Deciphering living networks

Perturbation strategies for functional genomics

Alberto de la Fuente

VRIJE UNIVERSITEIT

Deciphering living networks

Perturbation strategies for functional genomics

ACADEMISCH PROEFSCHIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Aard- en Levenswetenschappen
op maandag 3 juli 2006 om 13.45 uur
in het auditorium van de Universiteit,
De Boelelaan 1105

door

Alberto de la Fuente van Bentem

geboren te Amsterdam

promotoren:	prof.dr. H.V. Westerhoff
	prof.dr. J.L. Snoep
copromotor:	dr. P.J. Mendes

voor mama en papa

voor mijn broers

para Gaby

voor Alberto Xesús

The research described in this thesis was performed at the Virginia Bioinformatics Institute under the supervision of Dr. Pedro Mendes with co-supervision by Profs. Jacky Snoep and Hans Westerhoff. This work was supported financially by the Common Wealth of Virginia and by the National Science Foundation (NSF) of the United States of America.

Contents

1

Chapter 1: General introduction

Living systems are complex entities consisting of hundreds of thousands distinct molecules. It is the interaction between these molecules that generate the specific characteristics of living systems. In order to understand such complicated systems one has to integrate knowledge of the properties of their chemical constituents, make large-scale observations of these constituents and use theoretical approaches to analyze and interpret them. Here I introduce some concepts important for the study of complex biochemical systems. This will enable me to clearly state my thesis in the final section of this chapter.

Section 1.1: Systems biology

Molecular biology has recently reached a new stage, with completion of the sequencing of the full genomes of several organisms. Whereas a decade ago molecular biology was mainly concerned with the properties of molecules, now the attention is focused on understanding systems of interacting molecules and on how these systems give rise to properties, which we associate with life. Although systems approaches to biology systems have been suggested long ago [1-10], recently they have re-emerged under a common denominator, *i.e.* systems biology. The reason for the renewed interest in the systems approach is the new experimental technologies enabling the observation of hundreds or thousands of biochemical compounds such as RNAs, proteins and metabolites. Being able to do such large-scale parallel measurements of biochemical compounds calls for mathematical and information methodologies for the analysis and classification of those data, as well as for theory development in order to gain understanding of how those thousands of compounds are organized and function biologically [11].

Biologists have of course always known that living systems can not be understood solely by investigation the properties of the parts. Initially most biologists therefore restricted themselves to looking at the whole, *i.e.* at the physiology without necessarily relating that physiology to molecular constituents. The second half of the previous century, the reductionism approach to investigating complex biological systems has met with great success. In the corresponding disciplines biochemistry and molecular biology, the action mechanisms and structure of many macromolecules have been determined. Recently this has culminated in the determination of the sequence of the entire DNA and of most proteins of quite a few organisms.

That historical achievement for mankind also marked perhaps the culmination of the molecular biology era of Biology. The determined sequence does not only constitute the sequences of all proteins, but also the sequence of the entire genome, *i.e.* of the complete genetic material of the organisms, and therewith has an essential holistic aspect. The tendency of looking again at the whole, increased even further with subsequent breakthroughs in experimental technologies for measuring entire classes of chemical components of living organisms at the same time, *i.e.* in principle all mRNA's, all proteins in microorganisms and, soon, all metabolites. For the first time really one can go beyond studying the parts and study biology at a systems level without leaving open the issue that there could be more, immeasurable yet adjacent factors at play. One can study the properties of metabolic pathways rather than enzymes, study the properties of gene networks instead of single genes or operons, study the properties of signal transduction cascades instead of particular protein kinases, or in fact combine all levels of cellular organization in a single study, perhaps focused on a single cell function. Under the name systems biology, now efforts are taken to integrate the previously obtained knowledge into models of larger systems, accompanied with large-scale observation of the behavior of these parts in the *in vivo* context and the use of comprehensive tools such as mathematical modeling.

Mathematical models will provide additional insight into living systems and enable one to test hypotheses at rates that are impossible to do when only using wet-lab experimentation. The main goal of systems biology is to discover and mathematically formulate biological 'laws', which originate from the interactions between the parts of the biological system and can impossibly be discovered by studying the parts in isolation of the

biological context. Systems biology is not just a change from reductionist thinking to holistic thinking. The focus of systems biology is to understand the properties of systems in terms of the underlying mechanisms of component interactions.

A major change that is now pervading both molecular and cell biology is the shift from qualitative to quantitative biology. Whereas molecular biologists used to be satisfied with qualitative results, such as the presence/absence of a molecule, the binding/absence of binding of a repressor to its cognate piece of DNA, the 'new' biology requires more than that. It aims to quantify: how much is there, or how strong is the binding? Also the scale at which biochemical components are measured has changed. Typically, one biochemical component or a small set of components used to be investigated at a time, but systems biology aims to study large numbers of genes/proteins in parallel. Again, the possibility for systems level study of cell biology systems can be credited to the improvement of experimental techniques, which enable large scale as well as (semi) quantitative measurements of biochemical variables.

Section 1.2: Experimental technologies

The main constituents of living systems can be classified into three distinct categories: the hereditary information (*i.e.* genes and their activities: mRNA concentrations), proteins, which carry out the actual biochemical processes, and metabolites, *i.e.* small molecular compounds that provide free energy and material for maintaining the living state (although some carbohydrates are macromolecular, they still can be considered in this category). Since each of these categories contains molecules that are chemically very different from those of the others, three fields have emerged specialized in their measurements. Transcriptomics is the field concerning the genome scale measurement of gene activities, *i.e.* mRNA concentrations.

Microarrays [12-19] are becoming commonplace to measure mRNA concentration, although the technology is still plagued by a low signal-to-noise ratio. Quantitative RT-PCR [20] and related techniques may produce a much higher signal to noise ratio, be more accurate, but have the disadvantage that they are not genome wide.

Proteomics concerns genome-scale protein-concentration measurements. Proteomic technologies currently can measure one or two thousand different proteins and are therefore limited to organisms with small genomes, such as the prokaryotes. Most methods still rely on two-dimensional gels for separation [21], although higher throughput and resolution methods are being developed [22, 23]. Identification of the proteins is usually accomplished through mass spectrometry [24, 25]. As the concentrations of proteins are often less relevant for cell function than their actual activities, enzyme activity assays should still be upgraded towards high-throughput [26]. A good measure for protein activity is their degree of phosphorylation. Therefore, phosphoproteomics [27, 28], a currently rapidly developing field concerned with the large scale measurements of phosphorylated proteins, should show great potential for systems biology.

Metabolomics measures large-scale metabolite concentrations. Metabolites are harder to profile in a single run because they constitute many different chemical classes with widely different properties. The most promising technologies for this purpose are gas-chromatography coupled to mass spectrometry [29], liquid-chromatography mass spectrometry [30] and capillary electrophoresis with mass spectrometry [31]. It may be noted that the –omics suffix is often used, implying that the objective be to measure all, or a

large number (known and measurable) of the mRNAs, proteins and metabolites of the organisms. It should perhaps be added that the large scale aspect may not always be essential; networks that consist of only 15 components may already provide food for Systems Biology interest and may well be better served by more quantitative but less genome wide methodologies (*e.g.* [32, 33]).

Section 1.3: Mathematical modeling of biochemical systems

Science is an iterative process in which models are compared on the basis of how well they predict what is observed. Models are created to explain experimental observations that were made on a particular instantiation of a physical system, in much more generic terms. Because science is always after the understanding of the more general cases than of the particular unique instantiations in which the experiments were done, the results are discussed in terms of a virtual, more general model. By being a generalization away from the particular experimental case, a model needs to be a simplified representation of a real physical process; the characteristics of the particular experimental system on the particular day of the experiment need to be left out.

A 'best' or 'complete' model does not exist; some models may explain certain aspects of the physical process that is modeled, other models may explain other aspects. The process of modeling occurs by proposing a certain model to explain the observed phenomena and through a continuous process of validation of the model to new experimental observations. If the model can explain new experimental observations, and if it is internally consistent and if it is consistent with well-tested preexisting scientific theory such as thermodynamics, there are no reasons to reject it. If, however, the model is in disagreement with new experimental data, it has to be rejected or its use be restricted to its domain of validity, or modified until it does conform, if possible.

Cellular biochemistry can be viewed as a set of intricate networks composed of many and diverse interacting elements. Several types of networks are conceptually distinguished. Metabolic networks deal with small molecular substances, metabolites, which are connected (or maybe I should say 'wired' to emphasize the network analogy) by biochemical inter-conversions (enzymatic reactions). Signal transduction networks concern interactions between proteins, which can be complex formation or protein modifications, such as phosphorylations. Gene networks are high-level descriptions of gene regulatory processes in which only gene expression levels are considered, and genes are wired through their regulatory relationships. Biochemical networks (also called hierarchical networks in later chapters) are networks including all three types of molecules, mRNAs, proteins and metabolites, and interactions between them.

The dynamical behavior of these complicated networks is far from being clear intuitively and therefore we need to use mathematical models to help us understand how the systems behavior arises from the properties of the individual molecules. Except for a few metabolic pathways and just a few cases of gene expression, such mathematical models do not yet exist, however, primarily for lack of experimental data on the kinetics of their components. Ultimately such models should enable us to predict the effects of genetic mutations and of the environment on the system, thereby providing insight in the inner workings of living cells. Once properly validated against experimental observations, such predictive models, should be of great value in understanding the mode of action of drugs [34], identify drug targets [35, 36], identify the cause of diseases, and several biological

processes, such as infection and symbiosis. Ultimately they can be used as a design tools to engineer organisms [37, 38], or to act upon their environment to achieve other desirable outcomes [39].

Section 1.3.1: Mathematical frameworks

Different mathematical formalisms can be used for modeling. Perhaps the most common framework uses differential equations, either ordinary, when space is not taken into consideration, or partial, if spatial distribution is important. Modeling with ordinary differential equations (ODEs) usually occurs in the form of:

$$\frac{d}{dt}\mathbf{x} = \mathbf{N} \cdot \mathbf{v}\left(\mathbf{x}(\mathbf{p}), \mathbf{p}\right), \qquad \text{(Eq. 1.1)}$$

where $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is a vector of n state variables (the metabolite, protein or mRNA concentrations), $\mathbf{v} = \{v_1, v_2, \ldots, v_r\}$ is a vector of r kinetic rate functions (which depend on the state variables and a set of parameters given by vector **p**. These parameters include Michaelis constants, equilibrium constants, and maximal rates. **N** is an n x r matrix representing the chemical stoichiometries of all reactions.

This stoichiometry matrix effectively represents the structure of the chemical network, while any regulatory interactions are expressed in the functions of **v**. The functions in **v** are often non-linear in terms of **x** and **p**, however in special cases they could be linear. These non-linear functions correspond to the rate laws of enzyme kinetics, such as the Henri-Michaelis-Menten [40, 41], the Monod-Whyman-Changeaux [42] and even more complicated rate laws [43].

An alternative to such mechanistic models is to construct phenomenological equations for the right hand side of Eq. 1. Examples are linear models [7], lin-log kinetics [44] or power laws [8] that account for interaction between variables of the system without specifying a particular mechanism. The advantage of these phenomenological rate laws over the detailed mechanistic ones is that they depend on a smaller number of parameters and are mathematically easier to deal with. Linear models have the advantage that they can be solved analytically for any state of the system. A particular type of power law models, *i.e.* S-systems, can be solved analytically for steady states [8]. This is in general not possible with mechanistic rate laws, and therefore numerical analyses are required when dealing with models based on this type of equations. Mechanistic equations account for the molecular mechanisms of interactions, and are therefore more likely to be precise, though still approximations. Linear and power laws will be less suitable for extrapolation. On the other hand, the mechanistic models suffer from the fact that they require many parameters values to be known.

An approach that may combine some strength of the above approaches employs phenomenological equations that display the common characteristics of enzyme kinetics such as saturation and cooperativity. An example would be the use of the Hill equation [45, 46]. Using such equations keeps the number of parameters small, and yet common properties of enzyme kinetics, such as saturation, remain explicit.

Boolean functions have also been used, mostly in the modeling of gene networks (models describing genetic regulatory interaction, see Chapter 3) [4, 47], although in a few cases also for metabolic pathways [48, 49]. In this formalism variables can be in one of two

states. The state of each variable is determined by the states of the other variables that interact with it through Boolean functions. The Boolean approach is mathematically very tractable, but it is, however a very low-resolution method (cf. above). The Boolean approach has good educational features, since Boolean logic is easy to understand and the relevant concepts for biochemical modeling, such as point attractors, oscillations and stability, are readily demonstrated with these models. Therefore, this approach is useful to educate new systems biologist to gain insight into dynamical processes of large networks of interacting molecules.

Stochastic methods such as the Master Equation or Monte Carlo simulations have also been widely used; particularly for gene expression models [50-54] and for signal transduction models [55]. In this case the model is phrased in terms of numbers of molecules of each species rather than in terms of their thermodynamic potentials or ensembles averaged concentrations (as in ODEs). In stochastic simulations each particle has a certain probability to react with other particles, usually following distributions derived from statistical mechanics. Stochastic simulation is closer to the physical nature of biochemical processes, but is computationally expensive. The stochastic approach is particularly important when the number of particles of a specific species is low. At high numbers of molecules, the ODE approach is an accurate approximation and is then preferred, because of its lower computational cost. Other frameworks exist, but their use has been restricted to a few publications. For example Petri nets [56] have been employed [57-59], as well as process calculus [60].

The purpose of biochemical modeling is to understand and predict the dynamical properties of the system. There are two major ways to construct such models: the integrative approach assembles existing knowledge on isolated parts of the system into an integrated model, while the so-called global approach uses observations of the whole system to derive the model, perhaps arriving at mechanisms that detail the properties of the parts.

Section 1.3.2: Integrative approach

The integrative or 'bottom-up' modeling approach integrates knowledge acquired by means of reductionist methods. It assumes that the parts of the system (mostly enzymes) behave in the cell as they do in test tubes in dilute solutions, and as such it is possible to derive the properties of the whole system based on the properties of the isolated parts and their interactions. This follows the traditional approach of molecular biology and biochemistry, which is based on *in vitro* studies of purified components. It requires, by definition, that the parts be amenable to work with after purification. While the later is the case for most metabolic enzymes, and it is a well-established technique, there is much evidence that many enzymes work differently in the cell than *in vitro* [61, 62]. The integrative approach does deal with the differences between *in vivo* and *ex vivo*, by identifying the physical chemical origin for the difference in behavior, such as macromolecular crowding [63]. The problem is likely to be more important for pathways that depend on direct interactions between macromolecules, such as signal transduction (*cf.* [63]), transcription and translation, where gigantic molecular complexes have to assemble before the process can even begin. *In vitro* determination of parameters is laborious and the data appears scattered in hundreds of journal articles.

Despite these experimental difficulties, an international program led by Snoep and Westerhoff and named SiC!, for Silicon Cell [64], makes computer replicas for those

pathways for which sufficient kinetic information is available and these models are made available on the world wide web [65-68], ultimately to be integrated into whole-cell replica [64]. The goal is to construct a cell model based on precise quantitative description of the individual elements of the network. This contrasts with the *Mycoplasma genitalium* model [69-71], where only the interactions between the biochemical components are taken from literature sources, while kinetic functions and their parameters are set arbitrarily. Another large-scale project favoring the piecewise modeling approach is the Alliance for Cell Signaling [72] in which signal transduction pathways will be investigated experimentally in order to enhance quantitative modeling. Any molecular biology and biochemistry laboratory may contribute to these efforts in a piece-by-piece approach and indeed SiC now contains pathway models stemming from many different authors [65-68] (website: http://jjj.vbi.vt.edu/).

Validation of these models proceeds by comparing the model behavior with the behavior of the biochemical system *in vivo* under different conditions than those used for model calibration. In the case of glycolysis in *Trypanosoma brucei* the *in vivo* behavior of the pathway matched the behavior of the mathematical model, to the limited extent that this could be tested [33] showing the potential of this approach. For glycolysis in yeast, the *in vivo* behavior was quite different from the model's predictions when only the glycolytic enzymes were put in [32]. This led to the discovery of a role for trehalose phosphate synthethase in constituting a 'brake' on glycolysis. This clearly demonstrated that the SiC approach indeed can lead to discovery of new mechanisms.

Section 1.3.3: Global approach

In the global, or 'top-down modeling approach a different strategy is employed; it starts by collecting data from observations of the whole system (or large parts thereof) and then tries to infer a model consistent with those data. In this approach one must infer the interaction structure (stoichiometry and regulation) of the network, the kinetic functions of each step, and the values of the parameters of these functions. This may involve a three-step approach, or may be accomplished simultaneously.

The first step concerns the identification of the interaction structure of biochemical systems. There is much information about metabolic pathways, available in databases readily accessible through the World Wide Web, such as KEGG [73] and MPW [74] and similar for genetic interaction, such as GeneNet [75] and RegulonDB [76]. However, most of the structure of biochemical systems is unknown and yet to be discovered. Inferring the interaction structure of biochemical systems is currently a very popular research topic, which has resulted in a large body of literature. The idea here is to find out the network structure from measurements of the intact global system *in vivo*, for example using time series of mRNA concentrations after a change in the culture medium. Many approaches have now appeared for 'reverse engineering' gene networks from experimental data (see reviews by [39, 77, 78]). Most of these are based on multivariate statistical methods, such as Principal Component Analysis and Hierarchical Clustering Analysis. Others have used more challenging approaches like fitting data with genetic algorithms [79], graph theory methods [80], and Metabolic Control Analysis [81-84]. The main onclusion from the latter works is that one needs to perturb the expression of each single gene and measure the response of the whole system in order to be able to infer the complete structure of the network, which implies a huge experimental effort.

Once the structure of the system is known, the specific kinetic function of each reaction and its parameter values must be determined. A plausible approach to finding rate laws and their parameter values selects some fairly arbitrary rate functions, and then performs a least-squares fit to experimental data sets [85, 86]. If this were done for a number of rate laws and in different combinations, one would perhaps find a best set that explains the data. This process has been carried out with some success in an automated fashion by Koza and co-workers [87, 88], although they based their models on electric analogues. Provided the interaction structure of the model is correct, this process may select the appropriate kinetic functions as well as optimize for the parameter values.

In limited cases the form of the kinetic functions of enzymes is known and this information can be retrieved from current biochemical databases, such as BRENDA [89, 90] and MPW [74]. In this case the remaining task is to estimate the *in vivo* parameter values (vector **p** in Eq. 1.1). This is usually done through the same means as in the integrative method [85, 86], except for the larger scale in this case: more variables are considered simultaneously and thus there are many more dimensions in the parameter space. The increase in scale is, however, not trivial and large-scale fitting problems are hard to solve computationally.

Section 1.4: Software for biochemical calculations and simulations

Although some mathematical modeling can be done with just pencil and paper, computers are needed to solve the systems of equations. The main reason for having to us computers is that biochemical networks most realistically are described by non-linear differential equations for which no analytical solution exists. Therefore, the solutions to these equations are approximated by numerical means. To save pencils, paper and time it is therefore necessary to use computers.

All algorithms used in biochemical modeling can in principle be coded in C++, FORTRAN or any other programming language. Also general mathematical programs such as Mathematica, MLAB, and Matlab can be used. Most convenient are programs that have been developed specifically with biochemical network modeling in mind, such as GEPASI [85, 91, 92], SCAMP/Jarnac [93, 94], E-CELL [70], Virtual Cell [95], MIST [96], KINSIM [97], METAMODEL [98] and STOCHSIM [99]. Most of these use the ODE framework, but Virtual Cell uses partial differential equations and STOCHSIM uses a stochastic simulator. Recently, a large number of researchers involved in writing these software packages have agreed on a standard to allow each program to exchange models. This format is based on XML and has been named SBML, for Systems Biology Markup Language [100].

Section 1.5: Theoretical analyses for biochemical systems

There are several theoretical analyses to deal with biochemical systems, such as Biochemical Systems Analysis (BST) [8], Metabolic Control Analysis (MCA) [6, 7], Metabolic Pathway Analysis [101] and Flux Balance Analysis [102]. The former two are sensitivity analyses that express how the global system properties depend on the properties of the molecular components of the system. The latter two are analyses of the stoichiometry of metabolic networks. The types of biochemical systems that are dealt with in this thesis (hierarchical biochemical systems and gene networks) essentially consist of sub networks

that are not connected in terms of flux. The sub networks do interact through kinetic interactions, thus the latter two approaches are of less use. Since the differences between BST and MCA are subtle and MCA is a more intuitive approach (BST deals with phenomenological parameters that have no clear physical interpretation), I will further focus only on Metabolic Control Analysis. In the next chapter I will introduce an extension of MCA called Hierarchical Control Analysis, as developed by Kahn & Westerhoff [103], which explicitly deals with the flux disconnected models I consider here. I will also make extensive use of another variant of Metabolic Control Analysis, called Co-response Analysis [104, 105]. I will introduce the important concepts in the next section. Since I will often deal with matrices, I will conclude this chapter introducing the square matrix formulation of MCA [106].

Section 1.5.1: Metabolic Control Analysis (MCA) and Co-response Analysis

Metabolic control analysis (MCA) [6, 7] is a framework to describe biochemical networks and free-energy transduction systems. A central concept in MCA is the control coefficient, which is a measure of how sensitive a systemic variable (*e.g.*, the steady state concentration of an mRNA or metabolic intermediate) is to the change in the activity of any catalytic process in the network (*e.g.*, an enzymatic reaction or a transcription rate). That activity is often parameterized as its limiting rate, or as the effective concentration of the enzyme catalyzing the step [10]. Control coefficients are expressed as scaled derivatives of the steady-state value of variables with respect to the activities of the catalytic processes:

$$C_{v_i}^{A} = \frac{dA/A}{dv_i/v_i} = \frac{d\ln|A|}{d\ln|v_i|}, \qquad \text{(Eq. 1.2)}$$

Here, A refers to any system variable at steady state. Usually, intermediate concentrations, fluxes or free energies are considered. dv_i refers to the immediate (local) change in the activity of step i due to a change in a parameter specifically affecting that step. All such parameters for the other rates in the system remain unchanged. For each variable A, there are as many control coefficients, as there are steps in the biochemical system.

As is metabolic control itself, control coefficients are context dependent because, in general, biochemical kinetics are nonlinear and control coefficients are therefore derivatives of (double-logarithmically) nonlinear functions (see [107] for a higher order approach). Control coefficients reflect properties of the system as a whole. The concept of control coefficient is only meaningful in the context of an entire biochemical system; it is not a property of any individual reaction or its enzyme alone.

MCA also describes the kinetic properties of biochemical reactions in isolation, via elasticity coefficients, which are properties of the individual reactions in the network. The elasticity coefficients are defined by scaled partial derivatives of reaction rates with respect to the reaction effectors:

$$\varepsilon_{X_j}^{v_i} = \frac{\partial v_i / v_i}{\partial X_j / X_j} = \frac{\partial \ln|v_i|}{\partial \ln X_j} \qquad \text{(Eq. 1.3)}$$

X_j is the concentration of a certain effector and v_i is the rate of a certain reaction of the biochemical system. Local properties are properties of individual reactions and are independent of other reactions.

Co-control coefficients (Eq. 1.4) are ratios between control coefficients and represent how two system variables respond to a common rate perturbation [104, 105]:

$$^{v_m}O_{A_j}^{A_i} = \frac{C_{v_m}^i}{C_{v_m}^j} = \frac{dA_i / A_i}{dA_j / A_j} = \frac{d\ln|A_i|}{d\ln|A_j|} \qquad \text{(Eq. 1.4)}$$

As above, A_i and A_j are system variables and v_m is a reaction activity that has been perturbed. According to Eq. 1.4, the magnitude of the rate perturbation is not needed for calculating co-control coefficients (*cf.* control coefficients, Eq. 1). This greatly simplifies their experimental determination.

Regulatory Strengths [108] (Eq. 1.5) quantify the fractional changes in a system variable as a consequence of the change in another system variable through a specific reaction [108]. They quantify how perturbations propagate from one variable to another. Regulatory Strengths are also called partial internal response coefficients (see http://www.sun.ac.za/biochem/mcanom.html).

$$^{v_r}R_{A_j}^{A_i} = \varepsilon_j^r \cdot C_m^i = \frac{\partial v_r}{\partial A_j} \cdot \frac{A_j}{v_r} \cdot \frac{dA_i}{dv_r} \cdot \frac{v_r}{A_i} = \frac{\partial \ln|v_r|}{\partial \ln|A_j|} \cdot \frac{d\ln|A_i|}{d\ln|v_r|}. \qquad \text{(Eq. 1.5)}$$

The Regulatory Strength of the path through reaction r from variable A_j to A_i quantifies how variable A_i changes due to a change in v_m caused by a change in A_j. Summing up all the effects of A_j on A_i through all different network reactions (paths) connecting these two, one can obtain the connectivity theorems for concentration control [10, 108, 109] (see below).

It is important to realize that elasticity coefficients are local properties, *i.e.* properties of the individual reactions of the network; they quantify the effect of an effector concentration on the rate of reaction *while all other variables in the system are held constant* (or by isolating the reaction from the global system). This can be done for example by isolation of individual enzymes and evaluating their elasticity towards their effectors *in vitro*, as is common practice in enzyme kinetics. In each such experiment one of the modifiers is varied, while all other modifiers are kept at their *in vivo* concentrations. In fact, elasticity coefficients are scaled partial derivatives of the enzyme kinetic rate laws, evaluated at the physiological state. In contrast, control coefficients are global properties of the system that emerge from the collective action of all its elements. These global

properties have to be measured in the intact system. Several theorems of control analysis [6, 7, 103, 109] relate the local and global properties to each other. This will be explained in more detail in the next section in which the matrix formalism is used.

Figure 1.1 – *A simple two-step metabolic pathway. Source and product substances are omitted in this figure.*

For the two-step metabolic pathway depicted in Fig. 1.1, the control coefficient for the steady state concentration of the intermediate *M* can be written as:

$$C_{v_1}^{M} = \frac{1}{\varepsilon_M^{v_2} - \varepsilon_M^{v_1}}. \qquad \text{(Eq. 1.6)}$$

Here we see that the effect on the steady state concentration of M, of changing the rate of the first reaction in Fig 1.1 (for example, by increasing the concentration of the enzyme catalyzing this rate) can be expressed in terms of the local effects that the metabolite M would have on the rates of its production and its degradation in isolation from the rest of the system.

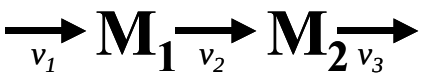

Figure 1.2 – *A simple three-step metabolic pathway. Source and product substances are omitted in this figure.*

If a linear three-step pathway is considered (Fig. 1.2) the equation becomes more complicated and depends not only on the sensitivities of the enzyme activities towards M_1 but also towards M_2:

$$C_{v_1}^{M_1} = \frac{\varepsilon_{M_2}^{v_2} - \varepsilon_{M_2}^{v_3}}{\left(\varepsilon_{M_1}^{v_2} - \varepsilon_{M_1}^{v_1}\right)\cdot\left(\varepsilon_{M_2}^{v_2} - \varepsilon_{M_2}^{v_3}\right) + \varepsilon_{M_1}^{v_2}\cdot\varepsilon_{M_2}^{v_2}}. \qquad \text{(Eq. 1.7)}$$

I show this to emphasize the fact that in a system of arbitrary size the control that each enzyme exerts is a function the properties of all other enzymes in the system.

Since these equations become very large even for systems of moderate size we will need to express these relationships in terms of matrices, which will compact the description and therefore make things more clear.

Section 1.5.2: Metabolic Control Analysis Matrix formulations

In Metabolic Control Analysis two different matrix formalisms are commonly used. One is the formalism by Reder [110], the other is the so called square matrix formalism [109]. Although the Reder formalism is mathematically more transparent, since it starts with differentiating the state equations, I prefer to use the square matrix formalism, because it starts directly by writing the summation and connectivity theorems of Metabolic Control Analysis in matrix format. Another advantage of this formalism is that it considers simultaneous the flux control and concentration control, so no separate proofs for each type of control coefficients are necessary (as needed for the Reder formalism). Furthermore, Co-response Analysis is originally formulated in this formalism as well as the generalization of Hierarchical Control Analysis [111] (which will be discussed in the next chapter).

1.5.2.1. The theorems of MCA; connectivity and summation

Ultimately it is the integration of all local properties of the biochemical steps in a pathway that sets the system's control properties, described by the control coefficients. The connectivity theorems for flux control [6] (Eq. 1.8) and concentration control [109] (Eq. 1.9) constitute the link between elasticities of individual enzymes and the control distribution in the pathway. For linear pathways, such as the ones depicted in Figures 1.1 and 1.2, the connectivity theorems are:

$$\sum_{r=1}^{n} C_{v_r}^{J_i} \cdot \varepsilon_{X_j}^{v_r} = 0\,, \qquad \text{(Eq. 1.8)}$$

$$\sum_{r=1}^{n} C_{v_r}^{X_i} \cdot \varepsilon_{X_j}^{v_r} = \begin{cases} -1, when\ i = j \\ 0, when\ i \neq j \end{cases}. \qquad \text{(Eq. 1.9)}$$

These theorems state that combining all regulatory effects of a metabolite through different paths on the flux will not modify the flux. Combining all regulatory effects of a metabolite through different paths will not modify any metabolite concentration, unless it is the effect of a metabolite on itself, in which case the effects sum up to –1, indicating that the metabolite will return to its steady state after a fluctuation [109].

Other important theorems are the summation theorems [6, 7] (Eqs. 1.10 and 1.11):

$$\sum_{r=1}^{n} C_{v_r}^{J_i} = 1, \qquad \text{(Eq. 1.10)}$$

$$\sum_{r=1}^{n} C_{v_r}^{X_i} = 0\,. \qquad \text{(Eq. 1.11)}$$

In words this simply means that if each enzyme in the metabolic pathway is increased by a factor ß, then the flux will increase a factor ß, while all metabolite concentrations will not change at all.

1.5.2.2. The square matrix expression

Equations 1.08 and 1.09 are not only valid for linear pathways, but for pathways of any structure in terms of their independent variables. It is always possible to reduce a system to one that only accounts for the independent fluxes and concentrations [110]. After reducing the system the theorems can be combined in a single matrix expression (Eq. 1.12).

$$\begin{bmatrix} \mathbf{C^J} \\ \mathbf{C^S} \end{bmatrix} \begin{bmatrix} \boldsymbol{\kappa} & -\boldsymbol{\varepsilon}\mathbf{L} \end{bmatrix} = \mathbf{I}. \qquad \text{(Eq. 1.12)}$$

Here $\boldsymbol{\kappa}$ is the scaled kernel of the stoichiometry matrix providing the relation between dependent and independent fluxes, $\mathbf{L}$ is a scaled link matrix, providing the relation between dependent and independent concentrations in the network, $\mathbf{I}$ is the identity matrix and $\boldsymbol{\varepsilon}$ is the matrix of elasticity coefficients. $\boldsymbol{\kappa}$ and $\mathbf{L}$ can be obtained directly from the stoichiometry matrix [110]. Both matrices in Eq. 1.12 are square and invertible (this assumption of invertibility is equivalent to the assumption of invertibility of the Jacobian matrix of a dynamical system, which is satisfied when the reference steady state is asymptotically stable). From Eq. 1.12 one can express the systemic properties of a biochemical system, *i.e*. the control coefficients, in terms of the kinetic properties of the individual enzymes, *i.e.* the elasticity coefficients:

$$\begin{bmatrix} \mathbf{C^J} \\ \mathbf{C^S} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\kappa} & -\boldsymbol{\varepsilon}\mathbf{L} \end{bmatrix}^{-1}. \qquad \text{(Eq. 1.13)}$$

Thus, with a complete knowledge of the stoichiometry and kinetics of all steps (in the form of elasticity coefficients) one can calculate the systemic properties in the form of control coefficients. Metabolic Control Analysis provides a means to integrate properties of single molecules to gain understanding of the properties of the whole system. As pointed out above, measuring elasticities can be done *in vitro*, by isolating steps from the network and studying them individually. For some single enzymes this can be done, but for complex processes like transcription, translation and other processes catalyzed by large protein complexes simulating the *in vivo* reaction *in vitro* is very difficult. Furthermore, it requires one to know all the effectors of all the rates. In addition, this approach is hampered by artifacts that derive from the harsh processes of isolating enzymes that may break enzyme complexes which work as a whole unit in the cell (see e.g. [112, 113]). Therefore, given these experimental difficulties, it might be interesting to consider the opposite direction: Eq. 1.12 can be inverted into Eq. 1.14, expressing local kinetic properties as a function of the global behavior of the system, which is more readily observable:

$$\begin{bmatrix} \boldsymbol{\kappa} & -\boldsymbol{\varepsilon}\mathbf{L} \end{bmatrix} = \begin{bmatrix} \mathbf{C^J} \\ \mathbf{C^S} \end{bmatrix}^{-1}. \qquad \text{(Eq. 1.14)}$$

This is an inverse problem, named "inverse MCA" [114], as one is using observations of effects to deduce the causes. Using Eq. 1.14 the values of the elasticities and pathway structure can be calculated after determination of all control coefficients [114]. Strictly speaking, Eq. 1.14 leads to $-\boldsymbol{\varepsilon}\mathbf{L}$, the product of the link matrix and the elasticity matrix, rather than to the elasticity matrix itself. When there are mass conservation relationships, *i.e.* $\mathbf{L} \neq \mathbf{I}$, one has to factorize the result, which can be done with the help of additional measurements, by perturbing the conserved moieties [105, 115].

1.5.2.3. Matrices of co-control and Regulatory Strengths

Regulatory Strengths are related to co-control coefficients by the following transformation of Eq. 1.14 [104, 105]:

$$\mathbf{D}^{\mathbf{C}}\left[\boldsymbol{\kappa} \quad -\boldsymbol{\varepsilon}\mathbf{L}\right] = \left(\begin{bmatrix}\mathbf{C}^{\mathbf{J}}\\ \mathbf{C}^{\mathbf{S}}\end{bmatrix}\left(\mathbf{D}^{\mathbf{C}}\right)^{-1}\right)^{-1}. \qquad \text{(Eq. 1.15)}$$

$\mathbf{D}^{\mathbf{C}}$ is a diagonal matrix that contains control coefficients on its diagonal. Its inverse $(\mathbf{D}^{\mathbf{C}})^{-1}$ is a diagonal matrix containing the reciprocals of these control coefficients.
If we take

$$\mathbf{D}^{\mathbf{C}}\left[\boldsymbol{\kappa} \quad -\boldsymbol{\varepsilon}\mathbf{L}\right]^{-1} \equiv \mathbf{R}^{\mathbf{C}} \text{ and } \begin{bmatrix}\mathbf{C}^{\mathbf{J}}\\ \mathbf{C}^{\mathbf{S}}\end{bmatrix}\left(\mathbf{D}^{\mathbf{C}}\right)^{-1} \equiv \mathbf{O}^{\mathbf{C}}, \qquad \text{(Eq. 1.16)}$$

Eq. 12 takes the form:

$$\mathbf{R}^{\mathbf{C}} = \left(\mathbf{O}^{\mathbf{C}}\right)^{-1}. \qquad \text{(Eq. 1.17)}$$

$\mathbf{O}^{\mathbf{C}}$ is a matrix of co-control coefficients and $\mathbf{R}^{\mathbf{C}}$ is a matrix containing control coefficients and Regulatory Strengths. Which particular set of control coefficients and Regulatory Strengths $\mathbf{R}^{\mathbf{C}}$ contains depends on which set of control coefficients is selected in $\mathbf{D}^{\mathbf{C}}$.

It has been shown that by using co-response analysis one can work out the values of all the control and elasticity coefficients [104, 105]. Since no knowledge of the perturbation size is needed (as long as it is small) Co-response Analysis may provide an experimentally simplified way for the measurements of control coefficients and elasticity coefficients. The disadvantage of this method is that all reactions need to be perturbed even if one wishes to determine a single control coefficient.

Section 1.6: Problem statement and summary

Currently, with modern experimental techniques, it is possible to measure the concentrations of a great many and ultimately all cellular constituents, such as mRNAs, proteins and metabolites. Given these experimental technologies, astronomical amounts of new data will appear. To enable us to see the forest for the trees, we need to find ways in which best to analyze the data so as to obtain better understanding of biochemical systems and predictive power. When those new ways of analyzing the data are found, this may

even lead to a preference for a certain type of data or certain experimental methodologies. This may then help direct experimentation towards the highest possible impact for understanding of the system. Ideally, the three levels of biochemical organization, *i.e.* mRNAs, proteins and metabolites, are studied all together in an integrated fashion. However, due to the number of components and complexity of such integrated systems it is reasonable to try to decompose the system and to study the subsystems or to use simplified descriptions of the whole system. It will be important to decompose the system into subsystems that behave in isolation in much the same way as they do when they are embedded in the whole system. This is exactly what I deal with in this dissertation; on the one hand I show how and when it is possible to study the systems properties of metabolism *in vivo*, ignoring the effects of gene and protein expression, and on the other hand I develop a quantitative concept in terms of Metabolic Control Analysis to describe the properties of the whole system in a simplified form, *i.e.* as a gene network a description of only the dynamics of gene expression without explicit account for metabolites and proteins. The latter reduces the system to one that only accounts for interactions between genes, and that can be used to model observations made by micro arrays. Furthermore, this concept enables the inference of the topology of such gene networks from experimental data. The analysis guides the experimenter towards the specific experiments that need to be done in order to be able to infer the interactions between genes on a genome scale.

In **Chapter 2** I will introduce the concept of hierarchical biochemical systems and show how to express their properties in terms of properties of the individual flux-disconnected modules it is composed of. I will propose several methods, which allow the quantification of properties of the individual modules as if they were isolated from the global system, from experimental data. In particular, I will focus on the study of metabolic systems. These analyses enable us to distinguish regulation that happens at the metabolic level only, from regulation that involves transcription or translation, thus quantifying the relative importance of each of these processes to the global systems behavior. Perturbation of the concentration of an enzyme enables us to measure the importance of the enzyme for the metabolic flux and metabolite concentrations, quantified by Control Coefficients. However, we don't have perfect experimental control on the concentration of the enzymes, as there may be metabolites that modify the expression of genes coding for the enzymes, thereby modifying the enzyme concentrations. When this is the case the initial intervention of the enzyme concentration has a primary effect on metabolism and successively a secondary effect due to the modification of concentration of enzymes through gene-expression regulation by that metabolism. It is interesting to decompose the total effect of the intervention into these two separate effects, because this enables to asses the relative importance of the metabolic system and gene–expression for metabolic regulation. The main train of thought behind the first method outlined in Chapter 2 is that metabolic processes generally occur at a much higher rate than processes of gene expression and protein synthesis and that the secondary effects can be neglected on a short timescale. On the short timescale the effects of the perturbation of the enzyme concentration are solely determined by the properties of the metabolic system. Two other methods rely on drastic experimental modifications of the system such as inhibition of transcription or translation or removing regulatory sequences from the enzyme coding genes thereby physically removing feedback loops from metabolism to gene-expression. The fourth method relies on experimental determinations of, in addition to the metabolic fluxes and metabolite concentrations, also the enzyme concentrations, which enables us to separate primary and secondary effects by a mathematical operation compensating for the secondary changes in

the enzyme concentration, thereby 'mathematically removing' feedback loops from metabolism to gene-expression. I will verify the experimental applicability of these methods by analyzing data obtained by simulation of a biochemical system.

In **Chapter 3** I introduce the concept of a gene network and briefly describe previously proposed modeling approaches. 'Gene networks' are network models in which the nodes represent genes or their activities (mRNAs) and the edges correspond to regulatory interactions between them. Such models are highly phenomenological because they do not represent explicitly the proteins and metabolites that mediate those interactions. I will show the use of Regulatory Strengths to quantify gene-gene interactions and show how to express these coefficients in terms of the biochemical system underlying these interactions. This approach establishes a clear and formal link between the phenomenological gene network modeling and more detailed approaches considering the hierarchical structuring of biochemical networks as introduced in Chapter 2.

Unraveling the structure of gene networks is an important step in the process of understanding the properties of living cells and therefore also a very active research program. In **Chapter 4** I will first review previously proposed approaches to infer gene networks from gene expression data. Then, I will briefly describe previously proposed approaches to use Metabolic Control Analysis to infer properties of the biochemical constituents from observations of the whole system, *i.e.* " the inverse MCA". In the "conventional MCA" it has been shown that when the *local* properties of the components of the system have been experimentally determined (elasticities) one can calculate the *systemic* properties of the system as a whole (control coefficients). Inverse MCA establishes the exact opposite; one measures the properties of the global system and calculates the properties of the components. I will propose a modification of one of these inverse MCA approaches to enable the inference of the regulatory structure of gene networks in terms of Regulatory Strengths from gene expression data. The proposed methodology relies on experiments in which the expression rates of individual genes are perturbed and the steady state responses in global gene expression levels are measured, which enables the calculation of co-control coefficients (system properties). When all genes in a network of interest are perturbed and the responses measured, the local properties of the components (mRNAs) such as the presence of direct regulatory effects (not mediated by any other gene) between genes can be calculated, thereby elucidating the interaction structure and quantifying it. The method is evaluated by applying it to data produced with several mathematical models of gene networks. Finally, a large set of simulated data on large gene networks is analyzed to thoroughly evaluate the proposed method.

Chapter 5 is the general discussion, in which I will describe the place of my work in current functional genomics and systems biology research. I will compare the gene network inference method described in Chapter 4 to recent variants that appeared in the literature and generalize the approach to deal with other types of biochemical systems. At the end, I will point out which improvements could be made to the method in particular in the light of the limited quality of currently produced data sets.

2

Chapter 2: Hierarchical biochemical systems

Traditional analyses of the control and regulation of steady-state concentrations and fluxes assume the concentrations of the enzymes to be constant. In living cells, a hierarchical control structure connects metabolic pathways to signal-transduction and gene-expression. Consequently, enzyme concentrations are not generally constant. This would seem to compromise analyses of control and regulation at the metabolic level. In this chapter the concept of hierarchical biochemical systems is introduced and it is explained under which conditions it is possible to study parts of the system in isolation, *i.e.* to study metabolism without taking changes in gene and protein expression into account. The analysis makes it possible to quantify both the control of a metabolic step exerted in the context of metabolism alone, and the control exerted by that step in the global system comprising regulated gene-expression and signal transduction.

Section 2.1: What are hierarchical systems?

The central dogma of molecular biology states that DNA specifies mRNA, which specifies protein, which specifies function. This conveys a 'dictatorial' view of the regulation and control of cell functioning, *i.e.* it suggests that the DNA dictates everything that happens inside the cell. Figure 2.1 shows such a dictatorial system in which mRNA (T, for transcript) is produced by transcription, and then protein P is produced by translation, and the protein has a function, in this case as an enzyme producing metabolite M. The communication between the levels runs downward only; transcription is not supposed to know anything about what is happening at the lower levels. Systems in where gene expression itself is not sensitive to anything that happens down in the cell are called "dictatorial hierarchical systems" [116] (in the terminology of Hierarchical Control Analysis, which will be introduced in a later section).

The above representation of biochemical regulation is highly simplified: transcription, which is depicted here as a single step, is actually a process consisting of hundreds of elementary steps. Similarly, the processes of transcript degradation, translocation, translation, protein translocation and degradation, and the metabolic processes consist of many elementary reactions. For simplicity, to omit unnecessary details, each such series of elementary steps is here summarized by an overall process.

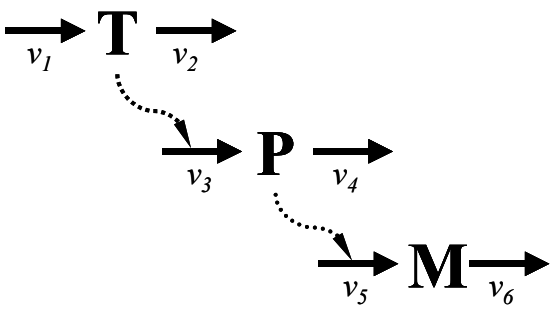

Figure 2.1 – *A scheme of a dictatorial hierarchical system. T stands for transcript (mRNA), P for Protein and M for metabolite. Thick arrows represent the fluxes of production and degradation processes. The dashed arrows stand for regulatory interactions; the transcript enhances protein production, the protein catalyzes the metabolite production. Input and output substances are omitted in the figure.*

In hierarchical control analysis it has been noted [103] that even in a dictatorial hierarchy not all control resides at the top level of the system, *i.e.* in transcription. Kinetic properties of processes lower in the hierarchy still partly control whatever happens at the same level of the hierarchy, plus what happens at levels lower in the hierarchy, but not what happens at levels higher in the hierarchy. If living cells would indeed consist of such dictatorial systems, understanding gene regulation plus the understanding of how mRNA levels regulate protein levels plus the understanding of how enzyme activities regulate function, would entail the understanding of how the biochemistry of the cell is regulated.

In many living cells there is feedback regulation of gene expression by the proteins and metabolites. Therefore, the central dogma should be modified slightly to

account for these additional interactions. Figure 2.2 shows such a scenario where gene expression regulates a certain metabolic process, but in turn the expression of that gene is subject to regulation by a corresponding metabolite. Such a system can produce the homeostatic behavior often observed in biochemical systems: gene expression increases transcript T concentration, which increases the rate of translation, leading to higher protein P concentration, which increases the synthesis rate of the metabolite M, leading to higher concentration of the metabolite. The metabolite communicates to the level of gene expression that enough is being produced and slows down the rate of transcription (by for example, binding certain transcription factors, which in turn bind to or dissociate from the gene's promoter region to shut off transcription). As a consequence, the increase in gene expression caused by the initial activation thereof is not only determined by that activation itself but also by its effects at the functional level.

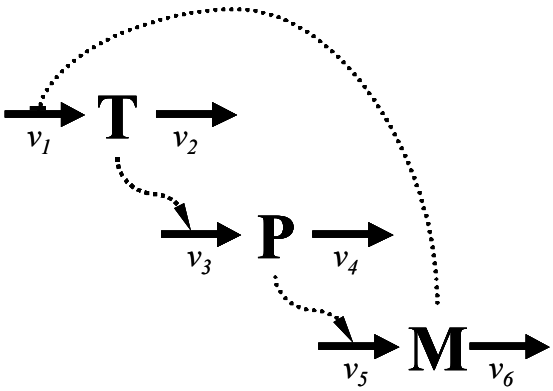

Figure 2.2 – *A schematic drawing of a hierarchical system. T stands for transcript (mRNA), P for Protein and M for metabolite. Thick arrows represent the fluxes of production and degradation processes. The dashed arrows stand for regulatory interactions, the transcript enhancing protein production, the protein catalyzing the metabolite production and the metabolite inhibiting mRNA synthesis*

The system in Fig. 2.2 is called a "democratic hierarchical system" [116], since the arrows do not point only from transcription down to metabolism, but also from metabolism up to transcription (such that the 'opinions' of all components matter). Other democratic scenarios include regulation of translation (or protein degradation) by metabolites and/or regulation of transcription (or transcript degradation) by proteins.

