CWI

Modeling and identification of a gene regulatory network programming erythropoiesis (1)

D.L. De Vos

MAS-E0910

Centrum Wiskunde & Informatica

Centrum Wiskunde & Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Science Park 123, 1098 XG Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3703

Modeling and Identification of a Gene Regulatory Network Programming Erythropoiesis (1)

Dirk De Vos

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
dirk.de.vos@cwi.nl

December 31, 2009

The development of mature blood cells of distinct lineages from the hematopoietic stem cells (hematopoiesis) involves a progressive restriction of differentiation potential and the establishment of lineage-specic gene expression profiles. The establishment of these profiles relies on lineage-specic transcription factors to modulate the expression of their target genes. This work is embedded in a wider ErasmusMC/CWI collaboration that develops the informatics and mathematics to underpin studies on gene expression regulation by mapping and analyzing the regulatory pathways and networks of transcription factors that control cellular functions (so called 'Gene Regulatory Networks' or 'GRNs'). This project is concerned with the mathematical part and concentrates on a GRN central to erythropoiesis. Among the many housekeeping and tissue-specic genes involved in the differentiation and the commitment of hematopoietic stem cells to erythrocytes (erythropoiesis), we focus on a small pool of genes (Gata-1, Gata-2, Pu.1, EKLF, FOG-1, α/β-globin) known to be critically involved in an intricate but well-less investigated regulatory circuit. Based on the regulatory interactions in the GRN we have developed models in the form of a system to account for the dynamics of gene expression and regulation involved in this process. Because of the lack of information about a signicant number of model parameters, our focus is on system identication. In this first report some preliminary results are presented based on synthetic data. However, time series of the levels of all relevant mRNAs are available from micro-array analysis of G1E cells, a murine cell line which recapitulates erythropoiesis. In the follow-up report a detailed account will be given of the parameter estimation and identifiability analysis with respect to these data. This will eventually allow for a thorough evaluation of the role of various characterized as well as hypothetical regulatory mechanisms.

In depth characterization of the necessary expression patterns and gene regulatory interactions responsible for the the set of commitments all along the erythroid lineage is essential to gain fundamental insight into the behaviour of these complex networks and to design further experiments. Ultimately, this may lead to ways to rescue erythroid differentiation in several anemic diseases.

AMS Subject Classification (2000): 93A30, 93B30, 93C10, 92C15, 92C37, 92C50.
Keywords and Phrases: Modeling, S-systems, identifiability, parameter estimation, Cell Biology, Gene Regulatory Network, erythropoiesis,.

Note: This research is financed by NBIC as project SP2.3.2 of the BioRange Program.

1 Introduction

The aim of this section is twofold. First, to give the necessary background information for what is presented in section 2. Second, to provide some insight on what will be the subject of the follow-up paper (part (2)), *i.e.* identifiability analysis and parameter estimation using real experimental data.

1.1 The investigation

Hematopoiesis is the formation of different types of mature blood cells starting from multipotent stem cells going through distinct intermediate stages. During this process the cells become more and more committed. This is called 'lineage specification'. Our investigation is concerned with erythropoiesis or red blood cell formation. Red blood cells or erythrocytes represent the bulk of the blood cells. In humans every day around 10^{11} to 10^{12} new erythrocytes are produced. subsequently they cycle for approximately 120 days in the blood stream and are finally degraded ([49] and references within). The maturation takes approximately one week. As the stem cells mature, they go though different stages (including several progenitor and erythroblast stages) finally extruding their nucleus as they slowly fill with hemoglobin until they are bright red reticulocytes ready to escape the bone marrow and squeeze into the blood capillaries to begin circulating around the body. Within a few days, the reticulocytes completely lose all their nuclear material and become full-fledged erythrocytes that are ready to serve the oxygen needs of the body. Hemoglobin (consisting of α and β globin chains) is their most important protein since it binds or releases the vital oxygen (depending on the conditions), and the most abundant protein as well (95% of its protein content) . Not surprisingly it plays an important role in our study. The development of mature blood cells of distinct lineages from the hematopoietic stem cells involves the establishment of lineage-specic gene expression profiles. The establishment of these profiles relies on lineage-specic transcription factors (TFs) to modulate the expression of their target genes (like hemoglobin). Our objective is to provide an in depth characterization of the regulatory interactions responsible for the set of commitments all along the erythroid lineage, based on a mathematical model. Since the GATA-1 transcription factor is known to be central in this process our models are centered on this important protein [17]. Our experimental model is the house mouse (*Mus musculus*), which is highly similar to humans with regard to erythropoiesis. This work is expected to give new insight into the regulation of stem cell differentiation, which is also an essential part of understanding the deregulation that occurs in carcinogenesis. One general question is whether any external input is needed to predict the time-dependent expression patterns obtained through experimentation. A more specific question is how the dynamics of expression of GATA transcription factors (GATA-1 and GATA-2 in this case) may be more important for the regulation of erythropoiesis than their identity [18]. Our models are also expected to be an aid in experimental design, suggesting new experiments to better characterize the regulatory interactions involved in the GATA-1 network. Finally, it may ultimately aid in finding ways to rescue erythroid differentiation in several anemic diseases.

