the contrary, ubiquitin and proteasome, that proceed to the recognition and degradation of anomalous proteins, as well as the SCF and anaphase promoting complexes, that cause the proteolysis of the cyclin-CDK complexes, have patterns of activation which are more delayed and broadened. Actually, this class of proteolytic processes (macrocategory "folding, sorting and degradation" in Fig. 7.3) has the highest values of phase, i.e. it has the slowest response to the transcription bursts.

The macrocategory "DNA replication and repair" (see Fig. 7.3) contains what remains of the "fast" responses to a large extent synchronous (protein complexes: DNA damage checkpoint, DNA repair, pre-replication, replication, replication fork, which includes all DNA polymerases, helicases and ligases, cyclin-CDK) or within a short time delay from the initial bursts of transcription. The peculiarity of this class is that the pulses are more long lived than in the "transcription" and (cytoplasmic) "translation" categories. Also the complexes regulating the cohesion and separation of sister chromatids during the S-phase (nuclear cohesion family of complexes) follow the same pattern (see Fig. 7.4).

Moving to the core of the cell's metabolic activity, the average phase increases further (see Fig. 7.3), but the main qualitative difference is on the shape of the pulses, which are now broader and often with an asymmetric rise/decay profile: still sufficiently fast activation but slower and less abrupt decay. This difference is likely to reflect the longer HL associated to these categories (all have average HL $\geqslant$ 30 min), and implies metabolic functions more overlapping than sequential. Along each metabolic pathway, the degree of correlation among enzymes catalyzing neighboring reactions is higher than it is expected (the "expected value" is inferred from a large collection of yeast microarray experiments, see Fig. 7.6) implying a coherent and coordinated temporal behavior along the metabolic routes.

7.2.1 Central metabolism

From Fig. 7.7, it seems that the long bursts of the citric acid cycle and oxidative phosphorylation genes could be composed of two distinct adjacent phases for each period. Similarly, the profiles of the anticorrelated isoenzyme pairs mentioned in section 7.4, show that of the two recurrent patterns described in Fig. 7.11(c), one resembles the mitochondrial transcription/translation burst (upregulation approximately in the interval 225÷375 min interval and periodically thereafter), the other is more delayed (interval 300÷450 min) and characterized by a deep downregulation during and after the main transcription bursts (200÷275 min).

The alcohol dehydrogenases isoenzymes are "prototypes" of the 2 patterns: *ADH1* and *ADH3* (respectively cytosolic and mitochondrial, both reducing acetaldehyde to ethanol) follow the first, while *ADH2* (using ethanol as substrate) follows the second. The first pattern (*ADH1*/*ADH3*) is synchronous with the hexokinases catalyzing the initial glucose phosphorylation: of the three isoenzymes, *HXK2* has the earliest response but is also more rapidly repressed, while *HXK1* is more long-lived and is expressed, together with the glucokinase *GLK1*, also in the

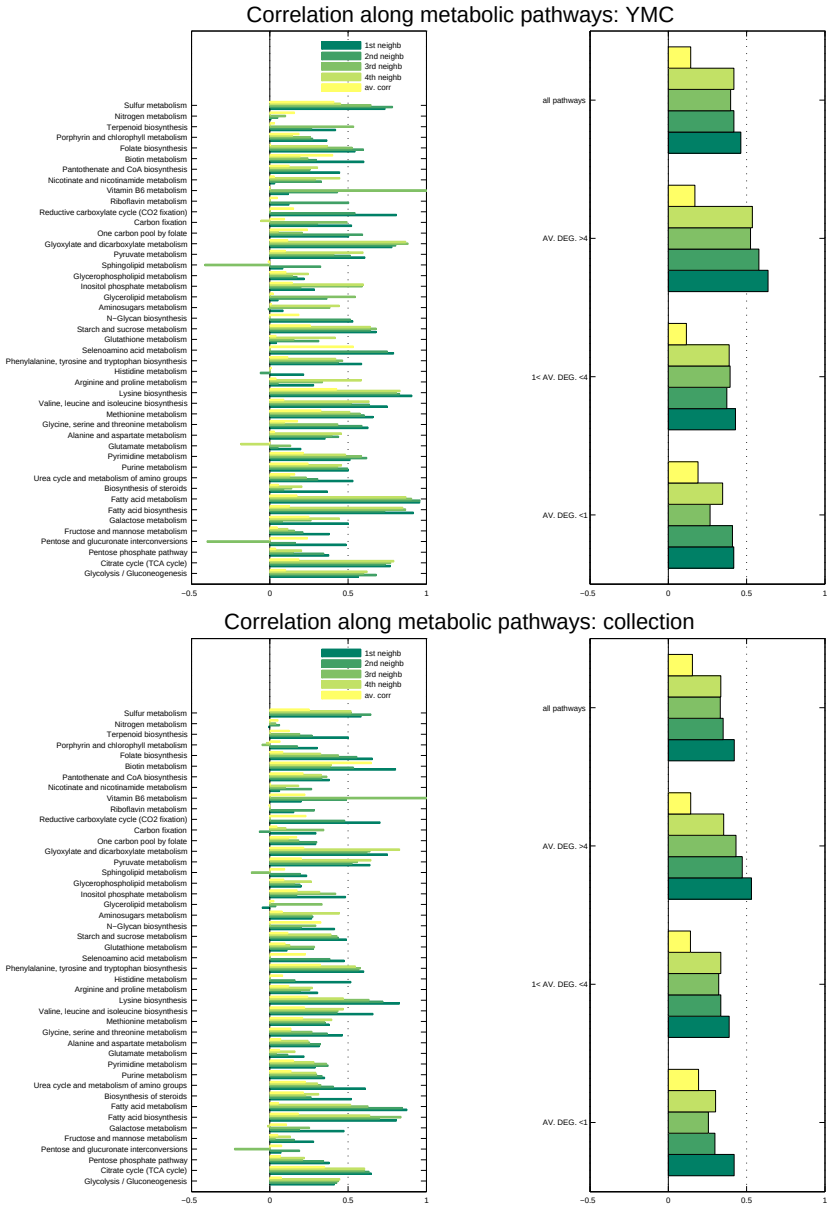


Figure 7.6: Mean correlations between genes along the KEGG metabolic pathways (genetic information categories such as transcription, translation, and DNA replication and repair are not included), computed for the yeast metabolic cycle time series (top plot) and for a collection of 790 yeast experiments (bottom plot). The correlations are computed between enzymes that are neighbors in terms of metabolic reactions: from adjacent genes, to genes separated by three intermediate reactions (green scales). Averages between all genes involved in a pathway is also shown in yellow. On the right panel of each plot are shown the average of those mean correlations along all pathways, grouped by their average enzyme connectivity degree. Correlations are higher for more tightly connected pathways than for those having a low connectivity degree. Comparing the right hand sides of the two plots, correlation among neighboring genes for the YMC is higher than for the collection, thus confirming the high level of functional coordination induced by the YMC along the metabolic pathways.

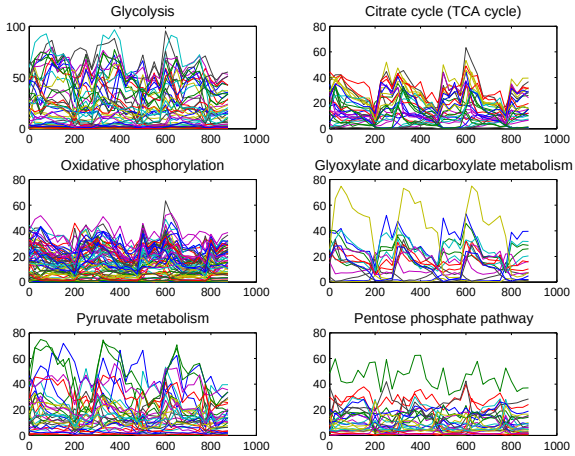
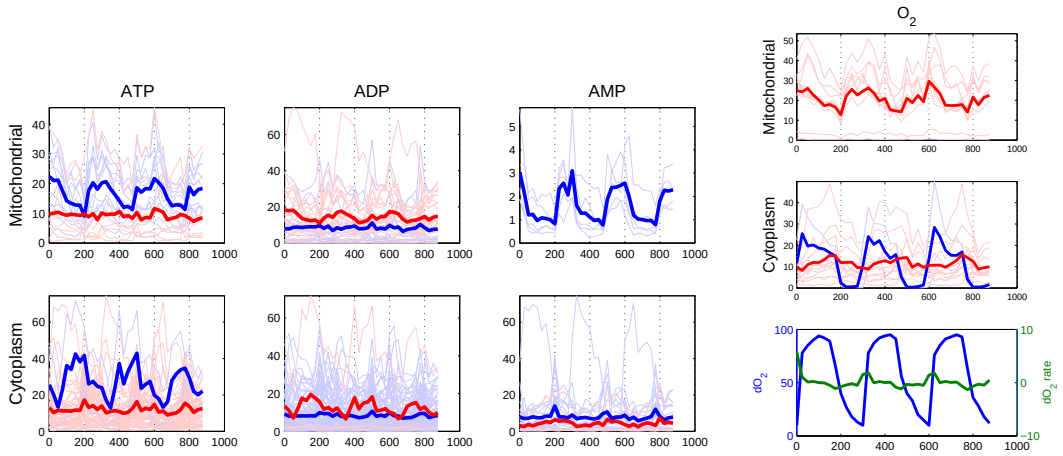


Figure 7.7: Time course of the central metabolic pathways.

other interval [102].

Quite unexpectedly, the enzyme for the final irreversible step of glycolysis, pyruvate kinase (*CDC19*, as the isoenzyme *PYK2* remains constantly basal), is neither synchronous with the *ADH1*/*ADH3* and *HXK2* profile, nor with the other one (*ADH2* and *GLK1*), but rather delayed with respect to both modes (in Fig. 7.4 pyruvate kinase has the highest phase delay). The high level of expression of alcohol dehydrogenase in all metabolic modes suggests that pyruvate production may not be the rate-limiting step of the pathway, and that a delayed pyruvate kinase action may help meeting cellular ATP demand by distributing uniformly ATP production along the cycle (see Fig. 7.8(a)). As a matter of fact, *CDC19* peaks always precede the transcription bursts (in correspondence of the downregulation of the mitochondrial genes) and fall right after that. Most of the enzymes for the intermediate steps of glycolysis do not show any significant periodic trend (see Fig. 7.9 for an overall view of the phase of the genes on the central metabolic pathways), although on the other irreversible reaction, phosphofructokinase (both genes *PFK1* and *PFK2*) has some degree of resemblance with *CDC19*. On the contrary, the gluconeogenesis enzymes pyruvate carboxylase (*PYC1*) and phosphoenolpyruvate carboxykinase (*PCK1*) show a strong correlation with the genes *ADH1*/*ADH3* and *HXK2*.

The acetaldehyde-ethanol exchange is part of the so-called “PDH bypass” (i.e. route alternative to the pyruvate dehydrogenase complex) for acetyl-CoA production, see [101]. The supply to this pathway (through *PDC5*) is almost continuous (except in the “valleys” of the pyruvate kinase) and the rest of the pathway, aldehyde dehydrogenase (mostly isoenzyme *ALD6*, mitochondrial) and acetyl-CoA synthase (*ACS1*, cytosolic) coordinated with *ADH2*. On the contrary, the pyruvate carboxylase branch follows *ADH1*/*ADH3*, while the direct route pyruvate/acetyl-CoA (PDH complex) is unclear (more synchronous with *ADH1*/*ADH3* though).



(a) (b)

Figure 7.8: (a): Time courses of the expression levels for the enzymatic genes catalyzing reactions involving ATP, ADP and AMP. The reactions are subdivided in mitochondrial and cytoplasmic ("cytoplasm" referring to the entire cell with the exclusion of the mitochondria) compartments and according on whether the metabolite is to be considered a substrate (red line) or a product (blue line) of the reaction. Thick curves represent the average over the mRNA expression of the corresponding enzymes. Information abound compartment and reaction direction is extrapolated from [49]. The expression of the enzymatic genes is taken as a measure of the flux of metabolites through the reaction node (scales are however not indicative). The peaks of consumption of ATP in the cytoplasm in correspondence of the main bursts are small but visible. More visible is the pattern of ATP-producing enzymes in both compartments. In the cytoplasm this is essentially due to the pyruvate kinase enzyme Cdc19 transforming phosphoenolpyruvate into pyruvate during anaerobic respiration, while in the mitochondria it is due to the oxidative phosphorylation pathway. The fermentative recharging of ATP in the cytoplasm is quite in antiphase with the respiratory mitochondrial one (scale here can be even misleading: aerobic ATP production is of course far more efficient than anaerobic one). Notice that during the bursts of transcription, ATP hydrolysis rather than peaks of ADP induces peaks in the production of AMP, as is expected for high energy demanding reactions such as RNA polynucleotide synthesis. (b): Time course of expression for enzymes of reactions involving O_2. Color, line thickness and compartment subdivision is the same as above. The third plot is the trace of dissolved O_2 (blue line, data reproduced from [119]), and the dO_2 ratio (green line). Its trend follows closely the cytoplasmic "oxygen production" (blue curve in the middle plot), which essentially is the time course of the catalases, enzymes detoxifying reactive oxygen species such as H_2O_2. Qualitatively the main discrepancy between the two curves occurs in the 50 min interval following the bursts (e.g. 200 ÷ 250 min.) where dO_2 keeps decreasing while the concentration of the Catalases mRNAs remains basically at zero level. From the top plot, the explanation could be that this is the interval in which mitochondrial respiration starts, thereby consuming O_2.