Section 2.2: Timescale separation

Traditionally, metabolism and gene expression have been studied and modeled separately. The justification for such a separation usually rests on the assumption that these two processes operate on widely different timescales, the rates of metabolic processes being much faster than the rates of mRNA and protein turnover. This assumption let investigators to study metabolism without taking account of changes in enzyme concentrations; the enzyme concentrations are assumed to be at a steady-state level that is unresponsive to changes at the metabolic level, at the timescale at which metabolic processes occur. The

enzyme concentrations thus appear as parameters rather than variables in mathematical models of metabolic systems. [In our formalism the term parameter is used for a property that is constant. It can only be changed by an action from outside, *i.e.* it does not vary when the system itself varies dynamically. In other formalisms, these parameters are called independent variables. In our formalism the term variable is used for any dependent variable, which is free to fluctuate and vary in time as the system changes dynamically. Only when the system is at steady state, a variable is also constant in time (*cf*. [10]).

The assumption of a well-defined time separation between metabolic and gene expression processes also underlies the formulation of gene-network models in which metabolism is not explicitly incorporated. These gene network models contain the hidden assumption that metabolism is so fast that it could be considered at steady state, such that all metabolic factors that could affect gene expression are included in constants regulatory terms. Formally, both assumptions amount to the application of the quasi-steady state assumption [117, 118]. Again, this assumption has rarely been tested and its validity should be questioned in actual cases of regulation of cell function.

Using mathematical models the validity of the assumption of time separation could be tested. Presence of a time-scale separation was evaluated by the ability to observe a quasi-steady state equal in the complete model (with gene regulation) to the steady state in the isolated metabolic model (without gene regulation). However, in the best case examined, the time scale of gene expression processes had to be some 3 orders of magnitude slower (time scales being expressed as transient times) than metabolism for the assumption to hold true [119-121]. Other examples (Mendes and Snoep, unpublished) showed that three orders of magnitude separation between time scales were not enough (five to six were required in that case). It seems that adopting the assumption of time separation between gene expression and metabolic processes blindly is rather dangerous, because the stability of specific mRNA and protein species, which mostly determine their time scales, vary widely [122, 123]. In general, it seems one would be best served by explicitly combining metabolic and gene expression processes in biochemical models.

Section 2.3: Modeling metabolism in combination with gene expression

Mathematical models describing metabolic pathways but not gene expression have been described as far back as 1960 [for example 32, 33, for example 124, 125-136]. In parallel to this, models of gene expression systems that did not include metabolism have also been abundant [4, 47, 137-143].

The first time metabolism and gene expression were combined in models was shortly after Jacob and Monod's presentation of their operon concept [144]. Goodwin proposed an abstract mathematical model that followed the characteristic of the operon and performed simulations with an analog computer [145]. The model consisted of three variables, a metabolite, a protein, and an mRNA, each being processed by two reactions, *i.e.* synthesis and degradation (see Fig. 2.2).

Later, Griffith performed analytical mathematical work on slightly modified Goodwin equations [146, 147]. These works were purely theoretical, and interesting for biochemical regulation from the fundamental point of view. Bliss *et al*. proposed a model with more biological reality [148]; experimentally determined parameter values were collected from the literature to formulate a model of the *Escherichia coli* tryptophan operon. The model considered both repression of gene expression by L-tryptophan and

feedback inhibition of the enzyme anthranilate synthase by L-tryptophan and included time delays in transcription and translation matching the observations. The simulation results obtained were in qualitative agreement with experimental observations [148]. Using most of the parameter values as in Bliss' model, Sinha refined the tryptophan operon model by taking into account, in a more detailed way, the interactions among the repressor molecules, the promoter, and tryptophan [149]. He investigated the behavior of the system over a wide range of parameter values. Sen further improved this model by setting the consumption rate of tryptophan to follow non-linear kinetics [150]. More recently, Santillan and Mackey [151] continued to improve this model by including metabolic feedback and transcriptional attenuation. Their model is in good qualitative agreement with experimental observations of the metabolic shift when cells are moved from growing in tryptophan-rich medium to minimal medium [151]. After 20 years, there finally seems to be a good model of the *trp*-operon of *E. coli*!

Lee and Bailey developed a model of the *lac* operon [152, 153]. They derived very detailed equations accounting for binding of several effectors to the operator and promoter. The behavior of the model was studied in response to DNA mutations, and plasmid copy number. Also Wong *et al.* [154] formulated a model of the *lac* operon that accounts for the induction by lactose and includes the mechanisms of catabolite repression and inducer exclusion. Kremling *et al*. [155, 156] incorporated metabolism, genetics and regulation in a very detailed manner. Their model was validated against wild type and different mutants, matching biomass production, glucose and lactose consumption and beta-galoctasidase synthesis during the diauxic shift. Although their model showed good agreement with the experiments, the validation focused on just 4 out of 30 variables in the model.

Van Dien and Keasling formulated a mathematical model of the *E. coli pho*-regulon, with which they modeled the phosphate starvation response [157, 158]. Their model reproduced the experimentally observed characteristics of the starvation response rather well. The single operon models and the *pho*-regulon model referred to above are very small compared to the real biochemical networks consisting of hundreds or thousands of genes and metabolites, and potentially millions of proteins, since alternative splicing is a common process in higher organisms. Tomita and co-workers have been pioneers in constructing genome scale models [69-71]. They formulated a large model of a *Mycoplasm genitalium* cell. Although this is the smallest non-viral genome for which a full genome sequence is known, the model is very large compared to any previous ones. The initial model describes the behavior of a fraction of all *Mycoplasm* genes (127), which include large parts of central metabolism, and contains a total of 495 reactions. The major problem facing this effort (and any other at this scale) is that there is very little information on the values of the kinetic constants and, worse, on what are the kinetic functions for each reaction. So far this model has not been founded on experimental observations, rather information from similar reactions in other organisms has been used or parameters have been set to arbitrary values. The model has also not been validated, at least not shown in their publications [69-71]. Nevertheless, this effort has had the quality of highlighting the difficulties with the construction of genome-scale models. Another large scale modeling effort combining gene expression and metabolism has been started for *E. coli* [159].

Section 2.3: Analysis of hierarchical systems; local versus global description

The MCA framework was formulated originally for metabolic systems, where the enzyme concentrations are taken as parameters. This restriction has been relaxed to be able to deal with systems that are hierarchical in nature, such as gene expression [103, 111, 116, 160] and signal transduction [103]. Hierarchical Control Analysis (HCA) [103, 111, 116, 160, 161] is an extension to MCA that explicitly accounts for the control exerted by subsystems not connected to the pathway by mass flow, only by kinetic effects. When it deals with networks that include gene expression, HCA considers the enzyme activities themselves as variables of the system since they change due to translation, proteolysis, binding to other proteins, and covalent modification. It is also possible to consider mRNA concentrations explicitly, which are also variables due to transcription and degradation. In this setting, it has been shown [116] that transcription and translation participate in the control of the metabolic flux.

The distribution of control in the full system, referred to as *integral control* [111], can be shown to be expressed in terms of the control of the modules in isolation, termed *intra-modular control* [111], and the sensitivity of the modules to each other, *inter-modular response* [111]. Integral or global control coefficients reflect the control when the entire system is allowed to adjust to a new steady state upon a modulation. They reflect total regulation, both within the module and within the system as a whole. Below, intra-modular control coefficients are indicated by lower case 'c' and integral control coefficients by capital 'C'.

In a similar fashion as MCA integrates properties of individual molecules (elasticities) into properties of the whole system (control), in HCA the local control properties of the individual modules are integrated into properties of the whole system. In Chapter 1 it was shown that the properties of a whole system (control) can be expressed in terms of the properties of the individual elements of the system (elasticities). I here reproduce some of what is explained in the Chapter 1 for reasons of clarity. Again, a small metabolic pathway is used as an example.

Figure 2.3 – *A simple two-step metabolic pathway*.

Recall that the control coefficient for the steady state concentration of the intermediate M can be written as:

$$c_{v_1}^{M} = \frac{1}{\varepsilon_M^{v_2} - \varepsilon_M^{v_1}}. \qquad \text{(Eq. 2.1)}$$

Again, it is seen that the effect of changing the rate of the first reaction in the model depicted in Fig 2.3 can be expressed by the local effects that the metabolite has on both the rates of its production and degradation.

If an additional level to the pathway in Fig. 2.3 is introduced, *i.e.* the synthesis of the enzyme catalyzing the first metabolic reaction, supposing that it is regulated by the concentration of the metabolite M, the system of Fig. 2.4 is obtained.

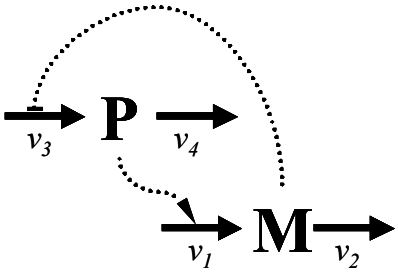

Figure 2.4 – *A simple model of a democratic gene-expression hierarchy. A protein regulates the metabolite production step; in turn the protein synthesis is inhibited by the metabolite.*

The hierarchical control can be expressed in terms of the intra level control (the control of each isolated subsystem, as Eq. 2.1 for the control in the metabolic subsystem) and the elasticities accounting for the interaction between the subsystems, *i.e.* the effect of the protein on the rate of metabolite production and the effect of the metabolite on the rate of protein synthesis:

$$C_{v_1}^{M} = \frac{c_{v_1}^{M}}{1 - \varepsilon_M^{v_3} c_{v_3}^{P} \varepsilon_P^{v_1} c_{v_1}^{M}} \qquad \text{(Eq. 2.2)}$$

Again, the capital C represents the control this step has in the whole integrated (hierarchical) system. Lower case c reflects the control that steps have if one were to isolate the subsystems from the global system. This framework is then capable of integrating the properties of the metabolic system with those of the gene-expression machinery and provides a detailed quantification of the importance of these in the whole system.

The hierarchical control can also be expressed in terms of elasticity coefficients only:

$$C_{v_1}^{M} = \frac{1}{\varepsilon_M^{v_2} - \varepsilon_M^{v_1} - \dfrac{\varepsilon_M^{v_3} \varepsilon_P^{v_1}}{\varepsilon_P^{v_4} - \varepsilon_P^{v_3}}} \qquad \text{(Eq. 2.3)}$$

The hierarchical control (Eqs. 2.2 and 2.3) is different from the metabolic control (Eq. 2.1) as a result of the interaction between gene expression and metabolism. Whether the hierarchical control is higher or lower than the metabolic control depends of the signs of $\varepsilon_M^{v_3}$ and $\varepsilon_P^{v_1}$, which quantify the activation (positive value) or inhibition (negative value) of reaction 3 by the metabolite and of reaction 1 by the enzyme, respectively. Although the precise values of elasticity coefficients vary between steady states, equations 2.1-2.3 are

valid for any kinetics and steady states of a system with this particular structure (Fig 2.4). Hierarchical Control Analysis is general and equations like Eq. 2.1-2.3 can be derived for any system of arbitrary complexity [111, 161]. Again, for even moderately large systems the equations of control coefficients in terms of elasticities become cumbersome. Matrix equations can be used to keep things clear [103, 111, 160, 161].

DNA supercoiling [162, 163] in living *E. coli* has been analyzed with Hierarchical Control Analysis [163]. DNA supercoiling is a mechanism by which bacteria compact their chromosome. The level of DNA supercoiling in the prokaryotic cell changes in response to changes in various extra-cellular conditions, such as temperature, osmolarity, pH and shifts between carbon or free-energy sources, or oxygen availability [164]. In turn, the supercoiling state of DNA affects the expression of numerous genes [165]. DNA topology may thus convert environmental signals into changes in gene expression. DNA gyrase introduces negative supercoiling in the DNA by coupling the reaction to ATP hydrolysis, and topoisomerase I relaxes negatively super-coiled DNA. The genes encoding topoisomerase I and DNA gyrase are among the genes that respond to changes in the level of DNA supercoiling. The expression of the DNA gyrase genes is highest when the level of DNA supercoiling is low and the expression of the topoisomerase I gene is stimulated by negative supercoiling. A surprising conclusion was that both proteins have low control over DNA supercoiling [163]. By blocking enzyme synthesis (*i.e.* effectively removing the feedback) intra level control could be measured, which was much higher than the control in the integrated system.

Section 2.5: Measuring local control in whole systems

The issue of the diverse mechanisms through which living cells are controlled is quite relevant in the realm of functional genomics. Whilst there has been an initial emphasis on the transcriptome as representative for function, more recent work [166] has begun to emphasize that the metabolome is where function resides. Rather than it being an issue of either-or, it is reasonable to assume that both metabolic and gene expression regulation are important. In every specific case one should quantify each one's contribution to regulation. Here, four methods are described that enable us to quantify, in addition to the control in the whole system, also the control in subsystems.

To illustrate the proposed methodology with maximum clarity, one of the simplest possible model systems that contains the essence of the problem is used, *i.e.* the simplest possible metabolic pathway that is subject to regulation by itself through the synthesis of a new enzyme (Fig. 2.5).

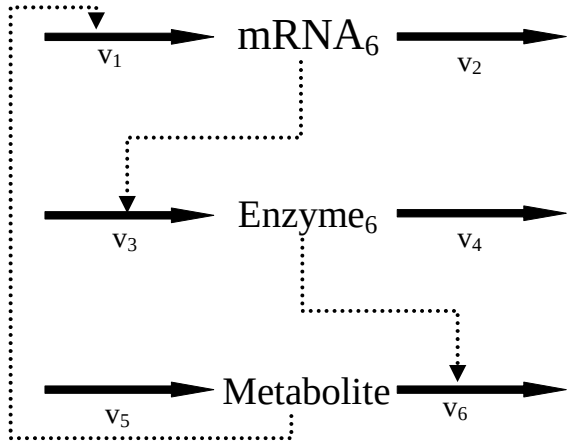

Figure 2.5 – *The model system. Interactions between the different levels (dotted arrows) run through the dependencies of (i) translation on mRNA concentration, (ii) the metabolic rate on enzyme concentration and (iii) the activation of transcription by the metabolite. Solid arrows indicate mass flow at the mRNA, protein and metabolic levels. Although both reactions 5 and 6 are catalyzed by mRNA-encoded proteins, this is only shown explicitly for reaction 6. This simplifies the model without detracting from the essence of hierarchical regulation. Accordingly, the model only takes into account this route for regulation through gene expression, effectively assuming that the gene encoding the enzyme of reaction 5 is expressed constitutively.*

In spite of its simplicity, the model of Fig. 2.5 should be sufficiently interesting because it mimics the basic structure of hierarchical biochemical systems including some routes along which the hierarchical levels communicate to each other. Rates of all 6 reactions of this model are given by Eqs. 2.4:

$$v_1 = \frac{V_1 \dfrac{[nucleotides]}{Km_1}}{1 + \dfrac{[nucleotides]}{Km_1} + \dfrac{K_a}{[Metabolite]}},$$

(Eqs. 2.4)

$$v_2 = k_2[mRNA], \quad v_3 = k_3[mRNA], \quad v_4 = k_4[Enzyme],$$

$$v_5 = \frac{V^f \dfrac{[S]}{K_{mS5}} - V^r \dfrac{[Metabolite]}{K_{mP5}}}{1 + \dfrac{[S]}{K_{mS5}} + \dfrac{[Metabolite]}{K_{mP5}}}, \quad v_6 = [Enzyme] \frac{k_{cat}{}^f \dfrac{[Metabolite]}{K_{mS6}} - k_{cat}{}^r \dfrac{[P]}{K_{mP6}}}{1 + \dfrac{[Metabolite]}{K_{mS6}} + \dfrac{[P]}{K_{mP6}}}.$$

Here [*S*] is the concentration of the pathway's substrate, [*Metabolite*] the concentration of the metabolite, [*mRNA*] the concentration of the messenger, [*Enzyme*] the concentration of the enzyme, [*nucleotides*] the concentration of nucleotide triphosphates, V_1 the limiting transcription rate, K_{m1} the Michaelis constant for nucleotide triphosphates, K_a the activation constant of transcription by the metabolite, k_2 the rate constant for mRNA degradation (and dilution due to cell growth), k_3 the translation rate-constant, k_4 the enzyme degradation rate-constant (and dilution due to cell growth), V the limiting rate of reaction 5, K_{eq5} the equilibrium constant for reaction 5, K_{mS5} the Michaelis constant for the substrate, K_{mP5} the Michaelis constant of reaction 5 for the metabolite, k_{cat} the catalytic rate constant of the enzyme of step 6, K_{eq6} the equilibrium constant of reaction 6 and K_{mP6} the Michaelis constant for the pathway's product. It should be noted that step 5 is enzyme-catalyzed and its enzyme concentration is implicit in V.

In order to determine the metabolic intra-modular control coefficients, the two upper modules or the feedbacks from metabolism to these, have to be ignored – thereby isolating the metabolic part from the global system. In this case the enzyme and mRNA concentrations are assumed constant (at their steady state levels). When the enzyme concentration becomes constant its product with k_{cat}, in the numerator, becomes a parameter itself (V, known as the limiting rate).

In this model, the units of the kinetic constants and time are arbitrary. Since it was not our intention to mimic any known system here, but rather to illustrate how the proposed methods work.

Table 2.1 lists the values of the rate constants used in the simulations. The values of the intra-modular control coefficients and the integral control coefficients under these conditions are listed in Table 2.2. Note that with respect to the simple metabolic control distribution, in the hierarchical system the flux control is distributed differently and the concentration control has reduced due to the homeostatic effect of the feedback loop.

Table 2.1 – *Parameter values used in the simulations of the model systems described in Fig. 1 and Eqs. 5-10*

Rate	Parameter	Value
v_1	V_1	0.01
	[*nucleotides*]	1
	K_{m1}	10
	K_a	100
v_2	k_2	0.01
v_3	k_3	0.1
v_4	k_4	0.01
v_5	V^f	100
	V^r	1
	[*S*]	1
	K_{mS5}	1
	K_{mP5}	10
v_6	k_{cat}^f	100
	k_{cat}^r	1
	[*P*]	0.1
	K_{mS6}	10
	K_{mP6}	1

Table 2.2 – *The values for the control coefficients according to Eq. 2.4, obtained using the elasticity coefficients calculated numerically at the standard parameter set in Table 2.1. Simulations were carried out with an Intel Pentium III 733 MHz computer with the biochemical simulation package Gepasi [85, 91, 92].*

Type of control	**Control coefficient**	**Value**
Intra-modular	$c_{v_5}^{J_{ss}}$	0.28
	$c_{v_6}^{J_{ss}}$	0.72
	$c_{v_5}^{[X]_{ss}}$	1.10
	$c_{v_6}^{[X]_{ss}}$	-1.10
Integral	$C_{v_5}^{J_{ss}}$	0.60
	$C_{v_6}^{J_{ss}}$	0.40
	$C_{v_5}^{[X]_{ss}}$	0.61
	$C_{v_6}^{[X]_{ss}}$	-0.61

The intra-modular control coefficients, to be indicated by lower case c's, can be written as functions of the elasticity coefficients:

$$c_{v_5}^{J} = \frac{\varepsilon_X^{v_6}}{\varepsilon_X^{v_6} - \varepsilon_X^{v_5}}, \qquad \text{(Eq. 2.5)}$$

$$c_{v_6}^{J} = \frac{\varepsilon_X^{v_5}}{\varepsilon_X^{v_5} - \varepsilon_X^{v_6}}, \qquad \text{(Eq. 2.6)}$$

$$c_{v_5}^{X} = \frac{1}{\varepsilon_X^{v_6} - \varepsilon_X^{v_5}}, \qquad \text{(Eq. 2.7)}$$

$$c_{v_6}^{X} = \frac{1}{\varepsilon_X^{v_5} - \varepsilon_X^{v_6}}. \qquad \text{(Eq. 2.8)}$$

Where X stands for the metabolite concentration, and J for metabolic flux. When considering the whole system, the integral control coefficients (indicated by capital C's) can be derived similarly:

$$C_{v_5}^{J} = \frac{\varepsilon_X^{v_6}}{\varepsilon_X^{v_6} - \varepsilon_X^{v_5}} + T, \qquad \text{(Eq. 2.9)}$$

$$C_{v_6}^{J} = \frac{\varepsilon_X^{v_5}}{\varepsilon_X^{v_5} - \varepsilon_X^{v_6}} - T, \qquad \text{(Eq. 2.10)}$$

where:

$$T = \frac{\varepsilon_X^{v_1}\varepsilon_X^{v_5}\varepsilon_N^{v_3}\varepsilon_E^{v_6}}{\left(\varepsilon_X^{v_6} - \varepsilon_X^{v_5}\right)\left(\left(\varepsilon_N^{v_2} - \varepsilon_N^{v_1}\right)\left(\varepsilon_E^{v_4} - \varepsilon_E^{v_3}\right)\left(\varepsilon_X^{v_5} - \varepsilon_X^{v_6}\right) - \varepsilon_X^{v_1}\varepsilon_N^{v_3}\varepsilon_E^{v_6}\right)}, \qquad \text{(Eq. 2.11)}$$

$$C_{v_5}^{X} = \frac{1}{\left(\varepsilon_X^{v_6} - \varepsilon_X^{v_5}\right)\left(1 - \frac{\varepsilon_X^{v_1}\varepsilon_N^{v_3}\varepsilon_E^{v_6}}{\left(\varepsilon_N^{v_2} - \varepsilon_N^{v_1}\right)\left(\varepsilon_E^{v_4} - \varepsilon_E^{v_3}\right)\left(\varepsilon_X^{v_5} - \varepsilon_X^{v_6}\right)}\right)}, \qquad \text{(Eq. 2.12)}$$

$$C_{v_6}^{X} = -C_{v_5}^{X}. \qquad \text{(Eq. 2.13)}$$

N stands for mRNA and *E* for enzyme. Calculating the control from elasticities using these equations result in the same value for the control coefficents as given in Table 2.2.

Comparison of Eq. 2.7, for the intra-modular control of enzyme 5 on the metabolite, to Eq. 2.12, for the integral control of enzyme 5 on the metabolite, reveals the difference. As compared to the intra-modular control, the integral control is attenuated by a rather complex factor involving inter-level elasticity coefficients. Since most actual systems have connections between regulatory levels, the question is *if* and *how* metabolic intra-modular control can be measured without interference form the inter-modular integral control.

There are several ways of measuring the local metabolic component of control. One of these relies on the metabolic response being faster than the gene-expression response and analyzes the system when the former has settled, while the latter has hardly changed. The second approach adds an inhibitor of transcription, so as to eliminate the non-metabolic response. Less obvious methods include one in which various modulations of the system are performed and global control is measured, after which intra-level control can be calculated. Each of these methods will now be illustrated in detail using the model of Fig. 2.5, Eqs. 2.4, and Table 2.1.

Section 2.5.1: Method 1, method based on metabolite time courses

This method requires one to follow the time evolution of the metabolite concentration after a perturbation has been introduced. The motivation comes from an anticipated wide difference in time scale between the metabolic reactions on the one hand, and the reactions of mRNA and protein levels, on the other [111, 119, 120]. After a perturbation of the limiting rate *V*, the concentration of the metabolite should first evolve to a metabolic quasi-steady state. This apparent steady state should be close to the one that the metabolic system would approach if decoupled from gene expression. Only subsequently should the system evolve, more slowly, towards the global steady state (Fig. 2.6, at the lower rate constants for transcription). When transcription, translation and metabolism operate at similar time scales, the concentration of the metabolite and its flux both move to the global steady state without exhibiting a metabolic quasi-steady state (Fig. 2.6, at high rate constants for transcription).

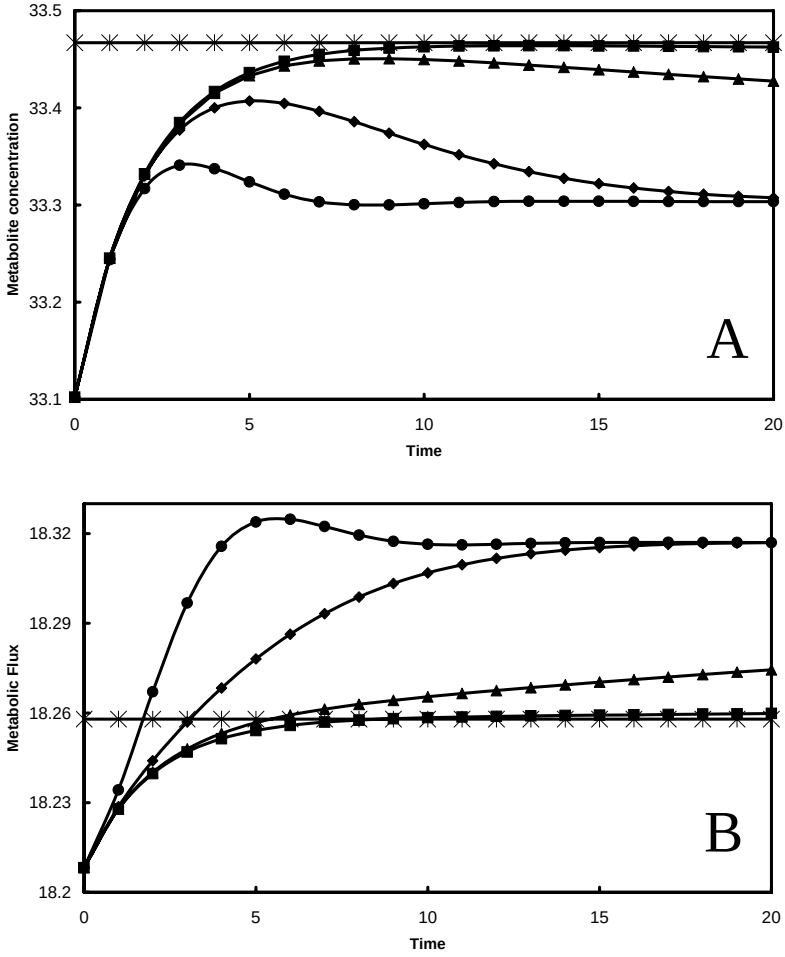

Figure 2.6 – *Time simulations of the model system of Fig. 2.5 at different magnitudes of transcription and mRNA-degradation rate constants. The concentrations and fluxes at time t=0 correspond to the initial steady state before perturbation. Parameter values are as indicated in Table 1, except that the rate constants at the translation level were k_3=10 and k_4=1, and the rates at the level of transcription were: squares: k_1=k_2=0.001, triangles 0.01, diamonds 0.1 and circles 1. At t=0, rate v_5 was perturbed by increasing V (Eq. 2.4) by one percent. The asterisks show the value of the intra-modular steady state after the perturbation. A) Metabolite concentration. B) Metabolic flux.*

In order to simulate a possible experimental determination of the values of the metabolic intra-modular control coefficients based on assumed time-scale separation, the model system was simulated for several values of the rate constants of transcription, mRNA degradation, translation, and protein degradation. The parameters were varied so as to obtain ratios of about 500, 50, 5 and 0.5 between the transition times [167] of

metabolism and the other levels. Simulations were performed such that the steady-state concentrations, steady-state fluxes, and global control coefficients were equal, so as to allow for meaningful comparisons. The parameter values corresponding to these operations are given in Table 1.1 and the legend of Fig. 2.6. The metabolic intra-modular control coefficients were calculated using the calculated time series, taking the highest point in metabolite concentration as the new metabolic intra-modular steady state after the perturbation:

$$c_{v_5}^{Y_{ss}} = \frac{Y_{ss(new)} - Y_{ss(initial)}}{v_{5(perturbed)} - v_{5(initial)}} \cdot \frac{v_{5(initial)}}{Y_{ss(initial)}}. \qquad \text{(Eq. 2.14)}$$

Y represents any system variable, *e.g.* the flux through the metabolic pathway. The modulation of v_5 was kept small, *i.e.* 1 %. If the value of the final (global) steady state is used in Eq. 2.14, then the global control coefficient is obtained.

When the integral control exceeds the intra-modular control, as is the case for the flux-control of reaction 5, another method needs to be applied, because the trajectory fails to exhibit an extremum (the global steady state would be an extremum, but it was not reached in the interval of the measurements). In Fig. 2.6B for instance, the transient flux rapidly increased towards the intra-modular steady state and then increased further towards the global steady state. A transient quasi-steady state has the characteristic that the first derivative of the time course is zero. Therefore, in order to locate the intra-modular steady state, first derivatives of the time course were estimated; the point at which the derivative was closest to zero was taken to be the quasi-steady state. This value was used as the new steady-state flux in Eq. 2.14 to calculate the intra-modular flux-control coefficient. Fig. 2.7 shows the concentration- and flux-control coefficients, calculated using this method for different combinations of parameter values at the levels of transcription and translation. Only for the smaller rate constants at the level of transcription, or for the smaller rate constants of translation, the control coefficients were estimated at an accuracy exceeding 95%.

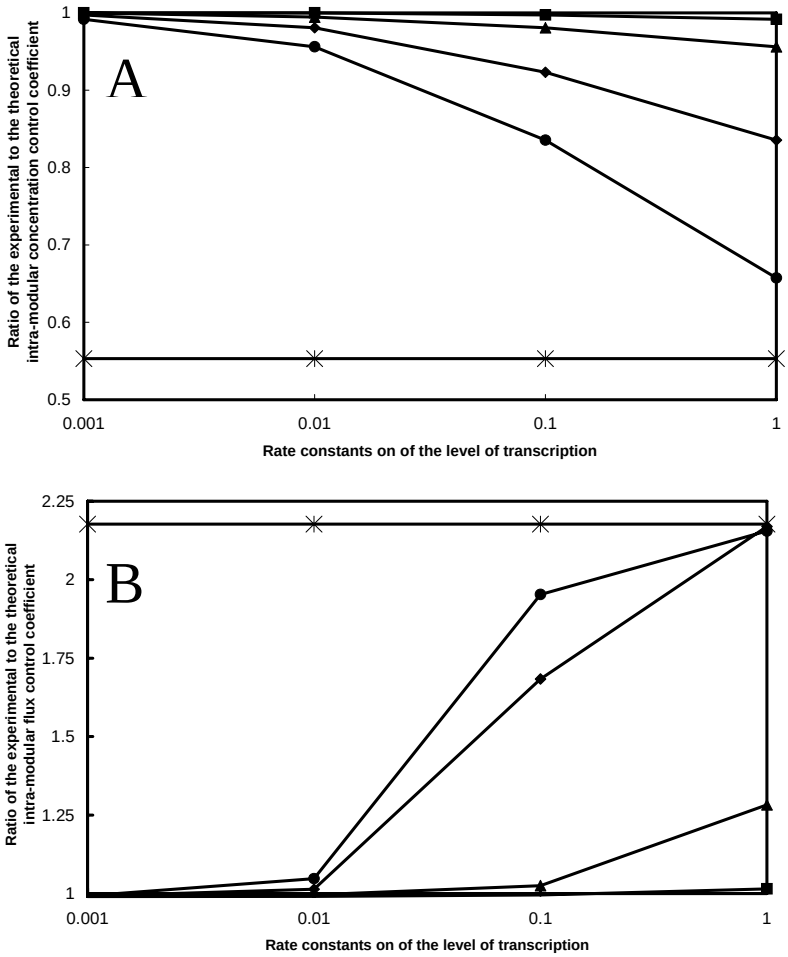

Figure 2.7 – *The metabolic intra-modular control coefficients as a function of transcription/mRNA degradation rate constant and translation/protein degradation rate constant using method 1. Measured control coefficients were scaled relative to the theoretical value for the intra-modular control coefficient. A value of 1 indicates a perfect determination of the intra-modular control coefficient. Rates on the level of translation were varied.* $k_3 \equiv 10 * k_4$ *and for squares* $k_4=0.01$*, for triangles* $k_4=0.1$*, for diamonds* $k_4=1$ *and for circles* $k_4=10$*. The asterisks indicate the analytical value for the integral control coefficient. A) Concentration-control coefficients. B) Flux-control coefficients.*

Section 2.5.2: Method 2, method based on inhibition of transcription

Inhibition of transcription or translation eliminates the feedback loops from metabolism to gene expression: metabolism will then behave as if isolated. Only in this case can one measure intra-modular control coefficients.

Global transcription can be inhibited by adding rifampicine to the medium and global translation by adding chloramphenicol. Here the action of a strong transcription inhibitor was mimicked by setting the rate of transcription to 10^{-25} in the simulations. Transcription should be inhibited at the same time as the metabolic perturbation is made. When transcription is abolished, this system cannot reach a finite global steady state as the concentrations of mRNA and protein decay to zero. As with method 1, it was studied how this would work at several values for the rate constants of transcript-degradation and translation/enzyme degradation differing over 3 orders of magnitude. Under conditions that lead to time separation, *i.e.* metabolic rates much higher than those of transcription and translation, the metabolite concentration first increased to the metabolic steady state and then slowly evolved to the global equilibrium. The flux first moved to the metabolic steady state and then decreased to zero, since the enzyme vanishes due to degradation. Without this separation in time scales no quasi steady state could be detected.

To calculate the intra-modular concentration-control coefficient in this example, the same procedure as described for method 1 was used, but determining the quasi-steady state point from estimates of the first derivatives. The metabolic flux-control coefficients were then calculated in the same way as described for method 1: the maximum in the time series was taken to be the quasi steady-state value and was used in Eq. 2.14. Fig. 2.8 shows the results of simulations for various values of the rate constants of transcription and translation. Again, the intra-modular control coefficients were only estimated accurately when the mRNA and/or enzyme degradation rate constants were small.

In our model, only one of the enzymes is variable in time. This assumes that the rate of degradation of the second enzyme (or its mRNA) is infinitely slower than the degradation rate of the first (or its mRNA). Simulations were done with a model system that is similar to the one described in Fig. 2.5, Eqs. 2.4 and Table 2.1, but where the transcription and translation of the gene coding for the enzyme producing the metabolite are explicit. Degradation and translation kinetics were taken identical to that of the gene for step 6. The transcription kinetics was assumed to be insensitive to the metabolite, and therefore its rate is constant, and set to 10^{-25} to mimic the effect of the transcription inhibitor. Simulations were performed with this system, the data was analyzed as described above, and accurate estimates of control coefficients were found. In this case both proteins decay to zero at the same rate so that both the production and consumption rates of the metabolite decrease in the same proportion, decoupling metabolism from gene expression (agreeing with the summation theorem for concentration control). Only when the time scales of metabolism and gene expression are close, were the estimates of concentration control coefficient poor (see Fig. 2.9A). The accuracy of the measured *flux* control coefficients was still low however (Fig. 2.9B), similar to the results of Fig. 2.8B: the simultaneous inhibition of the synthesis of the two enzymes will decrease all fluxes, even though the metabolite concentration remained almost constant.

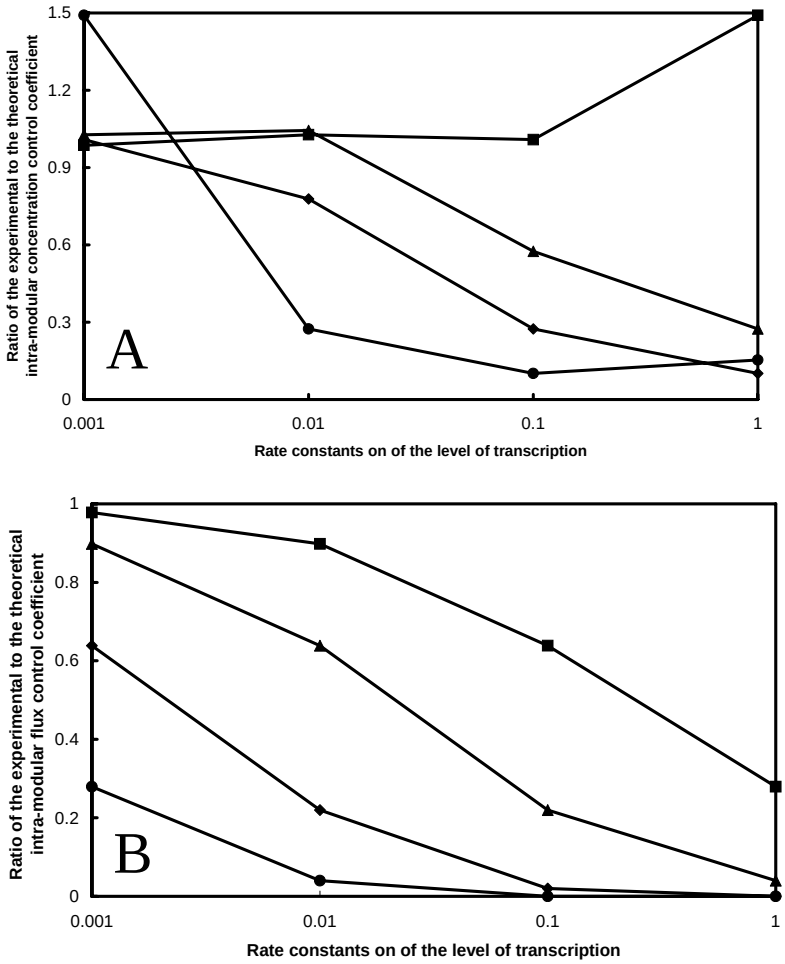

Figure 2.8 – *Intra-modular control coefficients as a function of mRNA-degradation rate and translation/protein degradation rate using method 2 for the case of only one variable enzyme (Fig 2.6). Measured control coefficients are scaled to the theoretical value of the intra-modular control coefficient (1 indicates a perfect determination). Rates on the level of translation differed as follows: squares: 10*k_4= k_3=0.01, triangles 0.1, diamonds 1 and circles 10. A) Concentration-control coefficients. B) Flux-control coefficients*.

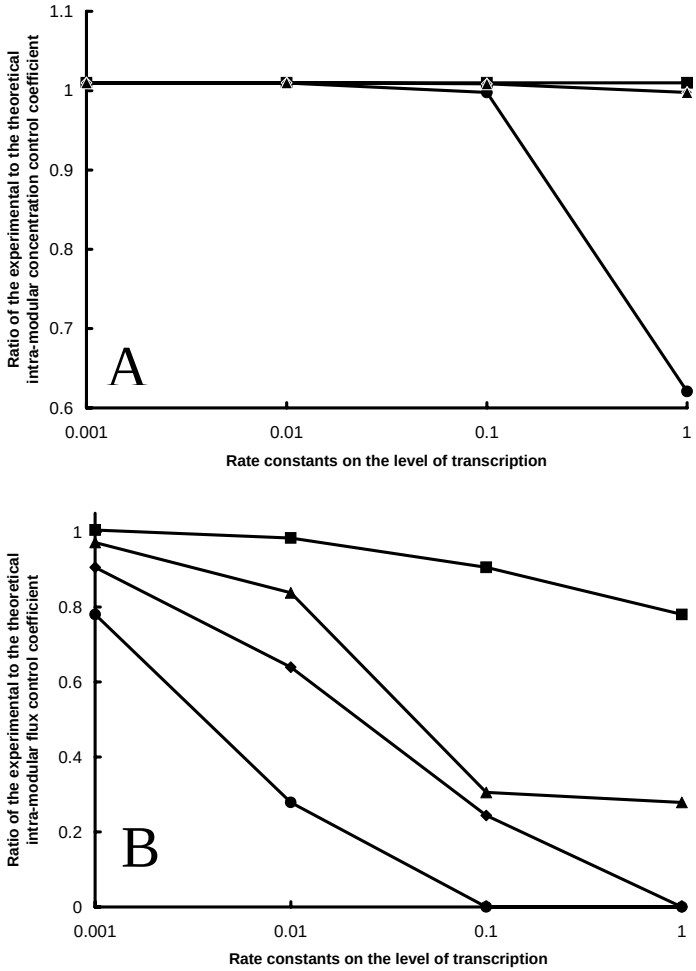

Figure 2.9 – *Intra-modular control coefficients as a function of the mRNA-degradation rate and translation/protein degradation rates brought about by adding corresponding inhibitors at time zero. This is method 2 for determining the intra-module control coefficients, but now with a system with both enzymes variable and with equal degradation rates. Measured control coefficients were scaled to the value of the theoretical value for the intra-modular control coefficient. Thereby a value of 1 indicates a perfect determination of the intra-modular control coefficient. Translation rates differed as follows: squares: 10*k_4 = k_3 = 0.01, triangles 0.1, diamonds 1 and circles 10. Rate constants for the expression of mRNA and protein for the metabolic step 5 are taken to vary identically to mRNA and Enzyme for step 6. A) Concentration-control coefficients. B) Flux-control coefficients.*

Proteins can have degradation rates varying over several orders of magnitudes. Therefore the systems studied here are special cases illustrating the extremes of behavior that can be observed. It is reasonable to expect that the results that can be obtained using this method will be somewhere between the results of these two extremes.

Section 2.5.3: Method 3, a method based on external gene induction

This method is based on replacing the gene promoter by another one whose activity does not depend on the metabolite concentration. A popular method is the replacement of the original promoter by the IPTG inducible *lac*-type promoter, as described in the context of metabolic control analysis by Jensen *et al.* [168]. By this substitution of promoters, the system is transformed to a dictatorial system, in which transcription is insensitive to the other levels. In our model, this substitution of promoters is represented by introducing a new parameter, *i.e.* the concentration of an external transcription activator, which is now the modifier v_1, instead of the pathway metabolite. This implies that there is no significant transcription without the presence of this external transcription activator, which is used to adjust the transcription rate independently from metabolism (just as IPTG has been used by Jensen *et al.* [168]). The activation constant of the external transcription activator was set to 100, and its concentration was adjusted such that the steady state would have the same concentrations of metabolite and enzyme as in the absence of the external activator. Without the feedback loop the response of metabolism to a perturbation in v_5 is purely intra-modular (*i.e.* at the level of metabolism alone). In simulations, it was found that the response is identical to the response that the metabolic pathway would have if considered in isolation. The rates of transcription were varied following the same methodology as in the previous two methods. An estimate of the intra-modular control coefficient was obtained by inserting the values of the new steady state variables in Eq. 2.14. In this case the ability to measure the intra-modular control coefficients was independent of the separation of time scales between metabolism and gene expression.

Section 2.5.4: Method 4, measuring and correcting for the altered enzyme activity

This method makes use of the fact that the rate equation for a metabolic step can be expressed by the product of three factors, one dependent only on the enzyme concentration (e_i), one representing the enzyme's specific activity (a_i) (i.e. that part of its activity that is not affected by the metabolic variables in the system, but only by the added noncompetitive inhibitor), and a term that fully displays the kinetic mechanism and fully contains the dependence of the rate on substrate, product concentrations and the concentrations of all the other metabolic variables that may affect the rate (u_i) [166]:

$$v_6 = e_6 \cdot a_6 \cdot u_6 \qquad \text{(Eq. 2.15)}$$

The specific activity can be perturbed independently of the concentration using non tight-binding inhibitors, such that total inhibitor concentration greatly exceeds total enzyme concentration. The concentration of the enzyme will change due to the change in metabolite through the regulatory feedback loop. All newly synthesized enzyme molecules will be inhibited to the same proportion as those originally present. Consequently a stays constant during the whole measurement. Due to metabolic changes u may vary between steady states. To obtain the global control coefficient one measures the change in flux or metabolite concentration and differentiates the logarithm of that change with respect to the logarithm of the perturbed process activity (for derivation, see [169]:

$$C^{J}_{v_6} = \frac{d\ln J}{d\ln a_6}, \qquad \text{(Eq. 2.16)}$$

$$C^{X}_{v_6} = \frac{d\ln X}{d\ln a_6}. \qquad \text{(Eq. 2.17)}$$

For the intra-modular control coefficients these expressions have to be corrected for the change in enzyme concentration (which could be seen as an additional perturbation to the metabolic level). The logarithm of the change in flux (or concentration) should then be:

$$c^{J}_{v_6} = \frac{d\ln J}{d\ln a_6 + d\ln e_6} = \frac{C^{J}_{v_6}}{1 + d\ln e_6 / d\ln a_6}, \qquad \text{(Eq. 2.18)}$$

$$c^{X}_{v_6} = \frac{d\ln X}{d\ln a_6 + d\ln e_6} = \frac{C^{X}_{v_6}}{1 + d\ln e_6 / d\ln a_6}. \qquad \text{(Eq. 2.19)}$$

In order to calculate the intra-modular control coefficient using this method, one needs to measure the enzyme concentration additionally to the fluxes and metabolite concentrations. Results of this method on the model system of Fig. 2.5 are given in the top rows of Table 2.3. It is seen that this method is reasonably accurate.

Eqs. 2.18 and 2.19 are only valid when the enzyme of the step under consideration was the only one that changed concentration in response to the perturbation. To remove this restriction, the model was extended by explicitly taking account of the mRNA and enzyme concentrations of the metabolic step 5. Degradation and translation kinetics are identical to that of the gene for step 6 (note that this choice of parameters is not necessary for this method to work). Transcription of this gene is supposed to be affected by the metabolite through competitive inhibition:

$$v_7 = \frac{V_7 \dfrac{[nucleotides]}{K_{m_7}}}{1 + \dfrac{[nucleotides]}{K_{m_7}} + \dfrac{[metabolite]}{K_{P_7}}}. \qquad \text{(Eq. 2.20)}$$

V_7 is the limiting transcription rate for this gene, K_{m7} the Michaelis constant for the nucleotide triphosphates and K_{P7} the product inhibition constant of the metabolite. The parameter values used were V_7 = 0.001, K_{m7} = 1 and K_{P7} =100, with [*nucleotides*] as in Table 1.1. In terms of the analytical equations obtained for the intra-modular flux control coefficient (correcting for the effect the altered expression of enzyme 5 has on the flux by using the flux control summation law, which remains valid also for the intra-modular control coefficients (*cf.* [103, 164]):

$$c_{v_6}^{J} = \frac{d\ln J - d\ln e_5}{d\ln a_6 + d\ln e_6 - d\ln e_5} = \frac{C_{v_6}^{J} - d\ln e_5 / d\ln a_6}{1 + d\ln e_6 / d\ln a_6 - d\ln e_5 / d\ln a_6} = \frac{C_{v_6}^{J} - C_{v_6}^{e_5}}{1 + C_{v_6}^{e_6} - C_{v_6}^{e_5}}, \qquad \text{(Eq. 2.21)}$$

and likewise for the intra-modular concentration control coefficient:

$$c_{v_6}^{X} = \frac{d\ln X}{d\ln a_6 + d\ln e_6 - d\ln e_5} = \frac{C_{v_6}^{X}}{1 + d\ln e_6 / d\ln a_6 - d\ln e_5 / d\ln a_6} = \frac{C_{v_6}^{X}}{1 + C_{v_6}^{e_6} - C_{v_6}^{e_5}}. \qquad \text{(Eq. 2.22)}$$

Interestingly, the quantity $d\ln e_i / d\ln a_j$ is the integral control of process *j* on the enzyme concentration, *i.e.* $C_j^{E_i}$. One can thus calculate the intra-modular control of a metabolic step on a metabolite from integrated control of that step on all components that have influence on that metabolite. Results of applying this method are given in the bottom rows of Table 2.3. Again, this proved to be an accurate method.