1.2 Modeling and identification of biochemical reaction networks

1.2.1 Modeling

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. The main steps of in this process are transcription and translation. Transcription refers to the copying of information contained in the DNA (more precisely the genes) to the so-called messenger RNA molecules (mRNA). Then, the protein-coding part of the mRNA is decoded into a peptide sequence (protein) during translation. In fact, typically, in eukaryotic cells after transcription the so-called pre-mRNA has to undergo several step including splicing, modification and transport over the nuclear membrane (translocation). Importantly, degradation of mRNA and protein, and post-translational modification are also part of the gene expression cascade.

The regulatory relationships between genes and their products are described by gene regulatory networks (GRNs). In practice, GRNs usually refer to transcription factor networks. Similarly, metabolic networks and signalling networks are defined. GRNs are usually concerned with the control of transcription, *i.e.* how genes are up and down regulated in response to signals. It was in the 1960s that experiments demonstrated the presence of regulatory sequences in the proximity of genes and the existence of proteins that are able to bind to those elements and to control the activity of genes by either activation or repression of transcription [29]. These regulatory proteins are themselves encoded by genes. This allows the formation of complex regulatory networks, including positive and negative feedback loops. During the years some models have been produced, however with the emergence of systems biology and high-throughput and genome-scale technologies, GRN modeling has increased its pace leading to a number of publications. Since they have been reviewed extensively elsewhere (for example [6, 12, 50, 53, 58]), we will not further dwell on this. Noteworthy, some reports are published on GRNs involved in hematopoiesis (*e.g.* [8, 32]). Moreover, a few studies on erythropoiesis have emerged [5, 27, 48]. A simple mathematical model for the PU.1–GATA-1 switch which did not focus on molecular details of the interaction, but served to elucidate bifurcation dynamics involved in lineage choice was reported by Huang *et al.* [27]. Roeder and Glauche [48] introduced a more detailed mathematical model for the PU.1–GATA-1 switch based on the then available biochemical information. However, in both the aforementioned models of the gene switch, the assumption of high cooperativity at the PU.1 and GATA-1 promoters was needed to obtain the desired bistable behaviour. In comparison to the previous approaches, the model of Bokes *et al.* [5] exhibits bistability even if this assumption is relaxed. However, these models are still too simple to predict the gene expression patterns recently obtained through microarray experiments.

In may biochemical systems space and the discrete nature of reactants play an important role. For fast reactions diffusion can be a limiting factor since typical cell dimensions are large compared to the molecules' size. Furthermore, the number of molecules involved can be low which is an additional source of stochasticity [14]. Possible approaches are discussed in [14, 42, 54]. To avoid over-complicating and because of the high computational burden these approaches were not adopted. We chose a ODE approach in which the system is continuous and deterministic. Still, different types of system were considered. Finally, to make the model suitable to produce reliable predictions, precise estimates of key parameters are required. How this can be done will be discussed in the next section.

1.2.2 Identifiability

The notion of identifiability of systems is fundamentally a problem of uniqueness of solutions for specific attributes of certain classes of mathematical models. The identifiability problem usually has meaning in the context of unknown parameters of the model. It is clearly a critical aspect of the modeling process, especially when the parameters are analogs of physical attributes of interest and the model is needed to quantify them [11].

We can state the identifiability problem as follows. *A parametrization of a subclass of dynamic systems will be called identiable if for any finite but sufficiently long time series of observed input-output trajectories there exists an unique element in the subclass of systems which represents those observations. In mathematical terms, if for a fixed input function the map from the parameters to the output is injective* (Jan van Schuppen, personal communication).

Verifying identifiability in principle preceeds determination of numerical values of the parameters. Methods for linear systems are well-described within the framework of linear time-invariant system theory. In biochemistry state variables (concentrations) typically can only take positive values. Interestingly, van den Hof [57] has developed procedures for the class of positive linear systems. Unfortunately, besides some specific cases (like for example compartmental models for farmacokinetic applications, cf. [11] biochemical reaction models are typically described by non-linear kinetic equations.