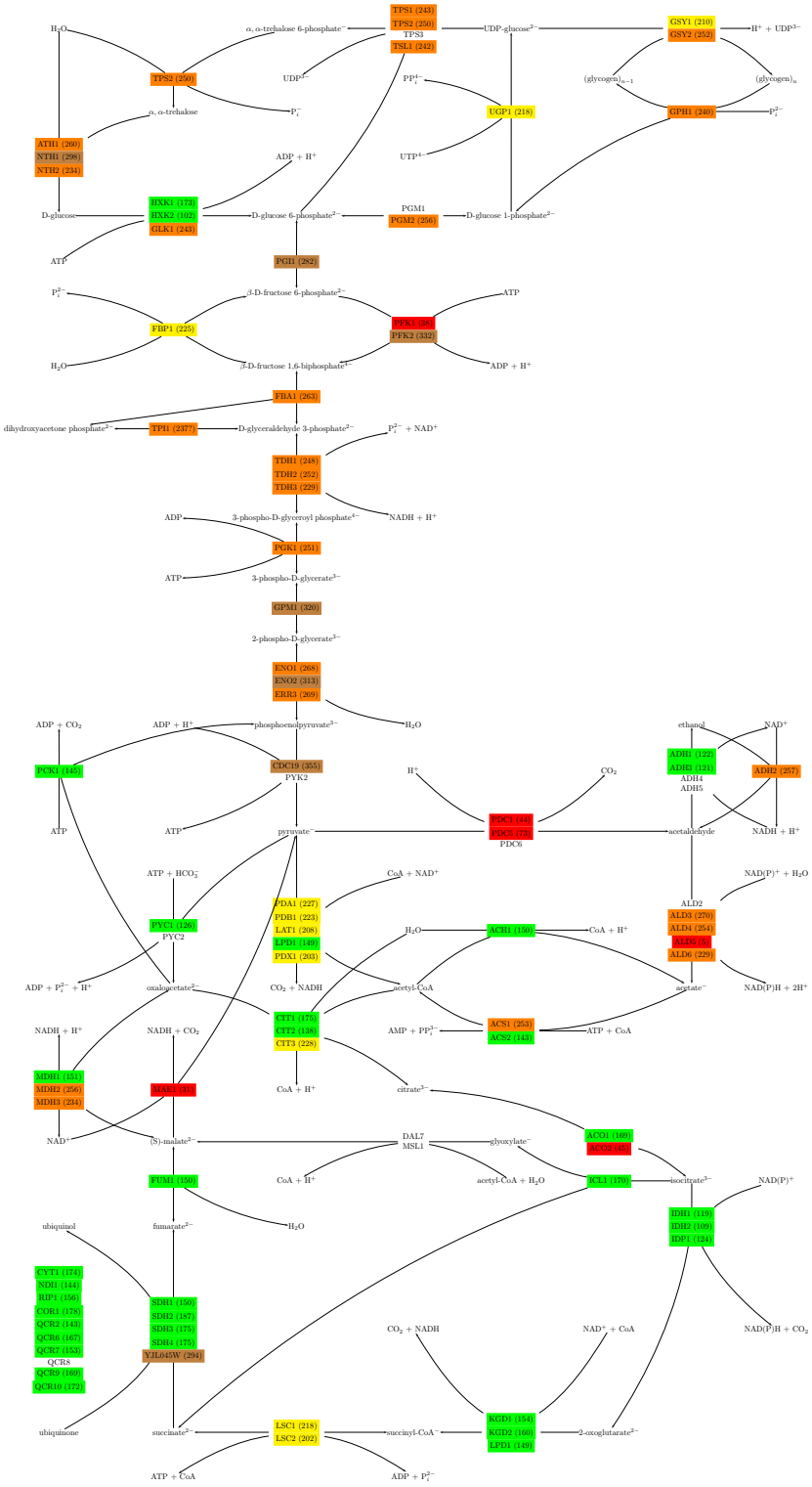


Figure 7.9: Phase of the periodic genes on the central metabolism. The color code (only for periodic genes) indicates the phase interval. Red: $0 \div 100$; green: $100 \div 200$; yellow: $200 \div 250$; orange: $250 \div 300$; brown: $300 \div 360$.

With the exclusion of the succinyl-CoA ligase (both *LSC1* and *LSC2*) all the steps of the citric acid cycle are more in agreement with the *ADH1*/*ADH3* pattern and are rigorously shut down during the transcription bursts. From Fig. 7.7, it seems that the long bursts of the oxidative phosphorylation genes overlap with both patterns. Looking at the trace of observed dO_2 (data taken from [119] and reproduced in Fig. 7.8(b)), citric acid cycle and oxidative phosphorylation activation seem to correspond to the maximum drop in dO_2 concentration ($200 \div 300$ min interval following the transcription burst), but they seem to persist also long after the recovery of dO_2. It must be noticed that the trace of dO_2 resembles closely the expression profile of the catalase enzymes that produce O_2 detoxifying reactive oxygen species.

7.2.2 Glucose-regulated carbon metabolism

There is a consistent literature on the influence of glucose abundance on gene expression [17, 32, 71, 103, 131]. On the YMC, the standard glucose activated and/or repressed signaling pathways are not expressed. For example the Snf1 serine/threonine protein kinase complex subunits *SNF1*, *SNF4*, *SIP1*, *SIP2*, *GAL83*, as well as the other regulated genes on the same pathway *MIG1*, *CAT8* and *ADR1*, do not show any significant pattern.

7.2.3 More compartmentalized categories

Other categories which can be associated with a particular cellular compartment emerge from the joint analysis of pathways and protein complexes. For example lipid biosynthesis, which is to a large extent localizable in the endoplasmic reticulum (ER) has a phase comparable to the translocon family of complexes (Fig. 7.4) which is composed of the Sec61 protein translocator, the signal recognition particle which binds the ER-specific sequence on the nascent polypeptide chain and the signal peptidase that cleaves it off.

7.2.4 Signaling proteins

Complexes with eminently intracellular signaling functions, such as the antagonistic cAMP-dependent protein kinase and serine/threonine phosphoprotein phosphatases (respectively phosphorylator and dephosphorylator of signaling proteins) have similar patterns of expression, similar timing during the YMC and high Pearson correlation (at least for what concerns periodic genes).

7.2.5 Weakly periodic categories

Several categories linked to transcriptional activation or RNA processing, like the histone acetyltransferase enzyme or the nuclear processing complex family (3'-end pre-RNA processing factors CFI and II and 3'-end polyadenylation factors PFI) or the chromatin assembly complex, seem to be evading the tight phase coordination.

However, this is mostly due to the evanescent periodic pattern, if any, of the corresponding genes. Likewise for the nuclear pore complex, which assists the export of mature mRNA through the nuclear envelope: most of its genes in fact show bursts which are synchronous with the initial pulses but of very small amplitude, thus ambiguous in terms of temporal classification.

7.3 Regulation via TFs versus RBPs

In terms of regulatory influence, while the importance of transcription initiation via TFs is widely studied and a large amount of data (computational and experimental) is available about the binding of TFs to target genes, similar post-transcriptional systematic data on the regulation by means of RBPs are still sporadic [62]. Notable examples are mRNAs associable to the nuclear export proteins Mex67 and Yra1 [61], the Puf family of RBPs [56], and the 3' UTR motif collection of [109].

Inspired by [93], we applied these RBP lists as well as the list of TF binding sites from [3, 86] to the YMC time series comparing the average correlation among genes being common targets of a TF or of a RBP. The two distributions are shown in Fig. 7.10. For both TFs and RBPs, only a few motifs emerge as having a significantly high correlation. The number of genes regulated by the same TF varies between 1 and 226 with a mean of 35.2, while the number of genes with a common target mRNA motif varies between 6 and 1138 with a mean of 81.7. If we draw from a null distribution representing random grouping, increasing the number of genes in a group the probability of finding a high mean correlation obviously decreases, so we expect the distribution for the second set to be tighter around 0. In our case, on the contrary, there are 6 groups out of 110 with a mean correlation $\geqslant 0.4$ for the TF target genes (versus an expected value of 1 for random groups of genes with the same cardinalities of these groups) and 7 groups out of 83 for the genes with a target mRNA motif (versus the expected 0 for random groups with same cardinalities). This suggests that post-transcriptional regulation is more significant that transcriptional regulation in the coordination of the metabolic cycle, although the evidence is not conclusive. When checking the groups of periodic genes with high correlation we found the following significant annotations:

- 44 genes out of 56 having Fhl1p as TF and 10 genes out of 12 having Sfp1p as TF are constituents of cytoplasmic ribosomes; notice that instead other cytoplasmic ribosomal TFs such as Rap1p do not correspond to a sufficiently high correlation;
- 22 genes out of 26 having Hap4p as TF code for subunits of respiration chain complexes;
- 62 out of 220 genes whose mRNA is bound by Puf3p are annotated for mitochondrial transcription/translation (56 are part of mitochondrial ribosomes, of which 47 are periodic), see Fig. 7.5.

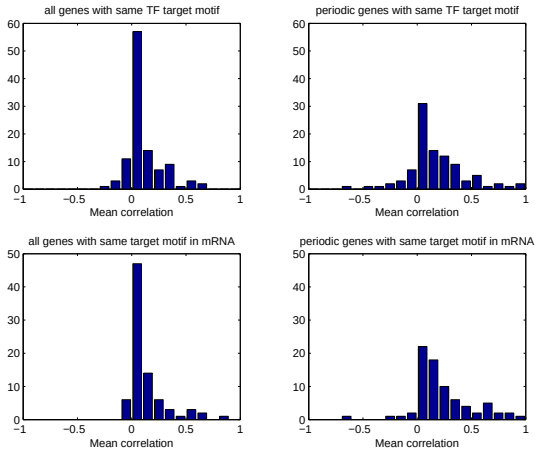


Figure 7.10: Top row: Distribution of the mean correlations for groups of genes having a common DNA motif likely to be the target of a TF [86]. Bottom row: Distribution of the mean correlations for groups of genes having a common mRNA motif likely to be the target of a RNA-binding protein (Yra1, Mex67 [61] or the five Puf proteins [56]) or having a common 3' UTR motif implicated in the stability or in the subcellular localization of the mRNA [109]. The mean correlation of a group of genes is defined as the average of the correlations between the expressions of each gene pair in the group. The mean correlations calculated for all the gene pairs are shown on the left, while on the right only the periodic genes of each group are considered.

7.4 Double peak and anticorrelated isoenzymes

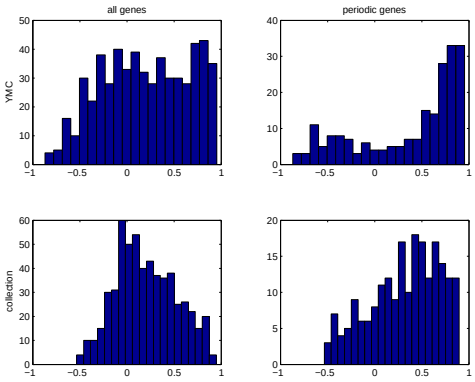
Especially for mitochondrially localized pathways, such as citric acid cycle and oxidative phosphorylation, the pulses are very broad, with a neat downregulation only in correspondence of the bursts of transcription and an overall profile often exhibiting a double peak on each period (occurring with a phase lag of $\sim 100^\circ$ one from the other, see Fig. 7.7). The four respiratory chain complexes for example follow this pattern in a fairly precise manner.

In order to investigate the meaning of this double peak characteristic, we consider genes whose products are classified as isoenzymes. If we look at the correlation for all pairs of isoenzymes, see Fig. 7.11(a), we see that restricting to periodic genes an almost bimodal distribution emerges, with a significant percentage (43 out of 210) of isoenzyme pairs being anticorrelated ($R < -0.3$). This behavior has no counterpart on the distribution of expected values (computed as above by means of a large collection of microarrays). In more than 50% of these anticorrelated pairs the pattern of activation in the time course is similar (see Fig. 7.11(c)), with one of the two isoenzymes exhibiting a deep and prolonged downregulation immediately following the transcription bursts. The majority of these pairs is involved in oxidoreductive processes, like, for example, *SDH1*-YJL045W, *SDH3*-YMR118C, *SDH4*-YLR164W (all subunits of succinate dehydrogenase), or the NADP-dependent isocitrate dehydrogenase pairs *IDH1-IDP3*, *IDH2-IDP3*, *IDP1-IDP3*, or the plasma membrane H+-ATPase isoenzymes *PMA1-PMA2*, or the NADH dehydrogenase pairs *NDE1-NDE2*. Three among the most anticorrelated pairs of isoenzymes showing this pattern are located along the pentose phosphate pathway, two on the cytosolic oxidative branch (*SOL3-SOL4* and *GND1-GND2*), the third (the transketolases *TKL1-TKL2*) downstream. The pentose phosphate pathway synthesizes NADPH, which is the major source of reducing equivalents and, according to [121, 127], plays a major role in the establishment of the cycle. Also the most anticorrelated isoenzymes in the glycolysis pathway, the alcohol dehydrogenases, have a similar pattern: *ADH1* and *ADH3* (reducing acetaldehyde to ethanol) versus *ADH2* (catalyzing the reverse reaction), see section 7.2.1 for a more detailed analysis of the periodic pattern in the central metabolism.

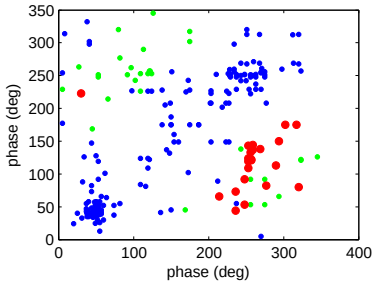
7.5 A minimal input-output dynamical model for the unfolding cycle

Possible origins of the sustained oscillations are discussed at length in the literature [20, 76, 83, 95, 100, 126, 127, 128]. Also Tu et al. explain the cycle and its time compartmentalization in terms of metabolism and redox balance [119, 120, 121].

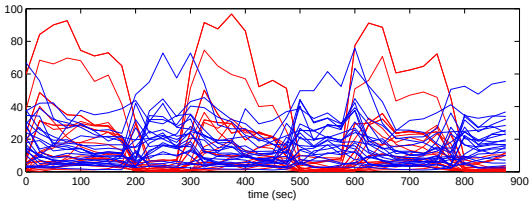
Rather than adding to the list of mechanisms for metabolic regulation, by viewing each cycle as the dynamical response to a burst of transcriptional activation, this work aims at providing a characterization of the dynamics of the unfolding of the cycle, i.e. of how these “impulse responses” are progressively delayed and



(a) Correlation between isoenzymes



(b) Scatter plot of the phase of the periodic isoenzymatic pairs (the pairs in blue have $R > -0.3$)



(c) Most anticorrelated isoenzymes

Figure 7.11: Anticorrelated isoenzymes for the YMC (top row in (a)) and for a collection of 790 yeast gene profiling experiments (bottom row (in (a)). The correlations between all pairs of isoenzymes in the two sets are shown on the left, while on the right only periodic pairs of genes are considered. For them, in the YMC the distribution of correlations tend to a bimodal distribution, i.e. a significant subset of isoenzymes is anticorrelated and oscillates with opposite phases. The same type of anticorrelation is not visible on the reference collection. The time series of the pairs in red in the scatter plot of the phases (b) are depicted in (c). One of the two genes of these pairs (in red) is characterized by a deep valley following the transcription bursts. Most of these pairs are involved in redox processes.

broadened with respect to the input pulses, and of how this correlates with the stability of the corresponding transcripts. The compactness in terms of phase and width of the early categories over repeated oscillatory cycles is an argument in favor of the existence of a single triggering event for each cycle, corresponding to the transcriptional activation bursts mentioned above. In fact, sharp, equispaced pulses are maintained in spite of the broader and less coordinated profiles of the events immediately preceding them. This hypothesis is not in contradiction with the observations about the metabolic origin of the YMC, neither with the observed alterations of the period following a genetic disruption [20, 120, 121] (which could in principle preserve the sequence of events described). On the contrary, it merges the metabolic control level described in [119] with an extra regulatory element which is known to play a role in dynamical contexts.

In fact, the mRNA stability reflects known properties of the corresponding gene products: while mRNAs encoding transcriptional machinery or regulatory components tend to be short-lived and to turn over more quickly, transcripts encoding core enzymatic proteins are typically more stable [91, 123, 125]. For what is known, protein synthesis tends to follow the concentration of the corresponding mRNA [99] and to be at least as stable, if not longer-lived [10, 60]. Hence, it is expected that the concentration of the gene products follows profiles that are similar to those of the mRNAs. The observation that the dynamics through a metabolic pathway can be considered as a timed and sequential process at the level of gene expression appears in several papers in the literature, see [19, 134]. The same principle seems to be reflected in the YMC, although it is not observable at the level of detail investigated, e.g. in [19], but more macroscopically and at genome-wide level.