Table 2.3 – *Values of global and intra-modular control coefficients calculated using method 4.*

Type of control	**Control coefficient**	**Real Value**	**Calculated**
System with one variable enzyme			
Integral	$C_{v_6}^{J_{ss}}$	0.40	0.40
	$C_{v_6}^{[X]_{ss}}$	-0.61	-0.62
Intra-modular	$c_{v_6}^{J_{ss}}$	0.72	0.73
	$c_{v_6}^{[X]_{ss}}$	-1.1	-1.12
System with two variable enzymes			
Integral	$C_{v_6}^{J_{ss}}$	0.43	0.43
	$C_{v_6}^{[X]_{ss}}$	-0.54	-0.54
Intra-modular	$c_{v_6}^{J_{ss}}$	0.68	0.69
	$c_{v_6}^{[X]_{ss}}$	-1.12	-1.14

A general matrix equation, of which 2.18, 2.19, 2.21 and 2.22 are special cases, can be derived. I start with the control matrix expression discussed in Chapter 1, but written as:

$$\mathbf{E} \cdot \mathbf{C} = \mathbf{I} \qquad \text{(Eq. 2.23)}$$

Matrix $\mathbf{C}$ contains control coefficients of the independent variables only and can be partitioned according to flux control coefficients, $\mathbf{C^J}$ and concentration control coefficients, $\mathbf{C^S}$:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C^J} \\ \mathbf{C^S} \end{bmatrix} \qquad \text{(Eq. 2.24)}$$

Matrix $\mathbf{E}$ contains structural and kinetic information,

$$\mathbf{E} = \begin{bmatrix} \boldsymbol{\kappa} & -\boldsymbol{\varepsilon}\mathbf{L} \end{bmatrix} \qquad \text{(Eq. 2.25)}$$

where $\boldsymbol{\kappa}$ is the scaled kernel, relating dependent to independent fluxes, $\boldsymbol{\varepsilon}$ the matrix of elasticity coefficients, and $\mathbf{L}$ the scaled link matrix relating dependent concentrations to independent ones. This relationship is general, as has been proven in [114]. It will pay off though to formulate it for the specific subset of cases that are organized in terms of three flux-disconnected systems. These will then pertain to the three functional levels of organization of cells, *i.e.* the transcriptome (T), the proteome (P) and the metabolome (M). Accordingly the matrices are partitioned as:

$$\begin{bmatrix} \mathbf{E}_{\mathbf{T}}^{\mathbf{v}_T} & \mathbf{E}_{\mathbf{P}}^{\mathbf{v}_T} & \mathbf{E}_{\mathbf{M}}^{\mathbf{v}_T} \\ \mathbf{E}_{\mathbf{T}}^{\mathbf{v}_P} & \mathbf{E}_{\mathbf{P}}^{\mathbf{v}_P} & \mathbf{E}_{\mathbf{M}}^{\mathbf{v}_P} \\ \mathbf{E}_{\mathbf{T}}^{\mathbf{v}_M} & \mathbf{E}_{\mathbf{P}}^{\mathbf{v}_M} & \mathbf{E}_{\mathbf{M}}^{\mathbf{v}_M} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\mathbf{v}_T}^{\mathbf{T}} & \mathbf{C}_{\mathbf{v}_P}^{\mathbf{T}} & \mathbf{C}_{\mathbf{v}_M}^{\mathbf{T}} \\ \mathbf{C}_{\mathbf{v}_T}^{\mathbf{P}} & \mathbf{C}_{\mathbf{v}_P}^{\mathbf{P}} & \mathbf{C}_{\mathbf{v}_M}^{\mathbf{P}} \\ \mathbf{C}_{\mathbf{v}_T}^{\mathbf{M}} & \mathbf{C}_{\mathbf{v}_P}^{\mathbf{M}} & \mathbf{C}_{\mathbf{v}_M}^{\mathbf{M}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad \text{(Eq. 2.26)}$$

Here, submatrices $\mathbf{C}_i^j$ give the control of the processes at level *i* on the fluxes and concentrations at level *j* (i, j can be T, P, or M). Similar, submatrices $\mathbf{E}_i^{\mathbf{v}_j}$ contain elasticities of the rates in level *i* towards the concentrations of level *j*. In Chapter 3, section 3.2.2, a biochemical explanation of each submatrix is given. Submatrices $\mathbf{C}_{\mathbf{v}_i}^i$ and $\mathbf{E}_i^{\mathbf{v}_i}$ are square and invertible. The submatrices in **E** are partitioned as in [170] and [111]:

$$\mathbf{E}_i^{\mathbf{v}_i} = \begin{bmatrix} \boldsymbol{\kappa}_i & -\boldsymbol{\varepsilon}_i^{\mathbf{v}_i}\mathbf{L}_i \end{bmatrix} \qquad \text{(Eq. 2.27)}$$

$$\mathbf{E}_i^{\mathbf{v}_j} = \begin{bmatrix} \mathbf{0} & -\boldsymbol{\varepsilon}_i^{\mathbf{v}_j}\mathbf{L}_i \end{bmatrix} \qquad \text{(Eq. 2.28)}$$

To be able to partition as in Eq. 2.26, the $\boldsymbol{\kappa}$ and $\mathbf{L}$ matrices have to be block diagonal. This is the case when there is no mass exchange between the three levels or when the mass exchange is negligibly small [111], such as in most materializations of the hierarchical systems considered here.

The inverse of the complete **E** matrix equals the integral (global) control matrix. Since $\mathbf{E}_i^{\mathbf{v}_i}$ contains only elasticities quantifying the effects of the level *i* on the rates on level *i*, inverting this sub matrix gives the intra-modular control matrix for level *i*. By inverting $\mathbf{E}_i^{\mathbf{v}_i}$ the interactions of level *i* with the other levels are ignored and thus yields a control matrix that would characterize level *i* in isolation. Rewriting Eq. 2.26 by substituting $\left(\mathbf{c}_{\mathbf{v}_i}^{i}\right)^{-1}$ for $\mathbf{E}_i^{\mathbf{v}_i}$ yields:

$$\begin{bmatrix} \left(\mathbf{c}_{\mathbf{v}_\mathbf{T}}^{\mathbf{T}}\right)^{-1} & \mathbf{E}_\mathbf{P}^{\mathbf{v}_\mathbf{T}} & \mathbf{E}_\mathbf{M}^{\mathbf{v}_\mathbf{T}} \\ \mathbf{E}_\mathbf{T}^{\mathbf{v}_\mathbf{P}} & \left(\mathbf{c}_{\mathbf{v}_\mathbf{P}}^{\mathbf{P}}\right)^{-1} & \mathbf{E}_\mathbf{M}^{\mathbf{v}_\mathbf{P}} \\ \mathbf{E}_\mathbf{T}^{\mathbf{v}_\mathbf{M}} & \mathbf{E}_\mathbf{P}^{\mathbf{v}_\mathbf{M}} & \left(\mathbf{c}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}}\right)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{C}_{\mathbf{v}_\mathbf{T}}^{\mathbf{T}} & \mathbf{C}_{\mathbf{v}_\mathbf{P}}^{\mathbf{T}} & \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{T}} \\ \mathbf{C}_{\mathbf{v}_\mathbf{T}}^{\mathbf{P}} & \mathbf{C}_{\mathbf{v}_\mathbf{P}}^{\mathbf{P}} & \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{P}} \\ \mathbf{C}_{\mathbf{v}_\mathbf{T}}^{\mathbf{M}} & \mathbf{C}_{\mathbf{v}_\mathbf{P}}^{\mathbf{M}} & \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad \text{(Eq. 2.29)}$$

In this chapter, the extent to which a measurement of the local control coefficients of metabolism may be confounded by interference by global control is examined, *i.e.* only the relationship between $\mathbf{c}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}}$ and $\mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}}$, $\mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{P}}$ and $\mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{T}}$ is of interest. Therefore the third row of **E** with the third column of **C** is multiplied giving:

$$\mathbf{E}_\mathbf{T}^{\mathbf{v}_\mathbf{M}} \cdot \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{T}} + \mathbf{E}_\mathbf{P}^{\mathbf{v}_\mathbf{M}} \cdot \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{P}} + \left(\mathbf{c}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}}\right)^{-1} \cdot \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}} = \mathbf{I} \qquad \text{(Eq. 2.30)}$$

Solving for $\mathbf{c}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}}$ in terms of $\mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}}$, $\mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{P}}$ and $\mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{T}}$ by going through some rearrangements

$$\mathbf{c}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}} = \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{M}} \left(\mathbf{I} - \mathbf{E}_\mathbf{T}^{\mathbf{v}_\mathbf{M}} \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{T}} + \mathbf{E}_\mathbf{P}^{\mathbf{v}_\mathbf{M}} \mathbf{C}_{\mathbf{v}_\mathbf{M}}^{\mathbf{P}}\right)^{-1} \qquad \text{(Eq. 2.31)}$$

assuming that the inverse matrices exist (again, this is now related to the assumption of stability of the reference steady state). The result is a general expression relating the intra-modular control to integral control. The intra-modular control can thus calculated from experimental measurements of the integral control.

In all the examples used in this dissertation, no direct effects from the transcriptome on the metabolome are considered, as these are unprecedented biochemically

(they would consist of direct effects of mRNA's on enzyme activities; they are not altogether impossible as some mRNAs have been shown to bind to enzymes, for unknown functional reasons). Therefore, $\mathbf{E}_{\mathbf{T}}^{\mathbf{v_M}}$ will be assumed to be a null matrix and Eq. 2.31 reduces to:

$$\mathbf{c}_{\mathbf{v_M}}^{\mathbf{M}} = \mathbf{C}_{\mathbf{v_M}}^{\mathbf{M}} \cdot \left(\mathbf{I} - \mathbf{E}_{\mathbf{P}}^{\mathbf{v_M}} \cdot \mathbf{C}_{\mathbf{v_M}}^{\mathbf{P}}\right)^{-1} \qquad \text{(Eq. 2.32)}$$

Using this result, equations 2.18, 2.19, 2.21 and 2.22 can be re-derived, but equation 2.31 is generally valid. For maximum clarity, I here write out these matrices explicitly for our example of Figure 2.5, with the additional modification of explicitly including the enzyme for the synthesis step of the metabolite, with the properties given in Eq. 2.20. Since the numbering of rates in Figure 2.5 is rather confusing, here the model is presented more clearly

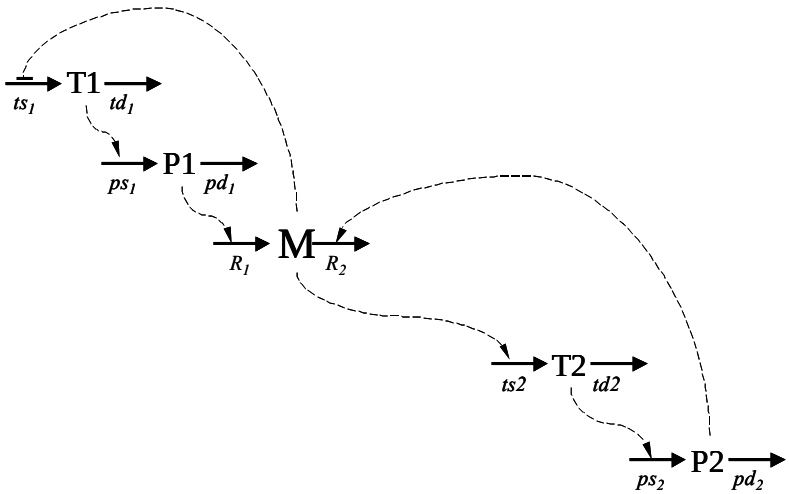

Figure 2.10 – *A system consisting of two mRNAs, two proteins and a two-step metabolic pathway. There is feedback from the metabolome to the transcriptome. T, P and M stand for transcript (mRNA), protein and metabolite, respectively. ts and td stand for rate of transcript synthesis and rate of transcript degradation, respectively; ps and pd stand for rate of protein synthesis and rate of transcript degradation, respectively; and R_1 and R_2 for metabolic rates*.

The corresponding elasticity matrix is:

$$\mathbf{E} = \begin{bmatrix} \mathbf{E_T^{v_T}} & \mathbf{0} & \mathbf{E_M^{v_T}} \\ \mathbf{E_T^{v_P}} & \mathbf{E_P^{v_P}} & \mathbf{0} \\ \mathbf{0} & \mathbf{E_P^{v_M}} & \mathbf{E_M^{v_M}} \end{bmatrix} = \left[\begin{array}{cccc|cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_M^{v_{ts1}} \\ 1 & -\varepsilon_{T1}^{v_{td1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_M^{v_{ts2}} \\ 0 & 0 & 1 & -\varepsilon_{T2}^{v_{td2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -\varepsilon_{T1}^{v_{ps1}} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P1}^{v_{pd1}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_{T2}^{v_{ps2}} & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P2}^{v_{pd2}} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P1}^{v_{R1}} & 0 & 0 & 1 & -\varepsilon_M^{v_{R1}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P2}^{v_{R2}} & 1 & -\varepsilon_M^{v_{R2}} \end{array}\right]$$

(Eq. 2.33)

Transcription was assumed not to be inhibited by its mRNA product, and that the protein product does not inhibit translation, and that metabolites only affect transcription not translation directly. The relevant integral control matrices are:

$$\mathbf{C}_{\mathbf{v_M}}^{\mathbf{M}} = \begin{pmatrix} C_{v_{R_1}}^{J_M} & C_{v_{R_2}}^{J_M} \\ C_{v_{R_1}}^{M} & C_{v_{R_2}}^{M} \end{pmatrix} \qquad \text{(Eq. 2.34)}$$

and

$$\mathbf{C}_{\mathbf{v_M}}^{\mathbf{P}} = \begin{pmatrix} C_{v_{R_1}}^{J_{P1}} & C_{v_{R_2}}^{J_{P1}} \\ C_{v_{R_1}}^{P1} & C_{v_{R_2}}^{P1} \\ C_{v_{R_1}}^{J_{P2}} & C_{v_{R_2}}^{J_{P2}} \\ C_{v_{R_1}}^{P2} & C_{v_{R_2}}^{P2} \end{pmatrix} \qquad \text{(Eq. 2.35)}$$

Writing out equation 2.32 in terms of these matrices one obtains:

$$\begin{pmatrix} c_{v_{R_1}}^{J_M} & c_{v_{R_2}}^{J_M} \\ c_{v_{R_1}}^{M} & c_{v_{R_2}}^{M} \end{pmatrix} = \begin{pmatrix} C_{v_{R_1}}^{J_M} & C_{v_{R_2}}^{J_M} \\ C_{v_{R_1}}^{M} & C_{v_{R_2}}^{M} \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & -\varepsilon_{P1}^{v_{R_1}} & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_{P2}^{v_{R_2}} \end{pmatrix} \begin{pmatrix} C_{v_{R_1}}^{J_{P1}} & C_{v_{R_2}}^{J_{P1}} \\ C_{v_{R_1}}^{P1} & C_{v_{R_2}}^{P1} \\ C_{v_{R_1}}^{J_{P2}} & C_{v_{R_2}}^{J_{P2}} \\ C_{v_{R_1}}^{P2} & C_{v_{R_2}}^{P2} \end{pmatrix} \right)^{-1}$$

(Eq. 2.36)

For the intra-modular control of step R2 (corresponding to v_6 in equation 2.21) on the metabolic flux one then obtains

$$c_{v_{R_2}}^{J_M} = \frac{C_{v_{R_2}}^{J_M} + C_{v_{R_2}}^{J_M} \varepsilon_{P1}^{v_{R_1}} C_{v_{R_1}}^{P1} - C_{v_{R_1}}^{J_M} \varepsilon_{P1}^{v_{R_1}} C_{v_{R_2}}^{P1}}{1 + \varepsilon_{P1}^{v_{R_1}} C_{v_{R_1}}^{P1} + \varepsilon_{P2}^{v_{R_2}} C_{v_{R_2}}^{P2} + \varepsilon_{P1}^{v_{R_1}} C_{v_{R_1}}^{P1} \varepsilon_{P2}^{v_{R_2}} C_{v_{R_2}}^{P2} - \varepsilon_{P1}^{v_{R_1}} C_{v_{R_2}}^{P1} \varepsilon_{P2}^{v_{R_2}} C_{v_{R_1}}^{P2}} \qquad \text{(Eq. 2.37)}$$

This expression is rather different from that in Eq. 2.21. However, because the elasticities towards the enzymes are taken to be equal to 1 (since most enzymes linearly increase the rate of metabolic reactions (see Eq. 2.15) the equations simplifies to:

$$c_{v_{R_2}}^{J_M} = \frac{C_{v_{R_2}}^{J_M} + C_{v_{R_2}}^{J_M} C_{v_{R_1}}^{P1} - C_{v_{R_1}}^{J_M} C_{v_{R_2}}^{P1}}{1 + C_{v_{R_1}}^{P1} + C_{v_{R_2}}^{P2} + C_{v_{R_1}}^{P1} C_{v_{R_2}}^{P2} - C_{v_{R_2}}^{P1} C_{v_{R_1}}^{P2}} \qquad \text{(Eq. 2.38)}$$

Some of the control coefficients are equal due to the hierarchical structure of the system and the validity of summation laws for control coefficients [cf. [103]] giving $C_{v_{R_1}}^{P1} C_{v_{R_2}}^{P2} = C_{v_{R_2}}^{P1} C_{v_{R_1}}^{P2}$ and simplifying the denominator. Again using such summation laws, the term $C_{v_{R_2}}^{J_M} C_{v_{R_1}}^{P1} - C_{v_{R_1}}^{J_M} C_{v_{R_2}}^{P1}$ in the numerator can be rewritten as $-\left(C_{v_{R_2}}^{J_M} + C_{v_{R_1}}^{J_M}\right) C_{v_{R_2}}^{P1}$ and since $C_{v_{R_2}}^{J_M} + C_{v_{R_1}}^{J_M} = 1$, one obtains:

$$c_{v_{R_2}}^{J_M} = \frac{C_{v_{R_2}}^{J_M} - C_{v_{R_2}}^{P1}}{1 + C_{v_{R_1}}^{P1} + C_{v_{R_2}}^{P2}} \qquad \text{(Eq. 2.39)}$$

which is the same as Eq. 2.21.

Similarly the equation for the intra-modular control of step 2 (Eq. 2.22) on the metabolite can be obtained from

$$c_{v_{R_2}}^{M} = \frac{C_{v_{R_2}}^{M} + C_{v_{R_2}}^{M} \varepsilon_{P1}^{v_{R_1}} C_{v_{R_1}}^{P1} - C_{v_{R_1}}^{M} \varepsilon_{P1}^{v_{R_1}} C_{v_{R_2}}^{P1}}{1 + \varepsilon_{P1}^{v_{R_1}} C_{v_{R_1}}^{P1} + \varepsilon_{P2}^{v_{R_2}} C_{v_{R_2}}^{P2} + \varepsilon_{P1}^{v_{R_1}} C_{v_{R_1}}^{P1} \varepsilon_{P2}^{v_{R_2}} C_{v_{R_2}}^{P2} - \varepsilon_{P1}^{v_{R_1}} C_{v_{R_2}}^{P1} \varepsilon_{P2}^{v_{R_2}} C_{v_{R_1}}^{P2}} \qquad \text{(Eq. 2.40)}$$

Again the elasticities towards the enzymes are taken to be equal to 1, the denominator is simplified as above, and since $C^M_{v_{R_2}} C^{P1}_{v_{R_1}} = C^M_{v_{R_1}} C^{P1}_{v_{R_2}}$ the numerator simplifies to

$$c^M_{v_{R_2}} = \frac{C^M_{v_{R_2}}}{1 + C^{P1}_{v_{R_1}} + C^{P2}_{v_{R_2}}} \qquad \text{(Eq. 2.41)}$$

which is the same as Eq. 2.22.

If it is assumed that the enzyme catalyzing the synthesis of the metabolite is not sensitive to the metabolite (as was done for the model described by Fig. 2.5, Eq. 2.4 and Table 2.1) Eqs. 2.26 and 2.38 reduce to 2.42 and 2.43 respectively,

$$c^{J_M}_{v_{R_2}} = \frac{C^{J_M}_{v_{R_2}}}{1 + C^{P2}_{v_{R_2}}} \qquad \text{(Eq. 2.42)}$$

$$c^M_{v_{R_2}} = \frac{C^M_{v_{R_2}}}{1 + C^{P2}_{v_{R_2}}} \qquad \text{(Eq. 2.43)}$$

which are the same as Eq. 2.18 and 2.19 used above. These equations are readily interpreted as the fact that the intra-level control equals global control corrected for the change in enzyme concentration.

Equation 2.32 is general and can be applied to any system of arbitrary complexity. Similar equations to Eq. 2.32 can be derived to calculate the intra-modular control on the proteome and the transcriptome from measurements of integral control coefficients on all three levels. The equations derived here are basically an inverse version of the approach described in [111], where the properties of the whole are described in terms of the properties of the individual levels.

Section 2.6: Discussion

In this chapter I have introduced the concept of hierarchical biochemical systems and their analysis. To simplify the analysis of these hierarchical systems the approach of decomposing the system into subsystems was taken. Mathematical models have been formulated for metabolism and for gene networks, based on the assumption that there is a large time scale difference between the rates at which processes at the different levels of the hierarchical system occur. I have shown under which conditions this assumption is justified.

This relates to the issue of when it is relevant to distinguish the control properties of metabolism as if isolated from those of the global hierarchical system, and in fact when it is useful to consider the hierarchical nature of the system at all. From the operational point of view, relevant for predicting behavior of the system, this may no longer be so if the time separation is too small. In practice, the three levels of the hierarchy will then change

virtually simultaneously anyway. However, from the conceptual point of view, it may remain useful always to distinguish regulation that happens at the metabolic level only, from regulation that involves transcription or translation. In addition, for engineering purposes knowing where the regulation resides should suggest strategies to enable the biological systems to produce at a higher efficiency or yield.

Quantifying the intra-modular control of metabolism in addition to the integral control thus provides insight into how regulation is partitioned between the different levels. Furthermore, if there is a time scale separation, the intra-modular control quantifies the short-term behavior of the system, while the integrated control the long-term behavior. If there is insufficient timescale separation, the concept of intra-modular control is somewhat academic, since it only exists in the isolated sub-system. In many cases there is a considerable time scale separation between metabolism and the mRNA and protein levels. For these cases the relevance of HCA and the present method is that they are able to distinguish between the control exerted all within one level (*e.g.* between the metabolic reactions) from the control of one level over another. Hierarchical Control Analysis allows one to describe these two types of control and has exact laws to relate them [cf. [103]]. The methods proposed here should facilitate their experimental implementation.

The methods to measure intra-modular control were illustrated using a kinetic model. Parameters were chosen so as to obtain a high ratio between the intra-modular and the integral control coefficients (also called A-coefficient in [111]). In real biochemical systems the values of two different types of control coefficients might be either closer or further apart. When the values of the two coefficients are closer, it will be more difficult to distinguish the two.

The first two methods for determining intra-level control coefficients were motivated by the time scale separation that might exist between the dynamics of intermediary metabolism and gene expression. Such time separation is mainly determined by the difference in the degradation rates of the different levels [111, 119, 120]. When this difference is sufficiently large, the initial behavior of the system is determined by the intra-modular (metabolic) control, and the final behavior by the integral control. In our models a difference of two to three orders of magnitude proved sufficient to observe this effect. Smaller differences in time scales reduced the accuracy at which the intra-modular control coefficients could be measured. Estimating for major metabolic pathways of *E. coli* metabolite turnover times (concentration divided by flux) of a few seconds, whereas most enzymes last many cell cycles of longer than 30 minutes, the characteristic times may indeed be apart on the order of the required factor of 100-500. Similar estimates apply to yeast glycolysis.

Seemingly, evolution resulted in a temporal organization of hierarchical regulation: the more free-energy demanding mode of inter-level regulation – protein synthesis – is switched on if intra-level-metabolic regulation alone proves to be incapable to sufficiently sustain the physiological state. Furthermore, it appears that time scales reduce the complexity of the dynamics displayed by biochemical networks thereby preventing uncontrollable (complex) behavior that may compromise proper cell function [171].

The ability to discriminate between fast and slow control, as elaborated in the present manuscript, may help understand the function of signal transduction networks, which often have more than one characteristic time constant. Indeed, kinases and phosphatases have been shown differentially to affect the slow and the faster phases and

this may be important for the decision between differentiation and proliferation in some cell types [172].

For our model system, it must be noted that should the feedback interaction of the metabolite to transcription be stronger (lower K_a in Eq. 2.4), the time scales, expressed by eigenvalues of the Jacobian [173] come closer. Indeed, with decreasing K_a, the fast time scale decreased towards the slow time scale (results not shown). One should thus be cautious when reasoning solely on the basis of rate constants of transcription and metabolism, without knowledge of the strength of interactions between these levels.

In our demonstration of methods 1 and 2, the sampling frequency of the measurements was rather high because both methods require one to locate a minimum of the first-order derivative of the curve. In practice, it might be difficult to make measurements at this frequency and thus the quasi-steady state could be missed or misplaced, resulting in much larger error. It is advisable to fit the time course to an arbitrary fitting function first and then to locate the quasi-steady state from the zero or minimum of the derivative of this function.

Method 2 did not prove any better than method 1. This is because the method requires inhibition of transcription. Thereby the method itself perturbs the steady state, just when the relaxation phase sets in that separates intra-level from global control: Since the mRNA continues to be degraded, whereas its synthesis is being stopped, its steady-state balance is perturbed. An alternative method would inhibit both transcription and mRNA degradation, such that the level of mRNA should remain constant. This is difficult to achieve experimentally and was therefore not considered here.

Methods 3 and 4 are similar to each other in that both remove the feedback loop from metabolism to gene expression, either physically (by replacing the promoter) or mathematically. The problem with method 3 is that it only works when there is a single feedback loop. In living cells there are certainly more feedback loops from metabolism to gene expression, so one still measures the global control of the system, but without that particular feedback loop [162]. In order to measure intra-modular control, one would have to replace all promoters. Method 4 is applicable to systems consisting of many variable enzymes, provided that one measures control coefficients for all enzymes and metabolites in the system, severely limiting its applicability to larger networks.

The distinction between intra-modular metabolic and global integrated control is crucial for the understanding of the regulation of cell physiology. An example is catabolite repression by glucose, which abounds in biology. This works *via* metabolic effects, via signal transduction and via gene-expression. The implications of the three types of mechanism differ greatly for the dynamics and persistence of the regulation. For humans, gene-expression regulation of glucose uptake after a rich meal should result in a subsequent undershoot in glucose levels, unless compensated by additional insulin-dependent regulation. On the other hand, gene expression-mediated regulation is the one that permits the best homeostasis of intracellular metabolites, and may hence lead to the most optimal state.

This approach is fundamentally different from that in the work of Acerenza *et al.* [174] and Heinrich and Reder [175], who studied the time-dependent control analysis (*i.e.* quantifying control of reactions on the relaxation processes). Although based on observation of time courses, our methods do not extend MCA to the time domain. Simply, a way was described of locating a quasi-steady state on the time course, followed by analysis with the traditional MCA approach, as if it were a true steady state. This has led to an emphasis on small changes (perhaps smaller than may be experimentally feasible),

steady states, control, and regulation. Aspects of spatial heterogeneity, and experimental errors [cf. 176] deserve scrutiny in future work. It is important to note that the present results are essentially the same when perturbations of 10% rather than 1% were applied (data not shown).

The work described in this chapter is dedicated to dealing with hierarchical systems and their simplified description in terms of individual flux-disconnected sub-systems. The emphasis was on the study of the subsystem of metabolism, independently from the subsystems of gene transcription and protein synthesis. The enhanced ability to distinguish between metabolic and hierarchical regulation may further our understanding of living organisms. This becomes acute with the greatly enhanced abilities to measure gene expression (transcriptome [12] and proteome [177]) and metabolome [178] in parallel and quantitatively. Since cell function depends on both, and in many interconnected ways (e.g. [166]), progress may well depend on our ability to dissect intra-modular metabolic regulation from the hierarchical regulation that hovers over it.

3

Chapter 3: Gene networks

In the previous chapter I have shown one way to simplify the study of hierarchical systems, *i.e.* by decomposing the global system into subsystems and by then studying the properties of these subsystems in order to gain knowledge of their importance for regulation in the whole system. In this chapter I describe another way to simplify the study of hierarchical systems. This simplification yields a 'condensed' description of the hierarchical system in terms of a gene network [83]. Gene networks only account for transcript (mRNA) concentrations, while proteins and metabolites are only implicitly present mediating interactions between the genes [82, 83]. I will introduce the concept of gene networks and will show how to express quantitatively the properties of the gene network in terms of the properties of the biochemical reactions underlying the communication between genes.

Section 3.1: The gene network concept

An increasingly popular model of biochemical regulation is that of 'gene networks' in which the nodes represent genes or their activities (mRNAs) and the edges correspond to regulatory interactions between them. Such gene networks are phenomenological models because they do not represent explicitly the proteins and metabolites that mediate those interactions. Gene networks are a logical way of attempting to describe phenomena observed with transcription profiling, such as is done at a large scale with the popular microarray technology. Being able to create gene networks from experimental data and to use them to reason about their dynamics and design principles will contribute to increased understanding of cellular function.

The 'genes' in gene networks represent the gene activities (mRNA concentrations) of an organism in a particular physiological state. Each mRNA can be synthesized and degraded. Synthesis and degradation rates are then proposed to be regulated by the activity of the gene itself and by other gene-activities, *i.e.* by all mRNA levels.

In reality the synthesis rates of mRNAs are rarely influenced directly by the concentrations of mRNA molecules, but by their protein products, such as transcription factors and by metabolites such as transcriptional inducers and suppressors. Clearly regulation does not only involve the level of mRNAs but also the level of protein and the level of intermediary metabolism. Figure 3.1 represents a model of a global biochemical network in which the three levels are shown explicitly by planes. The arrows illustrate that in such a global biochemical network the gene activities do not interact directly with each other. Instead gene induction or repression occurs through the action of specific proteins, which are, in turn, products of certain genes. Gene expression can also be affected directly by metabolites, as they are a source of the material and free energy required for the process, or through regulatory protein-metabolite complexes.

However, it could be useful to abstract from this reality of the actions of proteins and metabolites, and to represent the system by a simplified model in which gene-activities (mRNAs) are the only explicit actors, acting on other genes in a gene network (also called genetic regulatory, transcription or expression networks). This simplification of going from the global biochemical network to a gene network is akin to a projection of all interactions to the "gene space" [39] is illustrated in Figure 3.1. Figure 3.2 is the resulting gene network representation.

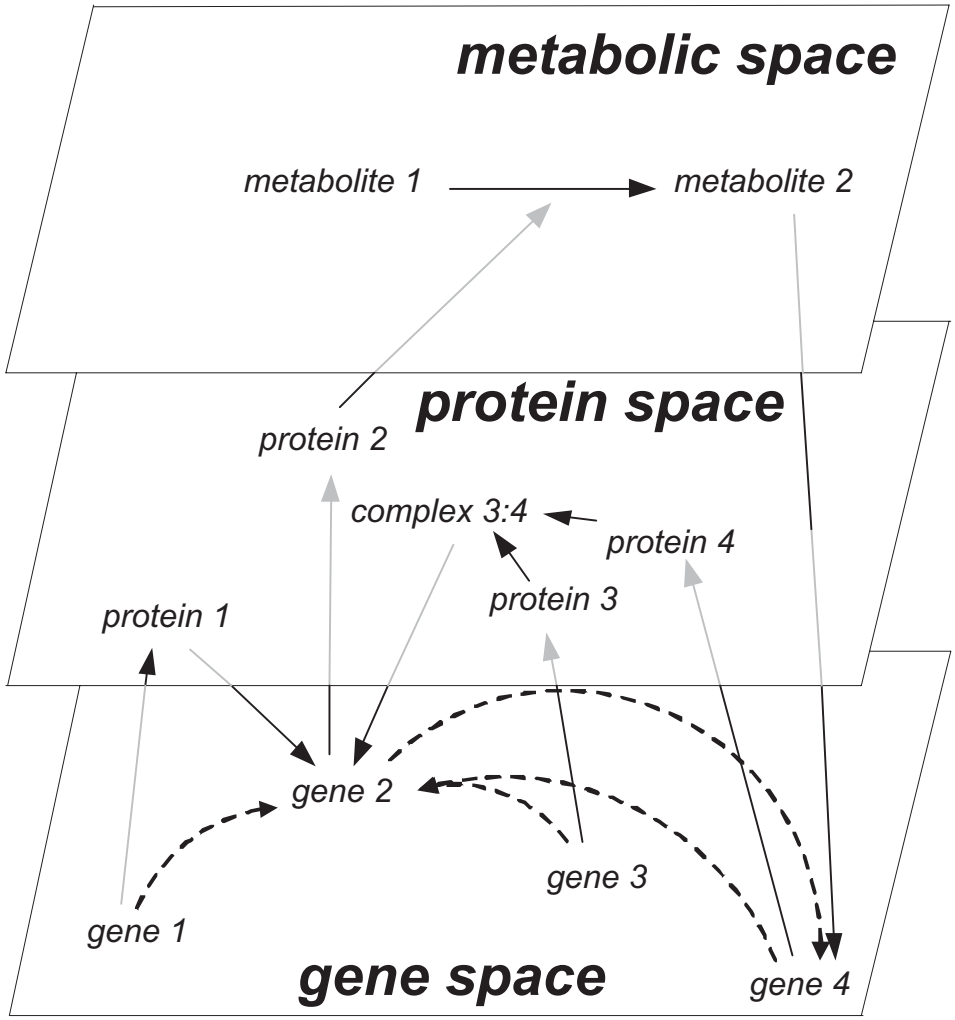

Figure 3.1 – *Example of a biochemical network. Molecular constituents (nodes of the network) are organized in three levels (spaces): mRNAs, proteins, and metabolites. Solid arrows indicate interactions. Three different mechanisms of regulation are shown: regulation of gene 2 by the product of the gene 1, protein 1; regulation of gene 2 by the complex 3:4 formed by the products of gene 3 and gene 4; and regulation of gene 4 by the metabolite 2, which in turn is produced by protein 2.* ***A:*** *Projections of these interactions into the "gene space", indicated by dashed arrows, constitute a corresponding gene network with many nodes not coinciding with actual genes (some arrows corresponding to synthesis and degradation were omitted to avoid cluttering).* ***B****: The approximate gene network used in most gene network analyses: only nodes that correspond to actual genes are considered and influences that run partly in parallel because they pass through common factors at the metabolic level are treated as if independent of one another.*

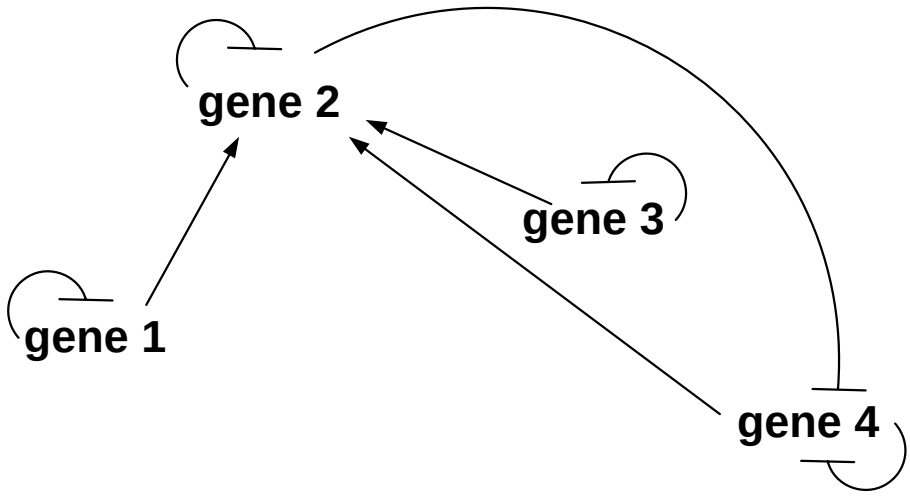

Figure 3.2 – *Simple graph representation of the gene network corresponding to the biochemical network in Figure 3.1. Lines show direct effects, with arrows standing for activation, and blunt ends for inhibition. The edges implicitly include the effects of the proteome and metabolome as shown in Figure 1. Most genes in gene networks will have a negative effect on their own concentration, because the degradation rate of their mRNA tends to be affected to the mRNA concentration.*

Section 3.1.1: Representations of gene networks

Gene networks are commonly represented by directed graphs as depicted in the example of Figure 3.2. The nodes of the graph are genes and the directed edges are direct causal relations between genes. A widely adopted norm is to use arrow tips on edges that represent positive interactions, where an increase in activity of the originating gene causes an increase in the target gene through that interaction, and blunt ends on edges that represent negative interactions, where an increase in activity of the originating gene causes a decrease in activity of the target gene through that interaction. Gene networks can also be represented through matrices, as illustrated in Eq. 3.1, which is the matrix representation of Figure 3.2 (for a specific condition). The matrix in Equation 3.1 was obtained applying the method of Regulatory Strengths [81-83, 179] (this will be described in more detail in the chapter 4) to simulated data. A row in this matrix corresponds to an effected gene, and the columns to the effector genes. Gene network matrices can be qualitative, in which positive interactions are represented by the number 1, negative interactions with the number –1 and 0 is written for the case of no interaction between genes or quantitative, in which case its elements take real values representing the strength and sign of the first order approximation of the interaction, such as in Eq 3.1.

$$\mathbf{R} = \begin{array}{c} \begin{matrix} G1 & G2 & G3 & G4 \end{matrix} \\ \begin{pmatrix} -1 & 0 & 0 & 0 \\ -0.4 & -0.92 & 0.44 & 0.55 \\ 0 & 0 & -1 & 0 \\ 0 & -0.14 & 0 & -0.92 \end{pmatrix} \end{array} \begin{matrix} \\ G1 \\ G2 \\ G3 \\ G4 \end{matrix} \qquad \text{(Eq. 3.1)}$$

An interaction between two genes is said to be *direct* if it does not run through any other genes in the network. For example, in Fig. 3.2, gene 1 directly affects gene 2. Gene 1 also affects gene 4, but only in an indirect way because this effect has to run through gene 2 first. Non-additive interactions are not captured in either of the two representations above, but the graph representation could be generalized to hyper-graphs, in which edges can connect more than two nodes: edges originating in all the cause genes and ending up in the effect gene (as is shown in Fig. 3.1).

Section 3.1.2: Relevance of studying gene networks

Knowledge of the structure of gene networks can be highly important for representing cell function with much more precision than is usually done with words [see (see 180]), even when these words are part of a controlled vocabulary, as they are in the Gene Ontology™ [181]. Gene networks will increase our knowledge of functions of genes in their context. For instance, from a sequence one may deduce that a certain gene codes for a protein kinase, but not the function of that kinase in the cellular context. If this gene is linked to genes involved in flagellum synthesis, it may conclude that it plays a role in the chemotaxis signal transduction pathway. In this sense, not only are gene networks (and especially their graphical representations) capable of describing a large number of interactions in a concise way but they may also come a long way towards representing the regulatory properties accompanying those interactions at a systems level. Cells exhibit complex interacting behavior that is usually not predictable from the properties of individual system components alone. Gene networks provide such a system view at the level of gene activities.

The detailed molecular mechanisms of how the products of one gene affect expression of another gene are often unknown but the effect itself may be observed in gene expression experiments. It is therefore appropriate and timely to use genome-wide gene expression data to identify gene networks, an important step towards uncovering the complete biochemical networks of cells. Research focused on developing methods for this identification of gene networks from microarray data is now an important part of bioinformatics.

Knowing the structure of gene networks and performing simulations of their behavior on computers will increase our fundamental understanding of living systems. Uncovering the structure of gene networks will then help us manipulate cells to our advantage, and may provide valuable clues for treating complex, devastating diseases. Knowledge about gene networks may help pharmaceutical research in discovering and prioritizing targets, eliminating toxic and ineffective compounds, tailoring drug therapy to the needs of a patient [182]. Indeed, single genes do not always affect genotype significantly, and most remaining major diseases result from a rather complex collective

interplay of the activities of a number of genes. It is the network between genes rather than the individual genes that matters. Metabolic engineering and drug targeting by manipulating single genes is limited [183]. Simultaneous manipulation of several genes may be needed in order to achieve a significant outcome. Gene networks will provide clues about which genes to manipulate simultaneously.

The number of genes in the human genome [184] may only be twice that of the worm *Caenorhabditis elegans* [185] (but see also[186]). There are several hypotheses pertaining to explain this relative "simplicity" of the human genome. One is that the average number of proteins encoded by each human gene is larger than that of proteins encoded by genes in other genomes [187]. Another one is that the proportion of regulatory genes (signaling proteins, transcription factors, etc.) in the human genome is higher than in other genomes. Yet another is that the human gene network has a higher average number of connections per gene than other genomes (which implies that the encoded proteins contain more binding sites). Both the latter two hypotheses could be tested by determining gene networks of various organisms and comparing them. Such studies could of course also reveal that the connections in the human gene network are not more, but just different and perhaps even fewer. Gene networks are then also well suited for comparative genomics.

Some studies [188, 189] indicate that the topology of gene networks might be largely responsible for the robustness shown by living organisms. A particular gene network topology might have been selected in evolution to originate the type of system robustness currently observed in many species against drastic perturbations at the genetic level (e.g., at least 40% of the genes of *Saccharomyces cerevisiae* can be removed without causing noticeable phenotypes under laboratory conditions). Comparative genomics done at the level of gene networks should be well suited to pursue this hypothesis.

Section 3.1.3: Connectivity of gene networks

Several authors [4, 47, 79, 80, 190] argue that gene networks are sparsely connected. However, there are simple arguments that suggest the opposite for gene networks that should describe true cell function of which a few will be listed here. The first one is a consequence of the connectivity of metabolic networks: if a certain metabolite is an effector of the transcription rate of a gene (usually by binding a transcription factor), then the genes coding for other enzymes that have large concentration control coefficients for that metabolite will also appear as interacting with the genes affected by that metabolite. One of the findings of Metabolic Control Analysis, i.e. concentration control is distributed, has major impact here: most metabolically regulated genes will be targeted by many other genes. The second argument comes from the fact that all transcription steps are dependent on metabolic energy. Consequently, genes that code for enzymes that have control on the energy level may interact with all genes. Third, the rates of transcription depend on the concentrations of nucleotides as they are the building blocks of nucleic acids; so all genes coding for enzymes involved in nucleotide synthesis may be inputs of all other genes. Fourth, if or when RNA polymerases exist at low concentrations, interactions between any two genes would arise by competition between the polymerase binding sites of these genes for the then scarce RNA polymerase molecules. In such a situation, increased expression of one gene would result in decreased expression of all others, generating inhibitions from any one gene to all others. A known corollary of this is the so-called protein-burden effect [191], in which it was shown that forced over-expression of one gene must compromise the

expression of all other genes and therewith compromise general cell functions such as growth rate. Fifth, any other genes that affect transcription, in some general way, will be inputs to all genes. For instance, genes that code for transporters that are responsible for transport of metabolites or of protein effectors of gene expression into the nucleus. There are many other examples of interactions that would arise from the complex interplay at the 'unobserved' proteome and metabolome. Except when a gene-gene interaction involves transcription factors, where it is clear that the protein product of one gene affects the expression of another, the meaning of gene-gene interactions in gene networks it is thus far from being obvious.

Whether these numerous potential interactions have a significant magnitude or not is still an open question. Certainly, some of these interactions may have small magnitude, for example in many situations there are plenty of nucleotides such that transcription rates are saturated with them, reducing the related interactions to values close to zero.

Section 3.2: Regulatory Strength as a quantitative measure for gene-gene interactions

Many mathematical frameworks have been proposed for describing gene networks, varying from Boolean logic to non-linear partial differential equations. Here I suggest describing gene networks in terms of Regulatory Strengths [108] (also called partial internal response coefficients, see http://www.sun.ac.za/biochem/mcanom.html) to quantify gene-gene interactions. It is a simple, but quantitative description, and has the benefit that there is a close link between the theoretical framework and the analysis of experimental data produced with current high scale experimental technologies.

Section 3.2.1: The Regulatory Strength

As explained in Chapter 1, Regulatory Strengths quantify the fractional changes in a system variable as a consequence of the change in another system variable through a specific path [108]. The meaning of the Regulatory Strength follows from the concentration connectivity theorem:

$$\sum_{r=1}^{n} C_{v_r}^{X_i} \cdot \varepsilon_{X_j}^{v_r} = \sum_{r=1}^{n} {}^{v_r}R_{X_j}^{X_i} = \begin{cases} -1, when\ i = j \\ 0, when\ i \neq j \end{cases}. \qquad \text{(Eq. 3.2)}$$

where X_i, X_j correspond to any independent concentration variable. Lets consider the gene network depicted in Figure 3.3, elaborated in terms of a hierarchical system [cf. [103]] in order to illustrate this.

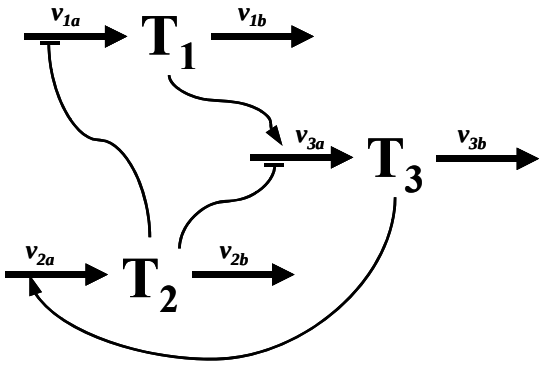

Figure 3.3 – *A simple three-gene network to illustrate the use of Regulatory Strengths. Thick arrows indicate mass flux and thin arrows regulatory interactions*

There are three ways for transcript T_2 to affect transcript T_3, one directly by inhibiting the synthesis rate v_{3a}, a second by affecting the synthesis rate of T_1, which stimulates the synthesis of T3, and a final one by affecting its own degradation rate, v_{2b}. Considering equation 3.2 with $i = T_3$ and $j = T_2$ one obtains:

$$C^{T_3}_{v_{1a}} \varepsilon^{v_{1a}}_{T_2} + C^{T_3}_{v_{2b}} \varepsilon^{v_{2b}}_{T_2} + C^{T_3}_{v_{3a}} \varepsilon^{v_{3a}}_{T_2} = {}^{v_{1a}}R^{T_3}_{T_2} + {}^{v_{2b}}R^{T_3}_{T_2} + {}^{v_{3a}}R^{T_3}_{T_2} = 0 \qquad \text{(Eq. 3.3)}$$

The three effects sum up to zero. Each individual term can be non-zero. The coefficients ${}^{v_{1a}}R^{T_3}_{T_2}$ and ${}^{v_{2b}}R^{T_3}_{T_2}$ correspond to indirect interactions, while ${}^{v_{3a}}R^{T_3}_{T_2}$ correspond to a direct interaction, since it quantifies the effect that T_2 has on T_3 by directly affecting the rate of synthesis of the latter.