Importantly, different concepts of identifiability exist. However, different definitions have been given for the same terms. To avoid confusion we will adhere to the definitions in [46] which, in turn, were adopted from Audoly *et al.* [3]. Then some of the mathematical concepts behind the computational procedures developed for identifiability analysis are presented, including computation of the Fisher information matrix, the covariance and correlation matrices, the confidence intervals and other related statistical measures.

A priori structural identifiability

The subject of *a priori* or structural identifiability analysis is whether the parameters for the mathematical model can be determined assuming that for all variables continuous and error-free data are available. Audoly *et al.* [3] state the identifiability problem as follows.

It is convenient to consider the output y as a function of time and of the observational parameter vector, $\Phi = [\Phi_1, \Phi_2, ..., \Phi_R]$. By definition, the components of Φ, are identifiable, since they can be evaluated from the designed experiment. Thus, each input-output experiment provides a particular value $\hat{\Phi}$ of Φ. The observational parameters $\Phi_i,\ i = 1, \ldots, R$, are algebraic functions of the basic parameters $p_i,\ i = 1, \ldots, P$, which may or may not be identifiable:

$$\Phi \ = \ \Phi(p)$$

In particular:

$$\hat{\Phi} \ = \ \Phi(\hat{p})$$

Considering the state-space formulation of a system of ODEs

$$\dot{x}(p,t) = f(x(p,t),u(t),p), \quad x(0) = x_0, \tag{1}$$
$$y(p,t) = g(x(p,t),u(t),p) \tag{2}$$

where x is the vector of N_x state variables and p is the vector of N_p model parameters. Note that f specifies the model, u the vector of inputs *i.e.* for a particular experiment and y specifies the vector of N_y measured states. An experiment is specified by the intial conditions $x(0)$, the inputs u chosen from among some set of possible inputs U and the observations y.

For the input class U and $p \in \mathbb{C}$ (the complex space), the single parameter p_i is a priori structurally:

- *globally* (or uniquely) identifiable if and only if, for almost any $\hat{p} \in P$, the system:

  $$y(\Phi(p),t) = y(\hat{\Phi},t) \tag{3}$$

  has the only solution $p_i = \hat{p}_i$ (in other words a parameter is globally identifiable if it can be uniquely determined given a particular input function);

- *locally* (or non-uniquely) identifiable if and only if, for almost any $\hat{p} \in P$, the system (1) has for p_i more than one, but finite number of solutions;

- *non-identifiable* if and only if, for almost any $\hat{p} \in P$, the system (3) has for p_i an infinite number of solutions.

The model is *a priori* structurally globally (or uniquely) identifiable, if all of its parameters are globally (or uniquely) identifiable; locally (or non-uniquely) identifiable, if all of its parameters are identifiable –either uniquely or non-uniquely– with at least one parameter non-uniquely identifiable; non-identifiable, if at least one of its parameters is non-identifiable.

Despite several techniques being available, for realistic models it is very difficult to obtain any results for global identifiability of a model. For linear models, the Laplace transform or transfer function approach can be applied [23, 22, 30]. However, when modelling biological systems, non-linear systems are ubiquitous, resulting, for example, from Michaelis Menten and Hill type kinetic equations. For non-linear models the oldest method is the Taylor or power series expansion [44]. Another classic method is the similarity transformation approach, based on the local state isomorphism theorem [15, 43, 56]. In Chappell *et al.* [7], the two latter methods are compared. Both methods have been successfully applied to some specific non-linear structures but they prove intractable for the general case, mainly when the non-linear system increases in size.

Techniques of differential algebra have also been applied to study this problem. Ollivier [41] and Ljung and Glad [35] have first proposed methods based on differential algebra. More recently, a new differential algebra algorithm has been developed by Audoly *et al.* [3] which improves the efficiency of the previous ones and enlarges their applicability domain. A software tool , DAISY [4], is publicly available. Very recently, in [40] an algebraic approach is presented to system identification for the classes of rational and Nash systems. However, although these methods greatly improve a priory identifiability analysis of non-linear models, the construction of an efficient algorithm applicable for the general case remains a difficult task.

A priori local identifiability

The limited applicability of the existing techniques for determining global structural identifiability, taken in conjunction with the need for practical methods, provides a key argument for emphasizing the use of *a priori* local identifiability despite its limitations derived from its local nature.

The output sensitivity functions are central to the evaluation of a priori local identifiability. If the sensitivity functions are linearly dependent, the model is not identifiable, and sensitivity functions that are nearly linearly dependent, result in parameter estimates that are highly correlated.

Zak *et al.* [64] have presented a numerical method for checking *a priori* local identifiability of the parameters at a given point $\hat{p}$ making use of the sensitivity equations (based on the method of Jacquez and Greif [30]).