In terms of dynamical models, the progressive broadening and smoothing of the response to a sequence of (transcriptional) pulses can be described by means of simple linear input-output models (i.e. transfer functions in the Laplace domain) of increasing order having “low-pass” characteristics. As the time constant of this low-pass filter is essentially given by the HL of the mRNA, this type of model naturally predicts the correlation HL–pulse width. In order to describe correctly also the phase along the cycle, a time delay is added to the response, see Methods for a thorough description and Fig. 7.12(a,b) for an example. If the order of such a fitted minimal dynamical model is used to sort the annotated macrocategories of Fig. 7.3, we still recover both the same expected cascade of events and the same direct proportionality with HL, see Fig. 7.12, meaning that even in terms of the simplest possible dynamical model the kernels providing the best fitting become increasingly complex as we progress through the cycle. This is of course expected as the mRNAs gradually pass from fast turnover to high stability.

7.6 A common dynamical gene expression program

As the YMC is obtained only in particular conditions (long-term continuous cultures in chemostats), an intriguing question is whether this highly organized un-

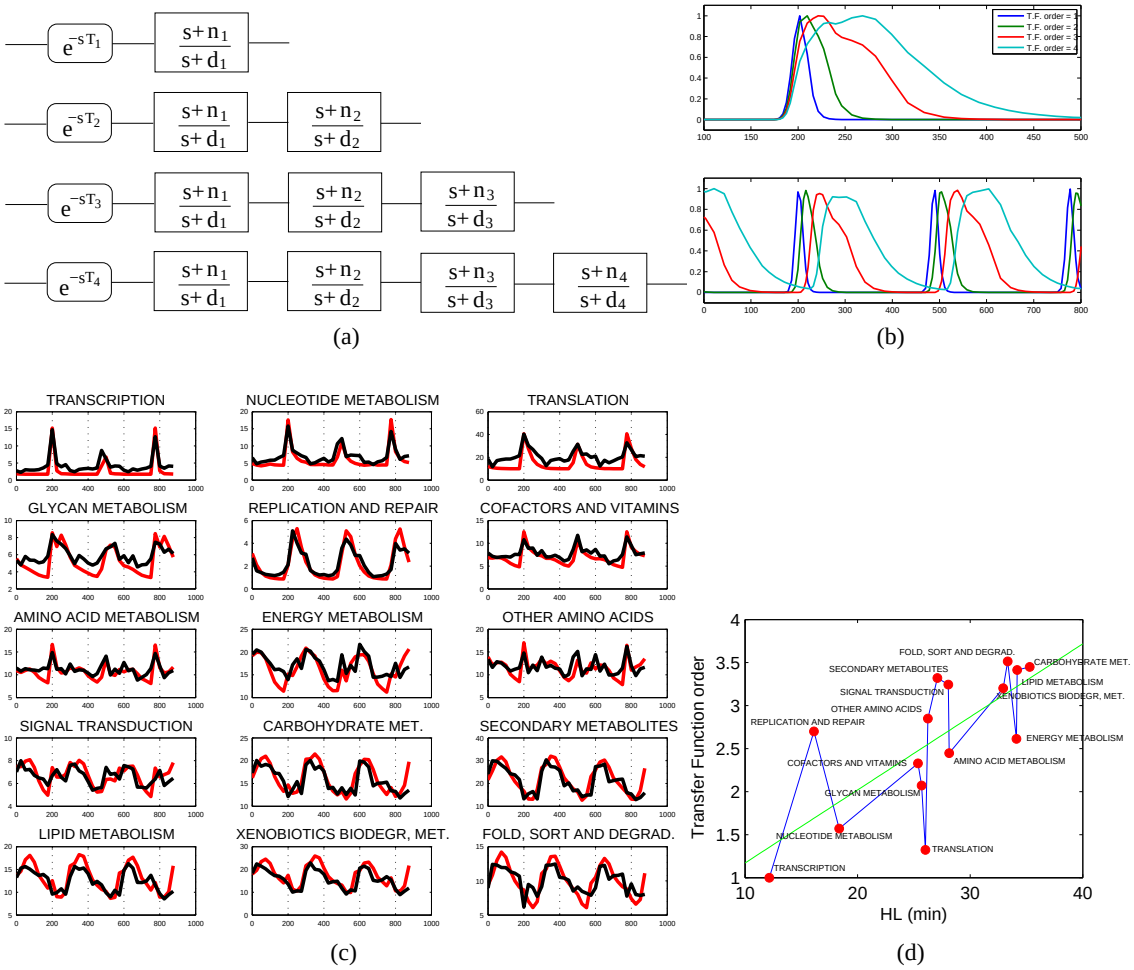
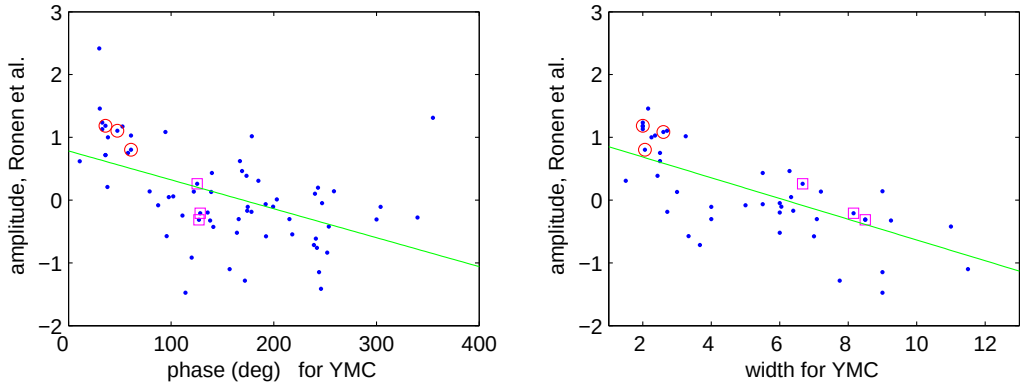


Figure 7.12: Dynamically, the response of the system to the sharp pulses of transcriptional activation can be modeled in terms of input-output transfer functions (i.e. convolution integrals in the Laplace domain, see Methods for details). The main feature of a simple zero-pole transfer function with low-pass characteristic is that in correspondence of an impulse-like input it yields an output which is a smoothed and broadened version of the input. Concatenations of such zero-pole transfer functions describe accurately the progressive broadening and delaying of the YMC gene expression time series. Typical time profiles obtained for transfer functions of order 1 to 4 sketched in (a) are shown in (b). The top plot in (b) shows the larger kernels obtained by concatenating up to 4 first order transfer function blocks. The lower plot in (b) shows how consecutive impulse responses look like for the various orders of transfer functions and an extra delay element as in eq. (6.4). A simple fitting of the n_i, d_i and T_i parameters and of the best model order for each gene allows to accurately reconstruct the average profiles for the 15 macrocategories of Fig. 7.3 (in (c) the model-based time courses are shown in red). With the usual exception of the category "translation", the best transfer function order is roughly proportional to the corresponding HL values, coherently with the other variables discussed in the text.

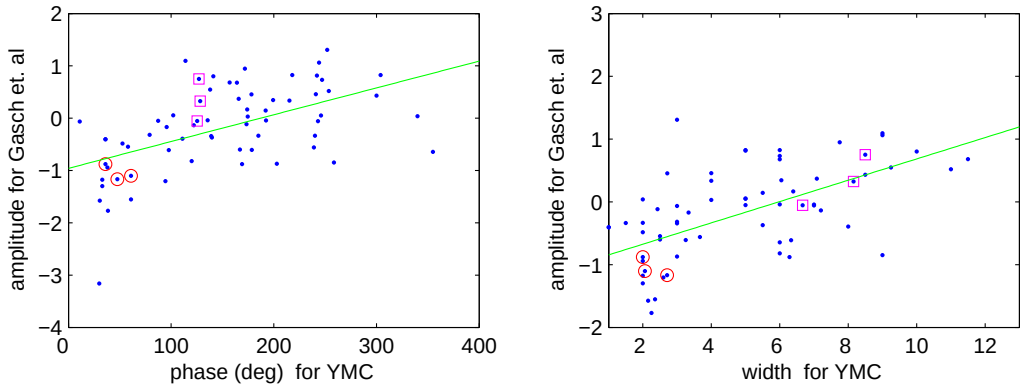
folding of the dynamical response to pulses of transcriptional activation is peculiar only of the YMC or can be observed also in other experimental conditions. For this purpose, we consider the gene expression response of steady-state yeast to pulses of glucose described in [104]. In this case, the yeast shows a transient dynamical response but no oscillatory behavior. Furthermore, the transient peaks are more or less synchronous for all genes, i.e. there is no time-ordering in the dynamics, unlike in the YMC. However, if for a gene we compare the maximal signed amplitude of each expression profile on these time series with the corresponding phase and pulse width in the YMC, a sizable anticorrelation emerges, see Fig. 7.13(a).

If, on the contrary, we consider the stress responses time series of [55], the YMC phase/pulse width turn out to be positively correlated (rather than anticorrelated) with amplitude, i.e. categories appearing early in the YMC tend to be downregulated in most stress responses, while "late phase" categories tend to be upregulated, see Fig. 7.13(b). It is known that in the stress responses genes annotated for ribosomal proteins and/or RNA metabolism are in general downregulated, while e.g. respiratory genes (such as those of the citric acid cycle and of the oxidative phosphorylation) become upregulated [55]. On Fig. 7.13, notice that also in all these responses cytoplasmic ribosomes (cluster 9 in Fig. 7.1) are aligned with the rest of the (cytoplasmic) transcriptional/translational machinery rather than with the assigned HL values.

The conclusion of this analysis is therefore that in intrinsically dynamical contexts some form of common response might indeed be taking place, although exerted by different means. Such genome-wide coordinated response shows a graded ordering which reflects the degree of stability of the genes involved.



(a)



(b)

Figure 7.13: The short-term responses of steady-state yeast to pulses of nutrient discussed in [104] and the stress responses of [55] show a transient peak of up/down regulation. The peaking times are substantially uniform on the genes. For each gene we compute the maximal signed amplitude at the peak and lump together genes belonging to each of the known protein complexes (see Fig. 7.4 in Additional file 1). If for [104] we compare this amplitude with the phase (left) and the pulse width (right) of the corresponding genes for the YMC, we can observe that both scatter plots have a consistent anticorrelation: complexes upregulated in the glucose stimulations of [104] correspond roughly to "early" complexes in the YMC and also to genes with a fast turnover. At the other end, complexes downregulated in [104] are late in the YMC and are more stable, see (a). This shows how, in spite of different growth and stimulation conditions, the gene expression program is substantially faithful. On the contrary stressful stimuli such as those described in [55] yield correlated pattern with phase/width of the YMC (b). Just like for the YMC, for both types of responses cytoplasmic translation behaves differently from the mitochondrial one. In red circles the first 3 complexes of Fig. 7.5(c) are highlighted, in magenta squares their mitochondrial counterparts. Hence the anomaly represented by cluster 9 of Fig. 7.1 with respect to the HL classification is confirmed by these other dynamical responses.

Chapter 8

Conclusion

In [119, 120] the time compartmentalization of the cycle is interpreted in terms of the need to accumulate sufficient products from the metabolic reactions in order to move on to the next phase of the cycle and to autoinduce further cycles of oscillations. This picture is not contradicted by our observations.

It is also affirmed in [119] that broad profiles (like those associated here to "late" categories) may be due to loss of synchronization in the population of yeast cells as they progress through the cycle(s). Based on what we showed in this work, such an interpretation is problematic: loss of synchronization during a cycle would jeopardize the entire transcriptional program on the following cycles, while on the contrary, we still see thin and precisely coordinated pulses in the fast categories.

If, as we do in this work, rather than looking at the YMC merely as cyclic oscillations, we study it as a highly organized dynamical response to pulses of transcriptional activation, then this response can be analyzed in much more detail at genome-wide level and we can observe how an important role in the coordination seems to be played by the mRNA turnover rate. The self-sustained character of what we consider the most upstream event of the cycle, the transcriptional activation burst, can still be conditioned to the accumulation of the required metabolites, while the unfolding of the cycle, which from the analysis of [119] is already known to be functional to the distribution of e.g. the redox load of the cells, is enriched of an extra, intrinsically dynamical feature. This feature is a fine-graded detail of our notion that genes with a fast turnover are typically regulatory, while slow genes are enzymatic and metabolic [91, 125]. It can be used to describe the sequence of events occurring in the YMC as a "natural" gene expression program.

Extrapolating from the specific YMC context, the ordered pattern of events described for the YMC is to a good extent similar to that found on other intrinsically dynamical contexts such as the stress/stimuli responses. Whether the mRNA stability is the cause of this coherent behavior or is simply another effect of a more profound regulatory mechanism is a question to which we cannot provide a definitive answer at the moment.

Part III

Chemical reaction network theory and its applications

Chapter 9

Introduction

Chemical reaction network (CRN) theory has been developed since the early 1970s to study the dynamical evolution of the concentrations of the chemical species involved in a known set of reactions [43]. Under the assumptions of well mixedness and constant temperature, a system of ordinary differential equations for the species concentrations can be derived from the chemical reactions.

Important aspects in the study of this type of dynamical systems are the existence of equilibria, their number and stability. In particular, the capacity of a reaction system to exhibit more than one equilibrium is not only of interest for chemistry or chemical engineering, but has become a major topic also for biologists, in particular in the study of the switch-like behavior observed during intracellular signaling [69, 90, 97] or cell differentiation [79, 129] processes.

The main difficulty which often arises in the verification of the multistability property is the poor knowledge of the rate functions of the reactions, from the type of kinetics (like mass action, Michaelis–Menten or Hill) to the value of the parameters (called rate constants) involved in each reaction kinetics. The power of CRN theory lies in the fact that its results are based on the network structure alone, and so are independent of the values of the constants and in some cases also of the type of kinetics.

In fact, Martin Feinberg and colleagues found a number of conditions [37, 41, 43, 44], centered on the concept of network deficiency, under which the dynamical system for a CRN with mass action kinetics does not admit multiple equilibria, regardless of the rate constants. Other different conditions for monostationarity, verifiable on a graph representing the CRN, were then proved [23, 25, 108]. These latest conditions, based on system injectivity, were later extended [4, 5] also to other kinetics, provided that the system is non-autocatalytic.

An interesting application of these theoretical results to biology is model discrimination [21, 22, 38]: if a biological process is known to be multistable, and for it there are multiple candidate reaction models, it is possible to eliminate some of them by proving that they are always monostationary.

For small CRNs the various conditions can be verified manually, while for

larger networks the only so far available software tool was the *Chemical Reaction Network Toolbox*, version 1.1a [38], a closed source DOS program which implements only the criteria based on deficiency theory. We decided to implement a new, open and complete toolbox for all the previously mentioned approaches, in a modern environment with integrated support for system of differential equations and linear programming like MATLAB.

Another interesting property that can be studied on chemical reaction networks is monotonicity [116]. Monotone dynamical systems have very useful characteristics [112], like the tendency of their solution to converge to an equilibrium (a bounded trajectory generically converges to an equilibrium) and the lack of "chaotic" behavior.

The link with biological systems arise from the observation that, although nonlinear and complex, these systems typically show highly predictable and ordered dynamical behavior, and a tendency to remain at equilibrium or to robustly return to it when perturbed. It has been suggested that biological systems might have evolved so as to be, if not monotone, at least near monotone [85, 116].

Monotonicity in dynamical systems is a well-studied property [64, 111], and can be stated in several alternative ways. For biological networks, a very useful way to formulate/verify it is in terms of the sign of all possible feedback loops among the variables of the system. Given an arbitrary system of ordinary differential equations (ODEs), consider the undirected multigraph having an edge between two variables labelled with the sign of the corresponding entry in the Jacobian matrix. Looking only at the signs of the Jacobian gives a basic indication of the effect (activatory/inhibitory) of a variable on another variable. In most biological contexts, this information is the best one can hope to obtain, as too little is known of the functional form of the ODE and of its dependence on concentrations, parameters, external conditions, hidden (non-modeled) variables, etc.