In a similar way the connectivity relations can be used to express the effects of the other variables on each other and on themselves. It is important to distinguish between the direct Regulatory Strengths and the indirect ones. The direct Regulatory Strengths are of the form ${}^{v_j}R^{T_j}_{T_i}$, where v_j stands for the synthesis or degradation rate of transcript j, and can be used to describe the direct interaction structure of gene networks quantitatively. In chapter 4 I will describe a method that enables inference of these direct Regulatory Strengths from experimental data.

Section 3.2.2: Regulatory Strengths in terms of component control and elasticities

In the example in the previous section I have considered for simplicity that genes directly affected each other. In reality however, the interactions between genes run through proteins and metabolites. Here I will (i) show how the Regulatory Strengths quantifying the interactions between the genes depend on the properties of the underlying hierarchical system and (ii) generalize the concept of the Regulatory Strength of a gene-gene

interaction that only involve paths through a gene network to such a Regulatory Strength of a gene-gene interaction that involves the proteome or metabolome.

I start with the control matrix expression [106, 114] as introduced in Chapter 1. For definitions of the symbols refer to Chapter 2, Section 2.5.4.

$$\mathbf{CE} = \mathbf{I} \qquad \text{(Eq. 3.4)}$$

As in Chapter 2, Section 2.5.4 matrix $\mathbf{C}$ contains is partitioned according to control coefficients for all independent fluxes, $\mathbf{C^J}$ and all control coefficients for all independent concentrations, $\mathbf{C^S}$:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C^J} \\ \mathbf{C^S} \end{bmatrix} \qquad \text{(Eq. 3.5)}$$

Matrix $\mathbf{E}$ contains both structural and kinetic information,

$$\mathbf{E} = \begin{bmatrix} \boldsymbol{\kappa} & -\boldsymbol{\varepsilon}\mathbf{L} \end{bmatrix} \qquad \text{(Eq. 3.6)}$$

Eq. 3.4 is written explicitly in terms of the three functional levels of organization of cells: the transcriptome, proteome and metabolome

$$\begin{bmatrix} \mathbf{C}^{\mathbf{T}}_{\mathbf{v_T}} & \mathbf{C}^{\mathbf{T}}_{\mathbf{v_P}} & \mathbf{C}^{\mathbf{T}}_{\mathbf{v_M}} \\ \mathbf{C}^{\mathbf{P}}_{\mathbf{v_T}} & \mathbf{C}^{\mathbf{P}}_{\mathbf{v_P}} & \mathbf{C}^{\mathbf{P}}_{\mathbf{v_M}} \\ \mathbf{C}^{\mathbf{M}}_{\mathbf{v_T}} & \mathbf{C}^{\mathbf{M}}_{\mathbf{v_P}} & \mathbf{C}^{\mathbf{M}}_{\mathbf{v_M}} \end{bmatrix} \begin{bmatrix} \mathbf{E}^{\mathbf{v_T}}_{\mathbf{T}} & \mathbf{E}^{\mathbf{v_T}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_T}}_{\mathbf{M}} \\ \mathbf{E}^{\mathbf{v_P}}_{\mathbf{T}} & \mathbf{E}^{\mathbf{v_P}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_P}}_{\mathbf{M}} \\ \mathbf{E}^{\mathbf{v_M}}_{\mathbf{T}} & \mathbf{E}^{\mathbf{v_M}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_M}}_{\mathbf{M}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad \text{(Eq. 3.7)}$$

Submatrices $\mathbf{C}^{j}_{\mathbf{v}_i}$ give the control of the rates in level *i* on the fluxes and concentrations in level *j*. Similarly, submatrices $\mathbf{E}_i^{\mathbf{v}_j}$ contain elasticities of the rates in level *i* towards the concentrations of level *j*. In order not to interrupt the derivation here, I will give a biochemical interpretation of these submatrices at the end of this section. Submatrices $\mathbf{C}^{i}_{\mathbf{v}_i}$ and $\mathbf{E}_i^{\mathbf{v}_i}$ are square and assumed invertible. The submatrices in **E** are partitioned as in [170] and [111]:

$$\mathbf{E}_i^{\mathbf{v}_i} = \begin{bmatrix} \boldsymbol{\kappa}_i & -\boldsymbol{\varepsilon}_i^{\mathbf{v}_i}\mathbf{L}_i \end{bmatrix} \qquad \text{(Eq. 3.8)}$$

$$\mathbf{E}_i^{\mathbf{v}_j} = \begin{bmatrix} \mathbf{0} & -\boldsymbol{\varepsilon}_i^{\mathbf{v}_j}\mathbf{L}_i \end{bmatrix} \qquad \text{(Eq. 3.9)}$$

To be able to partition as in Eq. 3.9, the $\boldsymbol{\kappa}$ and $\mathbf{L}$ matrices have to be block diagonal. This is the case when there is no mass exchange between the three levels or when the mass exchange is negligibly small [111].

Control coefficients can be obtained by inverting matrix $\mathbf{E}$ [106] but the inverse is also true; the elasticities and pathway structure can be obtained from the control matrix [114].

Using the relationship for the inverse of block matrices [192], the inverse of a matrix can be expressed in terms of its blocks:

$$\begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \mathbf{U} \end{bmatrix}\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

(Eq. 3.10)

$$\Rightarrow \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \mathbf{U} \end{bmatrix} = \begin{bmatrix} \left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \left(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{bmatrix}$$

assuming that matrices A and D are square and non-singular.

Our aim here is to show that the activity of the protein and metabolite levels in the system do not compromise the ability to describe the steady-state behavior of the system in terms of the gene level alone. Accordingly it will suffice to divide the system into two parts, *i.e.* that of the gene network (the mRNA's) and the remainder, comprising the protein and the metabolite levels. Accordingly, using the relationship in Eq. 3.10, Eq. 3.7 is rewritten as:

$$\begin{bmatrix} \mathbf{C}^{\mathbf{T}}_{\mathbf{v}_T} & \mathbf{C}^{\mathbf{T}}_{\mathbf{v}_r} \\ \mathbf{C}^{\mathbf{r}}_{\mathbf{v}_T} & \mathbf{C}^{\mathbf{r}}_{\mathbf{v}_r} \end{bmatrix}\begin{bmatrix} \mathbf{E}^{\mathbf{v}_T}_{\mathbf{T}} & \mathbf{E}^{\mathbf{v}_T}_{\mathbf{r}} \\ \mathbf{E}^{\mathbf{v}_r}_{\mathbf{T}} & \mathbf{E}^{\mathbf{v}_r}_{\mathbf{r}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

(Eq. 3.11)

where

$$\mathbf{C}^{\mathbf{r}}_{\mathbf{v}_T} = \begin{bmatrix} \mathbf{C}^{\mathbf{P}}_{\mathbf{v}_T} \\ \mathbf{C}^{\mathbf{M}}_{\mathbf{v}_T} \end{bmatrix}, \mathbf{C}^{\mathbf{T}}_{\mathbf{v}_r} = \begin{bmatrix} \mathbf{C}^{\mathbf{T}}_{\mathbf{v}_P} & \mathbf{C}^{\mathbf{T}}_{\mathbf{v}_M} \end{bmatrix}, \mathbf{C}^{\mathbf{r}}_{\mathbf{v}_r} = \begin{bmatrix} \mathbf{C}^{\mathbf{P}}_{\mathbf{v}_P} & \mathbf{C}^{\mathbf{P}}_{\mathbf{v}_M} \\ \mathbf{C}^{\mathbf{M}}_{\mathbf{v}_P} & \mathbf{C}^{\mathbf{M}}_{\mathbf{v}_M} \end{bmatrix},$$
$$\mathbf{E}^{\mathbf{v}_T}_{\mathbf{r}} = \begin{bmatrix} \mathbf{E}^{\mathbf{v}_T}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v}_T}_{\mathbf{M}} \end{bmatrix}, \mathbf{E}^{\mathbf{v}_r}_{\mathbf{T}} = \begin{bmatrix} \mathbf{E}^{\mathbf{v}_P}_{\mathbf{T}} \\ \mathbf{E}^{\mathbf{v}_M}_{\mathbf{T}} \end{bmatrix} \text{and } \mathbf{E}^{\mathbf{v}_r}_{\mathbf{r}} = \begin{bmatrix} \mathbf{E}^{\mathbf{v}_P}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v}_P}_{\mathbf{M}} \\ \mathbf{E}^{\mathbf{v}_M}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v}_M}_{\mathbf{M}} \end{bmatrix}.$$

(Eq. 3.12)

Here *r* stands for the *remainder* of the system: in this case the proteome and metabolome together.

According to Eq. 3.10, the sub matrix $\mathbf{C}^{\mathbf{T}}_{\mathbf{v}_T}$ can be expressed as:

$$\mathbf{C}^{\mathbf{T}}_{\mathbf{v}_\mathbf{T}} = \left(\mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} - \mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{r}} \left(\mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{r}}\right)^{-1} \mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{T}}\right)^{-1} \qquad \text{(Eq. 3.13)}$$

The inverse of $\mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{r}}$ equals a matrix of control coefficients quantifying control on the proteome and metabolome if the concentrations on the transcriptome levels remain fixed. To distinguish between the real control and the control in subsystems, I use the symbols **C**, **€** and **c**. **C** stands for control in the whole system, **€** for the control in two subsystems united (in this case the proteome and metabolome) if the third is fixed (in this case the transcriptome) and **c** for the control within each subsystem with the other two fixed (cf. the intra-level control of Chapter 2).

In general, the control of the processes at the transcriptome level on the variables at that same level (i.e., mRNA concentrations and gene expression fluxes) can thus be expressed as:

$$\mathbf{C}^{\mathbf{T}}_{\mathbf{v}_\mathbf{T}} = \left(\mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} - \mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{r}} \boldsymbol{€}^{\mathbf{r}}_{\mathbf{v}_\mathbf{r}} \mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{T}}\right)^{-1}. \qquad \text{(Eq. 3.14)}$$

Matrix $\mathbf{C}^{\mathbf{T}}_{\mathbf{v}_\mathbf{T}}$ can be partitioned as:

$$\begin{bmatrix} \mathbf{C}^{\mathbf{J}_\mathbf{T}}_{\mathbf{v}_\mathbf{T}} \\ \mathbf{C}^{[\mathbf{T}]}_{\mathbf{v}_\mathbf{T}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\kappa}_\mathbf{T} & -\boldsymbol{\upsilon}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} \end{bmatrix}^{-1}. \qquad \text{(Eq. 3.15)}$$

$\boldsymbol{\kappa}_\mathbf{T}$ is the kernel of the gene network (which in Eq. 3.6 is a part of $\mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}}$ and is unaffected by the subtraction of $\mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{r}} \boldsymbol{€}^{\mathbf{r}}_{\mathbf{v}_\mathbf{r}} \mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{T}}$, since the structural parts of both $\mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{r}}$ and $\mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{T}}$ are zero sub matrices).

$\boldsymbol{\upsilon}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} = \mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} + \mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{r}} \left(\mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{r}}\right)^{-1} \mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{T}} = \mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} + \mathbf{E}^{\mathbf{v}_\mathbf{T}}_{\mathbf{r}} \boldsymbol{€}^{\mathbf{r}}_{\mathbf{v}_\mathbf{r}} \mathbf{E}^{\mathbf{v}_\mathbf{r}}_{\mathbf{T}}$ is a matrix containing 'apparent' elasticities (similar to global elasticity in [170], and overall elasticity in [193]) of rates of gene expression towards concentrations of mRNAs. The scaled link matrix of the transcriptome, $\mathbf{L}_\mathbf{T}$, is always an identity matrix since there are no conservation relations between mRNA concentrations.

Eq. 3.15 contains all the flux and concentrations control coefficients of the gene network. Accordingly, Eq. 3.15 shows that for steady state control, the mRNA levels can be described as a gene network, even if that network is not only connected internally but also through external loops that lead through proteins and metabolites. This proof is general, as it only requires Eq. 3.7 to be valid and the both the system as a whole and its subsystems to be in stable steady state.

Rewriting Eq. 3.15 in the form of Eq. 3.4 yields:

$$\begin{bmatrix} \mathbf{C}^{\mathbf{J}_\mathbf{T}}_{\mathbf{v}_\mathbf{T}} \\ \mathbf{C}^{[\mathbf{T}]}_{\mathbf{v}_\mathbf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\kappa}_\mathbf{T} & -\boldsymbol{\upsilon}^{\mathbf{v}_\mathbf{T}}_{\mathbf{T}} \end{bmatrix} = \mathbf{I} \qquad \text{(Eq. 3.16)}$$

Considering only concentration control coefficients leaves us with:

$$-\mathbf{C}^{[\mathbf{T}]}_{\mathbf{v_T}}\boldsymbol{\upsilon}^{\mathbf{v_T}}_{\mathbf{T}} = \mathbf{I} \qquad \text{(Eq. 3.17)}$$

Matrices $\mathbf{C}^{[\mathbf{T}]}_{\mathbf{v_T}}$ and $\boldsymbol{\upsilon}^{\mathbf{v_T}}_{\mathbf{T}}$ are in general not square and thus one cannot write one of the matrices in Eq. 3.17 as minus the inverse of the other. However, the specific stoichiometry of the transcriptome enables us to overcome this problem. Gene networks are composed of flux-decoupled modules, each consisting of two reactions: one producing a specific mRNA (transcription) and the other degrading this mRNA. Due to this particular stoichiometry, the summation theorems for concentration control coefficients from Hierarchical Control Analysis [103, 111, 116, 160] enable us to reduce the matrices. Writing out Eq. 3.17 for a three-gene network explicitly:

$$-\begin{bmatrix} C^{T1}_{v_{ts_1}} & C^{T1}_{v_{td_1}} & C^{T1}_{v_{ts_2}} & C^{T1}_{v_{td_2}} & C^{T1}_{v_{ts_3}} & C^{T1}_{v_{td_3}} \\ C^{T2}_{v_{ts_1}} & C^{T2}_{v_{td_1}} & C^{T2}_{v_{ts_2}} & C^{T2}_{v_{td_2}} & C^{T2}_{v_{ts_3}} & C^{T2}_{v_{td_3}} \\ C^{T3}_{v_{ts_1}} & C^{T3}_{v_{td_1}} & C^{T3}_{v_{ts_2}} & C^{T3}_{v_{td_2}} & C^{T3}_{v_{ts_3}} & C^{T3}_{v_{td_3}} \end{bmatrix} \begin{bmatrix} \vartheta^{v_{ts_1}}_{T1} & \vartheta^{v_{ts_1}}_{T2} & \vartheta^{v_{ts_1}}_{T3} \\ \vartheta^{v_{td_1}}_{T1} & \vartheta^{v_{td_1}}_{T2} & \vartheta^{v_{td_1}}_{T3} \\ \vartheta^{v_{ts_2}}_{T1} & \vartheta^{v_{ts_2}}_{T2} & \vartheta^{v_{ts_2}}_{T3} \\ \vartheta^{v_{td_2}}_{T1} & \vartheta^{v_{td_2}}_{T2} & \vartheta^{v_{td_2}}_{T3} \\ \vartheta^{v_{ts_3}}_{T1} & \vartheta^{v_{ts_3}}_{T2} & \vartheta^{v_{ts_3}}_{T3} \\ \vartheta^{v_{td_3}}_{T1} & \vartheta^{v_{td_3}}_{T2} & \vartheta^{v_{td_3}}_{T3} \end{bmatrix} = \mathbf{I}. \qquad \text{(Eq. 3.18)}$$

Here, v_{tsi} and v_{tdi} stand for rate of synthesis and rate of degradation of transcript T_i, respectively. Carrying out the matrix multiplication in Eq. 3.18 explicitly, one obtains expressions similar to the connectivity theorems for concentration control [109]:

$$\begin{aligned} &-C^{T1}_{v_{ts_1}}\vartheta^{v_{ts_1}}_{T1} - C^{T1}_{v_{td_1}}\vartheta^{v_{td_1}}_{T1} - C^{T1}_{v_{ts_2}}\vartheta^{v_{ts_2}}_{T1} - C^{T1}_{v_{td_2}}\vartheta^{v_{td_2}}_{T1} - C^{T1}_{v_{ts_3}}\vartheta^{v_{ts_3}}_{T1} - C^{T1}_{v_{td_3}}\vartheta^{v_{td_3}}_{T1} = 1 \\ &-C^{T1}_{v_{ts_1}}\vartheta^{v_{ts_1}}_{T2} - C^{T1}_{v_{td_1}}\vartheta^{v_{td_1}}_{T2} - C^{T1}_{v_{ts_2}}\vartheta^{v_{ts_2}}_{T2} - C^{T1}_{v_{td_2}}\vartheta^{v_{td_2}}_{T2} - C^{T1}_{v_{ts_3}}\vartheta^{v_{ts_3}}_{T1} - C^{T1}_{v_{td_3}}\vartheta^{v_{td_3}}_{T2} = 0 \end{aligned} \quad \textit{etc.} \text{ (Eq. 3.19)}$$

The summation theorems for concentration control in hierarchical systems [103] imply:

$$C^{i}_{v_{td_j}} = -C^{i}_{v_{ts_j}}, \qquad \text{(Eq. 3.20)}$$

where i is mRNA T_1, T_2 or T_3 and j refers to rate number. Using these relationships, Eq. 3.19 can be written in terms of half the number of control coefficients:

$$\begin{aligned} &-C^{T1}_{v_{ts_1}}\left(\vartheta^{v_{ts_1}}_{T1} - \vartheta^{v_{td_1}}_{T1}\right) - C^{T1}_{v_{ts_2}}\left(\vartheta^{v_{ts_2}}_{T1} - \vartheta^{v_{td_2}}_{T1}\right) - C^{T1}_{v_{ts_3}}\left(\vartheta^{v_{ts_3}}_{T1} - \vartheta^{v_{td_3}}_{T1}\right) = 1 \\ &-C^{T1}_{v_{ts_1}}\left(\vartheta^{v_{ts_1}}_{T2} - \vartheta^{v_{td_1}}_{T2}\right) - C^{T1}_{v_{ts_2}}\left(\vartheta^{v_{ts_2}}_{T2} - \vartheta^{v_{td_2}}_{T2}\right) - C^{T1}_{v_{ts_3}}\left(\vartheta^{v_{ts_3}}_{T1} - \vartheta^{v_{td_3}}_{T2}\right) = 0 \end{aligned}. \qquad \text{(Eq. 3.21)}$$

Which can be expressed in matrix format as:

$$-\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\boldsymbol{\upsilon}_\mathbf{T}^{\boldsymbol{\nu}_\mathbf{T}\,*}=\begin{pmatrix} C_{\nu_{T1}}^{T1} & C_{\nu_{T2}}^{T1} & C_{\nu_{T3}}^{T1} \\ C_{\nu_{T1}}^{T2} & C_{\nu_{T2}}^{T2} & C_{\nu_{T3}}^{T2} \\ C_{\nu_{T1}}^{T3} & C_{\nu_{T2}}^{T3} & C_{\nu_{T3}}^{T3} \end{pmatrix}\begin{pmatrix} \phi_{T1}^{\nu_{T1}} & \phi_{T2}^{\nu_{T1}} & \phi_{T3}^{\nu_{T1}} \\ \phi_{T1}^{\nu_{T2}} & \phi_{T2}^{\nu_{T2}} & \phi_{T3}^{\nu_{T2}} \\ \phi_{T1}^{\nu_{T3}} & \phi_{T2}^{\nu_{T3}} & \phi_{T3}^{\nu_{T3}} \end{pmatrix}=\mathbf{I}\,, \qquad \text{(Eq. 3.22)}$$

where $C_{\nu_j}^{i}=C_{\nu_{\text{production of }j}}^{i}=-C_{\nu_{\text{degradation of }j}}^{i}$ and $\phi_j^{\nu_i}=\vartheta_j^{\nu_{\text{production of }i}}-\vartheta_j^{\nu_{\text{degradation of }i}}$. Eq. 3.22 contains square matrices, which can be inverted. Thus, one can be expressed as the inverse of the other. In the elasticity-type matrix the individual transcription and degradation rates are not considered explicitly anymore in terms of their dependence on the concentrations of mRNA's, but only the ratio of synthesis to degradation rate Here, information is lost to distinguish between effectors that act on the transcription process or on the degradation of the mRNA. However, the number of control coefficients under consideration is reduced to 25% of the original number of control coefficients in Eq. 3.4, which greatly simplifies the experimental determination of these coefficients (see below). For the gene network representation it is thus only of interest that there is an effect of a gene on another, not about details such as whether it acts on transcription or degradation.

In general we thus have

$$-\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\boldsymbol{\upsilon}_\mathbf{T}^{\boldsymbol{\nu}_\mathbf{T}\,*}=\mathbf{I}\,. \qquad \text{(Eq. 3.23)}$$

The asterisk is used to indicate that now is dealt with reduced matrices. $\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}$ contains only a set of concentration control coefficients and $\boldsymbol{\upsilon}_\mathbf{T}^{\boldsymbol{\nu}_\mathbf{T}\,*}$ contains aggregated apparent elasticities.

Eq. 3.23 enables one to calculate the new set of concentration control coefficients for the gene network.

Because co-control coefficients are experimentally more feasible to determine than control coefficients, a transformation of Eq. 3.23 is used similar to the one suggested by Hofmeyr *et al.* [104, 105]:

$$-\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\left(Dg\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\right)^{-1}Dg\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\boldsymbol{\upsilon}_\mathbf{T}^{\boldsymbol{\nu}_\mathbf{T}\,*}=\mathbf{I} \qquad \text{(Eq. 3.24)}$$

where $Dg\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}$ is the diagonal matrix having the elements of $\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}$ on its diagonal. Defining

$$-\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\left(Dg\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\right)^{-1}\equiv\mathbf{O}$$
$$Dg\mathbf{C}_{\boldsymbol{\nu}_\mathbf{T}}^{[\mathbf{T}]^*}\boldsymbol{\upsilon}_\mathbf{T}^{\boldsymbol{\nu}_\mathbf{T}\,*}\equiv\mathbf{R} \qquad \text{(Eq. 3.25)}$$

One finds:

$$\mathbf{O} \cdot \mathbf{R} = \mathbf{I}, \qquad \text{(Eq. 3.26)}$$

where $\mathbf{O}$ is the co-control matrix [104, 105] and $\mathbf{R}$ a matrix of Regulatory Strengths [108].

By determining all co-control coefficients of the part of the network that consists of the genes and the using the above equation, one can calculate the Regulatory Strengths quantifying the interactions between the genes of that sub- network

$$\mathbf{R} = \mathbf{O}^{-1} \qquad \text{(Eq. 3.27)}$$

This means that the gene network can be used as a representation of the entire network. Eq. 3.25 also enables us to reverse engineer gene networks from microarray experiments, where rates of transcription are perturbed one by one. This will be dealt with in more detail in Chapter 4.

The elements of $\mathbf{E}_{\mathbf{T}}^{\mathbf{v}_T}$ show the effect of the transcript concentrations on the rates of change of transcript concentrations. These effects are mainly due to the degradation rates, since each transcript increases its own degradation rate, transcripts do not interfere with the synthesis or degradation of other transcripts and transcription is an irreversible process. In the simplest case $\mathbf{E}_{\mathbf{T}}^{\mathbf{v}_T}$ is merely a lower diagonal matrix with negative numbers. However, if the RNases are saturated with mRNA, or if there are regulatory processes of the transcriptome on itself through the mechanism of RNAi, in which the presence of short double-stranded RNA prevents translation of the corresponding mRNA and activates its degradation [194], $\mathbf{E}_{\mathbf{T}}^{\mathbf{v}_T}$ becomes more complicated.

The elements of $\mathbf{E}_{\mathbf{P}}^{\mathbf{v}_T}$ represent the effects of the protein concentrations on the rates of change of transcript concentrations. RNA-polymerases, transcription factors and RNases, for example, are some of the proteins involved in these effects. Also the proteins that make up the spliceosome and proteins that transport mRNA from the nucleus to the cytoplasm will appear in this sub matrix.

$\mathbf{E}_{\mathbf{M}}^{\mathbf{v}_T}$ describes the effect of the metabolites on the rate of change of transcript concentrations. Certain metabolites interfere with the transcription of genes by changing the binding affinities of regulating proteins, leading to a change in transcript formation rate. For example, in tryptophan synthesis in *E. coli*, the *trp*-operon is inhibited by the concentration of L-tryptophan.

$\mathbf{E}_{\mathbf{T}}^{\mathbf{v}_P}$ describes the effects of the transcriptome on the proteome. Each mRNA codes for a protein, thereby increasing the rate of its formation. The column belonging to rRNA will have positive values in almost every row, since they are part of the ribosomes and thus stimulate the formation rate of all proteins.

$\mathbf{E}_{\mathbf{P}}^{\mathbf{v}_P}$ contains information of many different types of interaction between proteins. The columns of proteases will have many negative elements; ribosomal proteins will have positive entries in almost all rows. The effects of phosphatases and kinases, and

other components of signaling cascades, appear in this sub matrix, as well as any other form of protein-protein interaction.

$\mathbf{E}_{\mathbf{M}}^{\mathbf{v}_P}$ shows the effects of metabolites on rate changes in the proteome. Some metabolites interfere with the synthesis or degradation of proteins. For example, protein synthesis and many post-translation modification reactions depend on ATP, GTP and other metabolite concentrations. The intra-lysosomal pH should be considered a metabolic variable and will affect the proteins that end up in the lysosome.

$\mathbf{E}_{\mathbf{T}}^{\mathbf{v}_M}$ would represent the rare cases of ribozymes catalyzing metabolic reactions, and most entries can be expected to be zero.

$\mathbf{E}_{\mathbf{P}}^{\mathbf{v}_M}$ mainly contains the effects of metabolic enzymes on the rates of change of substrates and products of the reactions it catalyses. Also contained are the effects of transporters that pump metabolites in and out the cell.

$\mathbf{E}_{\mathbf{M}}^{\mathbf{v}_M}$ describes the effects metabolites have on the rate of change of metabolite concentrations. These are substrates and products of metabolic reactions and modifiers of reaction rates.

Section 3.3: Examples

Here I explore the above-derived equations for a simple hierarchical system with different types of wiring structures. It will be shown systematically how a hierarchical system can be 'condensed' into a gene network.

Section 3.3.1: Example 1. Hierarchical system with feedback from metabolism to gene expression

In the following example, the Regulatory Strengths between genes in a simple model system is derived. The system consists of just two communicating genes. This enables us to explain clearly how the Regulatory Strengths are determined by the properties of the whole system.

Fig. 3.4 shows a small hierarchical biochemical system consisting of two mRNAs, two proteins and a short metabolic pathway of two steps and one metabolic intermediate. There is feedback from the metabolite to the rates of transcript synthesis.

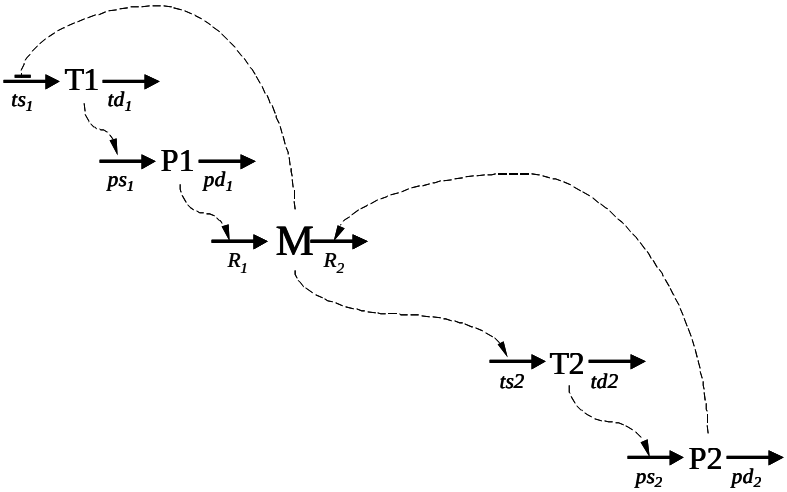

Figure 3.4 – *A system consisting of two mRNAs, two proteins and a short metabolic pathway of two steps and one metabolic intermediate. There is feedback from the metabolome to the transcriptome. T, P and M represent transcript (mRNA), protein and metabolite, respectively. ts and td stand for rate of transcript synthesis and rate of transcript degradation, respectively; ps and pd stand for rate of protein synthesis and rate of transcript degradation, respectively; and R_1 and R_2 for metabolic synthesis and degradation rates, respectively*.

The corresponding elasticity matrix is:

$$\mathbf{E}=\begin{bmatrix} \mathbf{E}_{\mathbf{T}}^{\mathbf{v_T}} & \mathbf{0} & \mathbf{E}_{\mathbf{M}}^{\mathbf{v_T}} \\ \mathbf{E}_{\mathbf{T}}^{\mathbf{v_P}} & \mathbf{E}_{\mathbf{P}}^{\mathbf{v_P}} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{\mathbf{P}}^{\mathbf{v_M}} & \mathbf{E}_{\mathbf{M}}^{\mathbf{v_M}} \end{bmatrix}= \left[\begin{array}{cccc|cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_M^{v_{ts_1}} \\ 1 & -\varepsilon_{T1}^{v_{td_1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_M^{v_{ts_2}} \\ 0 & 0 & 1 & -\varepsilon_{T2}^{v_{td_2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -\varepsilon_{T1}^{v_{ps_1}} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P1}^{v_{pd_1}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_{T2}^{v_{ps_2}} & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P2}^{v_{pd_2}} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P1}^{v_{R_1}} & 0 & 0 & 1 & -\varepsilon_M^{v_{R_1}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P2}^{v_{R_2}} & 1 & -\varepsilon_M^{v_{R_2}} \end{array}\right] \quad \text{(Eq. 3.28)}$$

As in Eq. 3.13 $\mathbf{C}_{\mathbf{v_T}}^{\mathbf{T}}$ is written as

$$\mathbf{C}^{\mathbf{T}}_{\nu_{\mathbf{T}}} = \left(\mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{T}} - \mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{r}}\left(\mathbf{E}^{\nu_{\mathbf{r}}}_{\mathbf{r}}\right)^{-1}\mathbf{E}^{\nu_{\mathbf{r}}}_{\mathbf{T}}\right)^{-1} \qquad \text{(Eq. 3.29)}$$

or explicit in the three subsystems:

$$\mathbf{C}^{\mathbf{T}}_{\nu_{\mathbf{T}}} = \left(\mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{T}} - \begin{bmatrix}\mathbf{0} & \mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{M}}\end{bmatrix}\begin{bmatrix}\mathbf{E}^{\nu_{\mathbf{P}}}_{\mathbf{P}} & \mathbf{0}\\ \mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{P}} & \mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{M}}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{E}^{\nu_{\mathbf{P}}}_{\mathbf{T}}\\ \mathbf{0}\end{bmatrix}\right)^{-1} \qquad \text{(Eq. 3.30)}$$

Because of the relation in Eq. 3.10:

$$\begin{bmatrix}\mathbf{E}^{\nu_{\mathbf{P}}}_{\mathbf{P}} & \mathbf{0}\\ \mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{P}} & \mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{M}}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{c}^{\mathbf{P}}_{\nu_{\mathbf{P}}} & \mathbf{0}\\ -\mathbf{c}^{\mathbf{M}}_{\nu_{\mathbf{M}}}\mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\nu_{\mathbf{P}}} & \mathbf{c}^{\mathbf{M}}_{\nu_{\mathbf{M}}}\end{bmatrix} \qquad \text{(Eq. 3.31)}$$

Substituting this result in Eq. 3.30, one obtains:

$$\mathbf{C}^{\mathbf{T}}_{\nu_{\mathbf{T}}} = \left(\mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{T}} - \begin{bmatrix}\mathbf{0} & \mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{M}}\end{bmatrix}\begin{bmatrix}\mathbf{c}^{\mathbf{P}}_{\nu_{\mathbf{P}}} & \mathbf{0}\\ -\mathbf{c}^{\mathbf{M}}_{\nu_{\mathbf{M}}}\mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\nu_{\mathbf{P}}} & \mathbf{c}^{\mathbf{M}}_{\nu_{\mathbf{M}}}\end{bmatrix}\begin{bmatrix}\mathbf{E}^{\nu_{\mathbf{P}}}_{\mathbf{T}}\\ \mathbf{0}\end{bmatrix}\right)^{-1} \qquad \text{(Eq. 3.32)}$$

Multiplying the sub matrices gives:

$$\mathbf{C}^{\mathbf{T}}_{\nu_{\mathbf{T}}} = \left(\mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{T}} + \mathbf{E}^{\nu_{\mathbf{T}}}_{\mathbf{M}}\mathbf{c}^{\mathbf{M}}_{\nu_{\mathbf{M}}}\mathbf{E}^{\nu_{\mathbf{M}}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\nu_{\mathbf{P}}}\mathbf{E}^{\nu_{\mathbf{P}}}_{\mathbf{T}}\right)^{-1} \qquad \text{(Eq. 3.33)}$$

The Regulatory Strengths are calculated by dropping flux control (Eqs. 3.17-3.22) and multiplying the matrix of apparent elasticities, $\boldsymbol{\upsilon}^{\nu_{\mathbf{T}}^{*}}_{\mathbf{T}}$, by $diag\mathbf{C}^{[\mathbf{T}]^{*}}_{\nu_{\mathbf{T}}}$ (Eq. 3.24):

$$\begin{bmatrix}R^{T1}_{T1} & R^{T1}_{T2}\\ R^{T2}_{T1} & R^{T2}_{T2}\end{bmatrix} =$$
$$\begin{bmatrix}\left(\varepsilon^{\nu_{ps_1}}_{T1}c^{P1}_{\nu_{ps_1}}\varepsilon^{\nu_{R_1}}_{P1}c^{M}_{\nu_{R_1}}\varepsilon^{\nu_{ts_1}}_{M} - \varepsilon^{\nu_{td_1}}_{T1}\right)C^{T1}_{\nu_{T1}} & \varepsilon^{\nu_{ps_2}}_{T2}c^{P2}_{\nu_{ps_2}}\varepsilon^{\nu_{R_1}}_{P2}c^{M}_{\nu_{R_1}}\varepsilon^{\nu_{ts_1}}_{M}C^{T1}_{\nu_{T1}}\\ \varepsilon^{\nu_{ps_1}}_{T1}c^{P1}_{\nu_{ps_1}}\varepsilon^{\nu_{R_1}}_{P1}c^{M}_{\nu_{R_1}}\varepsilon^{\nu_{ts_2}}_{M}C^{T2}_{\nu_{T2}} & \left(\varepsilon^{\nu_{ps_2}}_{T2}c^{P2}_{\nu_{ps_2}}\varepsilon^{\nu_{R_1}}_{P2}c^{M}_{\nu_{R_1}}\varepsilon^{\nu_{ts_2}}_{M} - \varepsilon^{\nu_{td_2}}_{T2}\right)C^{T2}_{\nu_{T2}}\end{bmatrix} \qquad \text{(Eq. 3.34)}$$

The effects of the genes on themselves consist of two parts. One part runs through degradation directly, since the degradation rate is dependent on the concentration of mRNA (thus expressed as a 'fundamental' elasticity). An indirect effect of an mRNA on its transcription through the protein and metabolite is quantified by an apparent elasticity, which is expressed as a product of elasticities and local control on the path of interaction. The interaction between the genes is similarly expressed in terms of the effects along the interaction path.

Section 3.3.2: Example 2. Hierarchical system with feedback from metabolism and proteins to gene expression

Here the system depicted in Fig. 3.5 is considered.

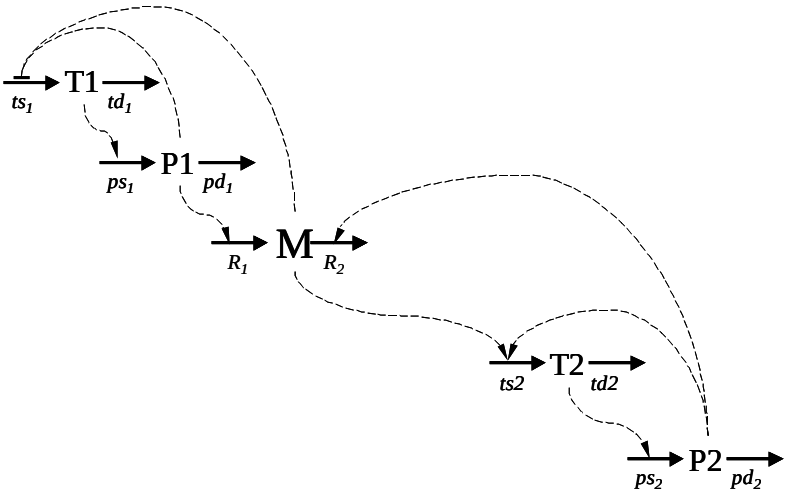

Figure 3.5 – *A system consisting of two mRNAs, two proteins and a short metabolic pathway of two steps and one metabolic intermediate. There is feedback from the metabolite and protein to the transcript synthesis rates. Abbreviations are as in the legend of Fig. 3.4.*

The corresponding elasticity matrix is:

$$\mathbf{E}=\begin{bmatrix} \mathbf{E}_{\mathbf{T}}^{\nu_{\mathbf{T}}} & \mathbf{E}_{\mathbf{P}}^{\nu_{\mathbf{T}}} & \mathbf{E}_{\mathbf{M}}^{\nu_{\mathbf{T}}} \\ \mathbf{E}_{\mathbf{T}}^{\nu_{\mathbf{P}}} & \mathbf{E}_{\mathbf{P}}^{\nu_{\mathbf{P}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{\mathbf{P}}^{\nu_{\mathbf{M}}} & \mathbf{E}_{\mathbf{M}}^{\nu_{\mathbf{M}}} \end{bmatrix} = \left[\begin{array}{cccc|cccc|cc} 1 & 0 & 0 & 0 & 0 & -\varepsilon_{P1}^{\nu_{ts_1}} & 0 & 0 & 0 & -\varepsilon_{M}^{\nu_{ts_1}} \\ 1 & -\varepsilon_{T1}^{\nu_{td_1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -\varepsilon_{P2}^{\nu_{ts_2}} & 0 & -\varepsilon_{M}^{\nu_{ts_2}} \\ 0 & 0 & 1 & -\varepsilon_{T2}^{\nu_{td_2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -\varepsilon_{T1}^{\nu_{ps_1}} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P1}^{\nu_{pd_1}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_{T2}^{\nu_{ps_2}} & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P2}^{\nu_{pd_2}} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P1}^{\nu_{R_1}} & 0 & 0 & 1 & -\varepsilon_{M}^{\nu_{R_1}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P2}^{\nu_{R_2}} & 1 & -\varepsilon_{M}^{\nu_{R_2}} \end{array}\right]$$

(Eq. 3.35)

As in Eq. 3.13, $\mathbf{C}^{\mathbf{T}}_{\mathbf{v_T}}$ is written as:

$$\mathbf{C}^{\mathbf{T}}_{\mathbf{v_T}} = \left(\mathbf{E}^{\mathbf{v_T}}_{\mathbf{T}} - \mathbf{E}^{\mathbf{v_T}}_{\mathbf{r}}\left(\mathbf{E}^{\mathbf{v_r}}_{\mathbf{r}}\right)^{-1}\mathbf{E}^{\mathbf{v_r}}_{\mathbf{T}}\right)^{-1}, \qquad \text{(Eq. 3.36)}$$

or explicitly in terms of the three subsystems:

$$\mathbf{C}^{\mathbf{T}}_{\mathbf{v_T}} = \left(\mathbf{E}^{\mathbf{v_T}}_{\mathbf{T}} - \begin{bmatrix}\mathbf{E}^{\mathbf{v_T}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_T}}_{\mathbf{M}}\end{bmatrix}\begin{bmatrix}\mathbf{E}^{\mathbf{v_P}}_{\mathbf{P}} & \mathbf{0} \\ \mathbf{E}^{\mathbf{v_M}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_M}}_{\mathbf{M}}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{E}^{\mathbf{v_P}}_{\mathbf{T}} \\ \mathbf{0}\end{bmatrix}\right)^{-1}. \qquad \text{(Eq. 3.37)}$$

Because of the relation in Eq.3.10:

$$\begin{bmatrix}\mathbf{E}^{\mathbf{v_P}}_{\mathbf{P}} & \mathbf{0} \\ \mathbf{E}^{\mathbf{v_M}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_M}}_{\mathbf{M}}\end{bmatrix}^{-1} = \begin{bmatrix}\mathbf{c}^{\mathbf{v_P}}_{\mathbf{P}} & \mathbf{0} \\ -\mathbf{c}^{\mathbf{M}}_{\mathbf{v_M}}\mathbf{E}^{\mathbf{v_M}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\mathbf{v_P}} & \mathbf{c}^{\mathbf{v_M}}_{\mathbf{M}}\end{bmatrix}. \qquad \text{(Eq. 3.38)}$$

Substituting this result in Eq. 3.37, one obtains:

$$\mathbf{C}^{\mathbf{T}}_{\mathbf{v_T}} = \left(\mathbf{E}^{\mathbf{v_T}}_{\mathbf{T}} - \begin{bmatrix}\mathbf{E}^{\mathbf{v_T}}_{\mathbf{P}} & \mathbf{E}^{\mathbf{v_T}}_{\mathbf{M}}\end{bmatrix}\begin{bmatrix}\mathbf{c}^{\mathbf{v_P}}_{\mathbf{P}} & \mathbf{0} \\ -\mathbf{c}^{\mathbf{M}}_{\mathbf{v_M}}\mathbf{E}^{\mathbf{v_M}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\mathbf{v_P}} & \mathbf{c}^{\mathbf{v_M}}_{\mathbf{M}}\end{bmatrix}\begin{bmatrix}\mathbf{E}^{\mathbf{v_P}}_{\mathbf{T}} \\ \mathbf{0}\end{bmatrix}\right)^{-1}. \qquad \text{(Eq. 3.39)}$$

Multiplying the submatrices gives:

$$\mathbf{C}^{\mathbf{T}}_{\mathbf{v_T}} = \left(\mathbf{E}^{\mathbf{v_T}}_{\mathbf{T}} - \mathbf{E}^{\mathbf{v_T}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\mathbf{v_P}}\mathbf{E}^{\mathbf{v_P}}_{\mathbf{T}} + \mathbf{E}^{\mathbf{v_T}}_{\mathbf{M}}\mathbf{c}^{\mathbf{M}}_{\mathbf{v_M}}\mathbf{E}^{\mathbf{v_M}}_{\mathbf{P}}\mathbf{c}^{\mathbf{P}}_{\mathbf{v_P}}\mathbf{E}^{\mathbf{v_P}}_{\mathbf{T}}\right)^{-1}. \qquad \text{(Eq. 3.40)}$$

The Regulatory Strengths are calculated as before.

$$\begin{aligned}
R^{T1}_{T1} &= \left(\varepsilon^{v_{ps_1}}_{T1}\, c^{P1}_{v_{ps_1}}\, \varepsilon^{v_{ts_1}}_{P1} + \varepsilon^{v_{ps_1}}_{T1}\, c^{P1}_{v_{ps_1}}\, \varepsilon^{v_{R_1}}_{P1}\, c^{M}_{v_{R_1}}\, \varepsilon^{v_{ts_1}}_{M} - \varepsilon^{v_{td_1}}_{T1}\right) C^{T1}_{v_{T1}} \\
R^{T1}_{T2} &= \varepsilon^{v_{ps_2}}_{T2}\, c^{P2}_{v_{ps_2}}\, \varepsilon^{v_{R_1}}_{P2}\, c^{M}_{v_{R_1}}\, \varepsilon^{v_{ts_1}}_{M}\, C^{T1}_{v_{T1}} \\
R^{T2}_{T1} &= \varepsilon^{v_{ps_1}}_{T1}\, c^{P1}_{v_{ps_1}}\, \varepsilon^{v_{R_1}}_{P1}\, c^{M}_{v_{R_1}}\, \varepsilon^{v_{ts_2}}_{M}\, C^{T2}_{v_{T2}} \\
R^{T2}_{T2} &= \left(\varepsilon^{v_{ps_2}}_{T2}\, c^{P2}_{v_{ps_2}}\, \varepsilon^{v_{ts_2}}_{P2} + \varepsilon^{v_{ps_2}}_{T2}\, c^{P2}_{v_{ps_2}}\, \varepsilon^{v_{R_1}}_{P2}\, c^{M}_{v_{R_1}}\, \varepsilon^{v_{ts_2}}_{M} - \varepsilon^{v_{td_2}}_{T2}\right) C^{T2}_{v_{T2}}
\end{aligned} \qquad \text{(Eq. 3.39)}$$

Here, there are three effects of the genes' activities on themselves. Again, one of these runs through degradation. A second effect of an mRNA on itself involves its own transcription and runs through the proteome. A third effect runs through the protein and metabolite.

Section 3.3.3: Example 3. Hierarchical system with feedback from metabolism to gene expression and protein synthesis

Now the system depicted in Fig. 3.6 is considered.

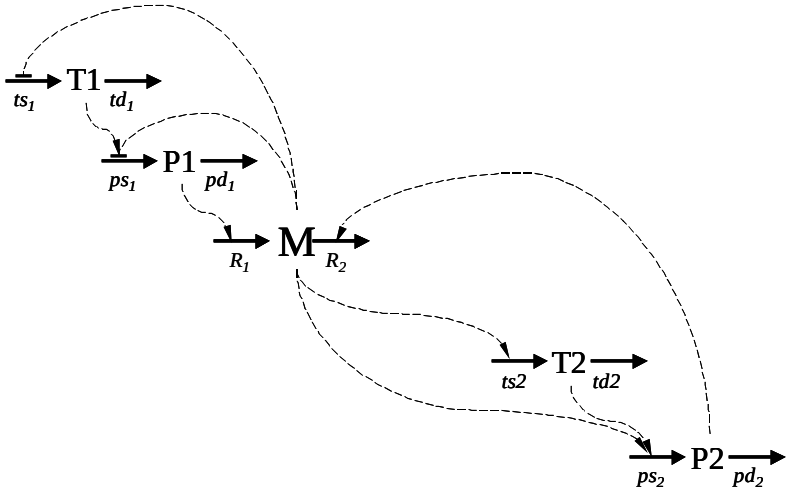

Figure 3.6 – *A system consisting of two mRNAs, two proteins and a short metabolic pathway of two steps and one metabolic intermediate. There is feedback from the metabolite to the transcription and protein synthesis rates. Abbreviations are as in the legend of Fig. 3.41.*

The corresponding elasticity matrix is:

$$\mathbf{E}=\begin{bmatrix} \mathbf{E}_{\mathbf{T}}^{\nu_{\mathbf{T}}} & \mathbf{0} & \mathbf{E}_{\mathbf{M}}^{\nu_{\mathbf{T}}} \\ \mathbf{E}_{\mathbf{T}}^{\nu_{\mathbf{P}}} & \mathbf{E}_{\mathbf{P}}^{\nu_{\mathbf{P}}} & \mathbf{E}_{\mathbf{M}}^{\nu_{\mathbf{P}}} \\ \mathbf{0} & \mathbf{E}_{\mathbf{P}}^{\nu_{\mathbf{M}}} & \mathbf{E}_{\mathbf{M}}^{\nu_{\mathbf{M}}} \end{bmatrix}= \left[\begin{array}{cccc|cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_M^{\nu_{ts_1}} \\ 1 & -\varepsilon_{T1}^{\nu_{td_1}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_M^{\nu_{ts_2}} \\ 0 & 0 & 1 & -\varepsilon_{T2}^{\nu_{td_2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -\varepsilon_{T1}^{\nu_{ps_1}} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -\varepsilon_M^{\nu_{ps_1}} \\ 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P1}^{\nu_{pd_1}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\varepsilon_{T2}^{\nu_{ps_2}} & 0 & 0 & 1 & 0 & 0 & -\varepsilon_M^{\nu_{ps_2}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -\varepsilon_{P2}^{\nu_{pd_2}} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P1}^{\nu_{R_1}} & 0 & 0 & 1 & -\varepsilon_M^{\nu_{R_1}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\varepsilon_{P2}^{\nu_{R_2}} & 1 & -\varepsilon_M^{\nu_{R_2}} \end{array}\right]$$

(Eq. 3.42)

Again, as in Eq. 3.12 $\mathbf{C}^{\mathbf{T}}_{\nu_T}$ is written as

$$\mathbf{C}^{\mathbf{T}}_{\nu_T} = \left(\mathbf{E}^{\nu_T}_{\mathbf{T}} - \mathbf{E}^{\nu_T}_{\mathbf{r}} \left(\mathbf{E}^{\nu_r}_{\mathbf{r}}\right)^{-1} \mathbf{E}^{\nu_r}_{\mathbf{T}}\right)^{-1} \qquad \text{(Eq. 3.43)}$$

or explicitly in the three subsystems:

$$\mathbf{C}^{\mathbf{T}}_{\nu_T} = \left(\mathbf{E}^{\nu_T}_{\mathbf{T}} - \begin{bmatrix} \mathbf{0} & \mathbf{E}^{\nu_T}_{\mathbf{M}} \end{bmatrix} \begin{bmatrix} \mathbf{E}^{\nu_P}_{\mathbf{P}} & \mathbf{E}^{\nu_P}_{\mathbf{M}} \\ \mathbf{E}^{\nu_M}_{\mathbf{P}} & \mathbf{E}^{\nu_M}_{\mathbf{M}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{E}^{\nu_P}_{\mathbf{T}} \\ \mathbf{0} \end{bmatrix} \right)^{-1} \qquad \text{(Eq. 3.44)}$$

Because of the relation in Eq. 3.8:

$$\begin{bmatrix} \mathbf{E}^{\nu_P}_{\mathbf{P}} & \mathbf{E}^{\nu_P}_{\mathbf{M}} \\ \mathbf{E}^{\nu_M}_{\mathbf{P}} & \mathbf{E}^{\nu_M}_{\mathbf{M}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{€}^{\mathbf{P}}_{\nu_P} & -\mathbf{c}^{\mathbf{P}}_{\nu_P} \mathbf{E}^{\nu_P}_{\mathbf{M}} \mathbf{€}^{\mathbf{M}}_{\nu_M} \\ -\mathbf{c}^{\mathbf{M}}_{\nu_M} \mathbf{E}^{\nu_M}_{\mathbf{P}} \mathbf{€}^{\mathbf{P}}_{\nu_P} & \mathbf{€}^{\mathbf{M}}_{\nu_M} \end{bmatrix} \qquad \text{(Eq. 3.45)}$$

By using relation 3.13 again it is possible to write:

$$\mathbf{€}^{\mathbf{P}}_{\nu_P} = \left(\mathbf{E}^{\nu_P}_{\mathbf{P}} - \mathbf{E}^{\nu_P}_{\mathbf{M}} \mathbf{c}^{\mathbf{M}}_{\nu_M} \mathbf{E}^{\nu_M}_{\mathbf{P}}\right)^{-1}$$

and

$$\mathbf{€}^{\mathbf{M}}_{\nu_M} = \left(\mathbf{E}^{\nu_M}_{\mathbf{M}} - \mathbf{E}^{\nu_M}_{\mathbf{P}} \mathbf{c}^{\mathbf{P}}_{\nu_P} \mathbf{E}^{\nu_P}_{\mathbf{M}}\right)^{-1}, \qquad \text{(Eq. 3.46)}$$

which is the control these steps should have if only the transcript levels were fixed. Substituting this result in Eq. 3.45, one obtains:

$$\mathbf{C}^{\mathbf{T}}_{\nu_T} = \left(\mathbf{E}^{\nu_T}_{\mathbf{T}} - \begin{bmatrix} \mathbf{0} & \mathbf{E}^{\nu_T}_{\mathbf{M}} \end{bmatrix} \begin{bmatrix} \mathbf{€}^{\mathbf{P}}_{\nu_P} & -\mathbf{c}^{\mathbf{P}}_{\nu_P} \mathbf{E}^{\nu_P}_{\mathbf{M}} \mathbf{€}^{\mathbf{M}}_{\nu_M} \\ -\mathbf{c}^{\mathbf{M}}_{\nu_M} \mathbf{E}^{\nu_M}_{\mathbf{P}} \mathbf{€}^{\mathbf{P}}_{\nu_P} & \mathbf{€}^{\mathbf{M}}_{\nu_M} \end{bmatrix} \begin{bmatrix} \mathbf{E}^{\nu_P}_{\mathbf{T}} \\ \mathbf{0} \end{bmatrix} \right)^{-1}. \qquad \text{(Eq. 3.47)}$$

Multiplying the sub matrices leads to:

$$\mathbf{C}^{\mathbf{T}}_{\nu_T} = \left(\mathbf{E}^{\nu_T}_{\mathbf{T}} + \mathbf{E}^{\nu_T}_{\mathbf{M}} \mathbf{c}^{\mathbf{M}}_{\nu_M} \mathbf{E}^{\nu_M}_{\mathbf{P}} \mathbf{€}^{\mathbf{P}}_{\nu_P} \mathbf{E}^{\nu_P}_{\mathbf{T}}\right)^{-1} \qquad \text{(Eq. 3.48)}$$

The Regulatory Strengths are calculated as before.

$$
\begin{aligned}
R_{T1}^{T1} &= \left(\varepsilon_{T1}^{v_{ps_1}} \in_{v_{ps_1}}^{P1} \varepsilon_{P1}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_1}} + \varepsilon_{T1}^{v_{ps_1}} \in_{v_{ps_1}}^{P2} \varepsilon_{P2}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_1}} - \varepsilon_{T1}^{v_{td_1}}\right) C_{v_{T1}}^{T1} \\
R_{T2}^{T1} &= \left(\varepsilon_{T2}^{v_{ps_2}} \in_{v_{ps_2}}^{P1} \varepsilon_{P1}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_1}} + \varepsilon_{T2}^{v_{ps_2}} \in_{v_{ps_2}}^{P2} \varepsilon_{P2}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_1}}\right) C_{v_{T1}}^{T1} \\
R_{T1}^{T2} &= \left(\varepsilon_{T1}^{v_{ps_1}} \in_{v_{ps_1}}^{P1} \varepsilon_{P1}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_2}} + \varepsilon_{T1}^{v_{ps_1}} \in_{v_{ps_1}}^{P2} \varepsilon_{P2}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_2}}\right) C_{v_{T2}}^{T2} \\
R_{T2}^{T2} &= \left(\varepsilon_{T2}^{v_{ps_2}} \in_{v_{ps_2}}^{P1} \varepsilon_{P1}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_2}} + \varepsilon_{T2}^{v_{ps_2}} \in_{v_{ps_2}}^{P2} \varepsilon_{P2}^{v_{R_1}} c_{v_{R_1}}^{M} \varepsilon_{M}^{v_{ts_2}} - \varepsilon_{T2}^{v_{td_2}}\right) C_{v_{ts_2}}^{T2}
\end{aligned}
\qquad \text{(Eq. 3.49)}
$$

Because of the feedback loops from the metabolite to the proteins, the expressions become more complicated. Now, control coefficients, $\in_{v_{psi}}^{Pj}$, appear that quantify the control of the reactions at the protein level on the concentrations in the protein level at a fixed transcripts but with a variable metabolite. The synthesis rate of each protein exercises control on the concentration of the other. Consequently the paths split in two and both paths affect the metabolite concentration, which in turn alters the transcription rates.

Section 3.4: Discussion

Gene networks will describe only the dynamics of gene activities. For a total understanding of living cells and even for the complete understanding of gene function in such cells, there is a need to study biochemical networks as a whole. The interest in gene networks was originally inspired by the "dictatorial" perception of the cell; it was assumed that the genes control everything in the cell, since they code for the proteins that actually establish processes in cells. Recent findings, however, show that a biochemical function (glycolytic flux) was regulated just for 30% by gene expression and for 70% by processes in metabolism itself [166]. It becomes obvious that in many actual cases regulation may be distributed over all levels of the cell, and that thereby the cellular hierarchy is a "democratic" one. Nevertheless, in the light of widespread and almost routinely whole-genome mRNA measurements, gene networks are an excellent first approach to describe cellular organization. Although only gene activity is explicitly taken into account in such an analysis, all other biochemical processes contribute to what is observed at the gene expression level. They are present implicitly in the gene network model.

Gene networks are thereby not only collections of gene-gene regulatory relations in a genome (or a subset thereof), but also a reflection of additional metabolic and other regulation. Gene networks are useful to rationalize phenomena in terms of how external perturbations propagate through the expression of genes. Starting from a high-level description of gene regulation in cells provided by the gene network, one could systematically add details of the mechanism of physical interaction and expand the network to include proteins and metabolites explicitly.

I have proposed here a quantitative description of gene networks in terms of quantities of Metabolic Control Analysis, specifically the Regulatory Strength [108]. Regulatory Strengths quantify the effect of variables onto each other, and are therefore perfectly suited to describe complex networks of interacting genes. In addition, a general equation was derived to express these Regulatory Strengths in terms of the properties of the

whole system, *i.e.* 'apparent elasticities' quantifying the sensitivity of net gene expression rates towards mRNA concentrations.

In the gene network representation, proteins and metabolites are 'hidden' variables. If certain genes are omitted in the representation (because they are unknown or for reasons of convenience) they will have the same roles as proteins and metabolites in the above treatment: they will give rise to phenomenological interactions between genes that are mediated by the unknown genes. The gene network representation can thus be made as simple as one would like, including an arbitrary number of genes. The more genes are included the higher the resolution and thus the realism.

The gene network model based on Regulatory Strengths is a purely additive one, based on a first-order Taylor approximation [7]. However, in many cases the regulation of the expression of a gene depends on a combination of several other genes. For example, the products of gene A and gene B may bind to each other before being capable of activating gene C. This framework is not able to represent such higher order scenarios explicitly, but rather an additive equivalent of it.

Although many mathematical frameworks have been proposed to describe gene networks, no such framework provided a solid base for experimental determination of the relevant quantities. The theory of Regulatory Strengths to quantify gene-gene interactions may provide a comprehensive framework accompanied by an experimental recipe of how to measure and calculate these measures from data that can be produced with current high scale experimental technologies in a straightforward way (e.g. Eq. 3.26). The next chapter will detail the procedure of the experimental determination of the Regulatory Strengths and the method will be studied using simulated data.

4

Chapter 4: Inferring gene networks with Regulatory Strength Analysis

With modern microarray technology, one can simultaneously measure expression levels of thousands of genes. This gene expression information has long been expected to provide all the necessary means to unravel the interactions between genes. Several methods have already been developed for this purpose, but in no case was experimental design explicitly taken into consideration. In contrast, I here describe a method based on a well-defined experimental setup and designed with the sole purpose of quantifying how much the expression of one gene affects other genes. This method makes use of results from Metabolic Control Analysis, i.e. Co-Response analysis and the concept of Regulatory Strength, and is able to use relative expression levels as measured with microarrays directly. Examples of application of the method are presented using *in silico* experiments.

Section 4.1: Existing methods for gene network inference

Research on gene networks has been geared towards two major goals: (1) to understand the dynamics and design principles of gene regulation, and (2) to reverse engineer gene networks from experimental measurements. Activities in gene-network modeling started with the pioneering work of Kauffman [4] on random Boolean [181] networks (but see also [195, 196]). More recently it has been questioned if the topology of gene networks is random. Arguments can be invoked in support of the alternate view that gene networks follow a "small-world" [197] and "scale-free" [198] topology, with a power law distribution for node connectivity. It has been shown that metabolic and protein interaction networks have such topologies. However, as pointed out in the previous chapter, gene networks are phenomenological models in which the actions of proteins and metabolites are 'projected' onto the gene space [39, 83]. Although the 'hard-wired' structure of a biochemical network may be small world and scale free, there is no reason to believe that their phenomenological descriptions, *i.e.* gene networks, should obey the same characteristics. Research in the dynamics and structure of gene networks is still active and much more can be expected to happen there. However, in the past five years or so, the majority of research in gene networks focused on methodologies for reconstructing gene networks from experimental observations, perhaps owing to the abundance of microarray data. The remainder of this section is devoted exclusively to this problem.

Many interactions between genes have been discovered through traditional molecular biology approaches. Gene networks can be obtained by combining knowledge about these interactions. The GeNet database [199] (http://www.csa.ru/Inst/gorb_dep/inbios/genet/genet.htm) is a convenient electronic repository for such information. Ideker et al. [200] constructed a gene network of 348 genes of *S. cerevisiae* based on information of 2709 protein-protein interactions [201] combined with 317 known protein-DNA interactions collected from the databases TRANSFAC [202] (http://transfac.gbf.de/TRANSFAC) and SCPD [203] (http://cgsigma.cshl.org/jian/). In a similar way, a network of 10 genes was proposed for the control of flower morphogenesis in Arabidopsis thaliana [142]. In a recent approach, the gene network determining sea urchin development was proposed [204]. The approach had been to knock out single genes and measure the response of the whole network. The authors considered that the perturbation data in itself was not sufficient to distinguish between direct and indirect effects. However, using prior knowledge about cis-regulatory elements in this genome [205], a network specified by direct effects was proposed. Genes were taken to be directly affected if they responded in an experiment in which a certain transcription factor was perturbed and they also contained the specific target sites of that transcription factor in their cis-regulatory elements. Similar approaches correlate gene expression data with the DNA sequence at the promoter sites of genes and also discover new transcription regulatory elements [206, 207].

The process of establishing cause-effect relationships between genes on the basis of observations of the whole system by measuring expression levels is referred to as "reverse engineering" and is a traditional inverse problem akin to many others in science. A number of approaches (see [39, 77] for reviews) have been proposed for inferring gene networks from experimental data. A popular method, sometimes called "guilt by association", assumes that genes with similar expression patterns are functionally related to each other, and is usually put to practice through clustering algorithms [208] and principal

component analysis [209]. There are other methods that are based on more sophisticated statistical analysis, like Bayesian belief networks [210, 211]. Following the classic work of Kauffman [4] on Boolean networks, there are several methods that include the simplification of considering genes to be either expressed at a fixed rate, or not at all [212-214]. These methods also consider time to be a discrete process, and the methods are based on constructing rules that govern if genes at one time step are on or off based on the values taken by the genes that are connected to it at the previous time step. Boolean approaches suffer from their inability to capture intermediate levels of gene expression, and can easily generate spurious results owing to their discrete nature [215]. Both Boolean and Bayesian approaches suffer from the need for data discretization. There are many strategies for data discretization, but it is unknown which is the best. In addition, discretization of continuous data implies the loss of information, suggesting that ultimately this cannot be the best approach.

The use of continuous functions is more challenging but potentially more accurate. Expression levels are allowed to take any positive value. These approaches are mathematically implemented by difference or differential equations, either linear or non-linear. In linear additive models the expression level [216, 217] or rate of change [218] of each gene transcript depends linearly on the expression level of other genes. Each interaction is characterized by one parameter that is positive for activation, negative for inhibition, or zero for no interaction. More realistic, but also more difficult, approaches use nonlinear kinetics to represent the rates of transcription. Wahde and Hertz used neural network-like sigmoidal functions [79], while Mendes [86] used empirical rate laws similar to those of enzyme kinetics. In both cases nonlinear optimization methods were used to fit the model equations to the observed data. In general these methods require much larger amounts of data than linear or Boolean methods, but would have the advantage of being predictive over a much larger domain.

Wagner proposed a graph theoretical approach to analyze gene expression data obtained with null mutants [80]. This method is promising because it uses the most abundant type of data currently available. Unfortunately the method is unable to distinguish between different gene networks of the same class. Wagner proposes to adopt the most parsimonious network [80]. Evidence from molecular biology is that the underlying gene networks will not necessarily be parsimonious. In addition, this approach is only applicable to acyclic graphs, a feature the approach has in common with Bayesian belief networks [210, 211]. By contrast, gene networks tend to have circular dependencies that originate from feedback loops (gene A affects gene B but gene B also affects gene A, directly or through a path of other genes) (Becskei and Serrano highlight the importance of feedback loops in this context [219, 220]).

Evaluations of the performance of some of the above methods on data produced by simulations [221, 222] have shown that they perform badly when asked to infer the gene network structure. Even though the fit to the data is acceptable, the networks obtained were not similar to the network that had actually produced the data [222]. This could well be a consequence of under-determination of the fit and suggests that more data may be needed to find the correct network.

A conclusion that arises from the descriptions in this section is that many more experiments are needed in order to infer gene networks with high accuracy. Furthermore, methods that are based on specific experimental designs [80-83, 179] are expected to perform better than those that disregard how the data was obtained (*e.g.* most applications of clustering).

Section 4.2: The Regulatory Strength Analysis

As I have shown in the Introduction to this dissertation, Metabolic Control Analysis provides an approach to calculate specific global systems properties given specific properties of the individual catalytic elements. Thus, individual elements can be studied in isolation and the knowledge of the elements obtained can then be used to understand the system as a whole. Also the opposite is possible; by observing systems properties one can gain knowledge of the elements of the system. Within the framework of Metabolic Control Analysis several of these 'inverse' strategies have been proposed. In these approaches one requires some knowledge of the interaction structure (stoichiometry of reactions) of the system for the elasticities to be calculated from measured control coefficients or co-response coefficients.

Co-response Analysis [104, 105] is such a framework in which flux and concentration co-response coefficients are measured and elasticities can be calculated. The advantage of this method is that in order to measure co-response coefficients one doesn't need to know the exact magnitude of the perturbation made to the system, as is necessary for the calculation of control coefficients.

Giersch [223, 224] has shown that for a particular stoichiometry and feedback structure not all enzymes need to be perturbed in order to infer the values of all elasticities. In his extension to the double modulation method of Kacser and Burns [225], he systematically derived the minimum number of perturbations needed, and which enzymes need to be perturbed, in order to calculate all elasticities of the system. This method does require measurements of flux control coefficients and concentration control coefficients, but fewer perturbations need to be made than with the co-response analysis method.

Yet another approach is Metabolic Control Design, proposed by Acerenza [115], and extended with the help of Ortega [226]. These authors make an additional step, showing that after elasticities are inferred one can propose the mechanistic equations complying with the observed elasticity pattern. Again, this method requires measurements of flux control coefficients and concentration control coefficients.

Westerhoff *et al.* [114] have shown that it is possible to infer the structure of a metabolic pathway by measuring flux control and concentration control. The assumption here is that the number of branches in the pathway is known, but not where the branches are positioned. Using inverse Metabolic Control Analysis one can then infer the pathway structure.

Here previous work is extended to deal with a currently widespread and popular experimental technique, the microarray determination of mRNA abundances. An extension of 'inverse Metabolic Control Analysis' is developed to infer the interaction structure of gene networks. In the previous chapter I derived a theoretical framework to quantify the regulatory structure of gene networks based on measurements of co responses of mRNA concentrations to small perturbations.

The formalism was stated as

$$\mathbf{R} = \mathbf{O}^{-1} \qquad \text{(Eq. 4.1)}$$

Thus, by determining the matrix of Co-Control Coefficients experimentally, one can calculate the matrix of Regulatory Strengths. The latter shows which genes interact with which other genes and to what extent.

Measuring **O** requires a set of experiments in each of which the rate of expression of a single gene is perturbed to a small extent. The experiments start with the collection of mRNA from a reference steady state. Then, the small perturbation is applied to a single gene and, once the system has settled to a new steady state, the gene expression levels are compared to against the reference levels. These measurements of gene expression are best carried out using microarrays to assess as many genes as possible, but quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR) would also be appropriate. Further perturbations are applied in a systematic way to all other genes, and their effects measured. When it is not possible to perturb all the genes, one should at least include all those that are already suspected of taking part in the phenomenon of interest, and perhaps others that have been associated with these (*e.g.* through clustering of data from other gene-expression experiments).

The perturbations could be made, for instance, through the use of anti-sense RNA, dsRNA/RNAi [227], by adding extra gene copies, or by engineering promoter sequences [228]. Another strategy would be to compare the expression profiles of wild type strains with heterozygous knockout strains, in which one chromosome contains an intact copy of the gene while the other had the gene deleted. Having 50% of the wild type gene dosage should lead to a 50% reduction in the rate of gene expression, if the gene expression rate depends linearly on the gene dosage. Obviously, this is only applicable to polyploid organisms. This type of experimentation differs from published microarray experiments where very drastic changes to transcription rates are made such as completely knocking out genes, or perturbing a group of many genes simultaneously [13-19].

Microarray experiments usually result in ratios of mRNA concentrations in a perturbed state, $[mRNA]$, to their concentrations in a reference state, $[mRNA]^0$, or more precisely, a ratio of fluorescence intensities, FR, that is equivalent to this ratio of concentrations ($FR = [mRNA]/[mRNA]^0$). Such relative measures, as opposed to absolute concentrations, are often seen as an inconvenience. But the proposed method takes advantage of this, because the relative change of the concentration $[mRNA]/[mRNA]^0$, needed to calculate co-control coefficients, can be directly expressed by fluorescence ratios. Indeed, microarray experiments quantify gene expression levels essentially as a ratio of the abundance of mRNA in response to a stimulus to its abundance in a reference state (as determined from the ratio of fluorescence intensities of two fluorophores):

$$FR_i = \frac{F'}{F} = \frac{[mRNA_i]'}{[mRNA_i]} \quad , \qquad \text{(Eq. 4.2)}$$

where F' and F are respectively the fluorescence intensities of the stimulated and reference state, $[mRNA_i]$ is the reference concentration of the message of gene i ($i = 1,\ldots,n$; n being the total number of genes analyzed) and $[mRNA_i]'$ is the concentration of the same message in the new steady state reached after the stimulus has been applied.

All coefficients in MCA (including the co-control coefficients) are by definition the result of infinitesimal calculations, but in real experiments one can only make finite changes. It is thus useful to reformulate the approximation explicitly using finite changes. Using a central finite differences approximation, $\Delta C/C$, to the scaled derivative C/C gives:

$$\frac{\Delta C}{C} = \frac{\left(C - C^0\right)}{\left(C + C^0\right)/2}, \qquad \text{(Eq. 4.3)}$$

where C^0 is the reference concentration and C is the concentration after perturbation. In microarray experiments, however, one does not determine absolute values but rather a ratio of fluorescence intensities that is equivalent to the ratio of concentrations ($FR=C/C^0$, C^0 being the reference concentration). The use of central finite differences is important, as it is free from the bias that left or right finite differences would introduce. This is especially important when the perturbations are large (as will be illustrated later on). Eq. 4.3 can then be expressed in terms of the fluorescence ratio FR by dividing denominator and numerator by the same factor C^0:

$$\frac{\Delta C}{C} = \frac{2\left(\frac{C}{C^0} - 1\right)}{\frac{C}{C^0} + 1} = \frac{2(FR - 1)}{FR + 1}. \qquad \text{(Eq. 4.4)}$$

Using this result to replace infinitesimal changes by finite changes in the definition of the co-control coefficient (Eq. 1.4 in Chapter 1), an approximation of the co-control coefficients in terms of fluorescence ratios can be written as:

$${}^{v_m}O_j^i = \frac{\Delta mRNA_i / mRNA_i}{\Delta mRNA_j / mRNA_j} = \frac{(FR_i - 1)(FR_j + 1)}{(FR_j - 1)(FR_i + 1)}, \qquad \text{(Eq. 4.5)}$$

${}^{v_m}O_j^i$ is thus a measurable quantity obtainable directly from microarray data.

From the data obtained in each perturbation experiment one column of the co-control matrix **O** can be calculated. When all experiments are completed one obtains of a full matrix

$$\mathbf{O} = \begin{bmatrix} {}^{v_1}O_1^1 & \dots & {}^{v_n}O_n^1 \\ \dots & \dots & \dots \\ {}^{v_1}O_1^n & \dots & {}^{v_n}O_n^n \end{bmatrix} \qquad \text{(Eq. 4.6)}$$

which can be inverted to get the Regulatory Strength matrix **R** which is a model for the gene network.

$$\mathbf{R} = \begin{bmatrix} {}^{v_1}R_1^1 & \dots & {}^{v_1}R_n^1 \\ \dots & \dots & \dots \\ {}^{v_n}R_1^n & \dots & {}^{v_n}R_n^n \end{bmatrix} \qquad \text{(Eq. 4.7)}$$

Summary of the procedure:

1) Allow a system of n genes to reach a reference steady state and use it in all iterations of step 3 as the reference state.
2) Perturb the rate of transcription of a single gene and allow the system to reach a new steady state.
3) Measure the mRNA concentrations of this new state relative to the mRNA concentrations from step 1 using microarrays.
4) Repeat steps 2—3, until all transcription rates have been perturbed.
5) Use the fluorescence ratios, *FR*, determined in the experiments above, and calculate the co-control coefficients (Eq. 4.5), filling a column of the co-control matrix **O** (Eq. 4.6).
6) Invert the co-control matrix to obtain the Regulatory Strength matrix **R** (Eq. 4.7), which quantitatively represents the gene network.

Section 4.3: Application of Regulatory Strength Analysis to simulated data

To illustrate the proposed method, I will apply it to data produced by artificial gene regulatory networks (computer models). These models were defined and run with the biochemical kinetics simulator Gepasi [85, 91, 92] (available at http://www.gepasi.org). Nonlinear kinetics was intentionally used for transcription to show that the inherent non-linearity of the system does not invalidate this linear method. To apply perturbations the rate of transcription of each transcription step was modified and a resulting new steady state calculated. From this data the co-control matrix was calculated via the fluorescence ratios. One advantage in using a computer model is the ability to judge how well the method performed. The models used represent several scenarios, including one in which there are hidden variables and another one where there are non-additive effects (complex formation between different gene products).

Section 4.3.1: Simple example

A small regulatory network of three genes is considered and simulated on the computer the experiments described above. The following system of ordinary differential equations describes the model:

$$\frac{d[A]}{dt} = \frac{V_a}{1+\frac{[B]}{K_{iB}}} - k_a[A]$$
$$\frac{d[B]}{dt} = \frac{V_b}{1+\frac{K_{aC}}{[C]}} - k_b[B] \quad . \qquad \text{(Eq. 4.8)}$$
$$\frac{d[C]}{dt} = \frac{V_c}{\left(1+\frac{[B]}{K_{iB}'}\right)\left(1+\frac{K_{aA}}{[A]}\right)} - k_c[C]$$

Here [*A*], [*B*] and [*C*] are concentrations of the mRNA species; V_a, V_b and V_c are basal rates of transcription; K_{iB} and $K_{iB'}$ are inhibition constants; K_{aC} and K_{aA} are activation constants; k_a, k_b, and k_c are first-order degradation constants. First *[A]*, *[B]* and *[C]* were calculated for a reference steady state setting all parameter values arbitrarily to unity, except for K_{iB}, $K_{iB'}$, K_{aA}, and K_{aC} which were set to 0.1.

Perturbations of transcription rates were then made by changing the value of the basal rates by 10%, followed by calculation of the new steady state concentrations. The co-control matrix was determined solely from those data:

$$\begin{pmatrix} {}^{v_a}O_A^A & {}^{v_b}O_B^A & {}^{v_c}O_C^A \\ {}^{v_a}O_A^B & {}^{v_b}O_B^B & {}^{v_c}O_C^B \\ {}^{v_a}O_A^C & {}^{v_b}O_B^C & {}^{v_c}O_C^C \end{pmatrix} = \begin{pmatrix} -1 & 0.834 & 0.407 \\ -0.128 & -1 & -0.486 \\ -0.261 & 1.16 & -1 \end{pmatrix} . \qquad \text{(Eq. 4.9)}$$

The regulatory strength matrix was then obtained by inverting this matrix:

$$\begin{pmatrix} {}^{v_a}R_A^A & {}^{v_a}R_B^A & {}^{v_a}R_C^A \\ {}^{v_b}R_A^B & {}^{v_b}R_B^B & {}^{v_b}R_C^B \\ {}^{v_c}R_A^C & {}^{v_c}R_B^C & {}^{v_c}R_C^C \end{pmatrix} = \begin{pmatrix} -0.903 & -0.757 & 0 \\ 0 & -0.639 & 0.311 \\ 0.236 & -0.544 & -0.639 \end{pmatrix} . \qquad \text{(Eq. 4.10)}$$

The matrix in Eq. 4.10 represents the gene network corresponding to equation 4.8. A graph can be drawn by making the correspondence of elements of this matrix to edges in the

graph, which can be labeled with the numeric values quantifying the strengths of interactions (Figure 4.1). The regulatory strength quantifies the change in a variable caused by another through a specific path of interaction. For example, ${}^{v_C}R_A^C$ indicates that gene C's activity would change by 0.236 times the changes in gene A due to this particular interaction. The total gene C response to a change in gene A is determined by all regulatory paths starting in gene A and ending in gene C.

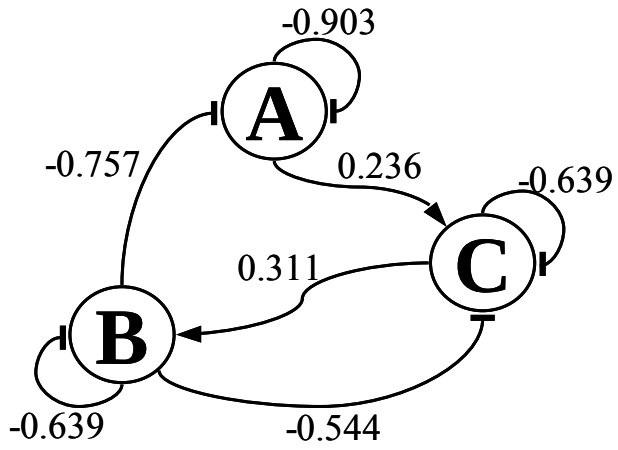

Figure 4.1 – *A model gene regulatory network of three genes. Arrows represent activation interactions while lines with a blunt end inhibitory interactions. The numbers next to the lines are the values determined for the corresponding regulatory strengths*.

Keeping in mind that small changes in transcription rates are difficult to achieve in practice and their effects are even harder to measure, the performance of the method with larger perturbations was explored (under-expression by 50% and over-expression by 200%). The values of the Regulatory Strengths obtained with these larger perturbations are compared to the theoretical values (calculated through the definition of the Regulatory Strength (Eq. 1.5 Chapter 1) using elasticity and control coefficients obtained with Gepasi) in Table 4.1. Table 4.1 shows that the error due to the finite differences approximation in our method is relatively small for a small perturbation (1.1x) but grows with larger perturbations (0.5x and 2x). Nevertheless, even with the larger perturbations the absolute error is less than 0.075 (17%), which may be well below the measurement noise, and is thus acceptable.

Table 4.1 – Effect of the size of perturbations used to estimation Regulatory Strengths. *Theoretical values of Regulatory Strengths for the model of Eq. 4.9 were calculated using Eq. 4.2 and the simulation's values of elasticity and control coefficients. "Experimental" values were calculated by applying different perturbations on rate of transcription (last three columns) and following the method described in the text.*

	Theoretical value	**1.1x perturbation**	**0.5x perturbation**	**2x perturbation**
$^{v_a}R_A^A$	-0.901	-0.903	-0.879	-0.916
$^{v_a}R_B^A$	-0.752	-0.757	-0.711	-0.787
$^{v_a}R_C^A$	0	0.001	-0.005	0.004
$^{v_b}R_A^B$	0	0.001	-0.006	0.005
$^{v_b}R_B^B$	-0.638	-0.639	-0.635	-0.646
$^{v_b}R_C^B$	0.315	0.311	0.348	0.288
$^{v_c}R_A^C$	0.241	0.236	0.279	0.213
$^{v_c}R_B^C$	-0.533	-0.544	-0.444	-0.607
$^{v_c}R_C^C$	-0.638	-0.639	-0.628	-0.651

Section 4.3.2: Hidden variables

In cases where one cannot measure the expression of all the genes, when it is not easy to perturb the transcription rate of some genes, or when it is not convenient to analyze the full network, the method must be applied only to a subset of the genes. In those cases only the rates of transcription of "available" genes and their mRNA concentrations are measured. Since a larger network is responsible for the observations one can no longer be sure if the interactions detected by the method are really direct or are a result of the hidden variables. To explore such a scenario the five-gene network shown in Fig. 4.2A was constructed (the kinetics were similar to the previous example and are available as supplementary information at http://www.vbi.vt.edu/~mendes/icsb01-supp.html).

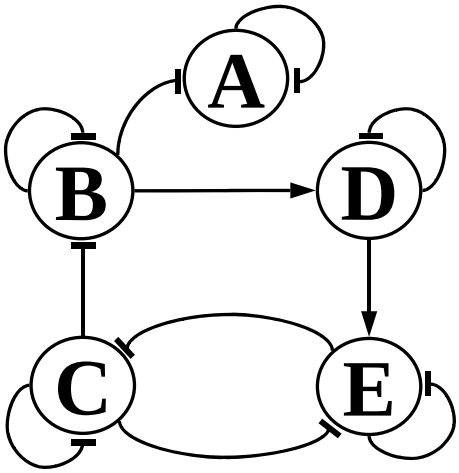

Figure 4.2A – *A model gene regulatory network of five genes. The full mathematical model is supplied as supplementary material at http://www.vbi.vt.edu/~mendes/icsb01-supp.html.*

Then the transcription rates of genes C, D and E were perturbed by 10% over-expression and their corresponding relative mRNA responses were "measured". All the direct-effect Regulatory Strengths for those three genes were calculated and used to reconstruct the network shown in Fig. 4.2B.

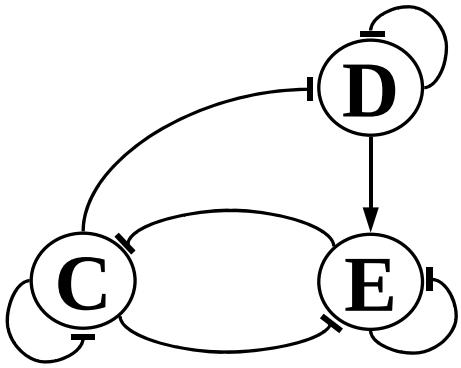

Figure 4.2B – T*he network reverse engineered with the method proposed when only genes C, D and E were perturbed and observed.*

Comparing the networks of Fig. 4.2 (the original and the one reconstructed by our method), it is clear that all the direct interactions on the original network between genes C, D and E were recovered from the Regulatory Strengths. In addition, there is a new arrow from C to D in Fig. 2B that does not exist in the original network (Fig. 4.2A).

In the original system, C influenced B and B influenced D, but because B was not included in the analysis these two interactions collapse into a single one which is, of course, only apparent. What this reveals is that if only a subset of genes is considered in the analysis, then the interactions identified with this method are not necessarily direct but can

also include indirect effects (through hidden variables). This is the same situation as the one that arises because in most cases it is not the mRNAs that interact with transcription, but rather their protein products and metabolic intermediates. Because proteins and metabolites are not represented explicitly here, they can be thought of as being hidden variables and their action is included in the arrows of the gene regulatory network. Chapter 3 explicitly dealt with hidden proteins and metabolites and showed how the Regulatory Strengths are a function of elasticities and inter-modular control coefficients of the underlying paths of genetic communication.

Section 4.3.3: Non-additive effects

The regulation of the expression of a gene can depend on a combination of several other genes. For example, the products of gene A and gene B may have to bind to each other in order to be capable of activating gene C. Because these interactions may be frequent, it is relevant to see how the method, which is linear and additive, performs with such networks. To that purpose the gene network depicted in Fig. 4.3A is analyzed.

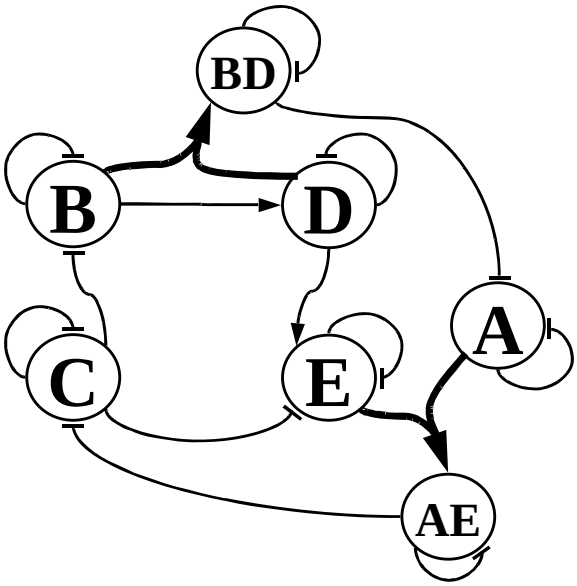

Figure 4.3A – *A gene network where some of the regulation occurs through binary complexes. The full mathematical model is supplied as supplementary material at http://www.vbi.vt.edu/~mendes/icsb01-supp.html.*

The main feature of this network is the formation of two complexes, AE and BD that are regulators of other genes. Perturbations were applied on all the transcription rates by increasing them by 10% and responses of the expression levels of A, B, C, D and E observed (the levels of AE and BD were not monitored, these variables were hidden). The calculated Regulatory Strengths describe the gene regulatory network presented in Fig. 4.3B. The network shows interactions from B to A, from D to A, from A on C and from E on C – these correspond to the interactions that complexes AE and BD have on target genes in the original model, but are recovered by our approach as an additive effect of individual

genes. This situation is similar to the previous case of hidden variables: as the complexes were not considered explicitly. Their effect though was recovered as separate additive effects of the constituents of the complexes. This results in extra arrows leading directly from the complex components to the target genes. There are also additional interactions of A on E and E on A, and B on D and D on B which are all negative. This could be interpreted in the following way: in the original network, increasing A, will tend to increase AE and therefore decrease E. In summary, although the method does not account directly for complex formation, the effect of such interactions does not get lost in our treatment, but is reflected in the resulting networks as coming from the individual constituents of the complexes.

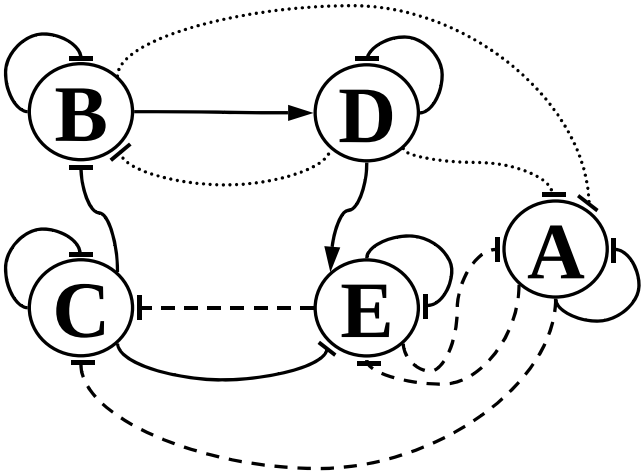

Figure 4.3B – *The gene network reverse engineered with the proposed method. The non-additive regulatory interactions that occur through the complexes in the original system are lost but an equivalent set of additive interactions appears in their place*.

Section 4.3.4: Arabidopsis flowering gene network

To apply the method to a more biologically relevant gene network structure, simulations were performed using the gene network proposed by Mendoza *et al*. [142] to control flower morphogenesis in *Arabidopsis thaliana* (Fig. 4.4). It is irrelevant for our purposes whether this network is indeed correct or what the molecular details behind it might be. For this illustration, one should assume that the model network is the real system. The gene network was modeled with equations that follows principles of biochemical kinetics and is formulated in terms of ordinary differential equations describing the rate of change of the mRNA concentrations. Actual parameter values are not important here, provided that they are consistent with the relationships shown in Fig. 4.4.

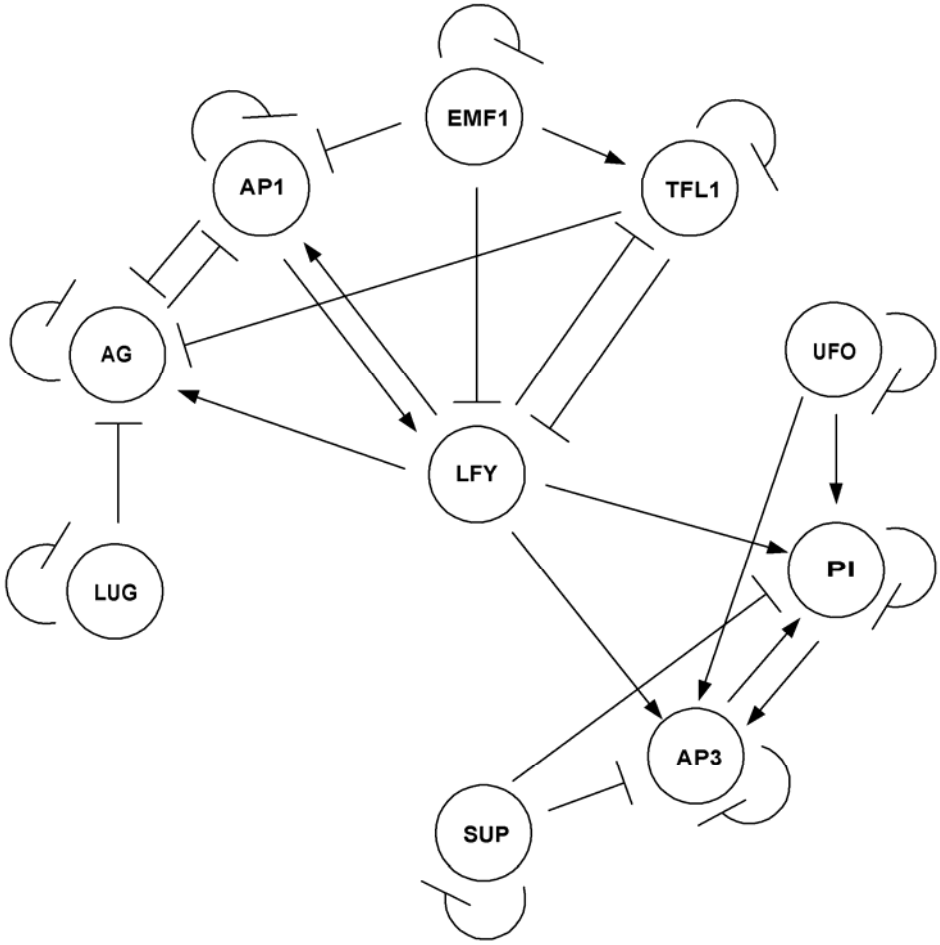

Figure 4.4 – *Graphical representation of the gene network controlling flower morphogenesis in Arabidopsis thaliana as proposed by Mendoza et al [142]. Circles represent genes; lines represent direct effects of one gene onto another, with arrows standing for activation, and blunt ends for inhibition. Meanings of the gene abbreviations and further details can be found in the original reference.*

Additional information about this model can be found at http://www.vbi.vt.edu/~mendes/tig02.html. The *in silico* experiments performed followed the method described in the previous sections by applying 10% perturbations on the rates of transcription of each gene. Equation 4.12 depicts the matrix of direct Regulatory Strengths that is the result of this exercise. By assuming that absolute values of Regulatory Strengths below 0.01 were below the noise level, and were probably zero, the diagram in Fig. 4.4 is reproduced exactly from the matrix of Eq. 4.12. This demonstrates that the method can recover a gene network from observations of relative mRNA concentrations.

The finite approximation used in the method could be a source of error. To assess how well this approximation (e.g. Equation 4.12) compared with reality, Equation 4.13

depicts the solution calculated through the definition of Regulatory Strength as in Eq. 1.5 in Chapter 1. The estimate obtained by the *in silico* experiment is indeed rather good when compared with the more precise solution of Eq. 4.13, indicating that 10% perturbations give almost perfect results, suggesting that it may be possible to apply larger perturbations.

$$\mathrm{R(inferred)} = \begin{array}{c} \begin{array}{cccccccccc} \textit{LUG} & \textit{AG} & \textit{AP}1 & \textit{EMF}1 & \textit{TFL}1 & \textit{LFY} & \textit{SUP} & \textit{AP}3 & \textit{PI} & \textit{UFO} \end{array} \\ \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.579 & -1.14 & -0.184 & -0.002 & -0.112 & 0.114 & 0 & 0 & 0 & 0 \\ -0.009 & -0.894 & -1.14 & -0.109 & -0.002 & 0.124 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.094 & -1.09 & -0.973 & 0 & 0 & 0 & 0 \\ -0.002 & -0.005 & 0.053 & -0.103 & -0.107 & -1.09 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.001 & -0.001 & 0.135 & -0.119 & -1.04 & 0.192 & 0.109 \\ 0 & 0 & 0 & -0.001 & -0.001 & 0.135 & -0.119 & 0.192 & -1.04 & 0.109 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix} \end{array} \begin{array}{c} \\ \textit{LUG} \\ \textit{AG} \\ \textit{AP}1 \\ \textit{EMF}1 \\ \textit{TFL}1 \\ \textit{LFY} \\ \textit{SUP} \\ \textit{AP}3 \\ \textit{PI} \\ \textit{UFO} \end{array}$$

(Eq. 4.11)

$$\mathrm{R(theoretical)} = \begin{array}{c} \begin{array}{cccccccccc} \textit{LUG} & \textit{AG} & \textit{AP}1 & \textit{EMF}1 & \textit{TFL}1 & \textit{LFY} & \textit{SUP} & \textit{AP}3 & \textit{PI} & \textit{UFO} \end{array} \\ \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.563 & -1.13 & -0.174 & 0 & -0.105 & 0.128 & 0 & 0 & 0 & 0 \\ 0 & -0.875 & -1.13 & -0.103 & 0 & 0.129 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.099 & -1.09 & -0.963 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.056 & -0.099 & -0.102 & -1.09 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.141 & -0.113 & -1.04 & 0.203 & 0.113 \\ 0 & 0 & 0 & 0 & 0 & 0.141 & -0.113 & 0.203 & -1.04 & 0.113 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix} \end{array} \begin{array}{c} \\ \textit{LUG} \\ \textit{AG} \\ \textit{AP}1 \\ \textit{EMF}1 \\ \textit{TFL}1 \\ \textit{LFY} \\ \textit{SUP} \\ \textit{AP}3 \\ \textit{PI} \\ \textit{UFO} \end{array}$$

(Eq. 4.12)

Section 4.4: Application of Regulatory Strength Analysis to simulated data of a biochemical network

It is well known that genes do not interact with each other through their mRNAs. Rather, the mRNAs give rise to proteins, and the proteins being a transcription factor, polymerase or mRNase, interfere with the expression of genes. The proteins might need to form complexes, or need to be phosphorylated, in order to fulfill their regulating function. These phenomena on the proteome level complicate the interpretation of the results of our strategy to infer gene networks, because only the gene level is observed. Even more complicating is the fact that metabolites can have a regulatory function on gene expression. The metabolome is complicated and can give rise to less obvious interactions between genes.