For this undirected multigraph, the monotonicity property corresponds to all cycles having positive sign, where the sign of a cycle is computed as the product of the signs of the edges forming the cycle. Undirected cycles may correspond to "true" oriented feedback loop or to e.g. distinct paths connecting pairs of vertices [116].

It is argued in [116] that biological networks are "near monotone" in the sense that a relatively small number of sign changes in edges is enough to make the graph monotone. Closely related to this idea is the intuition that biological networks may have many more positive cycles than negative ones, which is the approach taken in [85]. While the simple verification of whether a network is monotone or not is feasible in polynomial-time, the problem of testing how distant a given network is from monotonicity (i.e. estimating the "consistency deficit" in the terminology of [116]) is an NP-hard one [28]. As the size of a network becomes of the order of the thousands of nodes, like for example in any gene regulatory network, testing exhaustively the sign of all cycles quickly becomes an untreatable problem, because the number of cycles grows exponentially. In fact, in [85] only short cycles were tested for large networks, while in [28] approximation algorithms based on

semidefinite programming ideas were introduced.

Our purpose is to tackle this problem from a different perspective, using tools from graph theory, namely the notion of fundamental cycles. The concept of a fundamental cycle was introduced by Kirchhoff [72]. What Kirchhoff showed is that no matter how many cycles an undirected graph contains, considering only fundamental cycles with respect to a spanning tree is enough as the rest of the cycles are obtained as linear combinations of some fundamental cycles. In terms of linear algebra, fundamental cycles form a basis of a vector space whose elements are cycles and disjoint unions of cycles.

In this work we show that fundamental cycles of positive sign form a subspace which is invariant to the positivity property: any cycle of this subspace must have a positive sign, and the cycle subspaces obtained in this way correspond to monotone subsystems. The number of negative fundamental cycles corresponds to the number of sign changes that are required to render the network monotone. In fact, each fundamental cycle is uniquely associated to a chord not shared with any other fundamental cycle. By changing sign to the chords of all negative fundamental cycles we obtain a monotone graph.

As an easy byproduct, we get an upper bound on the number of inconsistencies of a network: *any* network can be rendered monotone by at most a number of sign changes equal to the cardinality of a basis of fundamental cycles. Unrelated (and usually sharper) upper bounds can also be obtained from the theory of signed graphs [115]. These bounds are quite helpful in defining a proper metric to test whether a given network can be classified as “near-monotone”.

All bases of fundamental cycles have the same cardinality, however the number of positive/negative fundamental cycles in a given basis depends on the choice of spanning tree (and can vary widely with it). Needless to say, testing all spanning trees requires computational time that grows exponentially with the size of the graph. In order to simplify the choice of a “good” spanning tree (with fewest possible negative fundamental cycles, hence as near as possible to monotonicity), we show that is possible to maximize the overall number of positive edges on the graph while maintaining unaltered the sign of each cycle using cut sets. The *rationale* of the method is that changes of sign through a cut set leave the consistency deficit invariant. In the theory of monotone systems [111], this operation corresponds to changing sign to the order relationship in one or more orthants; in the theory of signed graphs [135] this corresponds to changing the representative element in a “switching class of equivalence” where the consistency deficit is an invariant of the equivalence relation.

Chapter 10

Background material

10.1 Multigraphs

A basic reference for this Section is [34].

A *directed multigraph* is an ordered pair $G = (V, E)$ where V is a finite set of *vertices* and E is a finite set of ordered triples (a, b, l), called *edges*, where $a, b \in V$, and l is a string referred as the *label* of the edge. a and b are called the *endpoints* of (a, b, l).

An *undirected multigraph* $G = (V, E)$ is a directed multigraph where for every edge $(a, b, l) \in E$, also $(b, a, l) \in E$. For this type of graphs we consider (a, b, l) and (b, a, l) to be the same edge.

A *self-loop* is an edge (a, a, l) with equal endpoints. A multigraph without self-loops is called *simple*.

A *graph* is a multigraph where for every ordered pair (a, b) of vertices there is at most one edge $(a, b, l) \in E$.

A multigraph (resp. graph) $G_1 = (V_1, E_1)$ is a *submultigraph* (resp. *subgraph*) of a multigraph $G = (V, E)$ if $V_1 \subseteq V$ and $E_1 \subseteq E$.

A *walk* in a multigraph $G = (V, E)$ is an alternating sequence $\langle v_0, l_0, v_1, l_1, \ldots, l_{k-1}, v_k \rangle$ of vertices and labels, beginning and ending with vertices, such that $(v_i, v_{i+1}, l_i) \in E$ for any $i = 0, 1, \ldots, k-1$. The walk *contains* or *traverses* these edges. The *length* of a walk is the number of edges it traverses, counting multiple edges multiple times. To a walk in G we associate the submultigraph of G with same set of vertices and as edges the ones contained in the path.

A walk in which all vertices are distinct is called a (simple) *path*. Two vertices a and b are said to be *connected* if there is a path in G from a to b. A (simple) *cycle* in a multigraph is a walk that starts and ends at the same vertex and includes other vertices and edges at most once. A multigraph is *acyclic* if there are no cycles in it.

A multigraph is *signed* if each edge label is $+1$ or -1. A walk in a signed multigraph is positive (resp. negative) if it has an even (resp. odd) number of negative edge labels.

An undirected (resp. directed) multigraph G is called *connected* (resp. *strongly*

connected) if any two of its vertices are connected by a path in G.

In an undirected (resp. directed) multigraph G, a *connected component* (resp. *strongly connected component*) of G is a maximal connected submultigraph of G.

A *forest* is an acyclic undirected graph. A *tree* is a connected forest. Every tree $T = (V, E)$ has exactly $|V| - 1$ edges.

A *spanning forest* of an undirected multigraph $G = (V, E)$ is an acyclic subgraph T of G with the same set of vertices and as edges a maximal subset of E preserving acyclicity. The number of edges of every spanning forest of G is equal to $|V|$ minus the number of connected components of G.

With respect to a given spanning forest $T = (V, E_T)$, an edge of the multigraph that is not in E_T is called a *chord*. Adding a chord to T creates precisely one cycle, and we say that the chord *generates* the cycle. Obviously, each chord generates a different cycle.

A *spanning tree* is a connected spanning forest. Obviously an undirected multigraph has a spanning tree if and only if it is connected. A spanning forest of an undirected multigraph is composed by a spanning tree for each connected component of the multigraph.

A *cut set* for a multigraph G is a set of edges whose removal from G increases the number of connected components.

Theorem 1. *Every cycle in a multigraph has an even number of edges in common with any cut set.*

10.2 Cycle spaces

The *power set* of a set S, written $\mathcal{P}(S)$, is the set of all subsets of S. The *symmetric difference* is a binary operation on a power set $\mathcal{P}(S)$ defined as $A \ominus B = (A \setminus B) \cup (B \setminus A)$ for every $A, B \subseteq S$.

Proposition 1. *Symmetric difference on a power set is associative and commutative. Moreover, the empty set is an identity with respect to it ($A \ominus \emptyset = A$) and every set is its own inverse ($A \ominus A = \emptyset$).*

Consider the set of all submultigraphs of a multigraph $G = (V, E)$ which have the same vertices of G, or equivalently the power set $\mathcal{P}(E)$ of the edges of G. The symmetric difference of two submultigraphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ of G is the submultigraph $G_1 \ominus G_2 = (V, E_1 \ominus E_2)$.

A union of cycles of a multigraph G is a submultigraph of G.

Proposition 2. *The set of all unions of disjoint cycles of an undirected multigraph is closed under the symmetric difference.*

Theorem 2. *The set of all unions of disjoint cycles of an undirected multigraph G is a vector space over the Galois field GF(2), using the symmetric difference operation as vector addition and $\cdot$ as scalar multiplication. It is called the* cycle space *of G.*

Consider an undirected multigraph $G = (V, E)$ with k connected components and a spanning forest $T = (V, E_T)$. The set of *fundamental cycles* of G with respect to T is the set of cycles generated by all the chords. Since each chord generates exactly one fundamental cycle, the number of fundamental cycles of G is equal to the number of chords, i.e. $|E| - |E_T| = |E| - |V| + k$, which is independent from the choice of the spanning forest and is called the *nullity* of G.

Theorem 3. *The set of fundamental cycles of an undirected multigraph G with respect to a spanning forest is a basis of the cycle space of G (called a* strictly fundamental cycle basis *of G).*

Corollary 1. *The dimension of the cycle space of an undirected multigraph is equal to its nullity.*

With respect to a stricly fundamental cycle basis F_i, $i = 1, \ldots, n$, every cycle C can be written as the symmetric difference $\ominus_{i \in I} F_i$ of a set of fundamental cycles. Moreover C contains all and only the chords which generates the fundamental cycles F_i, $i \in I$.

Fact 1. *For some undirected multigraph G, there are bases of the cycle space that do not form sets of fundamental cycles with respect to any spanning tree of G. An example is the sunflower graph SF(3) [82].*

10.3 Injectivity

Let f be a function whose domain is a set D. The function f is injective if for all a and b in D, if $a \neq b$, then $f(a) \neq f(b)$.

10.3.1 P-matrices

Let M be a matrix $m \times n$ and let $I \subseteq \{1, \ldots, m\}, J \subseteq \{1, \ldots, n\}$ be a row and column index sets respectively. A *submatrix* $M[I, J]$ of M is the matrix obtained by selecting the rows in I and the columns in J from M.

A *minor* of M is the determinant of a square submatrix of M. If $I \subseteq \{1, \ldots, min(m, n)\}$, then $M[I, I]$ is called a *principal submatrix* of M and its determinant is called a *principal minor* of M.

A square real matrix is a *P-matrix* if all of its principal minors are positive. A square real matrix M is a *$P^{(-)}$ matrix* if $-M$ is a P-matrix, i.e. if all of its principal minors of size $k \times k$ have sign $(-1)^k$. P- and $P^{(-)}$-matrices are obviously nonsingular.

A square real matrix is a *P_0-matrix* if all of its principal minors are nonnegative. A square real matrix M is a *$P_0^{(-)}$ matrix* if $-M$ is a P_0-matrix.

The following lemma is needed for the proof of the next theorem, we provide our own proof of the lemma since we were not able to find it in the literature.

Lemma 1. *Let M be a $n \times n$ square matrix, and D a $n \times n$ diagonal matrix. Then*

$$\det(M + D) = \sum_{L \subseteq \{1,\dots,n\}} \left(\prod_{l \in L} D_{l,l} \right) \det(M[L^c, L^c]) \tag{10.1}$$

where the complement is with respect to the set $\{1, \dots, n\}$.

Proof. Define the series of $n \times n$ diagonal matrices $D^{(k)}$ such that $D^{(k)}_{i,i} = D_{i,i}$ if $i \leq k$ and 0 otherwise. We want to prove by induction over k that

$$\det(M + D^{(k)}) = \sum_{L \subseteq \{1,\dots,k\}} \left(\prod_{l \in L} D_{l,l} \right) \det(M[L^c, L^c]). \tag{10.2}$$

For $k = 0$, $D^{(k)}$ is an empty matrix and the summation is simply $\det(M)$. Now, suppose the previous formula is true for $k - 1$. For $k > 0$, using the Laplace expansion along column k, we have

$$\det(M + D^{(k)}) = \sum_{i=1}^{n} (-1)^{i+k} (M + D^{(k)})_{i,k} \det((M + D^{(k)})[\{i\}^c, \{k\}^c])$$

Since $D^{(k)}[\{i\}^c, \{k\}^c] = D^{(k-1)}[\{i\}^c, \{k\}^c]$ and

$$(M + D^{(k)})_{i,k} = \begin{cases} M_{i,k} = (M + D^{(k-1)})_{i,k} & \text{if } i \neq k \\ M_{k,k} + D_{k,k} = (M + D^{(k-1)})_{i,k} + D_{k,k} & \text{if } i = k \end{cases}$$

the expansion becomes

$$\left(\sum_{i=1}^{n} (-1)^{i+k} (M + D^{(k-1)})_{i,k} \det((M + D^{(k-1)})[\{i\}^c, \{k\}^c]) \right) + \\ + (-1)^{2k} D_{k,k} \det((M + D^{(k-1)})[\{k\}^c, \{k\}^c]) =$$

$$\det(M + D^{(k-1)}) + D_{k,k} \det(M[\{k\}^c, \{k\}^c] + D^{(k-1)}[\{k\}^c, \{k\}^c]).$$

It is possible to apply the inductive hypothesis to both these determinants, which gives:

$$\left(\sum_{L \subseteq \{1,\dots,k-1\}} \left(\prod_{l \in L} D_{l,l} \right) \det(M[L^c, L^c]) \right) + \\ + D_{k,k} \sum_{L \subseteq \{1,\dots,k-1\}} \left(\prod_{l \in L} D[\{k\}^c, \{k\}^c]_{l,l} \right) \det(M[\{k\}^c, \{k\}^c][L^c, L^c])$$

The second term in this sum can be rewritten as

$$\sum_{L \subseteq \{1,\dots,k-1\}} \left(\prod_{l \in L \cup \{k\}} D_{l,l} \right) \det(M[(L \cup \{k\})^c, (L \cup \{k\})^c])$$

and eq. 10.2 easily follows. For $k = n$ this proves eq. 10.1. □

Theorem 4 ([48])**.** *Let M be a square matrix. M is a P-matrix if and only if for every $x \neq 0$ there exists an index i such that $x_i(Mx)_i > 0$.*

Proof. Let M be a P-matrix and suppose by contradiction that there exists $x \neq 0$ such that $x_i(Mx)_i \leq 0$ for all i. Now, let $I = \{i \mid x_i \neq 0\}$, $\hat{M} = M[I, I]$ and $\hat{x} = x[I]$, and consider the $|I| \times |I|$ diagonal matrix D for which $D_{i,i} = -\frac{(\hat{M}\hat{x})_i}{\hat{x}_i} \geq 0$ for all $i = 1, \ldots, |I|$. It is easy to see that $(\hat{M} + D)\hat{x} = 0$, so, given that $\hat{x} \neq 0$, $\hat{M} + D$ is singular and $\det(\hat{M} + D) = 0$.

Using lemma 1, we have

$$\det(\hat{M} + D) = \det(\hat{M}) + \sum_{L \subset \{1,\ldots,|I|\}} \left(\prod_{i \in L} D_{i,i} \right) \det(\hat{M}[L^c, L^c]) \geq \det(\hat{M})$$

because $D_{i,i} \geq 0$ for all i, and the minor $\det(\hat{M}[L^c, L^c])$ is positive since $\hat{M}$ is also a P-matrix. Therefore $\det(\hat{M}) \leq 0$, a nonpositive principal minor of M, which contradicts the hypothesis of M being a P-matrix.

Conversely, suppose that for every $x \neq 0$ there exists an index i such that $x_i(Mx)_i > 0$. Let $M[I, I]$ be a principal submatrix of M, λ one of the real eigenvalues of $M[I, I]$ and $\hat{x}$ a corresponding eigenvector. Consider x such that $x[I] = \hat{x}$ and $x_i = 0$ for $i \notin I$. Since $x \neq 0$, let i be the index for which $x_i(Mx)_i > 0$. Then, if j is the index in $\hat{x}$ corrisponding to i, we have that $0 < \hat{x}_j(M[I, I]\hat{x})_j = \hat{x}_j(\lambda\hat{x})_j = \lambda\hat{x}_j^2$, which implies $\lambda > 0$. So, every real eigenvalue of $M[I, I]$ is positive.