The interplay between the proteome and the metabolome gives rise to and determine the strengths of many interactions between genes. Therefore it is useful to analyze the hierarchical system depicted in Fig. 4.5, in with the genes influence each other, through proteins (like the transcription factor, P_4) or through metabolism (M_1, M_2, M_3, M_4). A gene network representation of the biochemical network in Figure 4.5(A) is depicted in Figure 4.5(B).

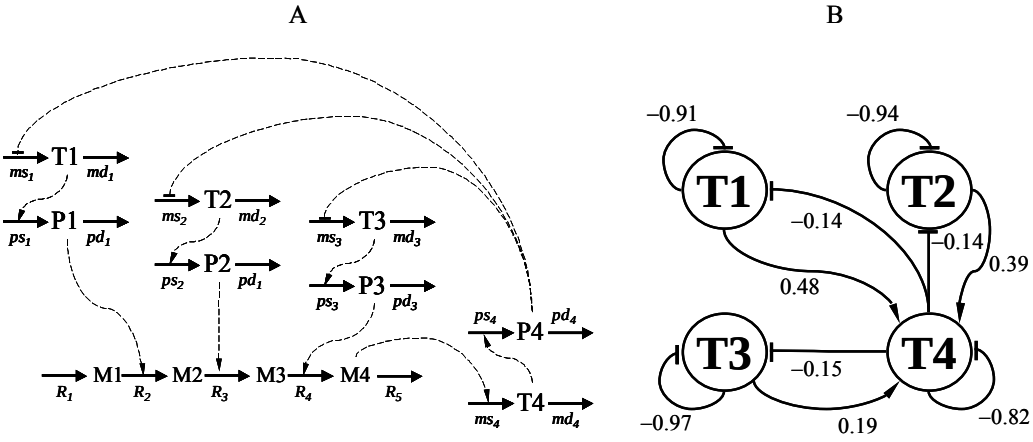

Figure 4.5 – ***(A)*** *A biochemical network consisting of 4 genes (T1, T2, T3 and T4), three enzymes (P1, P2 and P3), a transcription repressor (P4), and a short metabolic pathway of 5 reactions, containing 4 internal metabolites (M1, M2, M3 and M4). Abbreviations used are: ms = messenger synthesis, md = messenger decay, ps = protein synthesis, pd = protein decay and R stands for metabolic reaction. Solid arrows depict flow of mass and dashed arrows indicate interactions.* ***(B)*** *The genetic network representation of the biochemical network in (A). Lines show direct regulatory effects, arrows denote activation, and blunt ends inhibition.*

The details about the rate equations and model parameters are given in Table 4.2. Again, perturbations were simulated by increasing the transcription rates by 10%. Like in microarray experiments, only the changes in transcript levels were considered.

Table 4.2 – *The kinetic functions and parameter values used to simulate the network depicted in Figure 4.5(A).*

Level	Reaction	Parameter	Value
Transcription	$v_{\text{transcription } T_i} = \frac{V_i}{\left(1+\frac{[P4]}{K_{I_i}}\right)\left(1+\frac{K_{A_i}}{[M4]}\right)}$	V	0.01
		K_I	1, if interaction ∞, if no interaction
		K_A	1, if interaction 0, if no interaction
	$v_{\text{degradation } T_i} = k_i[T_i]$	k	0.1
Translation	$v_{\text{translation } P_i} = \frac{V_i[T_i]}{K_{0.5_i}+[T_i]}$	V	0.01
		$K_{0.5}$	1
	$v_{\text{degradation } P_i} = k_i[P_i]$	k	0.001
Metabolism	$v_{\text{metabolic}_i} = E_i \frac{k_{f_i}\frac{[M_{i-1}]}{K_{S_i}} - \frac{k_{r_i}[M_i]}{K_{P_i}}}{1+\frac{[M_{i-1}]}{K_{S_i}}+\frac{k_{r_i}[M_i]}{K_{P_i}}}$	k_{catf}	10
		k_{catr}	1
		K_{mS}	1
		K_{mP}	1
	External concentrations	M_0	1
		M_5	10^{-5}
		E_1	1
		E_5	1
		Variables	
		T	mRNA
		P	Enzyme
		M	Metabolite

The resulting co-control matrix is:

$$\mathbf{O} = \begin{pmatrix} -1 & 0.064 & 0.031 & 0.158 \\ 0.01 & -1 & 0.031 & 0.158 \\ 0.01 & 0.064 & -1 & 0.158 \\ -0.635 & -0.412 & -0.198 & -1 \end{pmatrix}. \qquad \text{(Eq. 4.13)}$$

Inversion gives the Regulatory Strength matrix:

$$\mathbf{R} = \begin{pmatrix} -0.909 & 0.001 & 0.001 & -0.143 \\ 0.001 & -0.939 & 0.001 & -0.148 \\ 0.001 & 0.001 & -0.97 & -0.153 \\ 0.477 & 0.386 & 0.191 & -0.818 \end{pmatrix}. \qquad \text{(Eq. 4.14)}$$

Only Regulatory Strengths with values exceeding 0.01 are considered significant, because smaller values are assumed erroneous due to the application of 10% perturbations rather than the infinitesimal perturbations that the theory requires. With this assumption the Regulatory Strength matrix exactly represents the network structure depicted in Figure 4.5(B). In the diagram in Figure 4.5(B) it is seen that, for example, gene transcript T1 has a positive effect on transcript T4, meaning that an increase in the concentration of T1 tends to increase the concentration of T4. Looking at the structure of the biochemical network in Figure 4.5(A), it is seen that the effect of T1 on T4 runs through the proteome and the metabolome; an increase in mRNA T1 leads to an increase in enzyme P1 which increases in metabolite M4, which in turn activates the transcription of T4 and this leads to an increase in mRNA T4. All reactions beyond the transcriptome are thus collapsed into a single arrow in the gene network representation. In a similar fashion, all other connections in the gene network could be explained by tracing paths in the hierarchical biochemical network or by using formal framework described in Chapter 3.

Section 4.5: Large scale simulation study

Thus far the effectiveness of the method was shown on data simulated with small mathematical models of several types of networks. In this section the performance of the algorithm on a large set of simulated data of artificial gene networks [229] is evaluated. 50 networks of 100 genes and 200 connections for each of three different architectures were tested. Since it is unknown which topology actual gene networks have, the data was generated by gene network models with three previously proposed topologies for biochemical networks: 'random' [4, 230], 'scale free' [198] and 'small world' [197]. Details of the networks can be found on the web at: http://staff.vbi.vt.edu/mendes/AGN/Century/index.html. The interactions in these models are defined with non-linear kinetics. In this study perturbations were applied by 10%, 25% and 50% reductions in gene expression. The performance is evaluated using two measures: 1) the False Discovery Rate (defined as the wrongly predicted edges as a percentage of the total number of predicted edges) and 2) Power (defined as the number of correct predicted edges as a percentage of the total number of edges in the network).

Section 4.5.1: Qualitative evaluation

Results are summarized in Figure 4.6 and Table 4.3. For comparison, for each network the theoretical Regulatory Strengths were calculated by means of Eq. 1.5. I considered Regulatory Strengths smaller than 0.05 to be zero (too small to be detected) in the theoretically obtained networks. For the inferred networks I used several cut off values for

presence/absence of an interaction: 0.05, 0.1 and 0.25. In Figure 4.6 it can be seen that at higher cut off values fewer wrong connections were proposed, but at a price of discovering less of the right connections. Still, at a cut-off of 0.25, about 80% of all connections were correctly inferred and almost no incorrect connections were found, even in the case for the large perturbation size of 0.5 (50% reduction in gene expression).

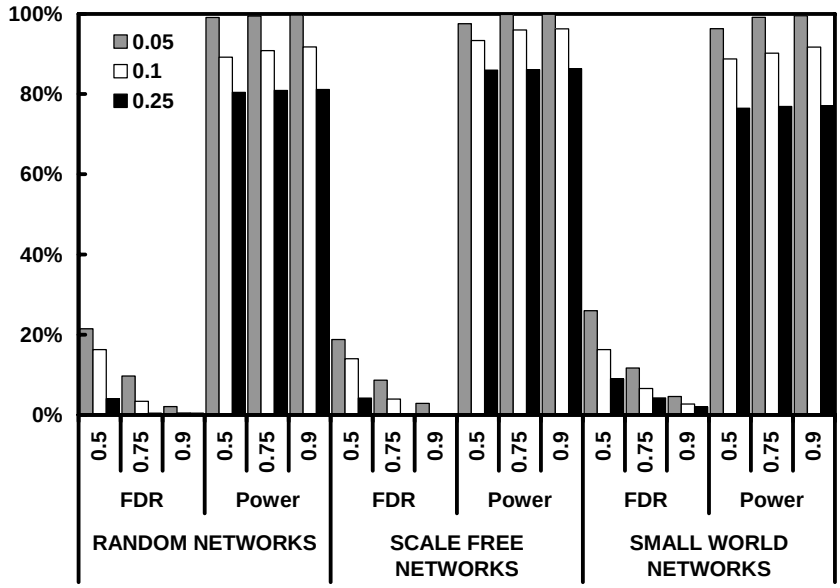

Figure 4.6 – *The results of the large scale simulated data study. Average values are given for the FDR and power for different perturbation sizes, network topology (x-axis) and different cut off values (legend).*

Table 4.3 – *The results of the large scale simulated data study. Average values and standard deviations are given for the FDR and power for different perturbation sizes, network topology and different cut off values.*

	False Discovery Rate			**Predictive Power**		
Perturbation size	**0.5**	**0.75**	**0.9**	**0.5**	**0.75**	**0.9**
Cut off value	**RANDOM NETWORKS**					
0.05	21.47%± 4.2%	9.68% ± 3.23%	2.08% ± 2.08%	99.11% ± 4.2%	99.50% ± 0.76%	99.71% ± 0.6%
0.1	16.24% ± 3.96469	3.40% ± 2.29776	0.48% ± 1.41467	89.20% ± 4.29198	90.87% ± 3.61277	91.76% ± 3.28708
0.25	4.01% ± 3.23317	0.45% ± 1.31386	0.42% ± 1.33426	80.41% ± 3.85336	80.94% ± 3.64326	81.16% ± 3.53236
	SCALE FREE NETWORKS					
0.05	18.78% ± 13.3995	8.66% ± 3.94669	2.88% ± 2.31134	97.56% ± 14.2399	99.87% ± 0.412182	99.92% ± 0.317724
0.1	13.98% ± 4.2325	3.93% ± 2.97896	0.26% ± 0.878423	93.37% ± 3.08506	95.99% ± 2.79924	96.28% ± 2.78399
0.25	4.18% ± 3.15305	0.33% ± 1.06124	0.00% ± 0	85.96% ± 4.23295	86.08% ± 4.30678	86.33% ± 4.20243
	SMALL WORLD NETWORKS					
0.05	25.95% ± 7.70312	11.68% ± 4.00742	4.57% ± 3.29901	96.32% ± 1.5879	99.17% ± 0.711446	99.57% ± 0.484208
0.1	16.27% ± 5.54405	6.55% ± 3.49321	2.71% ± 2.28987	88.75% ± 6.51032	90.21% ± 5.77863	91.72% ± 4.84278
0.25	8.99% ± 5.56196	4.22% ± 2.81074	1.97% ± 2.19155	76.46% ± 4.13616	76.89% ± 4.45847	77.09% ± 4.48288

In general, the qualitative predictions made were quite accurate. At the higher cut off levels predictive power was lost, but the edges identified in the networks were estimated with certainty (as expressed by a low FDR), even at the large perturbation size of 50% transcription reduction.

Section 4.5.2: Quantitative evaluation

Above, the inferred Regulatory Strengths matrices were compared to the theoretically obtained ones (Eq. 4.2) by scoring for absence or presence of connections. Here, the inferred matrices are quantitatively compared with the expected matrices. As an overall distance measure between matrices the Root Mean Squared Error (RMS) (Eq. 4.15) is used.

$$RMS = \left|\mathbf{R^E} - \mathbf{R^P}\right| = \sqrt{\frac{1}{n^2}\sum_{i,j}^{n}\left(R_{ij}^{E} - R_{ij}^{P}\right)^2} \quad . \qquad \text{(Eq. 4.15)}$$

Superscript *P* stands for 'predicted', indicating an element of the inferred Regulatory Strength matrix, and superscript *E* stands for 'expected', indicating an element for the expected (theoretically obtained) matrix. n^2 is the number of elements in the matrix. The RMS can be seen as the average deviation of a predicted Regulatory Strength from the expected Regulatory Strengths. Table 4.4 lists the average RMS and standard deviations for the different network topologies.

Table 4.4 – *Quantitative results of the large scale simulated data study. Average values and standard deviations are given for the RMS and for different perturbation sizes (Pert.) and network topologies*

	RMS		
Pert.	**RANDOM NETWORKS**	**SCALE FREE NETWORKS**	**SMALL WORLD NETWORKS**
0.5	0.155 ± 0.097	0.127 ± 0.041	9.044 ± 42.66
0.75	0.078 ± 0.103	0.052 ± 0.021	0.363 ± 0.271
0.9	0.073 ± 0.27	0.019 ± 0.008	0.272 ± 0.483

Given that the value of Regulatory Strengths usually ranges from 0.05 to 1.5 the RMS was in general very low, except for the small world topology. Especially at a perturbation size of 0.5 the RMS for the small words topology was extremely high. This was due to the fact that a small number of network predictions were completely wrong: two predictions had an RMS of around 200, while the other 48 predictions had an RMS comparable to the RMS of the random and scale free networks.

Another way of quantitatively evaluating the inferred matrices is by counting the number of errors of a particular size. Eq. 4.16 shows how this was done.

$$\mathrm{E} = \sum_{ij}^{n} e_{ij}$$
$$e_{ij} = \begin{cases} 1, \text{if } \left|R_{ij}^{E} - R_{ij}^{P}\right| > \delta \\ 0, \text{otherwise} \end{cases} \qquad \text{(Eq. 4.16)}$$

Several values for δ were selected and the errors made counted. Table 4.5 lists the numbers of errors made at these values for the different network topologies.

Table 4.5 – *Quantitative results of the large scale simulated data study. Average values and standard deviations are given for the number of errors of certain size made and for different perturbation sizes and network topologies*

Perturbation size	**0.5**	**0.75**	**0.9**
Error size (δ)	**RANDOM NETWORKS**		
>0.01	115.9 ± 33.13	127.02 ± 20.95	49.26 ± 16.05
>0.05	64.82 ± 12.85	15.34 ± 7.9	3.4 ± 2.94
>0.1	34.34 ± 9.86	6 ± 4.06	0.56 ± 1.54
>0.5	0.86 ± 2.4	0.46 ± 1.34	0.56 ± 1.78
	SCALE FREE NETWORKS		
>0.01	60.48 ± 35.11	81.1 ± 22.96	33.56 ± 16.19
>0.05	43.24 ± 12.24	10.74 ± 6.95	4.66 ± 4.05
>0.1	28.2 ± 12.6	6.82 ± 5.75	0.44 ± 1.49
>0.5	0.5 ± 1.52	0.02 ± 0.14	0
	SMALL WORLD NETWORKS		
>0.01	142.14 ± 29.21	104.44 ± 25.97	53.04 ± 12.76
>0.05	49.66 ± 18.32	22.56 ± 6.84	10.86 ± 8.57
>0.1	45.68 ± 16.24	19.58 ± 12.32	8.08 ± 6.87
>0.5	15.74 ± 11.94	8.1 ± 4.48	4.5 ± 3.42

Most of the errors made were of a size between 0.01 and 0.05. Even at the large perturbation size almost no errors above 0.5 were made in the random and scale free topology. More errors were made for the small world topology: even for small perturbations several large errors were made. In general the predictions were quite accurate.

Section 4.6: Discussion

I presented a theoretical framework and a design for microarray experiments that will enable investigators to infer genetic networks. Because there are as yet no published experiments that conform to such a design, the method was illustrated using *in silico* experiments.

Presently, the signal-to-noise ratio of microarrays is too low to measure the effect of small responses in gene expression. However, the technology is constantly improving, and it is only a matter of time before this approach becomes more feasible (or perhaps a superior technology will appear in the meantime, such as a high-throughput implementation of the quantitative RT-PCR techniques [20] or the new RT-MLPA method [231]). This contribution may increase the incentives for further technological improvements. Our method requires genome-scale experimental effort, in the form of gene expression rate manipulations, which are currently laborious. This is comparable to the situation with whole-genome sequencing some 15 years ago, when the automated sequencing was not in place to carry out the Human Genome Project. However that methodology was quickly developed thereafter. It may be expected that, considering the increasing trend of laboratory

robotics, and the demonstrated usefulness of this method, high-throughput means of carrying out gene manipulations will become feasible in the not-too-distant future.

The global gene network of an organism would be uncovered if perturbations were applied to all genes in a genome. However, if only a subset of the genes is manipulated, the method is still capable of identifying a gene network, even though indirect interactions running through the genes that were omitted will appear as direct. Networks obtained by examining subsets of a genome are still useful representations of the underlying gene regulatory structure. The fewer genes are included the more phenomenological the inferred networks. However, independent of the number of included genes, the network is a valid quantitative description of genetic regulation. This gives the advantage that experiments can be done on just few genes first and extended with more experiments gene by gene to increase the resolution of the network, but each intermediate network is a valid quantitative description of genetic regulation. Although for the final goal (a global gene network) all experiments are needed, the intermediate results already provide good insight in cellular regulation.

It was suggested to use of co-control coefficients rather than control coefficients for this analysis. This carries a disadvantage, which is that one cannot separate the part due to the elasticity from the part due to control (see Eq. 4.2). To be able to do so would result in uncovering the mechanisms of regulation, while the present method is limited to a phenomenological view. The main justification for the use of co-control coefficients and Regulatory Strengths, rather than control coefficients and elasticities, is that it is experimentally much harder to measure control coefficients, since one needs to know the absolute magnitude of the perturbation. The determination of co-control coefficients does not require knowledge of the exact size of the perturbation, though it should be small. This is most convenient, as one cannot accurately predict the magnitude of changes in transcription rates when a new gene copy is added (either to a chromosome or in a plasmid). Another approach to infer gene networks using MCA is to consider flux control as well. By measuring all co-control coefficients, one would be able to work out all concentration and elasticity coefficients [104, 105]. But this requires double the number of experiments than the method described here, since there are as many independent fluxes as there are genes in the network. Furthermore, in order to measure the fluxes a time series, albeit short, is needed and this requires even more measurements.

Some representations of gene networks have been proposed [79, 86, 232-234] that use non-linear rate functions to represent the dynamics of mRNA concentrations, but they have the disadvantage of including large numbers of parameters to be fitted and thereby require much more data [79, 86] than the method described in this chapter. To obtain a representation based on Regulatory Strengths first could be very helpful to the application of such non-linear regression models by constraining the space of possible solutions.

This approach requires perturbations to be applied to each gene independently. Another application of MCA to functional genomics is the FANCY method (standing for 'functional analysis of co-responses in yeast') [235]. That method uses the co-response of metabolites to gene knockouts to uncover the functions of genes, and it has recently been demonstrated with great success [178, 236]. Although this approach has the use of co-responses in common with ours, the objectives of each are quite different and the similarity is indeed only superficial.

Our model of gene networks is a purely additive one, which stems from our use of metabolic control analysis, a formalism based on a first-order Taylor approximation [7]. However, in many cases the regulation of the expression of a gene depends on a

combination of several other genes. The method is unable to identify such interactions, but finds separate Regulatory Strength for all of the components of such a complex, effectively taking into account the interaction. With current methods, uncovering the non-additive character of interactions requires a much higher number of experiments [212, 213].

Another consequence of the linear approximation around steady states is that a specific network is valid only for the steady state and other operating points in its vicinity. Unless the kinetics of mRNA concentrations is really linear (which is most likely not the case), the description of the regulatory network with Regulatory Strengths is only valid for that specific physiological state; in another physiological state the network might be quantitatively very different. Furthermore, because the gene network structure depends on elements of the matrix being different from zero, it is possible that the structure may even be different between different states. This is caused by the fact that elasticities can be zero (or indistinguishable from zero) in two situations: *i*) when there is no interaction, and *ii*) when the kinetics of the interaction is saturated with respect to a particular effector (Figure 4.7).

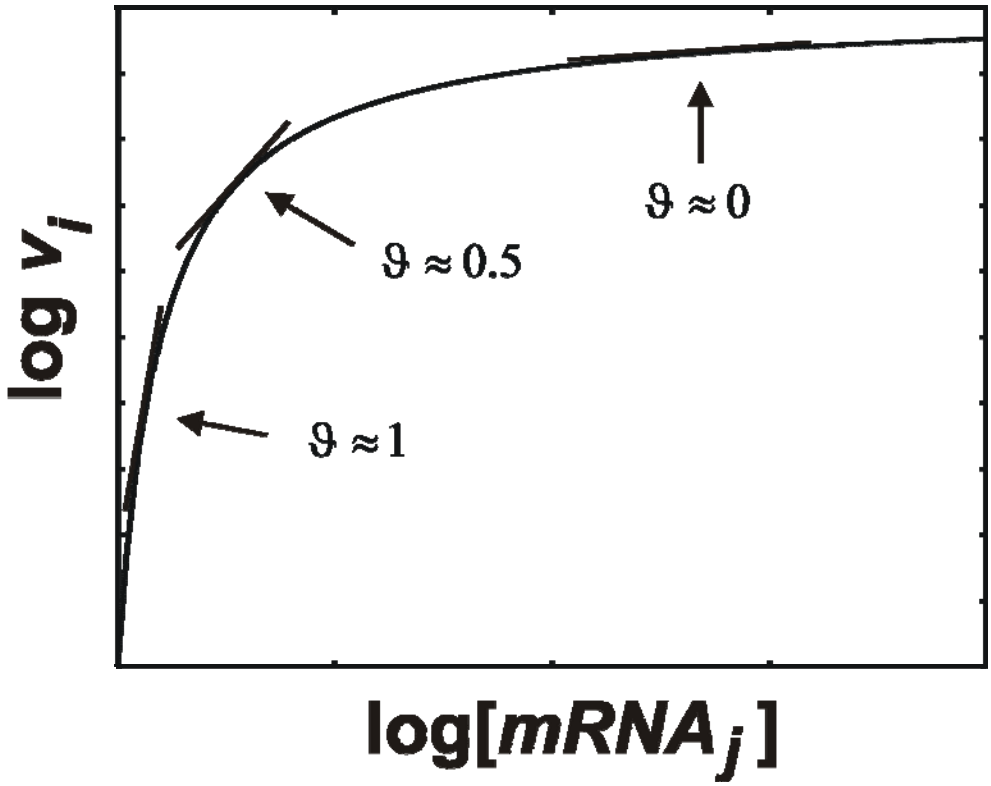

Figure 4.7 – *The rate of transcription v_i as a function of an interacting mRNA concentration, [$mRNA_j$]. The value for the apparent elasticity ϑ depends on the particular steady state of the system. In some certain steady states, the rate of expression is saturated for a particular effector; the corresponding elasticity will be zero (right tangent). In other steady states in which this rate is not saturated on that effector and the apparent elasticity will have a value (left two tangents).*

If, in a certain steady state, one rate of expression is saturated for a particular effector, the corresponding elasticity will be zero and this interaction will not appear in the network (and correctly since in effect it is not operating). There may be other steady states in which this rate is not saturated for that effector and then the interaction will be revealed by the analysis, resulting in a different network. Although less likely it might even be possible that the sign of the interaction is not the same in different steady states. Gene networks determined with Regulatory Strength analysis are, thus, phenomenological representations and may differ from state to state. Although Regulatory Strengths are phenomenological representations of interaction, they can be expressed in terms of the elasticities and control coefficients along the path of interaction through the biochemical network and thus be related to the properties of the global biochemical network [83] (see section 3.2.2 in Chapter 3). The presence or absence of interactions in a gene network, at a certain steady state, is determined by the kinetic properties of each step along the path of interaction that passes through the proteome and metabolome. Some of these interactions will reveal themselves only in certain steady states. In order to draw a more complete picture of the system, the system should be studied in several distinct states.

5

Chapter 5: General discussion

Living systems are complex entities consisting of hundreds of thousands distinct types of molecules organized in networks of interaction. These complex intra-cellular networks generate the defining characteristics of living systems. Understanding complex intra-cellular networks requires a methodology in which solid theory is combined with experimental observation and computation. In this dissertation, I proposed several methods for that purpose. One set of approaches enables experimental determination of the properties of metabolic systems in isolation from the global hierarchical network they are normally part of. Another one provides a theoretical framework for simplification of hierarchical networks into gene networks and provides an experimental protocol to infer gene networks from large-scale experimental observations. I will here put these results into more general perspectives.

Section 5.1: How does this work contribute to systems biology?

Modern experimental technologies enable us to do large-scale parallel measurements of biochemical compounds such as RNAs, proteins and metabolites. The large amount of data produced currently needs to be interpreted with appropriate tools. The interpretation of observations of these high-dimensional dynamical systems is far from obvious and might well benefit from mathematical theory with accompanying analyses. This dissertation extends a theoretical framework designed to study biochemical systems explicitly such that it becomes able to deal with the type of experimental data that results from multiple parallel technologies. The main objective of this thesis is to propose and validate formal ways to simplify the study of intricate biochemical systems, either by showing that it should be possible to study certain parts in isolation or by giving a compact description of the whole. The reason for the proposed simplifications is two-fold: 1) it facilitates the study of such complicated biochemical networks and 2) it deals with limitations in the richness of the available datasets: although there are high throughput technologies for measuring each level of biochemical organization, hardly any datasets exist that describe all levels for a single condition or variation. Mostly, only gene expression is measured and proteomic and metabolomics studies are done in different experimental setups, making it impossible to combine them in a complete hierarchical model. The simplifications are thus also made to 'suit' the state of current experimental technologies for data collection.

The possibility of studying metabolic pathways in isolation from other cellular biochemical processes, such as transcription, translation and signal transduction, rests on the presence of a time scale separation between them: turnover times of most proteins being hundred times slower than the turnover of most metabolites (Chapter 2). When such a time scale separation is present it is possible to study the metabolic behavior independently, which should then characterize the short-term response of the whole system. In that case MCA should be fit to describe the short-term behavior. For the long-term behavior of metabolism inside the living cell MCA can also be used, but fails to profit from advantages in description and conceptual understanding that is offered by HCA in which explicitly the modular structure of biochemical systems is taken into consideration.

Even when no such time-scale separation is present, it was predicted that it should be possible experimentally to measure the isolated metabolic behavior by making specific modifications to the hierarchical system. This could also be possible by measuring integral control coefficients for metabolites as well as proteins and then computing the intra-modular control of metabolism. When there is no time-scale separation the properties of metabolism in isolation will not describe the short-term behavior of metabolism embedded in the whole system. However, it is interesting to evaluate how much of the metabolic regulation is intrinsic to metabolism itself and how much is imposed by the hierarchical levels (gene expression), thereby evaluating their relative importance.

For the purpose of microarray data analysis, I developed a method to identify gene networks that may underlie those data. The identification of biochemical network structure, in particular gene networks, is taking a prominent place in systems biology. The main feature that differentiates the method that I developed (see Chapters 3 & 4) from most others is twofold: 1) it is based on a solid and comprehensive theoretical framework, and 2) it provides a clear experimental design. Additionally, this theoretical framework established for the first time a formal expression of the quantities describing genetic interactions in terms of the underlying biochemical systems (see Chapter 3). The first

experimental application of the perturbation strategy following network identification showed good results [36]. This result may constitute an incentive for large-scale application of the strategy. Gene networks are an intermediate step between the genotype (genome sequence) and the phenotype (dynamical function) of living organisms. Knowledge of the regulatory structure of gene networks will have an impact on biology of similar proportion to the impact of fully sequenced genomes. It opens the door to computational modeling and discovery of properties of such networks. Since gene networks provide a systemic description of the regulation in living cells, they also provide a means for rational experimental modification of organisms for biotechnological or pharmacological purposes. The consequences of 'genetic manipulation' of organisms by introducing non-native genes or removing genes from their genome will be much better understood having such a systemic description of genetic regulation. Knowledge of gene networks thus also might end the seemingly endless discussion about the ethics of genetic modification.

Ironically, throughout the dissertation I claim that these methods are applicable at genome scale data, while the examples studied are extremely simple (with exception of the large scale simulation study in section 4.5, Chapter 4). These simple examples were chosen for reasons of clarity, in order to explain the methods clearly. The matrix equations derived in this dissertation are valid for models of any size and complexity; indeed for genome-scale networks.

Section 5.2: Related methods for gene network inference

Recently, several variants of the Regulatory Strength Analysis (RSA) have appeared in the literature [36, 84]. To show the similarities between these newer methods and the original RSA that was developed in Chapters 3 and 4 and [81-83] it is useful to derive the relevant equations in a more general way than in Chapter 3, where they were developed from co-response analysis.

I start by assuming that gene networks (and hierarchical biochemical networks in general) can be modeled with a set of non-linear differential equations of the form:

$$\frac{d\mathbf{x}}{dt} = \mathbf{N}\mathbf{v}(\mathbf{x}(\mathbf{p}),\mathbf{p}) \qquad \text{(Eq. 5.1)}$$

where $\mathbf{x}$ is a vector with n elements (x_1, x_2,…, x_n) corresponding to mRNA concentrations, $\mathbf{v}$ a vector of r non-linear rate equations (v_1, v_2,…, v_r), depending on k parameters (p_1, p_2,…, p_k). The stoichiometry matrix $\mathbf{N}$, has dimensions n x r, and has as its elements the stoichiometric coefficients, usually integers, the ij^{th} element indicating the molar amount of variable i participates in reaction j (products and substrates of the reaction having positive and negative coefficients, respectively). These non-linear rate equations can be described by non-linear functions ranging from simple mass action kinetics to complicated mechanistic rate equations (for example the ones used in [152, 153]).

In steady state, by definition,

$$\mathbf{N}\mathbf{v}(\mathbf{x}(\mathbf{p}),\mathbf{p}) = 0\,. \qquad \text{(Eq. 5.2)}$$

Differentiating towards the parameter vector yields and assuming constant stoichiometries;

$$\mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{x}}\frac{d\mathbf{x}}{d\mathbf{p}} + \mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{p}} = 0 \qquad \text{(Eq. 5.3)}$$

Giving:

$$\mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{x}}\frac{d\mathbf{x}}{d\mathbf{p}} = -\mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{p}} \qquad \text{(Eq. 5.4)}$$

$\mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{x}}$ is the Jacobian of the system, from now on referred to as $\Im$. The Jacobian has *n* x *n* elements $\partial \frac{dx_i}{dt} \Big/ \partial x_j$, and can be interpreted as a linear description of the kinetics around a steady state that is a solution of Eq. 1. Since the Jacobian has non-zero elements wherever there is a direct interaction between two of the components in the system, it is a useful quantitative model for the structure of networks. $\frac{d\mathbf{x}}{d\mathbf{p}}$ is an *n* x *k* matrix of 'sensitivities' of the variables towards the parameters (the logarithmic equivalents of these are called response coefficients), each column corresponds to a different parameter p_i of a total number of *k* parameters $\mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{p}}$ is an *n* x *k* matrix with the sensitivities of the rates towards the parameters. They are the sensitivities that pertain to the rates in isolation from the system, *i.e.* they correspond to the dependence of the rates on the parameter at constant magnitudes of all other parameters and state variables.

Now *n* parameters p' are selected such that each affects only rates that specifically affect only one of the *n* variables. Then $\mathbf{N}\frac{\partial \mathbf{v}}{\partial \mathbf{p}'}$ is an *n* x *n* diagonal matrix and $\frac{d\mathbf{x}}{d\mathbf{p}'}$ is an *n* x *n* matrix. The parameters selected can affect several rates as long as these rates are specifically producing/consuming only one variable and do not directly affect the production or consumption rates of any other variable. Such unique rates exist for a variable if it has a non-zero element in a column of **N** while all other elements in that column are zero. If there are no such unique rates for a given variable, the variable cannot be included in the analysis, or one has to modify the system experimentally in order to create such rates (see below for an example).

Replacing the derivatives by finite differences Eq. 5.4 is written as $\Im\Delta\mathbf{x}\left(\Delta\mathbf{p}'\right)^{-1} \approx -\mathbf{N}\Delta\mathbf{v}\left(\Delta\mathbf{p}'\right)^{-1}$. Since $\Delta\mathbf{p}'$ appears on both sides it can be taken away resulting in:

$$\Im\mathbf{X} \approx -\mathbf{V} \qquad \text{(Eq. 5.5)}$$

Now, given that *n* perturbations have been made, $\mathbf{X}$ is an *n* x *n* matrix of the changes in the variables and $\mathbf{V}$ is an *n* x *n* diagonal matrix, resulting from $\mathbf{N}\Delta\mathbf{v}$ and contains the rate changes on its diagonal. We can now solve for the Jacobian: $\mathfrak{J} \approx -\mathbf{V}(\mathbf{X})^{-1}$

It is thus not necessary to evaluate derivatives towards the parameters; it suffices to measure just changes in the variables and in the rates, caused by the parameter changes.

Equation 5.5 prescribes a recipe for network inference: by measuring the changes in mRNA concentrations that appear as a response to specific rate perturbations, and then measuring the actual size of the rate perturbations one can infer the Jacobian of the system, *i.e.* infer the network structure.

This equation was previously suggested to infer gene networks [81], by experimentally determining matrix $-\mathbf{X}(\mathbf{V})^{-1}$ (the control matrix), and inverting it to get **J**, but although it gave good results, the need to measure the rate perturbations makes the approach experimentally unattractive. Despite the experimental difficulties, this relationship has been used to infer the SOS DNA-repair system gene network in *E. coli* [36]. These authors had in addition the ingenious idea that once a network **J** is known, it is also possible to infer the mode of action of unknown perturbations, i.e. to infer matrix $\mathbf{V}$ by measuring $\mathbf{X}$ and knowledge of **J** [36].

Again, the need to measure the rate perturbations makes the approach experimentally unattractive, because these are difficult to measure. Therefore the following transformations to Equation 5.5 were considered:

Pre-multiplying both sides of Eq. 5.5 with $-diag\left(\mathbf{J}^{-1}\right)= diag\left(\left(\mathbf{V}\mathbf{X}^{-1}\right)^{-1}\right)$ (the *diag*( ) operator returns a diagonal matrix with elements equal to the diagonal of the matrix given as the argument) gives:

$$-diag\left(\mathbf{J}^{-1}\right)\mathbf{J}\mathbf{X} \approx -diag\left(\left(\mathbf{V}\mathbf{X}^{-1}\right)^{-1}\right)\mathbf{V} \qquad \text{(Eq. 5.6)}$$

Since **V** is already a diagonal matrix $diag\left(\left(\mathbf{V}\mathbf{X}^{-1}\right)^{-1}\right)$ can be written as $diag(\mathbf{X})\mathbf{V}^{-1}$, simplifying Eq. 5.6 to:

$$-diag\left(\mathbf{J}^{-1}\right)\mathbf{J}\mathbf{X} \approx -diag(\mathbf{X})\mathbf{V}^{-1}\mathbf{V} \approx -diag(\mathbf{X}) \qquad \text{(Eq. 5.7)}$$

For the limit of infinitesimal changes I define

$$\mathbf{R} \equiv -diag\left(\mathbf{J}^{-1}\right)\mathbf{J} \equiv -diag(\mathbf{X})\mathbf{V}^{-1}\mathbf{J} \qquad \text{(Eq. 5.8)}$$

$$\mathbf{O} \equiv -\mathbf{X}\left(diag(\mathbf{X})\right)^{-1} \qquad \text{(Eq. 5.9)}$$

A simple way to write Equation 5.7 is thus [81-83, 179]:

$$\mathbf{R}\mathbf{O} = \mathbf{I} \qquad \text{(Eq. 5.10)}$$

(see also Chapter 3), or

$$\mathbf{R} = \mathbf{O}^{-1} \qquad \text{(Eq. 5.11)}$$

R is known as the 'regulatory strength matrix', with elements $dx_i / dv_i \cdot \partial v_i / \partial x_j$ (note that I here use unscaled versions of all measures I have described before; the derivation would be equivalent for the scaled quantities, but scaling was omitted for clarity). This is a global measure of the effects of each gene expression level on every other, partitioned as the effect that one gene expression level (concentration) has on the gene expression rate of the other (quantified by a Jacobian element) and the effect that the gene expression rate has on the gene expression level (quantified by a global control (sensitivity) coefficient dx_i / dv_i). **O** is known as the 'co-control matrix', with elements dx_i / dx_j, which can be calculated from gene expression data produced by applying systematic perturbations, but in this case without the need to measure the rate perturbations themselves!

To derive the framework for the 'interaction map' [84] both sides of Eq. 5.5 are pre-multiplied with $-\left(diag\left(\mathbf{J}\right)\right)^{-1} = \left(diag\left(\mathbf{V}\mathbf{X}^{-1}\right)\right)^{-1}$ giving:

$$-\left(diag\left(\mathbf{J}\right)\right)^{-1}\mathbf{J}\mathbf{X} = -\left(diag\left(\mathbf{V}\mathbf{X}^{-1}\right)\right)^{-1}\mathbf{V} \qquad \text{(Eq. 5.12)}$$

Again by simplify; this time by rewriting $\left(diag\left(\mathbf{V}\mathbf{X}^{-1}\right)\right)^{-1}$ as $\left(diag\left(\mathbf{X}^{-1}\right)\right)^{-1}\mathbf{V}^{-1}$, simplifying Eq. 5.12 to:

$$-\left(diag\left(\mathbf{J}\right)\right)^{-1}\mathbf{J}\mathbf{X} = -\left(diag\left(\mathbf{X}^{-1}\right)\right)^{-1} \qquad \text{(Eq. 5.13)}$$

which is equation 5 in [84] (note that in their terminology **X** is called $\mathbf{R_p}$):

$$\mathbf{r} \equiv -\left(diag\left(\mathbf{J}\right)\right)^{-1}\mathbf{J} = -\left(diag\left(\mathbf{X}^{-1}\right)\right)^{-1}\mathbf{X}^{-1} \qquad \text{(Eq. 5.14)}$$

Note that by Eq. 5.8 and 5.11:

$$\mathbf{R} \equiv -diag\left(\mathbf{J}^{-1}\right)\mathbf{J} = -diag\left(\mathbf{X}\right)\mathbf{X}^{-1} \qquad \text{(Eq. 5.15)}$$

This shows clearly the similarity between the Regulatory Strength Matrix and the Interaction Map.

Writing the equality:

$$-\left(diag\left(\mathbf{X}\right)\right)^{-1}\mathbf{R} = -diag\left(\mathbf{X}^{-1}\right)\mathbf{r} = \mathbf{X}^{-1} \qquad \text{(Eq. 5.16)}$$

leads to an expression for the Interaction Map in terms of the Regulatory Strength Matrix

$$\mathbf{r} = \left(diag\left(\mathbf{X}^{-1}\right)\right)^{-1}\left(diag\left(\mathbf{X}\right)\right)^{-1}\mathbf{R} \qquad \text{(Eq. 5.17)}$$

which can be rewritten as

$$\mathbf{r} = \left(diag\left(\mathbf{R}\right)\right)^{-1}\mathbf{R} \qquad \text{(Eq. 5.18)}$$

since $\left(diag\left(\mathbf{R}\right)\right)^{-1} = \left(-diag\left(\mathbf{X}\right)diag\left(\mathbf{X}^{-1}\right)\right)^{-1} = \left(diag\left(\mathbf{X}^{-1}\right)\right)^{-1}\left(diag\left(\mathbf{X}\right)\right)^{-1}$

This means that the Interaction Map is the Regulatory Strength matrix scaled to its diagonal elements, thereby inferring the strengths of the interactions as a fraction of the global self-effect of the genes.

Recently, it was stated in [237-239] that the RSA only is applicable to systems in which each variable has only one input and output flux. This is an incorrect statement. This misunderstanding probably resulted from the fact that the RSA was originally derived explicitly for gene networks, in which each variable has indeed one synthesis and degradation term. However, in the derivation above it became clear that this is not necessarily true. To emphasize this, I analyzed data with a model in which each of the variables had three input fluxes, of which two are affected by the other variable (Fig 5.1). In this case the Regulatory Strengths are partitioned into the two separate effects [108]. For example, the regulatory strength of T1 on T2 is expressed as

$$R_{T1}^{T2} = \varepsilon_{T1}^{v_3} C_{v_3}^{T2} + \varepsilon_{T1}^{v_4} C_{v_4}^{T2}, \qquad \text{(Eq. 5.19)}$$

and the regulatory strength of T2 on T1 as

$$R_{T2}^{T1} = \varepsilon_{T2}^{v_1} C_{v_1}^{T1} + \varepsilon_{T2}^{v_2} C_{v_2}^{T1}. \qquad \text{(Eq. 5.20)}$$

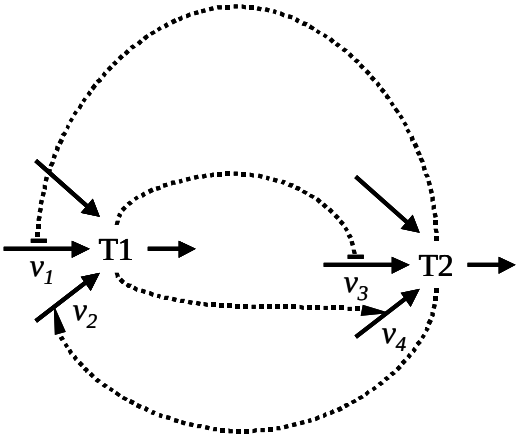

Figure 5.1 – *A simple network between two variables affecting each other through multiple input fluxes. Rate laws used were* $dT1/dt = 0.5 + 1/(1 + 5/T2) + 1/(1 + T2/0.1)$ *and* $dT2/dt = 1 + 1/(1 + 3/T1) + 1/(1 + T1/0.3)$. *Perturbations were made by increasing the rate parameters or multiple rates parameters by 10%.*

The Regulatory Strengths were calculated both by the expressions Eq 5.19 and 5.20, and by applying RSA to perturbation data from the model. The values for the Regulatory Strengths obtained by RSA indeed matched the values obtained through their definition (Eq 5.19 and 5.20) [108], independent of which of the fluxes was perturbed, and even when more than one of the input fluxes were affected by the perturbation. This example emphasizes that RSA can deal with such situations and also that flexible perturbations are possible. As long as the perturbations are specific to each of the variable RSA is able to infer the correct values for the Regulatory Strengths, just like the interaction map approach.

The RSA and the interaction map approach require the exact same type and number of experiments. Perturbations have to be specific to each variable. The authors of the interaction map method emphasize that a smaller number of perturbations is necessary to find regulatory connections between 'modules' consisting of any number of genes and that perturbations need not to be specific to one reaction [84]. When modules are considered as the nodes of the network, then one needs to make perturbations that specifically act on each module individually. This is indeed what is shown in [84]. Expressing the interactions between these modules, or more explicitly, between their communicating intermediates (the variables of the modules that interact kinetically with the rates of other modules) [161], can be equivalently done in terms of Regulatory Strengths, given that the interaction map is a scaled regulatory strength matrix. Note that knowing which variables are the communicating intermediates implies having a great deal of knowledge of the system under study already!! To illustrate that the RSA is able to deal with signal transduction networks I analyzed the same data of [84]. The data is produced using the model of a MAP Kinase signal transduction pathway (details can be found at http://www.pnas.org/cgi/data/192442699/DC1/4. This signal transduction cascade can be conceptually decomposed into modules (sets of reactions that don't share a common flux) (see figure 5.2). The goal in this exercise is to infer the interactions between the modules. Since there are three modules there are three perturbations needed. Perturbations were done as described in [84]. The responses of the communicating variables, *i.e.*

MAPKKKPP, MAPKKPP and MKPP, were taken from Figure 4 in [84] (case b was taken, in which large perturbations affecting several rates were made) and RSA was applied.

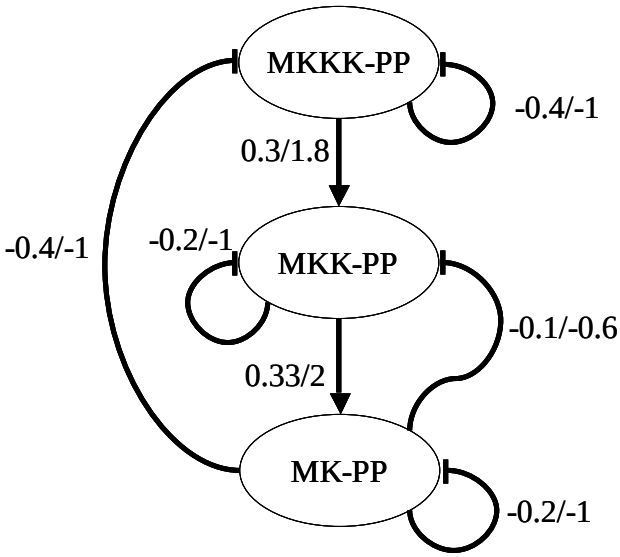

Figure 5.2 – *A simple cascade of signal transduction modules inferred with RSA, from data produced with a complicated signal transduction model of which details can be found at http://www.pnas.org/cgi/data/192442699/DC1/4. Edges correspond to non-zero (> 0.05) Regulatory Strengths. The numbers on the edges correspond to the inferred values of the Regulatory Strength and 'interaction map coefficients' taken from [84], respectively.*

The interaction map coefficients can also be calculated from the Regulatory Strengths, by dividing the Regulatory Strengths into a variable by the self-effect strength of that variable (note that the ratios are not consistent due to rounding off to one digit). Thus, regulatory strength analysis is not limited to only analysis of gene expression data, but also applicable to modular signal transduction. This exemplary result will soon be backed up formally (Bruggeman & de la Fuente *et al*., manuscript in preparation).

It is in principle even possible to apply regulatory strength analysis, and thus also the interaction map approach, for inferring the regulatory structure of metabolic networks. However, it is very hard experimentally to make the necessary perturbations. One has to make perturbations that specifically affect directly individual metabolites. This implies artificially adding a flux to each metabolite in the pathway in each experiment. Similar experiments are needed to infer complete signal transduction networks, and not just the interactions between modules. These conditions seem to be difficult to meet for those networks.

One possible (though hypothetical) way to be able to make the necessary perturbation experiments on metabolic pathway would be to use a Continuously Stirred Flow Reactor, using for example yeast cells [240]. In this setup one can control the inflow of the metabolites into the reaction vessel. If in each experiment a different metabolite were included in the inflow medium, this would impose the required added fluxes directly into a metabolite (making the assumption each metabolite can enter the cell from the medium).

Doing so for all metabolites and measuring the steady state responses to each perturbation allows calculation of matrix **O**, much like for gene expression data.

Here the application of RSA to the metabolic pathway depicted in Figure 5.3a is demonstrated. The reaction between M3 and M4 is described by reversible competitive inhibition kinetics and all other reactions by reversible Michaelis-Menten kinetics. All parameters are set to unity, except the inhibition constant, which was set to 0.1 and P1 and P2, were set to 10^{-5}. Metabolites M1, M2 and M5 could be specifically perturbed by changing the kinetic constants of the rates between M1, M2 and M5 and S, P1 and P2, respectively, since these rates are unique to these metabolites. M1 was perturbed by increasing S by 10%. M4 and M5 were perturbed by increasing their V_{max} by 10%. Since there are no rates unique to M2 and M3, in order to perturb these metabolites specifically, fluxes specifically into to these metabolites were added, each having a magnitude of 10% of the main flux through the pathway. After applying all perturbations **O** was calculated **R** obtained by inversion. The inferred matrix is:

$$\mathbf{R} = \begin{pmatrix} -1.42 & 0.49 & 0 & 0 & 0 \\ 1.26 & -2.55 & 0.71 & 0 & 0.45 \\ 0 & 1.61 & -1.93 & 0.22 & 0.24 \\ 0 & 0 & 1.14 & -1.18 & -0.4 \\ 0 & 1.31 & 0 & 0 & -1.46 \end{pmatrix} \qquad \text{(Eq. 5.21)}$$

The interaction map can simply be obtained by dividing each row by the negative value of the self-effect (diagonal element). The theoretical **R** matrix can be obtained from the Jacobian using $\mathbf{R} \equiv -diag\left(\mathbf{J}^{-1}\right)\mathbf{J}$. Since relative concentrations were used, matrix **R** should be scaled accordingly by the steady state concentrations of the metabolites. Diagonal matrix $\mathbf{D^M}$ has these metabolite concentrations on its diagonal. The scaled Regulatory Strength matrix is obtained by $\mathbf{R}^{scaled} = \left(\mathbf{D^M}\right)^{-1}\mathbf{R}\mathbf{D^M}$ [241]. Doing so yields the (scaled) theoretical $\mathbf{R_T}$ matrix in Eq. 5.21, which is indeed very close to the (scaled) **R** matrix obtained through the perturbation analysis Eq 5.20.

$$\mathbf{R_T} = \begin{pmatrix} -1.42 & 0.47 & 0 & 0 & 0 \\ 1.25 & -2.52 & 0.67 & 0 & 0.47 \\ 0 & 1.57 & -1.89 & 0.22 & 0.24 \\ 0 & 0 & 1.15 & -1.19 & -0.42 \\ 0 & 1.33 & 0 & 0 & -1.47 \end{pmatrix} \qquad \text{(Eq. 5.22)}$$

Figure 5.3b shows the resulting network. Obviously, the external metabolites don't appear in the network. In contrast to the arrows in the metabolic pathway, the arrows in the inferred network don't correspond to fluxes, but to regulatory effects.

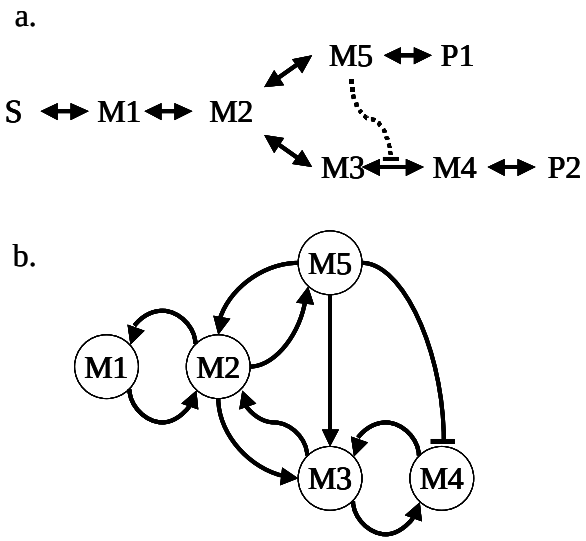

Figure 5.3 – *A) The metabolic pathway modeled studied with regulatory strength analysis and B) its inferred counterpart. Edges correspond to non-zero elements in the inferred R matrix. For clarity the negative self-interactions are omitted.*

Interestingly, the kinetic effect of M5 on the reaction between M3 and M4 shows up in the inferred network, but as separate regulatory effects. Since it inhibits the reaction the effect is negative on M4 and positive on M3. The regulatory description cannot be translated into the metabolic pathways structure, since different metabolic pathways could give rise to this regulatory network. For example, a pathway in which there is an irreversible reaction from M5 to M4 and in which M5 activates the degradation of M4 to P2 would give exactly the same regulatory network. The network inferred is a phenomenological description of the mechanistic pathway; this is similar to the case of gene networks inferred before. Again, it is important to mention that the perturbation experiments for metabolic pathways and signal transduction cascades are hard to realize experimentally given the present limited ability to perform such large scale specific perturbation experiments. The perturbation strategy proposed by the RSA is general and plausible to apply experimentally only in the case of gene networks. Regulatory strength analysis could be applied, however, to whole biochemical networks, consisting of genes, proteins and metabolites, as long as someone finds out a way to apply the required perturbations, which would necessarily be a complicated experiment. The presence of linear constraints between concentrations, due to moiety conserved cycles or equilibrium reactions, will complicate the application of this perturbation strategy, since such constraints won't allow for making the required unique rate perturbations and also cause the **O** matrix to be singular. How to deal with such constraints is a topic worth of further investigation.

Coming back to the use of the modular approach, there is no straightforward way for grouping genes into modules. In Hierarchical Control Analysis a module is defined as a series of reactions that have no flux exchange with other sets of reactions [103, 111, 116, 160]. In this way, signal transduction can be decomposed into such modules. However, in gene networks, using this definition for module, there are as many modules as there are

genes, since each gene is synthesized and degraded by its own 'private' fluxes; no flux is shared between genes, they solely communicate through kinetic effects. Furthermore, while the interaction between these supposed modules could be inferred, the interaction between the genes inside the module remains unknown. To infer the interaction structure inside the modules, all necessary perturbations have to be made.

One way to define a 'genetic module' is by applying a clustering algorithm to gene expression data and to assume that genes that end up in the same cluster belong to the same 'gene expression module'. Such an approach was applied to a yeast microarray dataset [242]. The genes were first clustered based on correlated time series and then knock-out mutants were used as perturbations to find the network of interactions between 'variables', of which the concentrations were equal to the average concentration of the genes in each cluster [242]. It remains questionable if it is valid to make the assumption that clustering of gene expression data yields 'functional genetic modules' or even groups genes of similar gene regulatory functions at all! It would be worthwhile to apply their method to simulated data in order to evaluate the results. The results of the application of the approach to yeast microarray data could not be thoroughly evaluated for correctness.

Section 5.3: How to deal with imperfect experimental data

There are three main concerns about using Regulatory Strength Analysis, as well as the methods described in Chapter 2, to real experimental biological data:

1) Biochemical interactions are non-linear
2) Experimental data are noisy
3) The problem of under-determination (insufficient number of experiments)

Section 5.3.1: Biochemical interactions are non-linear; linear approximation of non-linear kinetics

All analyses described in this dissertation originated from a linear description of biochemical systems. Therefore, the theory is only exact with infinitesimal perturbations, because the kinetics of such systems is nonlinear. However, infinitesimal perturbations are impossible to apply experimentally. One is forced to apply finite perturbations and the analyses become only approximate. The question now is how good the approximation for different perturbation sizes is.

Obviously, how well the linear approximation works will differ from system to system, depending on how non-linear their responses truly are. If the kinetics of a system is known one could analytically study the second order [107] or higher order terms in the Taylor expansion to check their relative importance to the responses. In Chapter 4 in section 4.5 this point has been addressed to some extend by analyzing a large set of data generated by large gene networks formulated with non-linear kinetic functions, and showed that even at perturbations as large as 50% the results were quite good. These results suggest that the need for very small perturbations can be relaxed. In the previous chapter a reasonable perturbation strategy was described: it was suggested to compare the expression profiles of wild type strains with heterozygous strains, in which one chromosome contains an intact copy of the gene while on the other the gene has been deleted. For a diploid organism, such as yeast, having 50% of the wild type gene dosage would lead to a reduction in the rate of

gene expression by about 50%, that is if the gene expression rate depends linearly on the gene dosage. It is reasonable to expect that the dependency of the gene expression rate on the gene dosage is linear, or maybe even asymptotic, and not with a faster growing function. Several programs are concerned with the creation of genomic scale libraries of knockout strains of various organisms [243, 244]). Gene expression data from knockout mutants can unfortunately not be used in our approach, since removing a gene completely is too drastic a perturbation and would lead to singularities in the analysis. However, it would experimentally feasible to create another library consisting of all the heterozygous mutants, each having only a functional copy of a particular gene on one of the two chromosomes, by backcrossing with the isogenic wild type. Gene expression data from these mutants would be very appropriately used with our approach to infer the network structure. By comparing the gene expression of these heterozygous mutants to the wild type gene expression, matrix **O** can be calculated and therefore matrix **R**, giving the network of genetic interactions.

Larger perturbations, although they introduce a certain amount of non-linearity to the data, are preferable to small perturbations in the light of the noisy nature of current microarray experiments. The responses towards small perturbations may easily get lost in the noise, while larger responses 'peak out' of the noise. Ideally, the size of the perturbations should be such that a good trade off is made between deviations from linearity and the measurement noise.

Section 5.3.2: Experimental data are noisy

A big concern is the quality of current microarray data in terms of reproducibility, and currently it is frequent to obtain such data that contains large variance. Matrix inversion is highly sensitive to noise in the matrix elements. Therefore, obtaining the **R** matrix directly by inverting the **O** matrix is likely to produce poor results. The method described in Section 2.5.4 in Chapter 2 also relies on matrix inversion. A way to solve a system of linear equations, which is less sensitive to noise in the data than just plain inversion, is regression with subset selection. This regression approach has been tested on noisy data for an approach equivalent to the Regulatory Strength Analysis (see section 5.2), and the results are very promising [36, 245]. Even at a noise level of 30% of the average responses, the False Discovery Rate (for definition see section 4.5.1 Chapter 4) was about 10% and 60% of the edges in the network were correctly identified.

To explain how the regression approach can be applied to solve for the regulatory strength matrix, I first rewrite the equation in order solve for individual rows **R**, *i.e.* R_i

$$R_i\mathbf{O} = I_i \qquad \text{(Eq. 5.23)}$$

where I_i is the corresponding row of the identity matrix.

Since there are *n* coefficients to estimate and *n* equations this equation is exactly identified and the solution is unique. Now, the assumption is made that gene networks are sparse and thus that each gene has a much lower number of input connections than there are genes in the network. A sparse matrix (network) has more zero elements than non-zero elements. If this sparseness in gene interactions assumption is correct, there is thus no need to solve for

all elements in row R_i, but just for the non-zero Regulatory Strengths. But the problem now is that one does not know which particular elements in R_i are non-zero and how many of them are non zero. Assuming that there are *k* non-zero elements in R_i, all (*n* choose *k*) = $n!/(n-k)!k!$ possible subsets of *k* variables have to be tested and select the subset that gives the best fit to the data (for large networks it is impossible to compute all possible combinations, but smart search strategies can be used, see below). 'Goodness of fit' can be expressed, for example**,** with the residual sum of squares (RSS): the lower the RSS the better the fit.

Solving for a particular subset *k*:

$$R_i^{subset(k)} \mathbf{O}^{subset(k)} = I_i \qquad \text{(Eq. 5.24)}$$

$R_i^{subset(k)}$ contains a certain subset of size *k* of assumed non zero Regulatory Strengths and $\mathbf{O}^{subset(k)}$ contains only the *k* corresponding rows. If the subset size *k* is smaller than the number of perturbations *n* the system is over-determined, which solution can be approximated using the least squares approach:

$$R_i^{sub} = \left(\mathbf{O}^{subset(k)} \mathbf{O}^{subset(k)^T} \right)^{-1} \mathbf{O}^{subset(k)} I_i \qquad \text{(Eq. 5.25)}$$

The subset that gives the best fit to Eq. 5.24 is the preferred solution.

There is yet another problem with this approach: it is not known how many inputs a certain gene has. A common problem in subset selection in regression is that a larger subset of regressors usually gives better fit than smaller ones [246]. Even if the correct subset *k* (the subset containing all true effectors of gene *i*) is identified, addition of another variable (subsets of *k*+1) will in general yield lower values for the RSS [246]. Therefore, there is need to score solutions with a function that gives penalty for including more variables, giving preference to simpler models (Occam's razor). Several such measures are widely used in statistics, such as the Akaike Information Criterion [247] and the Bayesian Information Criterion [248]. Both these criteria are simple functions of the RSS (for linear regression) and contain a penalty term for the number of variables in the subset.

Although this approach of regression by subset selection still needs to be thoroughly evaluated, these results indicate that application of Regulatory Strength Analysis to infer gene networks from real experimental biological data is possible. However, for the analysis to be quantitative, data of good quality should be used. Data with experimental noise and that displays non-linearity of interactions will only yield a qualitative description, at best.

Section 5.3.3: The problem of under-determination

The method requires a genome-scale experimental effort, in the form of gene expression rate manipulations, and measurements of genome-wide responses. It was shown that as many perturbations as there are genes in the network were needed to unambiguously infer

the complete interaction structure. Making all these perturbations is a huge experimental effort. When only *p* experiments out of *n* experiments are done on the *n*-gene system under study, the dimensions of **O** are *n* x *p* and the solution to the *n* x *n* **R** matrix is therefore not unique.

There are two distinct ways to deal with such underdetermined experimental data. The first is the obvious one of dimension reduction. Above I described two methods that reduce the number of variables in order to obtain a determined system. One proposes to consider 'modules' consistent of many genes, but provides no means of creating such modules [84], while the second proposes to group genes by means of hierarchical cluster analysis [242]. Effectively, a *p* x *p* **O** matrix is thus considered, and it is possible to find the interaction network between *p* modules or clusters. To infer the interactions inside each module or cluster and to discover which specific genes intermediate the communications between the modules or clusters, the additional *n* - *p* experiments must be done.

In the previous chapter a third strategy was suggested, alternative to the two above, which is more 'variable oriented', than 'module oriented', which doesn't require any previous knowledge of the network. As was argued in the previous chapter, a phenomenological network could obtained by considering $\mathbf{R}_{p\times p} = \mathbf{O}^{-1}_{p\times p}$, thus ignoring all *n* - *p* genes that have not been perturbed. The regulatory strength matrix obtained by this equation gives the interaction structure between the perturbed genes in which the connections may be direct or go through unperturbed genes. Again, although highly phenomenological, this **R** gives a valid quantitative description of the interaction between the perturbed genes. By including additional experiments the resolution of the network is increased. With this iterative process in mind, the responses of all *n* genes should be measured, not only the *p* perturbed genes. Although the unperturbed genes are not included in the present analysis, having them measured enables one to include them in the analysis whenever further experiments are done in which they are perturbed. It should be stressed that, although for the final goal (a global gene network), all experiments are needed; the intermediate results in this iterative process already provide good quantitative insight in cellular regulation.

Another way to uniquely identify an **R** matrix is to make assumptions about its structure. Recent papers have proposed strategies to deal with under-determination of data by using the assumption that gene networks are sparse. Again, the network being sparse implies that each gene only receives few inputs from other genes. Making this assumption enables, for example, to use Singular Value Decomposition of the non-square co-control matrix and to search for the sparsest solution, using robust regression [249], rather than the least square solution [250]. This approach has been implemented on a linear model to infer gene networks from time series gene expression data and it was demonstrated with simulated data that only log(*n*) measurements were needed to infer a network of *n* genes [249]. In has to be noted that in [249] the data was simulated with models of gene networks that were most sparse, *i.e.* there were only single paths between genes. It has been argued that biological networks should be as simple as possible, as they are optimized by evolution and simplicity is favorable for energetic reasons [80]. However, there are many more features on which organisms have been selected during evolution of which efficient regulation, homeostasis, and robustness are examples. Most important features of living cells require redundancy, *i.e.* many paths leading to the same 'goal'. Although the assumption that biochemical networks are sparse is reasonable, there is no reason to believe

that biochemical networks are the 'sparsest possible' networks explaining an observation. That said, the approach by [249] will probably not identify the correct gene network, but rather the simplest that conforms with the log(*n*) observations. This is still an interesting result, since the edges in the network found can be expected to be a subset of the edges in the real network. Their method will then identify a 'non-redundant backbone' of the gene network, a good starting point for further investigations. Another approach has been proposed by [251] which consists of searching, gene by gene, the space of solutions on subsets of inputs (similar to the approach outline in section 5.3.2) with computationally efficient search algorithms. They applied several distinct methods to search the enormous space of possible solutions, such as genetic algorithms, forward-, backward- and stepwise regression [246]. Yet another uses constrained least squares fitting, using the LASSO (least absolute shrinkage and selection operator) [252]. With the LASSO, in addition to the least squares constraint, a constraint is imposed that restricts the values of the regression coefficients by holding their absolute sum smaller than a preset value, thereby forcing most elements to be zero. When this preset value is small enough the solution is unique even for small numbers of observations. Using this approach, a unique solution to a regression problem of 6178 genes with only 73 observations (time points) was obtained [253].

The more assumptions are made and constraints imposed on a problem, the easier it is to solve it, but also the further it gets removed from the physical reality. It would be certainly interesting to apply these methods to solve for a unique **R** matrix from a non-square **O** matrix, at least before all experiments are done. Although the methods above have been demonstrated on linear models for time series analysis, the algorithms should work equivalently for our steady state perturbation approach, since they are equivalent in terms of matrix algebra.

Section 5.4: Final conclusion

In this dissertation, I have proposed several theoretical analyses accompanied by recipes to carry out the relevant experiments with the aim to decipher the regulatory structure of biochemical networks. One set of approaches enables experimental determination of the properties of metabolic systems in isolation from the global hierarchical network. Another provides a theoretical framework for simplification of hierarchical networks into gene networks and provides an experimental recipe to extract such networks from large-scale experimental observations. The methods have been thoroughly evaluated on data generated with mathematical models according to the proposed experimental designs. To be able to apply these methods with confidence to real biological experimental data, especially given the fact that the current experimental technologies suffer from low signal to noise ratios, further investigations and modifications to the methods are needed. On the other hand, this contribution will also increase the incentives for further technological improvements, improving the quality of such large-scale data. Therefore, in the light of datasets with higher quality and incorporation of robust computational techniques, I expect that it will be possible to apply these methods to large-scale genomics data in a not-too-distant a future.

Bibliography

1. Weiner, N., *Cybernetics or Control and Communication in the Animal and the Machine*. 1948, Cambridge, MA: MIT Press.
2. von Bertalanffy, L., *General System Theory*. 1968, New York: Braziller.
3. Katchalsky, A. and P.F. Curran, *Nonequilibrium Thermodynamics in Biophysics*. 1965, Cambridge, MA: Harvard University Press.
4. Kauffman, S., *Homeostasis and differentiation in random genetic control networks.* Nature, 1969. **224**(215): p. 177-178.
5. Glansdorff, P. and I. Prigogine, *Thermodynamic Theory of Structure, Stability, and Fluctuations*. 1971, London: Wiley.
6. Kacser, H. and J.A. Burns, *The control of flux.* Symp. Soc. Exp. Biol., 1973. **27**: p. 65-104.
7. Heinrich, R. and T.A. Rapoport, *A linear steady-state treatment of enzymatic chains. General properties, control and effector strength.* Eur. J. Biochem., 1974. **42**(1): p. 89-95.
8. Savageau, M.A., *Biochemical Systems Analysis.* 1976, Reading, MA: Addison-Wesley.
9. Reich, J.G. and E.E. Sel'kov, *Energy metabolism of the cell. A theoretical treatise*. 1981, London: Academic Press.
10. Westerhoff, H.V. and K. van Dam, *Thermodynamics and Control of Biological Free Energy Transduction*. 1987, Amsterdam: Elsevier.
11. de la Fuente, A. and P. Mendes, *Integrative modelling of gene expression and cell metabolism.* Appl Bioinformatics., 2003. **2**(2): p. 79-90.
12. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-470.
13. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-686.
14. Wodicka, L., et al., *Genome-wide expression monitoring in Saccharomyces cerevisiae.* Nature Biotechnology, 1997. **15**(13): p. 1359-1367.
15. Chu, S., et al., *The transcriptional program of sporulation in budding yeast.* Science, 1998. **282**(5389): p. 699-705.
16. Holstege, F.C., et al., *Dissecting the regulatory circuitry of a eukaryotic genome.* Cell, 1998. **95**(5): p. 717-728.
17. Cho, R.J., et al., *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 1998. **2**(1): p. 65-73.
18. Kehoe, D.M., P. Villand, and S. Somerville, *DNA microarrays for studies of higher plants and other photosynthetic organisms.* Trends Plant Sci., 1999. **4**(1): p. 38-41.
19. Wei, Y., et al., *High-density microarray-mediated gene expression profiling of Escherichia coli.* J. Bacteriol., 2001. **183**(2): p. 545-556.
20. Freeman, W.M., S.J. Walker, and K.E. Vrana, *Quantitative RT-PCR: pitfalls and potential.* Biotechniques, 1999. **26**(1): p. 112-122, 124-125.
21. O'Farrel, P.H., *High resolution two-dimensional electrophoresis of proteins.* Journal of Biological Chemistry, 1975. **250**: p. 4007-4021.

22. Gygi, S.P., et al., *Correlation between Proeitein and mRNA Abundance in Yeast.* Molecular and Cellular Biology, 1999. **19**(3): p. 1720-1730.
23. Wolters, D., M. Washburn, and J.r. Yates, *An automated multidimensional protein identification technology for shotgun proteomics.* Anal Chem., 2001. **73.**(23): p. 5683-5690.
24. Pappin, D., *Peptide mass fingerprinting using MALDI-TOF mass spectrometry.* Methods Mol Biol., 2003. **211**: p. 211-219.
25. Taylor, J. and R. Johnson, *Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry.* Anal Chem., 2001. **73**(11): p. 2594-2604.
26. Gibon, Y., et al., *A Robot-based platform to measure multiple enzyme activities in Arabidopsis using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness.* Plant Cell, 2004. **16**(12): p. 3304-3325.
27. Ficarro, S.B., et al., *Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae.* Nat Biotechnol, 2002. **20**(3): p. 301-5.
28. de la Fuente van Bentem, S., et al., *Phosphoproteomics as a tool to unravel plant regulatory mechanisms.* Physiologia Plantarum, 2006. **126**(10): p. 110 -119.
29. Roessner, U., et al., *Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry.* Plant J., 2000. **23(**(1): p. 131-142.
30. Kyranos, J., et al., *High-throughput high-performance liquid chromatography/mass spectrometry for modern drug discovery.* Curr Opin Biotechnol., 2001. **12**(1): p. 105-111.
31. Soga, T., et al., *Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry.* Anal Chem, 2002. **74**(10): p. 2233-9.
32. Teusink, B., et al., *Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry [In Process Citation].* Eur J Biochem, 2000. **267**(17): p. 5313-29.
33. Bakker, B.M., et al., *Glycolysis in bloodstream form Trypanosoma brucei can be understood in terms of the kinetics of the glycolytic enzymes.* Journal of Biological Chemistry, 1997. **272**(6): p. 3207-3215.
34. Jongsma, A.P., et al., *Evaluating limited specificity of drug pumps reduced relative resistance in human MDR phenotypes.* Eur J Biochem, 2000. **267**(17): p. 5369-5377.
35. Bakker, B.M., et al., *Network-based selectivity of antiparasitic inhibitors.* Mol Biol Rep, 2002. **29**(1-2): p. 1-5.
36. Gardner, T., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-105.
37. Koffas, M.A., G.Y. Jung, and G. Stephanopoulos, *Engineering metabolism and product formation in Corynebacterium glutamicum by coordinated gene overexpression.* Metab Eng, 2003. **5**(1): p. 32-41.
38. Hoefnagel, M.H., et al., *Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis.* Microbiology, 2002. **148(Pt 4**: p. 1003-1013.
39. Brazhnik, P., A. de la Fuente, and P. Mendes, *Gene networks: how to put the function in genomics.* Trends in Biotechnology, 2002. **20**(11): p. 467-472.

40. Henri, V., *Théorie générale de l'action de quelques diastases.* Compt. Rend. Hebd. Acad. Sci. Paris, 1902. **135**: p. 916-919.
41. Michaelis, L. and M.L. Menten, *Die kinetik der invertinwirkung.* Biochem. Z., 1913. **49**: p. 333-369.
42. Monod, J., J. Wyman, and J.-P. Changeux, *On the nature of allosteric transitions: a plausible model.* J. Mol. Biol., 1965. **12**: p. 88-118.
43. King, E.L.a.A., C., *A Schematic Method of Deriving the Rate Laws for Enzyme-Catalyzed Reactions.* J. Phys. Chem., 1956. **60**: p. 1375-1378.
44. Visser, D. and J.J. Heijnen, *Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics.* Metab Eng, 2003. **5**(3): p. 164-176.
45. Hill, A.V., *The possible effect of the aggregation of the molecules of hæmoglobin.* J. Physiol., 1910. **40**: p. iv-vii.
46. Hofmeyr, J.-H.S. and A. Cornish-Bowden, *The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models.* Computer Applications in the Biosciences, 1997. **13**: p. 377-385.
47. Thomas, R., *Boolean formalization of genetic control circuits.* Journal of Theoretical Biology, 1973. **42**(3): p. 563-585.
48. Hjelmfelt, A., E.D. Weinberger, and J. Ross, *Chemical implementation of finite-state machines.* Proc Natl Acad Sci U S A, 1992. **89**(1): p. 383-7.
49. Arkin, A. and J. Ross, *Computational functions in biochemical reaction networks.* Biophys J, 1994. **67**(2): p. 560-78.
50. Ko, M.S., *A stochastic model for gene induction.* Journal of Theoretical Biology, 1991. **153**(2): p. 181-194.
51. Carrier, T.A. and J.D. Keasling, *Mechanistic modeling of prokaryotic mRNA decay.* Journal of Theoretical Biology, 1997. **189**(2): p. 195-209.
52. McAdams, H.H. and A. Arkin, *Stochastic mechanisms in gene expression.* Proc Natl Acad Sci U S A, 1997. **94**(3): p. 814-9.
53. Arkin, A., J. Ross, and H.H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells.* Genetics, 1998. **149**(4): p. 1633-48.
54. McAdams, H.H. and A. Arkin, *It's a noisy business! Genetic regulation at the nanomolar scale.* Trends Genet, 1999. **15**(2): p. 65-9.
55. Morton-Firth, C.J. and D. Bray, *Predicting temporal fluctuations in an intracellular signalling pathway.* J Theor Biol, 1998. **192**(1): p. 117-28.
56. Petri, C.A., *Kommunikation mit Automaten.* 1962, Institut fur Instrumentelle Mathematik: Bonn.
57. Reddy, V.N., M.L. Mavrovouniotis, and M.N. Liebman, *Petri net representations in metabolic pathways.* Intelligent Systems for Molecular Biology, 1993. **1**: p. 328-336.
58. Goss, P.J. and J. Peccoud, *Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets.* Proc. Natl. Acad. Sci. USA, 1998. **95**(12): p. 6750-5.
59. Srivastava, R., M.S. Peterson, and W.E. Bentley, *Stochastic kinetic analysis of the Escherichia coli stress circuit using sigma(32)-targeted antisense.* Biotechnol Bioeng, 2001. **75**(1): p. 120-9.
60. Regev, A., W. Silverman, and E. Shapiro, *Representation and simulation of biochemical processes using the pi- calculus process algebra.* Pacific Symposium on Biocomputing, 2001. **7**: p. 459-70.

61. Srere, P.A., *Complexes of sequential metabolic enzymes.* Annual Review of Biochemistry, 1987. **56**: p. 89-124.
62. Clegg, J.S., *Cellular infrastructure and metabolic organization.* Current Topics in Cellular Regulation, 1992. **33**: p. 3-14.
63. Rohwer, J.M., et al., *Implications of macromolecular crowding for signal transduction and metabolite channelin.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10547-10552.
64. Westerhoff, H.V., *The silicon cell, not dead but live!* Metab Eng, 2001. **3**(3): p. 207-10.
65. Snoep, J.L. and B.G. Olivier, *Java Web Simulation (JWS); a web based database of kinetic models.* Mol Biol Rep, 2002. **29**(1-2): p. 259-63.
66. Snoep, J.L. and B.G. Olivier, *JWS online cellular systems modelling and microbiology.* Microbiology, 2003. **149**(Pt 11): p. 3045-7.
67. Olivier, B.G. and J.L. Snoep, *Web-based kinetic modelling using JWS Online.* Bioinformatics, 2004. **20**(13): p. 2143-4.
68. Snoep, J.L., et al., *Towards building the silicon cell: a modular approach.* Biosystems, 2006. **83**(2-3): p. 207-16.
69. Tomita, M., et al., *E-CELL: Software Environment for Whole Cell Simulation.* Genome Inform Ser Workshop Genome Inform, 1997. **8**: p. 147-155.
70. Tomita, M., et al., *E-CELL: software environment for whole-cell simulation.* Bioinformatics, 1999. **15**(1): p. 72-84.
71. Tomita, M., *Whole-cell simulation: a grand challenge of the 21st century.* Trends Biotechnol, 2001. **19**(6): p. 205-10.
72. Gilman, A.G., et al., *Overview of the alliance for cellular signaling.* Nature, 2002. **420**(6916): p. 703-6.
73. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* 2000. **28**(1): p. 27-30.
74. Selkov, E.J., et al., *MPW: the Metabolic Pathways Database.* Nucleic Acids Res, 1998. **26**(1): p. 43-5.
75. Kolchanov, N.A., et al., *GeneNet database: description and modeling of gene networks.* In Silico Biol, 2002. **2**(2): p. 97-110.
76. Huerta, A.M., et al., *RegulonDB: a database on transcriptional regulation in Escherichia coli.* Nucleic Acids Res, 1998. **26**(1): p. 55-59.
77. D'Haeseleer, P., S. Liang, and R. Somogyi, *Genetic network inference: from co-expression clustering to reverse engineering.* Bioinformatics, 2000. **16**(8): p. 707-726.
78. de Jong, H., *Modeling and simulation of genetic regulatory systems: a literature review.* J Comput Biol, 2002. **9**(1): p. 67-103.
79. Wahde, M. and J. Hertz, *Coarse-grained reverse engineering of genetic regulatory networks.* Biosystems, 2000. **55**(1-3): p. 129-136.
80. Wagner, A., *How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps.* Bioinformatics, 2001. **17**(12): p. 1183-1197.
81. de la Fuente, A., P. Brazhnik, and P. Mendes. *A quantitative method for reverse engineering gene networks from microarray experiments using regulatory strengths.* in *2nd International Conference on Systems Biology*. 2001. California Institute of Technology, Pasadena, CA: Omnipress.

82. de la Fuente, A., P. Brazhnik, and P. Mendes, *Linking the Genes: Inferring Quantitative Gene Networks from Microarray Data.* Trends Genet., 2002. **18**(8): p. 395-398.
83. de la Fuente, A. and P. Mendes, *Quantifying gene networks with regulatory strengths.* Mol Biol Rep, 2002. **29**(1-2): p. 73-7.
84. Kholodenko, B.N., et al., *Untangling the wires: A strategy to trace functional interactions in signaling and gene networks.* Proc Natl Acad Sci U S A, 2002. **99**(20): p. 12841-6.
85. Mendes, P. and D. Kell, *Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation.* Bioinformatics, 1998. **14**(10): p. 869-883.
86. Mendes, P., *Modeling large scale biological systems from functional genomic data: parameter estimation*, in *Foundations of Systems Biology*, H. Kitano, Editor. 2001, MIT Press: Cambridge, MA. p. 163-186.
87. Koza, J.R. and D. Andre, *A case study where biology inspired a solution to a computer science problem.* Pac Symp Biocomput, 1996: p. 500-11.
88. Koza, J.R., et al., *Reverse engineering of metabolic pathways from observed data using genetic programming.* Pac Symp Biocomput, 2001: p. 434-45.
89. Schomburg, I., et al., *BRENDA: a resource for enzyme data and metabolic information.* Trends Biochem Sci, 2002. **1**: p. 54-56.
90. Schomburg, I., A. Chang, and D. Schomburg, *BRENDA, enzyme data and metabolic information.* Nucleic Acids Res, 2002. **30**(1): p. 47-49.
91. Mendes, P., *GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems.* Computer Applications in the Biosciences, 1993. **9**(5): p. 563-571.
92. Mendes, P., *Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3.* Trends in Biochemical Sciences, 1997. **22**: p. 361-363.
93. Sauro, H.M., *SCAMP: a general-purpose simulator and metabolic control analysis program.* Computer Applications in the Biosciences, 1993. **9**(4): p. 441-450.
94. Sauro, H.M., *JARNAC; a system for interactive metabolic analysis*, in *BTK2000: animating the cellular map*, J.-H.S. Hofmeyr, J.H. Rohwer, and J.L. Snoep, Editors. 2000, Stellenbosch University Press: Stellenbosch. p. 221-228.
95. Schaff, J. and L.M. Loew, *The virtual cell.* Pac Symp Biocomput, 1999: p. 228-39.
96. Ehlde, M. and G. Zacchi, *MIST: a user-friendly metabolic simulator.* Computer Applications in the Biosciences, 1995. **11**(2): p. 201-207.
97. Dang, Q. and C. Frieden, *New PC versions of the kinetic-simulation and fitting programs, KINSIM and FITSIM.* Trends Biochem Sci, 1997. **22**(8): p. 317.
98. Cornish-Bowden, A. and J.H. Hofmeyr, *METAMODEL - A program for modeling and control analysis of metabolic pathways on the ibm pc and compatibles.* Computer Applications in the Biosciences, 1991. **7**(1): p. 89-93.
99. Le Novere, N. and T.S. Shimizu, *STOCHSIM: modelling of stochastic biomolecular processes.* Bioinformatics, 2001. **17**(6): p. 575-6.
100. Hucka, M., et al., *The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models.* Bioinformatics, 2002. **in press**.

101. Schuster, S., D.A. Fell, and T. Dandekar, *A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks.* Nature Biotechnology, 2000. **18**(3): p. 326-32.
102. Edwards, J. and B. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.* Proc Natl Acad Sci USA., 2000. **97**(10): p. 5528-5533.
103. Kahn, D. and H.V. Westerhoff, *Control theory of regulatory cascades.* J. Theor. Biol., 1991. **153**(2): p. 255-285.
104. Hofmeyr, J.H.S., A. Cornish-Bowden, and J.M. Rohwer, *Taking enzyme kinetics out of control - putting control into regulation.* European Journal of Biochemistry, 1993. **212**(3): p. 833-837.
105. Hofmeyr, J.H. and A. Cornish-Bowden, *Co-response analysis: a new experimental strategy for metabolic control analysis.* Journal of Theoretical Biology, 1996. **182**(3): p. 371-380.
106. Westerhoff, H.V. and D.B. Kell, *Matrix method for determining steps most rate-limiting to metabolic fluxes in biotechnological processes.* Biotechnol. Bioeng., 1987. **30**: p. 101-107.
107. Hofer, T. and R. Heinrich, *A 2nd-order approach to metabolic control analysis.* Journal of Theoretical Biology, 1993. **164**: p. 85-102.
108. Kahn, D. and H.V. Westerhoff, *The Regulatory strength: how to be precise about regulation and homeostasis.* Acta Biotheoretica, 1993. **41**: p. 85-96.
109. Westerhoff, H.V. and Y.-D. Chen, *How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control.* Eur. J. Biochem., 1984. **142**: p. 425-430.
110. Reder, C., *Metabolic control theory. A structural approach.* Journal of Theoretical Biology, 1988. **135**(2): p. 175-201.
111. Hofmeyr, J.H. and H.V. Westerhoff, *Building the cellular puzzle: control in multi-level reaction networks.* J. Theor. Biol., 2001. **208**(3): p. 261-285.
112. Welch, G.R., *Cytosociological aspects of enzyme action.*, in *The organization of cell metabolism.*, G.R. Welch and J.S. Clegg, Editors. 1987, Plenum Press: New York and London. p. 367-380.
113. Kell, D.B. and H.V. Westerhoff, *Control analysis of organised multienzyme systems*, in *Structural and organizational aspects of metabolic regulation.*, P.A. Srere, M.E. Jones, and C.K. Mathews, Editors. 1990, Alan Liss, Inc.: New York. p. 273-289.
114. Westerhoff, H.V., J.H.S. Hofmeyr, and B.N. Kholodenko, *Getting to the inside of cells using metabolic control analysis.* Biophysical Chemistry, 1994. **50**(3): p. 273-283.
115. Acerenza, L., *Metabolic control design.* J. Theor. Biol., 1993. **165**(1): p. 63-85.
116. Westerhoff, H.V. and M. van Workum, *Control of DNA structure and gene expression.* Biomed. Biochim. Acta, 1990. **49**: p. 839-853.
117. Schauer, M. and R. Heinrich, *Quasi-steady-state approximation in the mathematical modeling of biochemical reaction networks.* Mathematical Biosciences, 1983. **65**(2): p. 155-170.
118. Liao, J.C. and E.N. Lightfoot, Jr., *Extending the quasi-steady state concept to analysis of metabolic networks.* Journal of Theoretical Biology, 1987. **126**(3): p. 253-73.

119. de la Fuente, A., *The distribution of control on metabolic and genetic levels*, in *Molecular Cell Physiology*. 1998, University of Amsterdam: Amsterdam.
120. de la Fuente, A., et al., *Can metabolic control analysis be applied to hierarchical regulated metabolism? MCA versus HCA*, in *BTK2000: animating the cellular map*, J.-H.S. Hofmeyr, J.H. Rohwer, and J.L. Snoep, Editors. 2000, Stellenbosch University Press: Stellenbosch. p. 191-198.
121. de la Fuente, A., et al., *Metabolic control in integrated biochemical systems.* Eur J Biochem, 2002. **269**(18): p. 4399-408.
122. Bernstein, J.A., et al., *Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.* Proc Natl Acad Sci U S A, 2002. **99**(15): p. 9697-702.
123. Wang, Y., et al., *Precision and functional specificity in mRNA decay.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5860-5.
124. Chance, B., *Analogue and digital representations of enzyme kinetics.* J. Biol. Chem., 1960. **235**: p. 2440-2443.
125. Garfinkel, D., *A simulation study of the metabolism and compartmentation in brain of glutamate, aspartate, the Krebs cycle, and related metabolites.* J Biol Chem, 1966. **241**(17): p. 3918-29.
126. Achs, M.J. and D. Garfinkel, *Simulation of the detailed regulation of glycolytic oscillation in a heart supernatant preparation.* Comput Biomed Res, 1968. **2**(1): p. 92-110.
127. Park, D.J., *The hierarchical structure of metabolic networks and the construction of efficient metabolic simulators.* Journal of Theoretical Biology, 1974. **46**(1): p. 31-74.
128. Kohn, M.C. and D. Garfinkel, *Computer simulation of ischemic rat heart purine metabolism. I. Model construction.* Am J Physiol, 1977. **232**(4): p. H386-93.
129. Richter, O., H.J. Vohmann, and A. Betz, *A simulation study of oscillating glycolysis: a comparison between a model and experiments.* Chronobiologia, 1978. **5**(1): p. 56-65.
130. Reich, J.G. and E.E. Sel'kov, *Energy metabolism of the cell. A theoretical treatise.* 1981, London: Academic Press.
131. Brumen, M. and R. Heinrich, *A Metabolic osmotic model of human erythrocytes.* BioSystems, 1984. **17**(2): p. 155-169.
132. Franco, R. and E.I. Canela, *Computer simulation of purine metabolism.* European Journal of Biochemistry, 1984. **144**(2): p. 305-315.
133. Joshi, A. and B.O. Palsson, *Metabolic dynamics in the human red cell. Part I - A comprehensive kinetic model.* Journal of Theoretical Biology, 1989. **141**(4): p. 515-528.
134. Kuchel, P.W., et al., *Computer simulation of the pentose-phosphate pathway and associated metabolism used in conjunction with NMR experimental data from human erythrocytes.* Biomed Biochim Acta, 1990. **49**(8-9): p. 757-70.
135. Rizzi, M., et al., *In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae.2. Mathematical model.* Biotechnology and Bioengineering, 1997. **55**(4): p. 592-608.
136. Rohwer, J.M., et al., *Understanding glucose transport by the bacterial phosphoenolpyruvate:glycose phosphotransferase system on the basis of kinetic measurements in vitro.* J Biol Chem, 2000. **275**(45): p. 34909-21.

137. Womble, D.D. and R.H. Rownd, *Regulation of lambda dv plasmid DNA replication. A quantitative model for control of plasmid lambda dv replication in the bacterial cell division cycle.* J Mol Biol, 1986. **191**(3): p. 367-82.
138. Koster, J.G., H.V. Westerhoff, and O.H.J. Destree, *Kinetics of Histone Gene Expression during Early Development of Xenopus laevis.* J. Theor. Biol., 1988. **135**: p. 139-167.
139. McAdams, H.H. and L. Shapiro, *Circuit simulation of genetic networks.* Science, 1995. **269**(5224): p. 650-656.
140. Burstein, Z., *A network model of developmental gene hierarchy.* Journal of Theoretical Biology, 1995. **174**(1): p. 1-11.
141. Endy, D., D. Kong, and J. Yin, *Intracellular kinetics of a growing virus: a genetically structured simulation for bacteriophage T7.* Biotechnology and Bioengineering, 1997. **55**: p. 375-389.
142. Mendoza, L. and E.R. Alvarez-Buylla, *Genetic regulation of root hair development in Arabidopsis thaliana: a network model.* J Theor Biol, 2000. **204**(3): p. 311-26.
143. Kastner, J., J. Solomon, and S. Fraser, *Modeling a hox gene network in silico using a stochastic simulation algorithm.* Developmental Biology, 2002. **246**(1): p. 122-131.
144. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* Journal of Molecular Biology, 1961. **3**: p. 2415-2421.
145. Goodwin, B.C., *Oscillatory behavior in enyzmatic control processes.* Advances in Enzyme Regulation, 1965. **3**: p. 425-439.
146. Griffith, J.S., *Mathematics of cellular control processes I.* Journal of Theoretical Biology, 1968. **20**: p. 202-208.
147. Griffith, J.S., *Mathematics of cellular control processes II.* Journal of Theoretical Biology, 1968. **20**: p. 209-216.
148. Bliss, R.D., P.R. Painter, and A.G. Marr, *Role of feedback inhibition in stabilizing the classical operon.* J Theor Biol, 1982. **97**(2): p. 177-93.
149. Sinha, S., *Theoretical study of the tryptophan operon: Application in microbial technology.* Biotechnology and Bioengineering, 1988. **31**: p. 117-124.
150. Sen, A.K. and W. Liu, *Dynamic analysis of genetic control and regulation of amino acid synthesis: the tryptophan operon in Escherichia coli.* Biotechnology and Bioengineering, 1990. **35**: p. 185-194.
151. Santillan, M. and M.C. Mackey, *Dynamic regulation of the tryptophan operon: a modeling study and comparison with experimental data.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1364-9.
152. Lee, S.B. and J.E. Bailey, *Genetically structured models for lac promoter-operator function in the Escherichia coli chromosome and in multicopy plasmids: lac operator function.* Biotechnology and Bioengineering, 1984. **26**: p. 1372-1382.
153. Lee, S.B. and J.E. Bailey, *Promoter-operator function in the chromosome and in multicopy plasmids: lac promoter function.* Biotechnology and Bioengineering, 1984. **26**: p. 1383-1389.
154. Wong, P., S. Gladney, and J.D. Keasling, *Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose.* Biotechnol Prog, 1997. **13**(2): p. 132-43.