For general matrices, it is well known that

- the determinant, by Jordan canonical form theorem, is equal to the product of all eigenvalues repeated according to their multiplicity;
- for every complex eigenvalue, also its complex conjugate is an eigenvalue, and their product is obviously positive.

So $\det(M[I, I])$ must be positive, and as a consequence M is a P-matrix. □

Corollary 2. *Let M be a square matrix.*

- *M is a $P^{(-)}$-matrix if and only if for every $x \neq 0$ there exists an index i such that $x_i(Mx)_i < 0$;*
- *M is a P_0-matrix if and only if for every $x \neq 0$ there exists an index i such that $x_i \neq 0$ and $x_i(Mx)_i \geq 0$;*
- *M is a $P_0^{(-)}$-matrix if and only if for every $x \neq 0$ there exists an index i such that $x_i \neq 0$ and $x_i(Mx)_i \leq 0$.*

Note that the part of this corollary about P_0- and $P_0^{(-)}$-matrices is more precise than what found in [5], since they did not specify that x_i should be nonzero. However, this does not invalidate their results.

10.3.2 *P*-matrix Jacobian and injectivity

An *n-dimensional interval* is a subset of $\mathbb{R}^n$ that is the Cartesian product of n real intervals. An n-dimensional interval is *open* (resp. *closed*) if all the factors are open (resp. closed) intervals.

A function from $D \subseteq \mathbb{R}^n$ to $\mathbb{R}^m$ is *differentiable* if it has a total derivative at every point $x \in D$. If the total derivative of a function f exists at a point x, then all the partial derivatives (and so the Jacobian) of f exist at x.

Theorem 5 (Gale-Nikaidô [52])**.** *Let D be a closed n-dimensional interval and $f : D \to \mathbb{R}^n$ a differentiable function. If the Jacobian $J_f(x)$ of f is a P-matrix at every $x \in D$, then f is injective in D.*

Theorem 6 ([5])**.** *Let $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable function. If the Jacobian $J_f(x)$ of f is a $P^{(-)}$-matrix at every $x \in D$, then f is injective in D.*

10.4 Stability of dynamical systems

A dynamical system consists of a set of possible states and a deterministic rule that defines the evolution in time of the state in terms of past states.

Formally, a *continuous-time dynamical system* is a triplet (T, X, φ) where $T \subseteq \mathbb{R}$, X is a set called *state space* (or *phase space*) and $\varphi : U \subseteq T \times X \to X$ is the *evolution function* of the system, provided that $\varphi(0, x) = x$ for every $x \in X$, $\varphi(t_1 + t_2, x) = \varphi(t_2, \varphi(t_1, x))$ for every $(t_1, x), (t_1 + t_2, x) \in U$ (the *group property*) and $\varphi(t, x)$ is partially differentiable with respect to t.

For every continuous-time dynamical system, φ can be expressed as the solution of the initial value problem

$$\begin{aligned} \frac{dx}{dt}(t) &= f(x(t)) \\ x(0) &= x_0 \end{aligned}$$

where $f : X \to X$ is defined as $f(x) \doteq \frac{\partial \varphi}{\partial t}(0, x)$. Instead, not every system of ODEs define a dynamical system. If this is the case, we say that the vector field f *generates* the dynamical system.

For a state a, the set of states $\gamma_a = \{\varphi(t, a) \mid (t, a) \in U\}$ is called the *orbit* through a.

If $\varphi(t, a) = a$ for all $t \in T$, then a is called a *fixed point* (or *stationary point* or *equilibrium*) of the system. A state a is a fixed point for the continuous-time dynamical system generated by a vector field f if and only if $f(a) = 0$.

A set $A \subseteq X$ of states is said to be *invariant* if $\gamma_x \subseteq A$ for all $x \in A$.

Suppose from now on that X is a topological space and T is totally ordered (with order topology) and unbounded from below and above.

A state a is a *periodic point of period k* (or, in short, period-k point) for a dynamical system if $\varphi(k, a) = a$ and $\varphi(j, a) \neq a$ for $0 < j < k$. This is equivalent

to say that $\varphi(k+t,a) = \varphi(t,a)$ for all $t \in T$, and for every $0 < j < k$ there exists $t \in T$ such that $\varphi(j+t,a) \neq \varphi(t,a)$. The orbit through a periodic point is called a *periodic orbit* (or *closed orbit*).

The *ω-limit set* (or *forward limit set*) of a state x is the set $\omega(x) = \{z \in X \mid \exists \text{ a sequence } \{t_n\} \text{ such that } \lim_{n\to+\infty} t_n = +\infty \wedge \lim_{n\to+\infty} \varphi(t_n, x) = z\}$.

A set $A \subseteq X$ of states is said to be *attracting* if the set $B(A) = \{x \in X \mid \omega(x) \subseteq A\}$ is not empty. In this case $B(A)$ is called the *basin of attraction* for A. An *attractor* A for the dynamical system is a minimal attracting set, i.e. it has no proper subset $\emptyset \neq A_1 \subset A$ which is attracting. All fixed points and periodic orbits are attractors.

An attractor A is said to be *stable* (or *Lyapunov stable*) if, for every neighborhood N of A, there is a neighborhood $N' \subseteq N$ of A such that if $x \in N'$ then $\gamma_x \subseteq N$. Attractors that are not stable are called *unstable*. A stable attractor A is said to be *asymptotically stable* if its basin of attraction is a neighborhood of A. Stable attractors that are not asymptotycally stable are called *marginally stable*.

A dynamical system is called *multistable* if it has multiple stable attractors. It is instead called *multistationary* if it has multiple fixed points.

A fixed point a for a continuous-time dynamical system generated by a differentiable vector field f is called *hyperbolic* if the Jacobian matrix $J_f(a)$ of f at a has no eigenvalue with zero real part.

10.4.1 Monotone systems

The material of this Section is mainly taken from [116]. Given a partial order $\leq$ on X, a continuous-time dynamical system generated by the vector field f is said to be *monotone with respect to* $\leq$ if, for every pair of initial conditions $x_0, y_0 \in X$, the corresponding solutions $x(t)$ and $y(t)$ of the initial value problem satisfy $x(t) \leq y(t)$ for every $t \in T$. The system is said to be *monotone* if it is monotone with respect to a partial order.

Monotone systems have nice properties of convergency in their dynamical behavior. For example, they do not admit periodic orbits. Moreover, if there is just one equilibrium, under mild conditions on the variables and boundness of solutions, this equilibrium is globally asymptotically stable [27]. If, instead, there are multiple equilibria, Hirsch theorem [63] states that every bounded solution, except for a measure-zero set of initial conditions, converges to the set of equilibria, provided that the system if strongly monotone (i.e. when the inequalities in the previous definition of monotone system are strict instead of weak).

The *J-graph* (from Jacobian) associated to a dynamical system is a signed undirected multigraph with vertices $\{v_1, \ldots, v_n\}$, where n is the dimension of X. For every pair $i, j \in \{1, \ldots, n\}$, if the partial derivative $\partial f_i(x)/\partial x_j \geq 0$ (resp. ≤ 0) for some $x \in X$, we draw an edge labeled $+1$ (resp. -1) from v_j to v_i. No edge is drawn from vertex v_j to vertex v_i if $\partial f_i(x)/\partial x_j$ vanishes identically for all $x \in X$. Notice that, in principle, there could be two edges of different sign from v_j to v_i.

A *spin assignment* σ for a signed undirected graph is a labeling of each vertex v_i with a number σ_i equal to $+1$ or -1. An edge between vertices v_i and v_j is

consistent with the spin assignment σ provided that the edge label is equal to $\sigma_i\sigma_j$. We say that σ is a *consistent spin assignment for a signed undirected graph* if every edge of the graph is consistent with σ. In other words, if there is a positive edge between vertices v_i and v_j, then v_i and v_j must have the same spin, while if there is a negative edge connecting v_i and v_j, then v_i and v_j must have opposite spins. A signed undirected graph is said to be *consistent* if there exists a consistent spin assignment for it.

Lemma 2. *A signed undirected graph is consistent if and only if every cycle in it is positive.*

Going back to dynamical systems, if $X = \mathbb{R}^n$, particular partial orders can defined by the orthants. Let z be the only vector in the intersection of $\{1,-1\}^n$ and the chosen orthant, then the *orthant order* $\leq_z$ is such that $x \leq_z y$ if $z_i x_i \leq z_i y_i$ for every $i = 1,\ldots,n$.

Theorem 7 (Kamke's theorem [112])**.** *Consider an orthant order* $\leq_z$*. The continuous-time dynamical system generated by a vector field* f *is monotone with respect to* $\leq_z$ *if and only if*

$$z_i z_j \frac{\partial f_i}{\partial x_j}(x) \geq 0 \quad \text{for all } x \in \mathbb{R}^n, i, j = 1,\ldots,n \text{ such that } i \neq j. \tag{10.3}$$

Condition 10.3 is clearly equivalent to saying that the J-graph of the dynamical system is an undirected graph and z is a consistent spin assignment for it.

Corollary 3. *A dynamical system is monotone with respect to some orthant order if and only if its associated J-graph is consistent.*

10.5 Chemical reaction networks

A *chemical rection network* consists of:

1. a set $\mathcal{S} = \{s_1,\ldots,s_n\}$ of *n species*;
2. a set $\mathcal{C} \subseteq \mathbb{R}^n_{\geq 0}$ of *m complexes*;
3. a relation $\mathcal{R} \subseteq \mathcal{C} \times \mathcal{C}$ of *r reactions* such that $(y,y) \notin \mathcal{R}$ for every $y \in \mathcal{C}$.

For a vector $x \in \mathbb{R}^n_{\geq 0}$, let $supp(x) = \{i \mid x_i > 0\}$. A complex $y = (y_1,\ldots,y_n)$ is usually written as $\sum_{i \in supp(y)} y_i s_i$, and each y_i is called the *stoichiometric coefficient* of the species s_i in y.

To more clearly indicate a reaction $(y,y') \in \mathcal{R}$ we usually write $y \to y'$. The species indexed by $supp(y)$ are called the *reactants* of the reaction, and the species indexed by $supp(y')$ are called its *products*. Each reaction $y \to y'$ defines a *reaction vector* $y'-y \in \mathbb{R}^n$. Moreover, the *stoichiometric subspace* $S = span\{y'-y \mid (y,y') \in \mathcal{R}\} \subseteq \mathbb{R}^n$ is the linear span of the reaction vectors.

If we fix an order for the reactions, then the *stoichiometric matrix* N is the $n \times r$ matrix whose j-th column is the j-th reaction vector. So S can also be defined as the column space of N and $s = rank(N) = dim(S)$ is called the *rank* of the reaction network. Clearly $s \leq n$, in particular if $s = n$ then $S = \mathbb{R}^n$, otherwise S is an s-dimensional hyperplane passing through 0

Let $c(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ be the function of molar concentrations such that $c_i(t)$ is the concentration of species s_i at time t. A *composition* is the value $c \in \mathbb{R}^n_{\geq 0}$ of all molar concentrations at a particular instant of time.

A *reaction system* is a reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ with an associated *kinetics* $v(\cdot) : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^{\mathcal{R}}_{\geq 0}$ such that, for each $y \to y' \in \mathcal{R}$, $v_{y \to y'}(\cdot)$ is called the *rate function* of the reaction, and $v_{y \to y'}(c) > 0$ if and only if $supp(y) \subseteq supp(c)$. Each $v_{y \to y'}(c)$ describes the instantaneous probability of occurence of the reaction $y \to y'$ when the composition is c. The positivity condition requires that a reaction proceeds at non-zero rate exactly when all reactant species are present in the composition.

The dynamics of the reaction system is governed by the first order ODE

$$\dot{c}(t) = Nv(c(t)). \tag{10.4}$$

which, using block matrix multiplication, can be expressed as

$$\dot{c}(t) = \sum_{y \to y' \in \mathcal{R}} N_{\cdot, y \to y'} v_{y \to y'}(c(t)) = \sum_{y \to y' \in \mathcal{R}} (y' - y) v_{y \to y'}(c(t)). \tag{10.5}$$

This equation describes the rate of change of the concentration of a species s_i as the sum of the reaction rates for the present concentration, each weighted by the net number of molecules of s_i produced at every occurrence of the corresponding reaction.

The most common choice for rate functions is the *mass action* form, in which for a reaction $y \to y'$ we have:

$$v_{y \to y'}(c) = k_{y \to y'} \prod_{i=1}^{n} c_i^{y_i},$$

where $k_{y \to y'} \in \mathbb{R}_{>0}$ is *the rate constant* of $y \to y'$ and by convention $0^0 = 1$. This formulation amounts to assuming that the probability that reaction $y \to y'$ occurs in the next infinitesimal time interval dt is proportional to the concentration of each reactant, possibly elevated to some power if more molecules are needed for the reaction to take place. This is deemed to reflect the likelihood of an encounter among all molecules involved in the reaction.

If all reactions are of mass action type, the network kinetics is said to be mass action, and the ODE (10.5) becomes

$$\dot{c}(t) = \sum_{y \to y' \in \mathcal{R}} (y' - y) k_{y \to y'} \prod_{i=1}^{n} c_i(t)^{y_i}.$$

10.5.1 Stoichiometric compatibility and multiple equilibria

Let $c(\cdot)$ be a solution of (10.4) in the time interval $[0, T]$. For every $t \in [0, T]$, the fundamental theorem of calculus and (10.5) imply that

$$c(t) - c(0) = \int_0^t \dot{c}(\tau)d\tau = \int_0^t \sum_{y \to y' \in \mathcal{R}} (y' - y)v_{y \to y'}(c(\tau))d\tau$$
$$= \sum_{y \to y' \in \mathcal{R}} (y' - y) \int_0^t v_{y \to y'}(c(\tau))d\tau,$$

which is a linear combination of the reaction vectors, and thus

$$c(t) - c(0) \in S. \tag{10.6}$$

Two compositions $c_1, c_2 \in \mathbb{R}^n_{\geq 0}$ are said to be *stoichiometrically compatible* if $c_2 - c_1 \in S$ (i.e. if there exists $\alpha \in \mathbb{R}^r$ such that $c_2 - c_1 = N\alpha$). The "stoichiometrically compatible" relation is clearly an equivalence, its classes are called *stoichiometric compatibility classes* and are s-dimensional hyperplanes parallel to S. From (10.6), it follows that the orbit $\gamma_{c(0)}$ is entirely contained in the stoichiometric compatibility class of $c(0)$.

Moreover, every stoichiometric compatibility class $[c_1]$ is invariant, since for every $c_2 \in [c_1]$ the orbit $\gamma_{c_2} \subseteq [c_2] = [c_1]$.

Each stoichiometric compatibility class can have a different number of equilibria within itself. Thus, the problems of finding if a reaction system admits an equilibrium, if it presents multistationarity, or to determine the stability type of an equilibrium should be tackled within each stoichiometric compatibility class.

A reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ *has the capacity for multiple equilibria* (resp. *multiple positive equilibria*) if there exist a rate function $v(c)$ and two stoichiometrically compatible compositions $c_1, c_2 \in \mathbb{R}^n_{\geq 0}$ (resp. $\mathbb{R}^n_{>0}$) such that $c_1 \neq c_2$ and $Nv(c_1) = Nv(c_2) = 0$.

10.5.2 Conserved moieties

Let $S^\perp$ be the orthogonal complement of the stoichiometric subspace S in $\mathbb{R}^n$ relative to the scalar product $\cdot$, i.e. $S^\perp = \{g \in \mathbb{R}^n \mid g \cdot \mu = 0, \forall \mu \in S\}$. Alternatively, $S^\perp$ can be defined as the left null space of the stoichiometric matrix N, i.e. $S^\perp = \{g \in \mathbb{R}^n \mid g^T N = 0\}$. Linear algebra theorems show that $dim(S^\perp) = dim(\mathbb{R}^n) - dim(S) = n - s$, so $S^\perp \neq \{0\}$ if and only if $s < n$.

Consider $g \in S^\perp \setminus \{0\}$ and let $c(t)$ be a solution of the ODE (10.4) in the time interval $[0, T]$. From eq. (10.6) it follows that $g \cdot (c(t) - c(0)) = 0$, that is $g \cdot c(t) = g \cdot c(0)$. So, fixed an initial composition $c(0)$, $\sum_{i=1}^n g_i c_i(t)$ is constant for all $t \in [0, T]$, which is clearly a conservation relation for the species concentrations. Abstracting from the particular solution, we call the formula

$$\sum_{i \in supp(g)} g_i s_i$$

a *conserved moiety* of the reaction network. There are $n - s$ independent conserved moieties, which are defined by a basis of $S^{\perp}$.

A reaction network is said to be *conservative* if there exists a positive $g \in S^{\perp}$. From this definition, it follows that a conservative network does not exchange mass with the exterior.

10.5.3 Deficiency theory

The *complex graph* of a reaction network is the directed graph whose vertices are the complexes, and whose edges correspond to the reactions. A *linkage class* is a connected component of the undirected version of the complex graph. We indicate with l the number of linkage classes.

For every linkage class j, the rank s_j of the corresponding subnetwork must be less than the number of complexes ($r_j \leq m_j - 1$). Moreover, reaction vectors of different subnetworks are linearly independent, so $s = \sum_{j=1}^{l} s_j \leq \sum_{j=1}^{l}(m_j - 1) = m - l$.

The *deficiency* of the reaction network is $\delta = m - l - s$. For what said above, δ is a nonnegative integer.

A *strong-linkage class* is a strongly connected component of the complex graph. Thus, each linkage class is partitioned in one or more strong-linkage classes. A reaction network is *weakly reversible* if each linkage class is also a strong-linkage class. A strong-linkage class is *terminal* if it is not connected to other strong-linkage classes.

Theorem 8 (Deficiency zero theorem [41])**.** *For any reaction network of deficiency zero:*

1. *If the network is not weakly reversible, then for arbitrary kinetics the reaction network admits neither a positive equilibrium, nor a periodic orbit in* $\mathbb{R}^n_{>0}$.

2. *If the network is weakly reversible then, for mass action kinetics (regardless of the values of the rate constants), the reaction system admits precisely one equilibrium within each positive stoichiometric compatibility class, which is asymptotically stable.*

Proof. For a complete proof, see [45]. □

Theorem 9 (Deficiency one theorem [42, 43, 45])**.** *Let* $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ *be a reaction network of deficiency* δ *and let* δ_j *be the deficiencies of its linkage classes,* $j = 1, \ldots, l$*. Suppose that:*

1. $\delta_j \leq 1$*, for all* $j = 1, \ldots, l$

2. $\sum_{j=1}^{l} \delta_j = \delta$

3. *each linkage class contains just one terminal strong-linkage class.*

If, for particular values of the rate constants, the mass action system admits one equilibrium, then each positive stoichiometric compatibility class contains precisely one equilibrium. If the network is weakly reversible then, regardless of the values of the rate constants, the mass action system admits precisely one equilibrium within each positive stoichiometric compatibility class.

There is also an algorithm specific for networks of deficiency one [44], whose correctness was proved in [46].

10.5.4 Network injectivity

A reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ is *injective* if $Nv(c)$ is injective for all possible kinetics $v(c)$. In particular, a reaction system with mass action kinetics is injective if $Nv(c)$ is injective for all possible positive rate constants of the reactions.

If a reaction system is injective, then clearly it does not have the capacity for multiple equilibria. Therefore injectivity is a sufficient condition for monostationarity, but it is not necessary. In fact, even if there are two distinct composition c_1 and c_2 such that $Nv(c_1) = Nv(c_2)$, to have multiple equilibria it is needed also that they are fixed points, i.e. that $Nv(c_1) = Nv(c_2) = 0$.

The Jacobian matrix of a reaction system is the Jacobian of $Nv(c)$, i.e. $J_{Nv}(c) = \frac{\partial Nv(c)}{\partial c} = N\frac{\partial v(c)}{\partial c}$.

A reaction system is *nonautocatalytic* if for all $i = 1, \ldots, n$, $y \to y' \in \mathcal{R}$ and $c \in \mathbb{R}^n$, we have that

$$N_{i,y\to y'} \frac{\partial v_{y\to y'}(c)}{\partial c_i} \leq 0$$

and

$$N_{i,y\to y'} = 0 \Rightarrow \frac{\partial v_{y\to y'}(c)}{\partial c_i} = 0.$$

The first condition requires that if a species is a reactant (resp. product) in a reaction, then increasing its concentration cannot decrease (resp. increase) the reaction rate. The second condition instead states that if a species does not partecipate in a reaction, then it has no influence on the reaction rate.

Reaction systems with usual kinetics like mass action, Michaelis–Menten, etc., are nonautocatalytic provided that in each reaction a species appears only as a reactant or as a product, but not on both sides of the reaction. Usually, an autocatalytic reaction like $A + B \to 2A$ represents (and can be rewritten as) a set of subsequent reactions, e.g. $A + B \to C \to 2A$, where C is a short-lived intermediate molecule.

The *qualitative class* of a real matrix A, denoted $Q(A)$, is the set of all real matrix B such that $sign(B) = sign(A)$. So we can say the a reaction system is nonautocatalytic if $\frac{\partial v(c)}{\partial c} \in Q(-\tilde{N}^T)$, where $\tilde{N}$ is equal to N with some elements that may be changed to 0.

A square matrix is *sign-nonsingular (SNS)* if every matrix in its qualitative class is nonsingular.

Theorem 10 ([14]). *A real matrix A is SNS if and only if for every matrix $B \in Q(A)$ $sign(det(B)) = sign(det(A)) \neq 0$.*

Theorem 11 ([14]). *A real matrix A is SNS if and only if in the standard expansion (Leibniz formula) of $det(A)$ there is at least one nonzero term, and all the nonzero terms have the same sign.*

A real matrix A is an *L-matrix* if for every matrix $B \in Q(A)$ the rows of B are linearly independent. Clearly, a square matrix is SNS if and only if it is an L-matrix.

A *signing* of order k is a nonzero diagonal matrix of order k whose entries are in the set $\{-1, 0, 1\}$. A *row signing* of a $m \times n$ matrix A is a product matrix DA where D is a signing of order m. A vector is *unisigned* if it is nonzero and it is either nonnegative or nonpositive.

Theorem 12 ([75]). *A real matrix A is an L-matrix if and only if every row signing of A contains a unisigned column.*

A matrix M is *strongly sign determined (SSD)* if all square submatrices of M are either singular or SNS.

Theorem 13 ([5]). *If the stoichiometric matrix N of a non-autocatalytic reaction system is SSD, then the Jacobian matrix $J_{Nv}(c)$ is a $P_0^{(-)}$-matrix at all c.*

An *inflow* (resp. *outflow*) reaction for a species $s \in \mathcal{S}$ is a reaction $0 \to s$ (resp. $s \to 0$) with mass action kinetics, i.e. $v_{0\to s}(c) = k_{0\to s}$ (resp. $v_{s\to 0}(c) = k_{s\to 0}c_s$).

Theorem 14 ([5]). *Consider a non-autocatalytic reaction system having an outflow reaction for each species. If its stoichiometrix matrix N is SSD, then the Jacobian matrix $J_{Nv}(c)$ is a $P^{(-)}$-matrix at all c.*

If the Jacobian matrix $J_{Nv}(c)$ is a $P^{(-)}$-matrix at all c, then the reaction system is injective, and thus it does not have the capacity for multiple equilibria.

An equilibrium a is said to be *degenerate* if $null(J_{Nv}(a)) \cap S \neq \{0\}$. In this case there is a vector $v \in S \setminus \{0\}$ such that $J_{Nv}(a)v = 0 = 0v$. So, 0 is an eigenvalue of $J_{Nv}(a)$, S intersects the eigenspace of 0 and thus a is not hyperbolic in its stoichiometric compatibility class $[a]$.

Theorem 15 ([4, 24]). *Let $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ be a reaction network having an outflow reaction for all the species in the subset $\mathcal{M} \subseteq \mathcal{S}$. If the reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R} \cup \{s \to 0 \mid s \in \mathcal{S} \setminus \mathcal{M}\})$ is injective, then $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ does not have the capacity for multiple nondegenerate equilibria.*

The *SR graph* for a reaction network is a bipartite undirected multigraph with two kinds of vertices: species vertices and reaction vertices. There is a species vertex for each species in the network, and there is a reaction vertex for each reaction or reversible reaction pair in the network. If a species has a nonzero stoichiometric

coefficient in a complex which is at one side of a reaction, then there is an edge between the corresponding species vertex and reaction vertex with the complex name as the label and the stoichiometric coefficent as the edge weight.

Pairs of edges that meet at a reaction vertex and have the same complex label are called *c-pairs*. If a cycle contains an odd number of c-pairs is called an *o-cycle*, otherwise is called an *e-cycle*. Since the SR graph is bipartite, every cycle as an even lenght. Cycles for which alternately multiplying and dividing the edge labels along the cycle gives the final result 1 are called *s-cycles* (for stoichiometric cycles).

We say that two cycles in the SR graph have a *species-to-reaction (S-to-R) intersection* if the common edges of the two cycles constitute a path that begins at a species vertex and ends at a reaction vertex, or if they constitute a disjoint union of such paths.

Theorem 16 ([25])**.** *Consider a reaction network having an outflow reaction for each species. If in its SR graph:*

1. *each cycle is an o-cycle or an s-cycle,*

2. *no two e-cycles have an S-to-R intersection,*

then, for mass action kinetics, the reaction system is injective.

Theorem 17 ([4])**.** *Consider a reaction network such that in each reaction a species does not appear on both sides of the reaction. If its SR graph satisfy the two conditions of theorem 16, then its stoichiometric matrix is SSD.*

So, for a nonautocatalytic reaction system with mass action kinetics, checking if its stoichiometric matrix is SSD is enough, and checking the properties of the SR graph does not give extra information. Otherwise, if the system is autocatalytic, only the SR graph can be analyzed.

Chapter 11

ERNEST Toolbox

ERNEST Reaction Network Equilibria Study Toolbox is a software package structured as a set of MATLAB functions and classes. The software is available under the GNU GPL free software license and can be downloaded from http://people.sissa.it/~altafini/papers/SoAl09/ . It requires the MATLAB Optimization Toolbox.

The analysis is performed by the main function *model_analysis*, which needs as input a structure specifying the species and reactions of a CRN. The format of this structure is simply the one defined by the *TranslateSBML* function from libSBML [12], which imports SBML files in MATLAB. So a SBML model, after the standard import, can be directly analyzed by our toolbox, but all the extra information potentially contained in the file, like compartments, constraints, reaction modifiers and kinetic laws, will be ignored.

All the criteria implemented by ERNEST aim to verify conditions on the CRN structure which are sufficient for monostationarity of the relative dynamical system, i.e. to rule out the possibility of multiple equilibria regardless of the rate constants and the initial concentrations.

The *model_analysis* function operates in the following way:

(1) calculates complexes, stoichiometric matrix and rank, linkage classes, strong-linkage classes, network reversibility, weak reversibility and deficiency;

(2a) if the Deficiency Zero theorem [41] is applicable, prints out the relative response;

(2b.1) otherwise calculates terminal strong-linkage classes and deficiencies of the linkage classes;

(2b.2a) if the Deficiency One theorem [43] is applicable, prints out the relative response for mass action kinetics;

(2b.2b) otherwise verifies that the network is regular, and in case apply the Deficiency One algorithm [44] for mass action kinetics; if this algorithm verifies

that the network admits multiple positive equilibria within a stoichiometric compatibility class, it also prints out an example of reaction rate constants and two equilibria for the corresponding mass action system;

(3a) if the CRN is autocatalytic, then calculates the Species–Reaction graph, finds all its cycles and tries to verify the two conditions for monostationarity with mass action kinetics of [25];

(3b) otherwise verifies if the stoichiometric matrix is SSD and in case exclude multiple non-degenerate equilibria [5] for general kinetics.

Obviously the runtime of this function increases with the number of species and reactions of the network. The most complex part is the Deficiency One algorithm (point 2b.2b), which involves the solution of linear programming problems with additional sign constraints, but for medium/big reaction networks this code is usually not executed since their deficiency is typically greater than one.

We verified the correctness of our toolbox by successfully reproducing the results for all the examples of [2, 22, 25, 26], plus others selected from the cited Feinberg's papers.

One interesting example is the reaction network proposed in example 1.1 of [45]:

$$\begin{array}{c} A \leftrightarrows 2B \\ A + C \leftrightarrows D \\ \nwarrow \swarrow \\ B + E \end{array} \tag{11.1}$$

As explained in the original paper, this network has deficiency 0 and is weakly reversible, so by theorem 8, for mass action kinetics the reaction system is monostationary within each positive stoichiometric compatibility class. But what happens if the dynamics is not of mass action type?

If we analyze this network with ERNEST, this is the relative output:

> The reaction network is weakly reversible and has deficiency 0, so with mass action kinetics each positive stoichiometric compatibility class contains precisely one equilibrium, which is asymptotically stable.
> The reaction network with mass action or Michaelis–Menten kinetics is non-autocatalytic.
> The stoichiometric matrix is not SSD. The reaction network has the capacity for multiple equilibria.
> One set of species and reactions because of which the stoichiometric matrix is not SSD:
>
> ans =

'A' 'B'

reaction_string =

2 B → 1 A

reaction_string =

1 B + 1 E → 1 A + 1 C

So, the system behavior for mass action is confirmed, but for other type of kinetics ERNEST is not able to exclude the multistationarity and instead suggests a subset of species and reaction for further investigation. In fact, if we study the smaller network:

$$\begin{array}{c} A \leftrightarrows 2B \\ A \leftrightarrows B \end{array} \tag{11.2}$$

this has the same properties of network 11.1, and a similar output from ERNEST. For mass action kinetics, there is only one nontrivial equilibrium, that can be calculated very easily. To see if really this system with non mass action kinetics can have multiple equilibria, we can suppose that all the reaction are mass action except $B \to A$, which is of Michaelis–Menten type. In this case, the system of ODEs will be:

$$\begin{cases} \dot{c}_A = -k_1 c_A + k_2 c_B^2 - k_3 c_A + \dfrac{k_4 c_B}{k_M + c_B} \\ \dot{c}_B = 2k_1 c_A - 2k_2 c_B^2 + k_3 c_A - \dfrac{k_4 c_B}{k_M + c_B} \end{cases}$$

Imposing the equilibrium conditions $\dot{c}_A = \dot{c}_B = 0$, it is easy to see that, apart from the trivial equilibrium (0, 0), the solutions of the system are:

$$c_{B_{1,2}} = \frac{-k_M \pm \sqrt{k_M^2 + 4\frac{k_1 k_4}{k_2 k_3}}}{2}$$
$$c_A = \frac{k_2}{k_1} c_B^2$$

Therefore, there are two different equilibria in the same stoichiometric compatibility class, which in this case is simply $\mathbb{R}^2$, and the system is in fact multistationary. We have to remark that, since verifying that the stoichiometric matrix is SSD is only a sufficient condition for monostationarity, if ERNEST says that the reaction system has the *capacity* for multiple equilibria, it does not mean that this is always the case.