155. Kremling, A. and E.D. Gilles, *The organization of metabolic reaction networks. II. Signal processing in hierarchical structured functional units.* Metab Eng, 2001. **3**(2): p. 138-50.
156. Kremling, A., et al., *The organization of metabolic reaction networks. III. Application for diauxic growth on glucose and lactose.* Metab Eng, 2001. **3**(4): p. 362-79.
157. Van Dien, S.J. and J.D. Keasling, *A dynamic model of the Escherichia coli phosphate-starvation response.* J Theor Biol, 1998. **190**(1): p. 37-49.
158. Van Dien, S.J. and J.D. Keasling, *Effect of polyphosphate metabolism on the Escherichia coli phosphate- starvation response.* Biotechnol Prog, 1999. **15**(4): p. 587-93.
159. Juty, N.S., et al., *Simultaneous modelling of metabolic, genetic and product-interaction networks.* Brief Bioinform, 2001. **2**(3): p. 223-32.
160. Westerhoff, H.V. and D. Kahn, *Control involving metabolism and gene expression: the square-matrix method for modular decomposition.* Acta Biotheor, 1993. **41**(1-2): p. 75-83.
161. Bruggeman, F.J., et al., *Modular response analysis of cellular regulatory networks.* J Theor Biol., 2002. **218**(4): p. 507-520.
162. Jensen, P.R., et al., *Extensive regulation compromises the extent to which DNA gyrase controls DNA supercoiling and growth rate of Escherichia coli.* Eur J Biochem, 1999. **266**(3): p. 865-77.
163. Snoep, J.L., et al., *Hierarchical control of DNA supercoiling in Escherichia coli: how to study homeostatically controlled sytems using control analysis*, in *BTK2000: animating the cellular map*, J.-H.S. Hofmeyr, J.H. Rohwer, and J.L. Snoep, Editors. 2000, Stellenbosch University Press: Stellenbosch.
164. van Workum, M., et al., *DNA supercoiling depends on the phosphorylation potential in Escherichia coli.* Mol Microbiol, 1996. **20**(2): p. 351-60.
165. Yang, H.L., et al., *Differential sensitivity of gene expression in vitro to inhibitors of DNA gyrase.* Proc Natl Acad Sci U S A, 1979. **76**(7): p. 3304-3308.
166. ter Kuile, B. and H.V. Westerhoff, *Transcriptome meets metabolome.* FEBS Lett., 2001. **500**: p. 169-171.
167. Easterby, J.S., *A generalized theory of the transition time for sequential enzyme reactions.* Biochem. J., 1981. **199**(1): p. 155-161.
168. Jensen, P.R., H.V. Westerhoff, and O. Michelsen, *The use of lac-type promoters in control analysis.* Eur. J. Biochem., 1993. **211**: p. 181-191.
169. Westerhoff, H.V., et al. *Hierarchical Control of Electron-Transfer*. in *NATO/ESF Workshop Bioelectron-Transfer Chains*. 1998. Tomar, Portugal: Kluwer Academic Publishers.
170. van der Gugten, A.A. and H.V. Westerhoff, *Internal regulation of a modular system: the different faces of internal control.* Biosystems, 1997. **44**(2): p. 79-106.
171. Heinrich, R. and I. Sonntag, *Dynamics of non-linear biochemical systems and the evolutionary significance of time hierarchy.* Biosystems, 1982. **15**(4): p. 301-316.
172. Hornberg, J.J., et al., *Principles behind the multifarious control of signal transduction. ERK phosphorylation and kinase/phosphatase control.* FEBS J, 2005. **272**(1): p. 244-258.
173. Stucki, J.W., *Stability analysis of biochemical systems. A practical guide.* Progress in Biophysics and Molecular Biology, 1978. **33**: p. 99-187.

174. Acerenza, L., H.M. Sauro, and H. Kacser, *Control analysis of time dependent metabolic systems.* Journal of Theoretical Biology, 1989. **137**(4): p. 423-444.
175. Heinrich, R. and C. Reder, *Metabolic control analysis of relaxation processes.* J. Theoret. Biol., 1991. **151**(3): p. 343-350.
176. Sorribas, A., et al., *Metabolic pathway characterization from transient response data obtained in-situ. Parameter estimation in S-system models.* J. Theoret. Biol., 1993. **162**(1): p. 81-102.
177. Mann, M., R.C. Hendrickson, and A. Pandey, *Analysis of Proteins and Proteomes by Mass Spectrometry.* Annu Rev Biochem, 2001. **70**: p. 437-473.
178. Raamsdonk, L.M., et al., *A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.* Nature Biotechnology, 2001. **19**(1): p. 45-50.
179. de la Fuente, A., P. Brazhnik, and P. Mendes, *Regulatory Strength Analysis for Inferring Gene Networks*, in *Metabolic Engineering in the Post Genomic Era*, K. BN and W. HV, Editors. 2004, Horizon Bioscience: Wymondham, UK. p. 107-137.
180. Kell, D.B. and R.D. King, *On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning.* Trends in Biotechnology, 2000. **18**: p. 93-98.
181. Consortium, G.O., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res., 2004. **32**(Database issue:D258-61).
182. Somogyi, R. and L.D. Greller, *The dynamics of molecular networks: applications to therapeutic discovery.* Drug Discov Today, 2001. **6**(24): p. 1267-1277.
183. Bailey, J.E., *Lessons from metabolic engineering for functional genomics and drug discovery.* Nat Biotechnol, 1999. **17**(7): p. 616-8.
184. Ewing, B. and P. Green, *Analysis of expressed sequence tags indicates 35,000 human genes.* Nat Genet, 2000. **25**(2): p. 232-4.
185. Kim, S.K., et al., *A gene expression map for Caenorhabditis elegans.* Science, 2001. **293**(5537): p. 2087-92.
186. Das, M., et al., *Assessment of the total number of human transcription units.* Genomics., 2001. **77**(1-2): p. 71-78.
187. Roberts, G. and C. Smith, *Alternative splicing: combinatorial output from the genome.* Curr Opin Chem Biol., 2002. **6**(3): p. 375-383.
188. von Dassow, G., et al., *The segment polarity network is a robust developmental module.* Nature, 2000. **406**(6792): p. 188-92.
189. Little, J.W., D.P. Shepley, and D.W. Wert, *Robustness of a gene regulatory circuit.* Embo J, 1999. **18**(15): p. 4299-307.
190. Thieffry, D. and R. Thomas, *Qualitative analysis of gene networks.* Pac Symp Biocomput, 1998: p. 77-88.
191. Snoep, J.L., et al., *Protein burden in Zymomonas mobilis: negative flux and growth control due to the overproduction of glycolytic enzymes.* Microbiology, 1995. **141**: p. 2329-2337.
192. Gantmacher, F.R., *The Theory of Matrices.* The Theory of Matrices. Vol. II. 1960, New York: Chelsea Publishing Company.
193. Schuster, S., D. Kahn, and H.V. Westerhoff, *Modular analysis of the control of complex metabolic pathways.* Biophysical Chemistry, 1993. **48**: p. 1-17.
194. Cogoni, C. and G. Macino, *Post-transcriptional gene silencing across kingdoms.* Curr Opin Genet Dev, 2000. **10**(6): p. 638-43.

195. Sugita, M., *Functional analysis of chemical systems in vivo using a logical circuit equivalent. II. The idea of a molecular automation.* J Theor Biol, 1963. **4**(2): p. 179-92.
196. Sugita, M. and N. Fukuda, *Functional analysis of chemical systems in vivo using a logical circuit equivalent. 3. Analysis using a digital circuit combined with an analogue computer.* J Theor Biol, 1963. **5**(3): p. 412-25.
197. Watts, D. and S. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, 1998. **393**(6684): p. 440-442.
198. Barabasi, A. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-512.
199. Serov, V., A. Spirov, and M. Samsonova, *Graphical interface to the genetic network database GeNet.* Bioinformatics, 1998. **14**(6): p. 546-547.
200. Ideker, T., et al., *Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.* Science, 2001. **292**(5518): p. 929-34.
201. Schwikowski, B., P. Uetz, and S. Fields, *A network of protein-protein interactions in yeast.* Nat Biotechnol, 2000. **18**(12): p. 1257-61.
202. Wingender, E., et al., *TRANSFAC: an integrated system for gene expression regulation.* Nucleic Acids Res, 2000. **28**(1): p. 316-9.
203. Zhu, J. and M.Q. Zhang, *SCPD: a promoter database of the yeast Saccharomyces cerevisiae.* Bioinformatics, 1999. **15**(7-8): p. 607-11.
204. Davidson, E.H., et al., *A genomic regulatory network for development.* Science, 2002. **295**(5560): p. 1669-78.
205. Yuh, C.H., H. Bolouri, and E.H. Davidson, *Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.* Science, 1998. **279**(5358): p. 1896-902.
206. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nature Genetics, 1999. **22**(3): p. 281-285.
207. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
208. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences U.S.A., 1998. **95**(25): p. 14863-14868.
209. Hilsenbeck, S.G., et al., *Statistical analysis of array expression data as applied to the problem of tamoxifen resistance.* J Natl Cancer Inst, 1999. **91**(5): p. 453-9.
210. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* J Comput Biol, 2000. **7**(3-4): p. 601-20.
211. Pe'er, D., et al., *Inferring subnetworks from perturbed expression profiles.* Bioinformatics, 2001. **17**(Suppl 1): p. S215-24.
212. Liang, S., S. Fuhrman, and R. Somogyi, *REVEAL, a general reverse engineering algorithm for inference of genetic network architectures.* Pacific Symposium on Biocomputing, 1998. **3**: p. 18-29.
213. Akutsu, T., et al., *A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions.* Genome Informatics, 1998. **9**: p. 151-160.
214. Ideker, T.E., V. Thorsson, and R.M. Karp, *Discovery of regulatory interactions through perturbation: inference and experimental design.* Pacific Symposium on Biocomputing, 2000. **5**: p. 305-16.

215. Hatzimanikatis, V. and K.H. Lee, *Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information.* Metabolic Engeneering, 1999. **1**(4): p. 275-281.
216. D'Haeseleer, P., et al., *Linear modeling of mRNA expression levels during CNS development and injury.* Pacific Symposium of Biocomputing, 1999. **4**: p. 41-52.
217. van Someren, E.P., L.F. Wessels, and M.J. Reinders, *Linear modeling of genetic networks from experimental data.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 355-66.
218. Holter, N.S., et al., *Dynamic modeling of gene expression data.* Proceedings of the National Academy of Sciences USA, 2001. **98**(4): p. 1693-1698.
219. Becskei, A. and L. Serrano, *Engineering stability in gene networks by autoregulation.* Nature, 2000. **405**(6786): p. 590-3.
220. Becskei, A., B. Seraphin, and L. Serrano, *Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion.* Embo J, 2001. **20**(10): p. 2528-35.
221. Mendes, P. *Metabolic simulation as an aid in understanding gene expression data.* in *Workshop on computation of Biochemical Pathways and Genetic Networks*. 1999. Heidelberg.
222. Wessels, L.F., E.P. van Someren, and M.J. Reinders, *A comparison of genetic network models.* Pac. Symp. Biocomput., 2001: p. 508-519.
223. Giersch, C., *Determining elasticities from multiple measurements of steady-state flux rates and metabolite concentrations. Theory.* Journal of Theoretical Biology, 1994. **169**(1): p. 89-99.
224. Giersch, C., *Determining elasticities from multiple measurements of flux rates and metabolite concentrations. Application of the multiple modulation method to a reconstituted pathway.* European Journal of Biochemistry, 1995. **227**(1-2): p. 194-201.
225. Kacser, H. and J.A. Burns, *Molecular democracy: who shares the controls?* Biochem. Soc. Trans., 1979. **7**: p. 1149-1160.
226. Ortega, F. and L. Acerenza, *Optimal metabolic control design.* J Theor Biol, 1998. **191**(4): p. 439-49.
227. Sharp, P.A., *RNAi and double-strand RNA.* Genes Dev., 1999. **13**(2): p. 139-141.
228. Snoep, J.L., et al., *DNA supercoiling in Escherichia coli is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase.* Eur. J. Biochem., 2002. **269**(6): p. 1662-1669.
229. Mendes, P., W. Sha, and K. Ye, *Artificial gene networks for objective comparison of analysis algorithms.* Bioinformatics, 2003. **19**: p. Suppl 2:II122-II129.
230. Erdös, P. and A. Rényi, *On the Evolution of Random Graphs.* Publ. Math. Inst. Hungar. Acad. Sci., 1960. **5**: p. 17-61.
231. Eldering, E., et al., *Expression profiling via novel multiplex assay allows rapid assessment of gene regulation in defined signalling pathways.* Nucleic Acids Res, 2003. **31**(23): p. e153.
232. Ando, S. and H. Iba, *Quantitative modeling of gene regulatory network: identifying the network by means of genetic algorithm.* Genome Informatics, 2000. **11**: p. 278-280.
233. Kyoda, K.M., et al., *A gene network inference method from continuous-value gene expression data of wild-type and mutants.* Genome Informatics, 2000. **11**: p. 196-204.

234. Maki, Y., et al., *Development of a system for the inference of large scale genetic networks.* Pacific Symposium on Biocomputing, 2001. **6**: p. 446-58.
235. Teusink, B., et al., *Metabolic control analysis as a tool in the elucidation of the function of novel genes.*, in *Methods in Microbiology: Yeast gene analysis*, M.F.T.a.A.J.P. Brown, Editor. 1998, Academic Press: London. p. 297-336.
236. Cornish-Bowden, A. and M.L. Cardenas, *Functional genomics. Silent genes given voice.* Nature, 2001. **409**(6820): p. 571-2.
237. Stark, J., R. Callard, and M. Hubank, *From the top down: towards a predictive biology of signalling networks.* Trends Biotechnol., 2003. **21**(7): p. 290-293.
238. Stark, J., et al., *Reconstructing gene networks: what are the limits?* Biochem Soc Trans., 2003. **31**(Pt 6): p. 1519-1525.
239. Sontag, E., A. Kiyatkin, and B.N. Kholodenko, *Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data.* Bioinformatics, 2000. **Epub ahead of print**.
240. Hynne, F., S. Dano, and P. Sorensen, *Full-scale model of glycolysis in Saccharomyces cerevisiae.* Biophys Chem., 2001. **94**(1-2): p. 121-163.
241. Hofmeyr, J.H. *MCA in a nutshell*. in *2nd International Conference on Systems Biology*. 2001. California Institute of Technology, Pasadena, CA: Omnipress.
242. Curtis, R. and M. Brand, *Control analysis of DNA microarray expression data.* Mol Biol Rep., 2002. **29**(1-2): p. 67-71.
243. Jacobs, M.A., et al., *Comprehensive transposon mutant library of Pseudomonas aeruginosa.* Proc Natl Acad Sci U S A, 2003. **100**(24): p.)14339-14344.
244. Winzeler, E.A., et al., *Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.* Science, 1999. **285**(5429): p. 901-906.
245. Di Bernardo, D., T. Gardner, and J. Collins, *Robust identification of large genetic networks.* Pac Symp Biocomput., 2004(9): p. 486-97.
246. Miller, A., *Subset Selection in Regression*. 1990, London: Chapman & Hall.
247. Akaike, H., *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 1974. **AC-19**(6): p. 716-723.
248. Schwarz, G., *Estimating the dimension of a model.* Ann Stat., 1978. **6**: p. 461-464.
249. Yeung, M., J. Tegner, and J. Collins, *Reverse engineering gene networks using singular value decomposition and robust regression.* Proc Natl Acad Sci U S A., 2002. **99**(9): p. 6163-6168.
250. Holter, N.S., et al., *Fundamental patterns underlying gene expression profiles: simplicity from complexity.* Proceedings of the National Academy of Sciences USA, 2000. **97**(15): p. 8409-8414.
251. van Someren, E.P., et al. *Searching for limited connectivity in genetic network models*. in *2nd International Conference on Systems Biology*. 2001. California Institute of Technology, Pasadena, CA: Omnipress.
252. Tibshirani, R., *Regression shrinkage and selection via the lasso.* J Royal Statist Soc B., 1996. **58**(1): p. 267-288.
253. Gustafsson, M., M. Hornquist, and A. Lombardi. *Large-scale reverse engineering by the Lasso*. 2004 [cited; Available from: http://arxiv.org/PS_cache/q-bio/pdf/0403/0403012.pdf.

Summary

Resumen

Samenvatting

Summary

Deciphering living networks

Perturbation strategies for functional genomics

Living beings are very complicated. We, humans, for example, consist of about 10^{14} (that's a 1 followed by 14 zeros!) cells. Every single cell contains between 30,000 and 100,000 different genes (no one knows how many exactly), most certainly more than 100,000 different proteins and thousands of different small molecules, called metabolites. Our genes, the hereditary information, produce proteins, the molecules that actually drive cell physiological processes. Metabolites are substances that we receive through food, from which we extract energy and make other metabolites with important functions. For example, certain metabolites determine our appearance; the color of our eyes and skin is determined by the presence or absence of a certain metabolite. The functioning of living cells rests on the coordinated interplay between all these distinct molecules.

Biology is the science concerned with understanding living creatures. Given that cells are so complicated it is a logical decision to take them apart and to study the molecules and processes individually, for instance in test tubes. This form of biological research has been practiced over the last 3 decennia with great success under the name Molecular Biology. The philosophy behind this approach is that if all components of the living cell are studied and understood, then the whole, the living cell, is understood. Probably this is not the case; the whole is more than the sum of its parts. However, this approach has been the most logical because, until recently, it was not possible to obtain experimental data simultaneously on all components of living cells. This ability does exist today. Recently, several experimental techniques have been developed allowing us to measure simultaneously all gene-activities, cellular concentrations of a large number of proteins and metabolites as well. The experimental data obtained through these technologies thus provide us, for the first time in the history of biology, with a means to study all components and processes simultaneously. The scientific discipline concerned with measurement and interpretation of such large scale biological data has been named 'functional genomics'.

But how do we study a living cell? How can we gather and analyze experimental data to obtain insight into the functioning of living cells? One approach is to create interventions (perturbations) in certain components in the cell, for instance to reduce the activity of a certain gene, and then measure the effect of that intervention on the other processes in the cell. In this way one can assess the importance of the component and learn something about its function in the context of the other processes in the cell. This information can only be obtained by studying the cell as a whole and not by the study of individual components.

In my thesis I describe several methods that I have developed during my doctoral research, based on the approach of making perturbations in cellular components and processes, followed by measuring their effects on all other cellular components and processes. These methods rest on a sound theoretical basis and provide recipes for experimentation aimed at optimal extraction of biological knowledge.

Metabolism is the total of biochemical reactions in which metabolites are inter-converted inside living cells. These metabolic reactions are catalyzed by enzymes (proteins with a metabolic function). A perturbation in the concentration of an enzyme enables us to measure the importance of the enzyme for the metabolic flux and metabolite concentrations. However, we don't have perfect experimental control on the concentration of the enzymes, since there are many processes inside the cell that determine the enzyme concentrations as well; for example, some metabolites inhibit or activate the expression of genes coding for enzymes, thereby modifying their concentrations. If this happens, the initial experimental perturbation of the enzyme concentration has a primary effect on metabolism and successively a secondary effect due to the modification of the enzyme concentration through gene-expression regulation by such metabolites. It is interesting to decompose the total effect of the perturbation into these two separate effects, because this enables to asses the relative importance of the metabolic system and gene-expression for metabolic regulation.

The first series of methods I developed in this thesis should enable one to do this experimentally. The main train of thought behind these methods is that metabolic processes generally occur at a much higher rate than processes of gene expression and protein synthesis and that the secondary effects can be neglected on the short timescale. On the short timescale the effects of the perturbation of the enzyme concentration are solely determined by the properties of the metabolic system. On the long run, the effects of gene-expression regulation become important. Another method relies on experimental determinations of metabolic fluxes, concentrations, and the enzyme concentrations. Having such data enables us to separate the primary and secondary effects by a mathematical operation that basically compensates for the secondary changes in the enzyme concentration.

Every cell in our body, with only few exceptions, contains the same genome (the complete collection of hereditary information). Still, our body consists of a large number of different types of cells and tissues with widely different properties and every cell seems to know what it is supposed to do. How the same genome can manifest itself in such a variety of ways is an important subject of current biological research. Certain genes produce proteins that inhibit or activate the activity of other genes, which, in their turn, regulate the activities of yet other genes. In this way, the genome can be visualized as a network of interacting genes, a 'gene network', and the activities of each individual gene depends on the state of the complete network, which in turn is determined by external factors, like hormone concentrations at the location of the cell in the body, nutrients, temperature, etc. The state of the gene network determines the properties of the cell. In this way the properties of our liver cells differ widely from the properties of our brain cells, in spite of them containing the exact same genome. This is due to the different states of their genomes; some genes are more active in liver cells than in brain cells and *vice versa*. For a small number of cases, for a small number of genes, the regulatory systems have been studied in detail, but the global structure of gene networks of all organisms is still unknown. Unraveling the structure of gene networks is an important step in the process of understanding the properties of living cells and therefore also a very active research program.

Recently a break-through on the experimental determination of gene activities has been made. A technology called 'microarrays' enables us to measure simultaneously the

activities of all genes in a genome of a specific cell. Using this technology it can thus be measured how differently genomes are manifested in different types of cells. In my thesis I develop a method to use this kind of experimental data in order to unravel the structure of gene networks. The method is as follows: first an experiment is carried out in which the activities of all genes in a specific type of cell are measured under 'standard' conditions. In the next experiment the activities of all genes are measured after having experimentally perturbed the activity of one particular gene. Comparing the genome wide activities of the genes provides knowledge on the regulatory properties of the perturbed gene; genes that changed activity after the perturbation are thus regulated by the perturbed gene. Such gene regulation can occur in two different ways. The gene could directly affect another gene, or it could do so indirectly: it first affects the activity of another gene which in turn regulates the activity of the next etc. Deducing the structure of gene networks thus boils down to the ability to distinguish direct from indirect effects. If successively identical experiments are executed for all genes in the gene network, data is obtained for which in this thesis I showed that the direct and indirect effects can be distinguished using a simple mathematical operation. This method thus provides a recipe for the experimental setup and the theory to extract the information to decipher gene networks. Because all genes in the gene network have to be perturbed in individual experiments, it seems that an enormous amount of data is required for this method. This is true, but methods that have been proposed previously for the same purpose require even more experimental data, varying from ten times as much to more than thousand times as much as the method proposed in this thesis. The method worked out in this thesis indicates the minimal data requirement for gene network inference without assumptions about the specific network structure.

In summary, whether metabolism or genetic processes are studied, insight into living cells can be obtained by making specific perturbations, measuring their effects and by applying mathematical analysis. In this thesis I described how metabolism can be studied by quantification of the importance of the primary metabolic effects with respect to the secondary genetic effects after a perturbation of a metabolic process. I also described a method that in addition to a theoretical framework provides a recipe for a specific experimental setup that enables deciphering gene networks. All these methods will provide more insight in the functioning of living cells.

Resumen

Descifrando redes vivas

Estrategias de perturbación para genómica funcional

Los seres vivos son muy complicados. Nosotros, los humanos, por ejemplo, estamos constituidos por cerca de 10^{14} (¡un 1 seguido por 14 ceros!) células. Cada célula contiene entre 30 mil y 100 mil genes diferentes (nadie sabe cuantos exactamente), ciertamente mas de 100 mil proteínas diferentes y miles de moléculas pequeñas llamadas metabolitos. Nuestros genes, la información hereditaria, producen proteínas, las moléculas que actualmente incitan los procesos fisiológicos celulares. Metabolitos son substancias que recibimos a través de la comida, de la cual extraemos energía y hacemos otros metabolitos con funciones importantes. Por ejemplo, ciertos metabolitos determinan nuestra apariencia; el color de nuestros ojos y piel esta determinado por la presencia o ausencia de un cierto metabolito. El funcionamiento de las células depende de la interacción coordinada entre todas esas moléculas distintas.

Biología es la ciencia que trata del entendimiento de los seres vivos. Dado que las células son muy complicadas, es lógica la decisión de descomponerlas y estudiar las moléculas y procesos individualmente, por una prueba de tubo de ensayo. Esta forma de investigación biológica ha sido practicada con gran éxito en los últimos 3 decenios bajo el nombre de Biología Molecular. La filosofía subyacente de esta investigación es que si todos los componentes de la célula están estudiados y entendidos, entonces el todo, la célula viviente, está entendido. Probablemente esto no es el caso; el todo es más que la suma de sus partes. No obstante, este estudio ha sido el mas lógico porque, hasta recientemente, no era posible obtener datos experimentales de todos los componentes de las células vivientes simultáneamente.

Esta habilidad existe en estos tiempos. Recientemente, se han desarrollado diversas técnicas experimentales permitiéndonos medir simultáneamente todas las actividades de los genes, concentraciones celulares de grandes números de proteínas, así como de los metabolitos. Los datos experimentales obtenidos a través de estas técnicas nos proveen, por primera vez en la historia de la biología, con los medios para estudiar todos los componentes y procesos simultáneamente. 'Genómica funcional' es la disciplina científica que su ocupa con obtener e interpretar largas cuantidades de datos experimentales biológicos.

¿Pero cómo investigamos una célula viviente? ¿Cómo obtenemos y analizamos datos experimentales que nos ayude a entender el funcionamiento de las células vivientes? Uno de estos estudios es la creación de perturbaciones en ciertos componentes en la célula, por ejemplo reducir la actividad de un cierto gen y después medir el efecto de esa intervención en los otros procesos en la célula. Esta información solo puede ser obtenida por el estudio de la célula como un todo y no por el estudio de componentes individuales.

En mi tesis describo diversos métodos que he desarrollado durante mi investigación doctoral, basados en el estudio de hacer intervenciones en los componentes celulares y sus procesos, y midiendo sus efectos en todos los otros componentes celulares y

sus procesos. Estos métodos se basan en una base teórica sonada y provee instrucciones para experimentos con el fin de poder extraer óptimamente conocimientos biológicos.

El metabolismo es el total de las reacciones bioquímicas en las que metabolitos son convertidos dentro de las células vivientes. Estas reacciones metabólicas son catalizadas por enzimas (proteínas con una función metabólica). Una perturbación en la concentración de una enzima nos ayuda a medir la importancia de la enzima por el flujo metabólico y concentraciones metabólicas. Sin embargo no tenemos un perfecto control experimental en la concentración de las enzimas, ya que hay muchos procesos celulares que determinan esas concentraciones; por ejemplo hay metabolitos que inhiben o activan la expresión del código genético por la enzima, así modificando su concentración. En este caso, la perturbación inicial tiene un efecto primario en el metabolismo y sucesivamente un efecto secundario debido a la modificación de la expresión genética. Es interesante descomponer el efecto total de la perturbación en estos dos efectos separados porque esto nos ayuda a obtener la importancia relativa del sistema metabólico y la expresión genética a través de la regulación metabólica.

Varios métodos que he desarrollado en esta tesis deben ayudar a hacer esto experimentalmente. La idea principal detrás de estos métodos es que los procesos metabólicos generalmente ocurren en una velocidad mucha más alta que la de los procesos de expresión genética y la síntesis de proteínas, y que los efectos secundarios pueden ser ignorados en una corta escala de tiempo. En la corta escala de tiempo los efectos de la perturbación de la concentración de la enzima son determinados solamente por las propiedades del sistema metabólico. En la larga escala de tiempo los efectos de la expresión genética comienzan a ser importantes. Otro método necesita determinaciones experimentales de los flujos metabólicos, concentraciones metabólicas y de las concentraciones encímales. Esos datos nos permitan a separar efectos primarios y secundarios por medio de una operación matemática, que compensa por los cambios secundarios en la concentración de la enzima.

Cada célula en nuestro cuerpo, con algunas cuantas excepciones, contiene el mismo genoma (toda la información hereditaria). Aun así, nuestro cuerpo consiste en un gran número de diferentes tipos de células y tejidos con propiedades muy distintos y cada célula parece saber que es lo qué se supone que debe hacer. Como el mismo genoma puede manifestarse el mismo en tan variadas formas es un tema importante en la investigación biológica actual. Ciertos genes producen proteínas que inhiben o activan la actividad de otros genes, los cuales a su vez, regulan las actividades de otros genes. De esta manera, el genoma puede ser visualizado como una 'red de genes' interactuando y las actividades de cada gen individual dependen del estado de la red completa, la cual depende también de factores externos como concentraciones hormonales en donde se encuentre la célula en el cuerpo, nutrientes, temperatura, etc. El estado de la red de genes determina las propiedades de la célula. De esta manera, las propiedades de las células de nuestro hígado, difieren en gran proporción de las propiedades de las células de nuestro cerebro, a pesar de que contienen exactamente el mismo genoma. Esto es debido a los diferentes estados de sus genomas; algunos genes son más activos en las células del hígado que en las células del cerebro y *viceversa*. Para un pequeño número de casos, para un pequeño número de genes, los sistemas regulatorios han sido estudiados detalladamente, pero la estructura global de las redes genéticas de todos los organismos es aun desconocida. Desenredar la estructura de

las redes genéticas es un paso importante para el entendimiento de las propiedades de las células vivientes y así también un programa de investigación bastante activo.

Recientemente, se ha dado un paso adelante a la determinación experimental de las actividades genéticas. Una tecnología llamada 'microarrays' nos ayuda a medir simultáneamente las actividades de todos los genes en un genoma de una célula específica. Usando esta tecnología se puede medir qué tan diferentemente es manifestado el genoma en distintos tipos de células. En mi tesis desarrollo un método para utilizar este tipo de datos experimentales con el fin de determinar la estructura de las redes genéticas. El método es como sigue: primero un experimento es llevado a cabo en el cual las actividades de todos los genes en un tipo de célula específica son medidos bajo condiciones 'estándar'. En el siguiente experimento, las actividades de todos los genes son medidas después de haber perturbado experimentalmente la actividad de un gen particular. Comparando las variadas actividades de los genes del genoma se obtienen conocimientos de las propiedades regulatorias del gen perturbado; los genes que cambiaron su actividad después de la perturbación son entonces regulados por el gen perturbado. Esa regulación puede ocurrir de dos diferentes maneras. Ya sea que el gen afecta directamente a otro gen, o lo hace indirectamente, por lo tanto el gen primero afecta la actividad de otro gen el cual a su vez regula la actividad del siguiente etc. Desenredar las redes genéticas equivale a distinguir los efectos directos de los indirectos. Si experimentos idénticos son llevados a cabo exitosamente para todos los genes en la red genética se obtienen datos, los cuales en esta tesis mostré que usando una simple operación matemática los efectos directos e indirectos pueden ser distinguidos. Así este método provee de una receta para experimentos específicos y la teoría de extraer información para poder descifrar las redes genéticas. Porque todos los genes en la red genética tienen que ser perturbados en experimentos individuales, parece que una enorme cantidad de datos es requerida para este método. Esto es verdad, pero los métodos que han sido propuestos previamente por la misma razón requieren aun más datos experimentales, variando desde diez veces más hasta miles de veces más que el método propuesto en esta tesis. El método que se utilizo en esta tesis indica que el requerimiento de datos mínimos para deducir una red genética sin hacer suposiciones acerca de la estructura especifica de la red.

En resumen, que si estamos investigando del metabolismo o los procesos genéticos, comprensión de las células vivientes puede llevarse a cabo haciendo perturbaciones específicas, midiendo los efectos y aplicando análisis matemático. En esta tesis, describo como el metabolismo puede ser estudiado por la cuantificación de la importancia de los efectos metabólicos primarios con respecto a los efectos genéticos secundarios después de una perturbación a los procesos metabólicos. También describo un método, que en adición a un marco teórico, provee una propuesta para hacer experimentos específicos, que ayuda a descifrar redes genéticas. Todos estos métodos proveerán mayor comprensión en el funcionamiento de las células vivientes.

Samenvatting

Het ontcijferen van levende netwerken

Verstorings-strategieën voor functionele genomica

Levende wezens zijn heel erg gecompliceerd. Wij mensen, bijvoorbeeld, bestaan uit ongeveer 10^{14} (da's een 1 met 14 nullen!) cellen. Elke cel bevat tussen de 30.000 tot 100.000 verschillende genen (niemand weet precies hoeveel), welzeker meer dan 100.000 verschillende eiwitten en duizenden verschillende kleine moleculen, genaamd metabolieten. Onze genen, het erfelijk materiaal, produceren eiwitten, de moleculen die daadwerkelijk celfysiologische processen besturen. Metabolieten zijn stoffen die we binnen krijgen door ons voedsel, waar we energie aan onttrekken en andere metabolieten van maken met belangrijke functies. Bijvoorbeeld, sommige metabolieten bepalen ons uiterlijk; de kleur van onze ogen en huid wordt bepaald door de aan- of afwezigheid van een bepaald metaboliet. De werking van levende cellen berust op een gecoördineerde samenwerking tussen al deze verschillende moleculen.

Biologie is de wetenschap die zich bezighoudt met het bestuderen van levende wezens. Aangezien cellen zo gecompliceerd zijn is het een logisch besluit om ze uit elkaar te halen en moleculen of processen individueel gedetailleerd te bestuderen in bijvoorbeeld reageerbuisjes. Deze vorm van biologisch onderzoek heeft de laatste 3 decennia zeer succesvol plaatsgevonden onder de naam Moleculaire Biologie. De filosofie achter deze werkwijze is dat als alle componenten van de levende cel zijn bestudeerd en begrepen, dan is het geheel, de levende cel, begrepen. Dit is waarschijnlijk niet het geval; het geheel is groter dan de som der delen. Toch was deze aanpak de meest logische aangezien tot voor kort niet de mogelijkheid bestond om experimentele gegevens te vergaren simultaan voor alle componenten in de levende cel.

Deze mogelijkheid bestaat nu wel. Recentelijk zijn experimentele technieken ontwikkeld die ons in staat stellen om de activiteiten van grote hoeveelheden genen, en concentraties van eiwitten en metabolieten te meten. De experimentele gegevens verkregen door middel van deze technieken stellen ons dus in staat om, voor het eerst in de geschiedenis van de biologie, alle componenten en processen in levende cellen simultaan te bestuderen. Het vakgebied 'funcionele genomica' houdt zich bezig met de productie en interpretatie van grote hoeveelheden biologische gegevens.

Maar hoe bestudeer men nou zo'n levende cel? Hoe kunnen we experimenten doen en analyseren die inzicht geven in het functioneren van levende cellen? Een manier is om verstoringen (perturbaties) aan te brengen in bepaalde componenten in de cel, bijvoorbeeld het reduceren van de activiteit van een bepaald gen, en dan de effecten die deze verstoring heeft op de andere processen in de cel te meten. Op deze manier kan men de belangrijkheid van de component evalueren en iets leren over zijn functie in samenhang met de rest van de processen in de cel. Deze informatie kan dus alleen verkregen worden door de cel als geheel te bestuderen en niet door middel van het bestuderen van individuele componenten.

In mijn proefschrift beschrijf ik methoden die ik heb ontwikkeld tijdens mijn promotieonderzoek, gebaseerd op de aanpak van verstoringen van specifieke cellulaire componenten en processen en metingen van de effecten daarvan op alle cellulaire componenten en processen. Deze methoden zijn voorzien een grondige theoretische fundering en leveren een recept voor het uitvoeren van experimenten om maximale biologische kennis te kunnen onttrekken.

Metabolisme is het geheel van biochemische reacties waar metabolieten in elkaar worden omgezet binnenin levende cellen. Deze metabole reacties worden gekatalyseerd door enzymen (eiwitten met een metabole functie). Verstoringen in de concentratie van een enzym stelt ons instaat om te meten hoe belangrijk het enzym is voor de metabole flux en voor de concentraties van de metaboliten. Enzym concentraties staan echter niet perfect onder onze experimentele controle staan, aangezien veel processen in levende cellen ze beinvloeden; bijvoorbeeld, sommige metabolieten kunnen genen remmen of activeren die coderen voor enzymen en zo de enzymconcentraties beïnvloeden. Op deze manier heeft de perturbatie van het enzym een direct effect op het metabolisme en dan, door middel van het effect van de metabolieten op de concentratie van het enzym, ook een secundair effect. Het is interessant om het totale effect van de enzym-perturbatie te ontleden in deze twee individuele effecten, want dat laat zien hoe belangrijk het metabole systeem is voor de regulatie van metabolisme en hoe belangrijk gen-expressie is voor de regulatie van metabolisme.

De eerste reeks methoden die ik heb ontwikkeld in dit proefschrift stelt ons in staat om dit experimenteel te doen. De algemene gedachtegang is dat metabole processen in het algemeen veel sneller verlopen dan de processen als gen-expressie en eiwit synthese en daarom kan op de korte duur het secundaire effect worden genegeerd. Op korte tijdschaal worden de effecten van de enzym perturbatie dan geheel bepaald door de eigenschappen van het metabole systeem. Op de lange duur beginnen de gene expressie effecten een rol te spelen. Een andere methode berust op het experimenteel bepalen van de metabole fluxen, de concentraties van metaboliten en enzymen. Deze gegevens stellen ons in staat om de primaire en secundaire effecten te scheiden, door wiskundig te corrigeren voor de secundaire veranderingen in de enzym concentraties.

Met slechts enkele uitzonderingen bevat elke cel in ons lichaam hetzelfde genoom (=het gehele erfelijke materiaal). Toch bestaat ons lichaam uit een grote hoeveelheid verschillende cellen en weefsels met heel verschillende eigenschappen en elke cel blijkt te weten wat hij geacht wordt te doen. Hoe hetzelfde genoom zich kan uiten op zulke verschillende manieren is een belangrijk onderwerp van hedendaags biologisch onderzoek. Sommige genen produceren eiwitten die de activiteit van andere genen remmen of activeren, die op hun beurt weer de activiteit van andere genen reguleren. Op deze manier kan het genoom worden gezien als een netwerk van interagerende genen, een ‘gen-netwerk’, en de individuele activiteit van elk gen hangt af van de toestand van het gehele netwerk, welke naast de interne regulatie ook wordt bepaald door omgevingsfactoren, zoals de concentraties van hormonen op de locatie van de cel in het lichaam, voedingsstoffen en temperatuur. De toestand van het genoom bepaaldt de eigenschappen van de cel.

Op deze manier hebben onze levercellen eigenschappen die sterk verschillen van die van onze hersencellen. Ondanks dat hun genoom identieke informatie bevat, verschilt de toestand van hun genoom: sommige genen zijn meer actief in levercellen dan in

hersencellen en *vice versa*. Nu is er voor een klein aantal genen in detail onderzocht hoe ze worden gereguleerd, maar de totale structuur van deze gen-netwerken is onbekend voor alle organismen. Het ontrafelen van de structuren van gen-netwerken is een belangrijke stap in de richting van het begrijpen van de eigenschappen van levende cellen en is dus ook deel van een actief onderzoeksprogramma.

Vrij recentelijk heeft er een doorbraak plaatsgevonden op het gebied van de experimentele bepaling van genactiviteiten. De zogenaamde 'microarray' technologie stelt in staat om in één experiment alle genactiviteiten in bepaald celtype te meten. Op deze manier kan dus gevolgd worden hoe verschillend het genoom wordt geuit in verschillende types cel. In mijn proefschrift beschrijf ik een methode die ik heb ontwikkeld om dit soort experimentele gegevens te gebruiken om de structuur van gennetwerken te ontrafelen. Het principe van de methode is als volgt: in het eerste experiment worden de genactiviteiten in een type cel onder 'standaard' condities gemeten. In het volgende experiment worden de genactiviteiten gemeten nadat de activiteit van een bepaald gen experimenteel is verstoord. Vergelijking van de genoomactiviteit tussen deze twee condities geeft inzicht in de regulatieeigenschappen van het gen wiens activiteit experimenteel was verstoord; genen wiens activiteit is veranderd als gevolg van de experimentele verstoring worden gereguleerd door het verstoorde gen. Zulke regulatie kan op twee verschillende manieren verlopen: een gen kan een ander gen direct reguleren, of indirect, waar het gen eerst een gen reguleert dat op zijn beurt weer een ander gen reguleert etc. Het ontrafelen van de structuur van een gen-netwerk komt dus neer op het onderscheiden van deze directe regulatie van de indirecte regulatie. Als vervolgens identieke experimenten worden uitgevoerd voor elk gen in het genoom, dan worden gegevens verkregen waarvan ik in dit proefschrift laat zien dat ze met behulp van een eenvoudige wiskundige bewerking de directe van de indirecte effecten kunnen doen onderscheiden. Deze methode geeft dus aan wat voor soort experimenten nodig zijn en geeft de theorie om de informatie te verzamelen die nodig is voor het ontrafelen van gennetwerken. Aangezien elk gen in het netwerk in een individueel experiment moet worden verstoord, lijkt het dat deze methode een enorme hoeveelheid experimenten nodig heeft. Dit is juist, maar de methoden die eerder voor dit doel zijn voorgesteld in de wetenschappelijke literatuur hebben nog veel meer experimentele gegevens nodig, variërend van tientallen malen zoveel tot meer dan duizenden maal zoveel als de methode ontwikkeld in dit proefschrift. De methode in dit proefschrift geeft de theoretisch minimale hoeveelheid gegevens die nodig is om gennetwerken te kunnen ontrafelen zonder aannames te maken over hun specifieke structuur.

Samenvattend, of nu metabolisme of gen-expressie processen worden bestudeerd, inzicht in levende cellen kan verkregen worden door specifieke verstoringen aan te brengen, metingen van de effecten te verrichten, en vervolgens wiskundige analyse toe te passen. In dit proefschrift heb ik beschreven hoe metabolisme bestudeerd kan worden door te kwantificeren hoe belangrijk de primaire metabole effecten zijn ten opzichte van de secundaire gen-expressie effecten na een verstoring van een metabool proces. Ook beschreef ik een methode die voorziet in zowel een specifiek experimentele als een theoretische opzet met als doel de structuur van gennetwerken te ontrafelen. Al deze methodes zullen meer inzicht bieden in hoe levende cellen functioneren.

Dankwoord

Dankwoord

The six year USA adventure is over. It was quite an experience. My life has changed a lot during this period. I left home with one suitcase and my bicycle, I return now with a family (I sold the bike). The first 4 months I was in the 'land of enchantment' New Mexico working at the National Center for Genome Resources in Santa Fé. After that I worked at the Virginia Bioinformatics Institute in Blacksburg, Virginia. It was a rough transition from NM to VA, but soon the state-slogan 'Virginia is for lovers' made itself true and that was good! I have gained a lot of scientific knowledge; I have become much less naïve than I was 6 years ago. However, now that I know more, I realize much better how naïve I truly am. Here I would like to express my thanks to everybody that really made these past six years such an enjoyable time.

Most people know either a little bit about many things, or a lot about a small number of things. But there are some people that know a lot about many things! I had the privilege and pleasure to do my research under supervision of such a person of the third kind: **Pedro Mendes**. Dear Pedro, thank you very much for your sublime mentoring and friendship during these years. I have learned so much from you! I also would like to express thanks to **Angela Mendes**. Dear Angela, thank you very much for your friendship, care and concern during these years. Muitíssimo obrigado! Grote dank aan **Hans Westerhoff** and **Jacky Snoep**. Jullie twee hebben het mogelijk gemaakt voor mij om te kunnen promoveren via het ingenieuze plan om mij als jullie student uit te lenen aan Pedro! We hebben elkaar weinig gezien de afgelopen 6 jaar, maar toch heb ik over het algemeen veel wetenschappelijke input van jullie gekregen. Ook **Jeannet Wijker** ben ik heel dankbaar. Dank je wel, Jeannet, zonder jou had dit ook niet mogelijk geweest. Zonder jou had namelijk niemand ooit geweten dat ik daadwerkelijk een student aan de VU ben. Thanks to Pedro's group members, the BNMers (Biochemical Network Modelers). Foremost, I would like to thank **Stefan Hoops** who helped me a lot with programming issues in C++. Stefan, whether I came to you with a worthy question or a really simple stupid little issue, you always took the time and patience to explain things to me. Also, I want to thank **Aejaaz Kamal** in that respect. Aejaaz, you were the first one I would bother if my code was doing something funny or didn't run at all, mostly for some really trivial reason (poor Aejaaz, his cubicle was right next to mine). Thanks to the members of the Portuguese conquista at VBI, **Diogo Camacho** and **Ana Martins**. Diogo, though our collaboration didn't go always so smoothly, in general I enjoyed it very much and eventually it resulted in a very nice paper on correlations in metabolomics data. Not to mention all the conference posters on this work! You should contact Guinness book of records! Ana, I want to thank you for….well, just for being Ana!! I thank **Paul Braznik** for drawing several really excellent figures of gene networks for our papers. There are so many people at VBI I want to thank! Thanks to **Dustin Machi**, for being such a joyful 'smoke partner' these years! Too bad I am too stingy to pay for a full color cover otherwise I would have used one of your great 3-D drawings. I'll use one for my next book! I thank **Ina Hoeschele** for giving my first pre-graduation post-doc job. I very much enjoyed learning about statistical genetics the past two years! Also to Ina's student and my close colleague, **Bing Liu**, always so friendly accompanying me on my smoke-breaks and answering my many questions about statistics: thanks! I really enjoy working with the both of you! I also thank Ina's former students **Nan Bing** and **David Henderson**. David was my first collaborator ever and we did some nice work on

factor analysis of gene expression data. The collaboration with Nan resulted in a really nice publication about partial correlation analysis of gene expression data. Thanks to **SalsaTech** for organizing salsa dancing on Wednesday nights from 9.30pm to 11.30pm, yippy! Blacksburg party town….Van mijn (ex)medestudenten van Hans' kant, ken ik vrijwel alleen **Frank Bruggeman** en **Rogier Stuger**. Frank, onze discussies over allerlei zaken hebben mij veel inzicht opgeleverd. Een aantal zaken in mijn proefschrift zijn direct gebaseerd op zulk inzicht. Ik hoop dat we in de toekomst nog veel discussies hebben en ook onze samenwerking voortzetten (ook al moeten we toch wel eens gaan denken om iets sneller onze artikelen af te krijgen!) Rogier, dankzij jou ben ik niet het zwarte schaap van Hans' groep! Bedankt! Also thanks to **Peter Jackson** for making the Lord of the Rings movie trilogy. Thanks to **Valerie Robinson** and **Edwin Robinson**. Val and Ed, thanks for your friendship during these years! I'll think about you whenever I'll hear the sound of the 'gaita', whenever I am clipping the nails of a sheep, and whenever I'll have a rich Thanksgiving meal! I thank my good friend **Dhaval Mackhecha**. Dhaval, I very much enjoyed the work we did together. I know that it must have been challenging for an aerospace-engineer to do some biology! It is really nice to see how the work we started as our little 'Sunday mornings at the coffee shop hobby project' now just got published! Also thanks to **Ralf Steuer,** for very useful conversations about a lot of interesting scientific topics, almost on a daily basis (thanks to e-messengers!). We have very nice ideas, the most brilliant plans I'd say…too bad we are both too busy (or should I say too lazy?) to work them out!

¡Gracias, mis suegros, **Donmi** y **Lety**, para darme la mano de tu preciosa hija! Ik dank mijn broers, **Andrés** en **Sergio**. Ik heb jullie moeten missen al deze jaren. Slechts twee weken per jaar waren we samen, maar als we dan samen waren dan was het altijd alsof er geen 11,5 maanden tussen had gezeten! Ok, we chatten dan ook bijna dagelijks. Andrés, ik heb wel al die Albariños moeten missen, maar we maken het deze zomer goed! Sergio, we moeten het nog effe hebben over Fuentomics!! Weinig woorden, maar uit het diepste van mijn hart: **papa** en **mama**, aan jullie heb ik alles te danken. Jullie liefde en aanmoediging is de grootste reden voor hetgeen dat ik nu bereikt heb. Ik hoop dat mijn volgende baan iets dichterbij is zodat we vaker samen kunnen zijn. **Gaby**, mi amor, gracias por todo. Los últimos 4 años fueron los más felices de mi vida. Gracias por tu paciencia conmigo, por darme tanto amor y por darme el regalo más bonito y hermoso del mundo...Albertito!! **Alberto Xesús**, mijn kleine schatje, jij bent het allermooiste wat mij ooit is overkomen in mijn leven. Jouw geboorte heeft mijn proefschrift enkele jaren vertraagd, want spelen en knuffelen met jou is natuurlijk een veel leukere bezigheid dan een proefschrift schrijven. Van jou heb ik iets geleerd wat niet geleerd kan worden uit boeken, experimenten, modellen, algorithmen of theorieën: 'de zin van het leven'. ☺