This example is clearly very simple, but shows how ERNEST can be useful not only for model discrimination, but also for exploring the behavior of a reaction network with different types of kinetics.

Chapter 12

Fundamental cycles and monotonicity

Consider a continous-time dynamical system generated by a vector field $f(\cdot)$. Checking whether the system is monotone, i.e. whether its associated J-graph is consistent, is a simple task, verifiable in polynomial time with a dynamic programming algorithm.

In general, for any given dynamical system, the corresponding J-graph will not be consistent, although in a biological context it might be "near-monotone", i.e. closer to monotone than expected by random edge assignments, as claimed in [116]. Our goal is to identify the smallest number of edges such that if we change their signs the obtained graph is consistent, and the tool we use for this scope is an extension of the theory of fundamental cycles for signed graphs.

Theorem 18. *Let F_i, $i = 1, \dots, n$, be a stricly fundamental cycle basis of an undirected multigraph G with respect to a spanning forest T. For every non-fundamental cycle $C = \ominus_{i \in I} F_i$ of G, there is a partition $\{I_1, I_2\}$ of I such that $C_1 = \ominus_{i \in I_1} F_i$ and $C_2 = \ominus_{i \in I_2} F_i$ are cycles.*

Proof. Let F_k, $k \in I$, be one of the fundamental cycles generating C. $C \cap F_k$ contains at least one edge, the chord of F_k.
Suppose now that $C \cap F_k$ is a unique path. Then also $C \setminus F_k$ and $F_k \setminus C$ must be disjoint paths with the same endpoints of $C \cap F_k$, so together they form the cycle $(C \setminus F_k) \cup (F_k \setminus C) = C \ominus F_k = \ominus_{i \in I \setminus \{k\}} F_i$. Therefore, $\{\{k\}, I \setminus \{k\}\}$ is the desired partition of I.
If instead $C \cap F_k$ consists of two or more disconnected paths, we can choose two of these paths which are "near" in F_k, i.e. connected by a path p_1 in $F_k \setminus C$. The endpoints of p_1 are also vertices of C, so they divide C in two paths p_2 and p_3 such that $C = p_2 \cdot p_3$. Note that $p_1 \cap p_2 = p_1 \cap p_3 = \emptyset$.
Now, $C_1 = p_1 \cdot p_2$ and $C_2 = p_1 \cdot p_3$ are clearly two cycles of G. p_1 does not contain any chord, since $p_1 \subset F_k \setminus C \subseteq T$. The paths p_2 and p_3 instead determine a partition of the set of chords contained in C, which corresponds to a partition $\{I_1, I_2\}$ of I such that $C_1 = \ominus_{i \in I_1} F_i$ and $C_2 = \ominus_{i \in I_2} F_i$. □

Theorem 19. *Consider a signed undirected multigraph. If the symmetric difference of two cycles C_1, C_2 is a unique cycle, then $sign(C_1 \ominus C_2) = sign(C_1) \cdot sign(C_2)$.*

Proof. C_1 can be divided in two disjoint sets of edges:

$$C_1 = (C_1 \setminus C_2) \cup (C_1 \cap C_2),$$

which implies that

$$sign(C_1) = sign(C_1 \setminus C_2) \cdot sign(C_1 \cap C_2),$$

and symmetrically for $sign(C_2)$. For the definition of symmetric difference

$$C_1 \ominus C_2 = (C_1 \setminus C_2) \cup (C_2 \setminus C_1),$$

which implies that

$$sign(C_1 \ominus C_2) = sign(C_1 \setminus C_2) \cdot sign(C_2 \setminus C_1) = \frac{sign(C_1) \cdot sign(C_2)}{sign(C_1 \cap C_2)^2}$$
$$= \frac{sign(C_1) \cdot sign(C_2)}{+} = sign(C_1) \cdot sign(C_2).$$

□

Theorem 20. *Let F_i, $i = 1, \ldots, n$, be the fundamental cycles of a signed undirected multigraph G with respect to a spanning forest, and let $C = \ominus_{i \in I} F_i$ be a cycle of G. Then*

$$sign(C) = \prod_{i \in I} sign(F_i).$$

Proof. By induction on $|I|$. If $|I| = 1$, then C is a fundamental cycle F_i, so they have the same sign.
Now let $|I| > 1$ and assume the theorem true for every I' with $|I'| < |I|$. For theorem 18 there is a partition $\{I_1, I_2\}$ of I such that $C_1 = \ominus_{i \in I_1} F_i$ and $C_2 = \ominus_{i \in I_2} F_i$ are cycles. $I_1, I_2 \subset I$ implies $|I_1|, |I_2| < |I|$, so using the inductive hypothesis we have that $sign(C_1) = \prod_{i \in I_1} sign(F_i)$ and $sign(C_2) = \prod_{i \in I_2} sign(F_i)$. Finally, applying theorem 19 to C_1, C_2, we can conclude that $sign(C) = sign(C_1 \ominus C_2) = sign(C_1) \cdot sign(C_2) = \prod_{i \in I_1} sign(F_i) \cdot \prod_{i \in I_2} sign(F_i) = \prod_{i \in I} sign(F_i)$. □

Corollary 4. *A signed undirected graph G is consistent if and only if, for an arbitrarily chosen spanning forest, all corresponding fundamental cycles of G are positive.*

The minimum number of edges whose sign should be changed in order that the multigraph becomes consistent is called the *consistency deficit* of the multigraph. This value measures how close a given signed multigraph is to a consistent graph. Computing the consistency deficit is an NP-hard problem, equivalent to the well-known MAX-CUT problem [28] or to the problem of finding the ground state of a frustrated spin system in statistical physics [116].

Given a signed undirected multigraph with μ fundamental cycles of which ν have positive sign and $\mu - \nu$ have negative sign, one way to render the entire multigraph consistent is to change the sign of the last $\mu - \nu$ fundamental cycles.

Corollary 5. *A signed undirected multigraph having a μ-dimensional fundamental cycle basis characterized by $\mu - \nu$ negative fundamental cycles, can be rendered consistent by exchanging the signs of the $\mu - \nu$ chords generating the fundamental cycles having negative sign.*

Of course, the worst case is when all fundamental cycles are negative, i.e. any multigraph can be rendered consistent with at most μ sign changes. From the theory of signed graphs [115], we also have another worst-case upper bound on the consistency deficit, $\eta = (|E| - \sqrt{|E|})/2$. Hence we have the following Proposition.

Proposition 3. *Any signed multigraph can be rendered consistent with at most* $\min(\mu, \eta)$ *sign changes in its edge labels.*

The two values for the upper bound are unrelated: for very sparse graphs (with average connectivity of a node < 2) then $\mu < \eta$, viceversa for more dense graphs. While the value of μ is always attainable in a graph, it is not clear from the literature in which cases η is achievable as a worst-case upper bound.

In general the sign associated with a basis of fundamental cycles is not invariant to changes of basis (i.e. of spanning tree). Therefore, if we can find a fundamental cycle basis with fewer fundamental negative cycles, we have to do fewer changes of sign in order to obtain a monotone system. The following Proposition is the starting point for "simplifying" the graph by changing its signs in a suitable equivalence class in which the monotonicity properties and the number of inconsistencies are preserved.

Proposition 4. *Exchanging the sign of the edges through a cut set preserves the sign of each cycle of a given signed multigraph.*

Proof. From Theorem 1, every cycle intersects a cut set in an even number of edges and hence a sign change through an entire cut set does not alter the sign of a cycle. □

Starting from this observations, it is possible to write (heuristic) algorithms to find an equivalent signing of the multigraph which minimizes the number of negative edges [67]. Provided we associate high weights to the edges having negative sign after the application of the heuristic, any minimum spanning tree algorithm [31] will select a spanning tree with a minimal number of minus signs. From Corollary 4, the cycle subspace associated with the set of fundamental cycles having positive sign corresponds to the monotone subsystem of the original system.

Chapter 13

Conclusion

Powerful tools to study biochemical networks are particularly needed in Systems Biology, where the number of (unknown) reaction parameter increase dramatically.

ERNEST can be quite useful if applied for model discrimination, as in the examples cited above. It has several advantages over the Chemical Reaction Network Toolbox since it verifies more criteria, it is applicable also to kinetics not of mass action type, it can be applied to SBML models, it is multiplatform and open source.

Two possible extensions of the toolbox features are the implementation of the advanced deficiency theory [37] (which generalizes the deficiency one algorithm to CRNs of deficiency greater than one), and the verification of some sufficient conditions for multistationarity, like those of section 4 of [23] or maybe others inspired by [117].

Another possible improvement would be to verify sufficient conditions for monotonicity, like the one recently proposed in [2]. This would give information on the stability of the equilibria, which is otherwise proved only for deficiency 0 networks.

Regarding instead the distance to monotonicity, a set of heuristic algorithms for its estimation has been proposed in [67] by our colleagues, based on the theoretical framework presented here. The output of these algorithms is an interval inside which the the consistency deficit must lie. These programs are able to treat also large networks in a limited computational time. Moreover, two gene regulatory networks (for *E. coli* and *S. cerevisiae*) have been analyzed with this algorithms and, as supposed by [116], they are indeed near-consistent, i.e. the corresponding dynamical systems are close to monotone.

Bibliography

[1] B. D. O. Anderson, M. Deistler, L. Farina, and L. Benvenuti. Nonnegative realization of a linear system with nonnegative impulse response. *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, 43(2):134–142, 1996.

[2] D. Angeli, P. De Leenheer, and E. D. Sontag. Graph-theoretic characterizations of monotonicity of chemical networks in reaction coordinates. submitted to J. Math. Biol., 2009.

[3] S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, 360(1):213–227, 2006.

[4] M. Banaji and G. Craciun. Graph-theoretic criteria for injectivity and unique equilibria in general chemical reaction systems. *Advances in Applied Mathematics*, in press.

[5] M. Banaji, P. Donnell, and S. Baigent. *P* matrix properties, injectivity, and stability in chemical reaction systems. *SIAM J. Appl. Math.*, 67(6):1523–1547, 2007.

[6] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, 3(78), 2007.

[7] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[8] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, 37:382–390, 2005.

[9] A. Becskei, M. G. Boselli, and A. van Oudenaarden. Amplitude control of cell-cycle waves by nuclear import. *Nat. Cell Biol.*, 6(5):451–457, 2004.

[10] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea. Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. U.S.A.*, 103(35):13004–13009, 2006.

[11] A. Beyer, J. Hollunder, H.-P. Nasheuer, and T. Wilhelm. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell. Proteomics*, 3(11):1083–1092, 2004.

[12] B. J. Bornstein, S. M. Keating, A. Jouraku, and M. Hucka. LibSBML: an API Library for SBML. *Bioinformatics*, 24(6):880–881, 2008.

[13] R. Brockmann, A. Beyer, J. J. Heinisch, and T. Wilhelm. Posttranscriptional expression regulation: What determines translation rates? *PLoS Comput. Biol.*, 3(3):e57, 2007.

[14] R. A. Brualdi and B. L. Shader. *Matrices of sign-solvable linear systems*, volume 116 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1995.

[15] A. J. Butte and I. S. Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. *Proc. AMIA Symp.*, pages 711–715, 1999.

[16] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 5:418–429, 2000.

[17] G. P. Cereghino and I. E. Scheffler. Genetic analysis of glucose regulation in *Saccharomyces cerevisiae*: control of transcription versus mRNA turnover. *EMBO J.*, 15(2):363–374, 1996.

[18] C. Cheadle, J. Fan, Y. S. Cho-Chung, T. Werner, J. Ray, L. Do, M. Gorospe, and K. G. Becker. Stability regulation of mRNA and the control of gene expression. *Ann. N. Y. Acad. Sci.*, 1058:196–204, 2005.

[19] G. Chechik, E. Oh, O. Rando, J. Weissman, A. Regev, and D. Koller. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat. Biotechnol.*, 26(11):1251–1259, 2008.

[20] Z. Chen, E. A. Odstrcil, B. P. Tu, and S. L. McKnight. Restriction of DNA replication to the reductive phase of the metabolic cycle protects genome integrity. *Science*, 316(5833):1916–1919, 2007.

[21] C. Conradi, D. Flockerzi, J. Raisch, and J. Stelling. Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proc. Natl. Acad. Sci. U.S.A.*, 104(49):19175–19180, 2007.

[22] C. Conradi, J. Saez-Rodriguez, E.-D. Gilles, and J. Raisch. Using chemical reaction network theory to discard a kinetic mechanism hypothesis. *IEE Proc. Syst. Biol.*, 152(4):243–248, 2005.

[23] G. Craciun and M. Feinberg. Multiple equilibria in complex chemical reaction networks: I. The injectivity property. *SIAM J. Appl. Math.*, 65(5):1526–1546, 2005.

[24] G. Craciun and M. Feinberg. Multiple equilibria in complex chemical reaction networks: extensions to entrapped species models. *IEE Proc. Syst. Biol.*, 153(4):179–186, 2006.

[25] G. Craciun and M. Feinberg. Multiple equilibria in complex chemical reaction networks: II. The species-reaction graph. *SIAM J. Appl. Math.*, 66(4):1321–1338, 2006.

[26] G. Craciun, Y. Tang, and M. Feinberg. Understanding bistability in complex enzyme-driven reaction networks. *Proc. Natl. Acad. Sci. U.S.A.*, 103(23):8697–8702, 2006.

[27] E. N. Dancer. Some remarks on a boundedness assumption for monotone dynamical systems. *Proc. Am. Math. Soc.*, 126(3):801–807, 1998.

[28] B. DasGupta, G. A. Enciso, E. D. Sontag, and Y. Zhang. Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. *Biosystems*, 90(1):161–178, 2007.

[29] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska. Estimating mutual information using B-spline functions – an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5:118, 2004.

[30] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[31] N. Deo. *Graph theory with applications to engineering and computer science*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1974.

[32] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.

[33] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Mining the gene expression matrix: inferring gene relationships from large scale gene expression data. In R. Paton and M. Holcombe, editors, *IPCAT '97: Proceedings of the second international workshop on Information processing in cell and tissues*, pages 203–212, New York, NY, USA, 1998. Plenum Publishing.

[34] R. Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, Heidelberg, 2005.

[35] J. C. Doyle, B. A. Francis, and A. R. Tannenbaum. *Feedback Control Theory*. Macmillan Publishing Co., New York, 1992.

[36] D. Edwards. *Introduction to Graphical Modelling*. Springer, New York, NY, USA, 2000.

[37] P. Ellison. *The advanced deficiency algorithm and its applications to mechanism discrimination*. PhD thesis, Department of Chemical Engineering, University of Rochester, Rochester, NY, 1998.

[38] P. Ellison and M. Feinberg. How catalytic mechanisms reveal themselves in multiple steady-state data: I. Basic principles. *J. Mol. Catal. A: Chem.*, 154(1-2):155–167, 2000.

[39] P. Erdös and A. Rényi. On random graphs. *Publ. Math. Debrecen*, 6:290–297, 1959.

[40] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):e8, 2007.

[41] M. Feinberg. Mathematical aspects of mass action kinetics. In L. Lapidus and N. R. Amundson, editors, *Chemical Reactor Theory: A Review*, chapter 1, pages 1–78. Prentice-Hall, Englewood Cliffs, NJ, 1977.

[42] M. Feinberg. Chemical oscillations, multiple equilibria and reaction network structure. In W. E. Stewart, editor, *Dynamics and modelling of reactive systems*, pages 59–130. Academic Press, New York, 1980.

[43] M. Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors–I. The deficiency zero and deficiency one theorems. *Chem. Eng. Sci.*, 42(10):2229–2268, 1987.

[44] M. Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors–II. Multiple steady states for networks of deficiency one. *Chem. Eng. Sci.*, 43(1):1–25, 1988.

[45] M. Feinberg. The existence and uniqueness of steady states for a class of chemical reaction networks. *Arch. Ration. Mech. Anal.*, 132(4):311–370, 1995.

[46] M. Feinberg. Multiple steady states for chemical reaction networks of deficiency one. *Arch. Ration. Mech. Anal.*, 132(4):371–406, 1995.

[47] M. Feinberg and F. J. M. Horn. Dynamics of open chemical systems and the algebraic structure of the underlying reaction network. *Chem. Eng. Sci.*, 29(3):775–787, 1974.

[48] M. Fiedler and V. Pták. On matrices with non-positive off-diagonal elements and positive principal minors. *Czech. Math. J.*, 12(3):382–400, 1962.

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

[50] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3-4):601–620, 2000.

[51] B. Futcher. Metabolic cycle, cell cycle, and the finishing kick to Start. *Genome Biol.*, 7(4):107, 2006.

[52] D. Gale and H. Nikaidô. The Jacobian matrix and global univalence of mappings. *Math. Ann.*, 159(2):81–93, 1965.

[53] J. García-Martínez, F. González-Candelas, and J. E. Pérez-Ortín. Common gene expression strategies revealed by genome-wide analysis in yeast. *Genome Biol.*, 8(10):R222, 2007.

[54] T. S. Gardner and J. J. Faith. Reverse-engineering transcription control networks. *Phys. Life Rev.*, 2(1):65–88, 2005.

[55] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–4257, 2000.

[56] A. P. Gerber, D. Herschlag, and P. O. Brown. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.*, 2(3):342–354, 2004.

[57] J. Grigull, S. Mnaimneh, J. Pootoolal, M. D. Robinson, and T. R. Hughes. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol. Cell. Biol.*, 24(12):5534–5547, 2004.

[58] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, and V. Stümpflen. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, 34(Database issue):D436–D441, 2006.

[59] D. A. Hall, H. Zhu, X. Zhu, M. Gerstein, and M. Snyder. Regulation of gene expression by a metabolic enzyme. *Science*, 306(5695):482–484, 2004.

[60] J. L. Hargrove and F. H. Schmidt. The role of mRNA and protein stability in gene expression. *FASEB J.*, 3(12):2360–2370, 1989.

[61] H. Hieronymus and P. A. Silver. Genome-wide analysis of RNA–protein interactions illustrates specificity of the mRNA export machinery. *Nat. Genet.*, 33:155–161, 2003.

[62] H. Hieronymus and P. A. Silver. A systems view of mRNP biology. *Genes Dev.*, 18(23):2845–2860, 2004.

[63] M. W. Hirsch. Differential equations and convergence almost everywhere in strongly monotone semiflows. In J. Smoller, editor, *Nonlinear partial differential equations (Durham, NH, 1982)*, volume 17 of *Contemp. Math.*, pages 267–285, Providence, RI, USA, 1983. Amer. Math. Soc.

[64] M. W. Hirsch and H. L. Smith. Monotone dynamical systems. In A. Cañada, P. Drábek, and A. Fonda, editors, *Handbook of Differential Equations: Ordinary Differential Equations*, volume 2, chapter 4, pages 239–357. North-Holland, 2006.

[65] F. Horn and R. Jackson. General mass action kinetics. *Arch. Ration. Mech. Anal.*, 47(2):81–116, 1972.

[66] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[67] G. Iacono, F. Ramezani, N. Soranzo, and C. Altafini. Determining the distance to monotonicity of a biological network: a graph-theoretical approach. *IET Syst. Biol.*, 4(3):223–235, 2010.

[68] P. D. Karp, M. Riley, S. M. Paley, A. Pellegrini-Toole, and M. Krummenacker. Eco Cyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, 27(1):55–58, 1999.

[69] M. Kaufman and R. Thomas. Model analysis of the bases of multistationarity in the humoral immune response. *J. Theor. Biol.*, 129(2):141–162, 1987.

[70] J. D. Keene. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.*, 8(7):533–543, 2007.

[71] J.-H. Kim, V. Brachet, H. Moriya, and M. Johnston. Integration of transcriptional and posttranslational regulation in a glucose signal transduction pathway in *Saccharomyces cerevisiae*. *Eukaryotic Cell*, 5(1):167–173, 2006.

[72] G. Kirchhoff. Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. *Ann. Phys. Chem.*, 72:497–508, 1847.

[73] H. Kishino and P. J. Waddell. Correspondence analysis of genes and tissue types and finding genetic links from microarray data. In A. K. Dunker, A. Konagaya, S. Miyano, and T. Takagi, editors, *Genome Informatics*, volume 11, pages 83–95, Tokyo, JPN, 2000. Universal Academy Press.

[74] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[75] V. Klee, R. Ladner, and R. Manber. Signsolvability revisited. *Linear Algebra Appl.*, 59:131–157, 1984.

[76] R. R. Klevecz, J. Bolen, G. Forrest, and D. B. Murray. A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc. Natl. Acad. Sci. U.S.A.*, 101(5):1200–1205, 2004.

[77] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice*. Wiley VCH, Weinheim, DEU, 2005.

[78] L. Kuai, B. Das, and F. Sherman. A nuclear degradation pathway controls the abundance of normal mRNAs in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, 102(39):13962–13967, 2005.

[79] M. Laurent and N. Kellershohn. Multistability: a major means of differentiation and evolution in biological systems. *Trends Biochem. Sci.*, 24(11):418–422, 1999.

[80] D.-S. Lee and H. Rieger. Comparative study of the transcriptional regulatory networks of *E. coli* and yeast: Structural characteristics leading to marginal dynamic stability. *J. Theor. Biol.*, 248(4):618–626, 2007.

[81] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.

[82] C. Liebchen and R. Rizzi. Classes of cycle bases. *Discrete Appl. Math.*, 155(3):337–355, 2007.

[83] D. Lloyd and D. B. Murray. Redox rhythmicity: clocks at the core of temporal coherence. *Bioessays*, 29(5):465–473, 2007.

[84] S. Ma, Q. Gong, and H. J. Bohnert. An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Res.*, 17(11):1614–1625, 2007.

[85] A. Ma'ayan, A. Lipshtat, R. Iyengar, and E. D. Sontag. Proximity of intracellular regulatory networks to monotone systems. *IET Syst. Biol.*, 2(3):103–112, 2008.

[86] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:113, 2006.

[87] P. M. Magwene and J. Kim. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol.*, 5(12):R100, 2004.

[88] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[89] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular networks. *Nat. Protoc.*, 1(2):663–672, 2006.

[90] N. I. Markevich, J. B. Hoek, and B. N. Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J. Cell Biol.*, 164(3):353–359, 2004.

[91] J. Mata, S. Marguerat, and J. Bähler. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem. Sci.*, 30(9):506–514, 2005.

[92] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl 2):ii122–ii129, 2003.

[93] M. C. Palumbo, L. Farina, A. De Santis, A. Giuliani, A. Colosimo, G. Morelli, and I. Ruberti. Collective behavior in gene regulation: Post-transcriptional regulation and the temporal compartmentalization of cellular cycles. *FEBS J.*, 275(10):2364–2371, 2007.

[94] S. J. Parulekar, G. B. Semones, M. J. Rolf, J. C. Lievense, and H. C. Lim. Induction and elimination of oscillations in continuous cultures of *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.*, 28(5):700–710, 1986.

[95] P. R. Patnaik. Oscillatory metabolism of *Saccharomyces cerevisiae*: an overview of mechanisms and models. *Biotechnol. Adv.*, 21(3):183–192, 2003.

[96] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, MA, USA, 2000.

[97] J. R. Pomerening, E. D. Sontag, and J. E. Ferrell, Jr. Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2. *Nat. Cell Biol.*, 5(4):346–351, 2003.

[98] D. Porro, E. Martegani, B. M. Ranzi, and L. Alberghina. Oscillations in continuous cultures of budding yeast: A segregated parameter analysis. *Biotechnol. Bioeng.*, 32(4):411–417, 1988.

[99] T. Preiss, J. Baron-Benhamou, W. Ansorge, and M. W. Hentze. Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. *Nat. Struct. Biol.*, 10(12):1039–1047, 2003.

[100] H. Reinke and D. Gatfield. Genome-wide oscillation of transcription in yeast. *Trends Biochem. Sci.*, 31(4):189–191, 2006.

[101] F. Rodrigues, P. Ludovico, and C. Leão. Sugar metabolism in yeasts: an overview of aerobic and anaerobic glucose catabolism. In G. Péter and C. Rosa, editors, *Biodiversity and Ecophysiology of Yeasts*, The Yeast Handbook, pages 101–121. Springer, Berlin Heidelberg, 2006.

[102] A. Rodríguez, T. de la Cera, P. Herrero, and F. Moreno. The hexokinase 2 protein regulates the expression of the *GLK1*, *HXK1* and *HXK2* genes of *Saccharomyces cerevisiae*. *Biochem. J.*, 355(3):625–631, 2001.

[103] F. Rolland, J. Winderickx, and J. M. Thevelein. Glucose-sensing and -signalling mechanisms in yeast. *FEMS Yeast Res.*, 2(2):183–201, 2002.

[104] M. Ronen and D. Botstein. Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. *Proc. Natl. Acad. Sci. U.S.A.*, 103(2):389–394, 2006.

[105] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stümpflen, and H. W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, 36(Database issue):D646–D650, 2008.

[106] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, 34(Database issue):D394–D397, 2006.

[107] J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[108] P. M. Schlosser and M. Feinberg. A theory of multiple steady states in isothermal homogeneous CFSTRs with many reactions. *Chem. Eng. Sci.*, 49(11):1749–1767, 1994.

[109] R. Shalgi, M. Lapidot, R. Shamir, and Y. Pilpel. A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. *Genome Biol.*, 6(10):R86, 2005.

[110] C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[111] H. L. Smith. Systems of ordinary differential equations which generate an order preserving flow. A survey of results. *SIAM Rev.*, 30(1):87–113, 1988.

[112] H. L. Smith. *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, volume 41 of *Mathematical Surveys and Monographs*. Amer. Math. Soc., Providence, RI, USA, 1995.

[113] V. A. Smith, E. D. Jarvis, and A. J. Hartemink. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, 18(Suppl 1):216S–224S, 2002.

[114] H.-Y. Sohn, D. B. Murray, and H. Kuriyama. Ultradian oscillation of *Saccharomyces cerevisiae* during aerobic continuous culture: hydrogen sulphide mediates population synchrony. *Yeast*, 16(13):1185–1190, 2000.

[115] P. Solé and T. Zaslavsky. A coding approach to signed graphs. *SIAM J. Discrete Math.*, 7(4):544–553, 1994.

[116] E. D. Sontag. Monotone and near-monotone biochemical networks. *Syst. Synth. Biol.*, 1(2):59–87, 2007.

[117] C. Soulé. Graphic requirements for multistationarity. *ComPlexUs*, 1(3):123–133, 2003.

[118] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.*, 100(21):12123–12128, 2003.

[119] B. P. Tu, A. Kudlicki, M. Rowicka, and S. L. McKnight. Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, 310(5751):1152–1158, 2005.

[120] B. P. Tu and S. L. McKnight. Metabolic cycles as an underlying basis of biological oscillations. *Nat. Rev. Mol. Cell Biol.*, 7(9):696–701, 2006.

[121] B. P. Tu, R. E. Mohler, J. C. Liu, K. M. Dombek, E. T. Young, R. E. Synovec, and S. L. McKnight. Cyclic changes in metabolic state during the life of a yeast cell. *Proc. Natl. Acad. Sci. U.S.A.*, 104(43):16886–16891, 2007.

[122] R. C. Vallari, W. J. Cook, D. C. Audino, M. J. Morgan, D. E. Jensen, A. P. Laudano, and C. L. Denis. Glucose repression of the yeast *ADH2* gene occurs through multiple mechanisms, including control of the protein synthesis of its transcriptional activator, ADR1. *Mol. Cell. Biol.*, 12(4):1663–1673, 1992.

[123] Y. Wang, C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown. Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U.S.A.*, 99(9):5860–5865, 2002.

[124] A. V. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531, 2006.

[125] C. J. Wilusz and J. Wilusz. Bringing the role of mRNA decay in the control of gene expression into focus. *Trends Genet.*, 20(10):491–497, 2004.

[126] J. Wolf, H.-Y. Sohn, R. Heinrich, and H. Kuriyama. Mathematical analysis of a mechanism for autonomous metabolic oscillations in continuous culture of *Saccharomyces cerevisiae*. *FEBS Lett.*, 499(3):230–234, 2001.

[127] Z. Xu and K. Tsurugi. A potential mechanism of energy-metabolism oscillation in an aerobic chemostat culture of the yeast *Saccharomyces cerevisiae*. *FEBS J.*, 273(8):1696–1709, 2006.

[128] Z. Xu and K. Tsurugi. Role of Gts1p in regulation of energy-metabolism oscillation in continuous cultures of the yeast *Saccharomyces cerevisiae*. *Yeast*, 24(3):161–170, 2007.

[129] S.-J. Yan, J. J. Zartman, M. Zhang, A. Scott, S. Y. Shvartsman, and W. X. Li. Bistability coordinates activation of the EGFR and DPP pathways in *Drosophila* vein differentiation. *Mol. Syst. Biol.*, 5(278), 2009.

[130] M. K. S. Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. U.S.A.*, 99(9):6163–6168, 2002.

[131] Z. Yin, S. Wilson, N. C. Hauser, H. Tournu, J. D. Hoheisel, and A. J. P. Brown. Glucose triggers different global responses in yeast, depending on the strength of the signal, and transiently stabilizes ribosomal protein mRNAs. *Mol. Microbiol.*, 48(3):713–724, 2003.

[132] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, 2006.

[133] D. E. Zak, F. J. Doyle, G. E. Gonye, and J. S. Schwaber. Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In *Proceedings of the Second International Conference on Systems Biology*, pages 231–238, 2001.

[134] A. Zaslaver, A. E. Mayo, R. Rosenberg, P. Bashkin, H. Sberro, M. Tsalyuk, M. G. Surette, and U. Alon. Just-in-time transcription program in metabolic pathways. *Nat. Genet.*, 36(5):486–491, 2004.

[135] T. Zaslavsky. Signed graphs. *Discrete Appl. Math.*, 4(1):47–74, 1982.