Consider the system-experiment model described by Eqs. (1) and (2). Taking the values of the parameter set $\hat{p}$ as true values, the $N_y \times N_p$ sensitivity matrices of the measured states S_y are calculated at some large enough number of points N where:

$$S_{y_{ij}} = \left(\frac{\partial y_i}{\partial p_j} \right)_{y=y(t,\hat{p}),p=\hat{p}}$$

The matrix G is then constructed stacking the matrix of sensitivities at those points:

$$G = \begin{bmatrix} S_y(t_1) \\ S_y(t_2) \\ \vdots \\ S_y(t_N) \end{bmatrix}$$

Finally the $N_p \times N_p$ correlation matrix of the parameters (M_c) is calculated:

$$M_c = \text{correlation}(G)$$

Parameters that are locally identifiable have correlations between -1 and $+1$ with all others parameters. Parameters that are not locally identifiable have correlations of exactly $+1$ or -1 with at least one other parameter. That means that these parameters influence the measured variables in exactly the same or exactly the opposite manner. The original parameter set, p, can be reduced to the identifiable parameter set, p_I, of length N_I, by calculating M_c, removing one unidentifiable parameter, recalculating M_c, removing another unidentifiable parameter, etc., until no more unidentifiable parameters remain.

In order to apply the a priori local identifiability analysis proposed by Zak *et al.* [64] to the nominal point, we need to calculate the sensitivity matrices of the measured states at some large enough number of time points. It should be verified that this gives enough points *i.e.* diminishing this sampling time has no significant effect on the resulting correlation matrix. A robust general method for the numerical calculation of local sensitivities is ODESSA developed by Leis and Kramer [33].

Numerous methods have been presented to deal with a priori identifiability analysis of linear and non-linear models. In terms of local identifiability Yao *et al.* [63] propose an algorithm to assess wether individual model parameters will be estimable from existing or proposed experimental data. Farina *et al.* [16] have derived for a particular class of non-linear systems, *i.e.* based on mass-action

kinetics, sufficient conditions for *a priori* local parameter identifiability. With this method the problem of identifiability can be recast as the question of observability of a specific system expansion. Hengl *et al.* [25] present the method of *mean optimal transformations*, a non-parametric bootstrap-based algorithm for identifiability testing, capable of identifying linear and non-linear relations of arbitrarily many parameters, regardless of model size or complexity.

A posteriori or practical identifiability

The question addressed is the following: with the available experimental data, can the parameters be uniquely estimated? Or, in other words, if a small deviation of the parameter set occurs, does this have a great impact on the quality of the fit? Mathematically, this can be formally expressed as the maximization of the so-called Fisher Information Matrix which expressed the information content of the experimental data [34, 39].

$$FIM = \sum_{i=1}^{N} \left(\left. \frac{\partial y}{\partial p} \right|_{t_i} \right)^T Q_i \left(\left. \frac{\partial y}{\partial p} \right|_{t_i} \right)$$

where $\partial y/\partial p$ are the output sensitivity functions at times t_i $(i = 1 \ldots N)$, and Q_i is a square matrix with user-supplied weighting coefficients. The problem of analyzing practical identifiability is similar to that of analyzing *a priori* local identifiability but now the evaluation points of the functions are limited to the experimental data points. If the sensitivity equations show linear dependence at the experimental data points, the covariance matrix becomes singular and the model is not identifiable. A singular FIM indicates the presence of unidentifiable parameters, and correlations between parameters that are greater than 0.99 may lead to a singular FIM.

Similar methods are used in Zak *et al.* [64] for an *in silico* genetic regulatory network. A formal identifiability analysis was performed that considered the accuracy with which the parameters in the network could be estimated using gene expression data and *a priori* structural knowledge as a function of the input perturbation and stochastic gene expression. In their analysis they first determined the set of *a priori* identifiable parameters. Then the Fisher information matrix was used to determine which parameters were practically identifiable, following the methods of Landaw and DiStefano [31] and Delforge *et al.* [13] The Fisher information matrix is also an approximation of the inverse of the parameter estimation error covariance matrix of the best linear unbiased estimator (BLUE):

$$C = FIM^{-1} \tag{4}$$

Useful information about the correlation between estimated parameters can be also obtained from the covariance matrix. The correlation matrix, whose elements are the approximate correlation coefficients between the ith and the jth parameter, is defined by:

$$\begin{aligned} R_{ij} &= \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}, \quad i \neq j, \\ R_{ij} &= 1, \quad i = j. \end{aligned}$$

The correlation matrix measures the interrelationship between the parameters and gives an idea of the compensation effects of changes in the parameter values on the model output. If two parameters

are highly correlated, a change in the model output caused by a change in a model parameter can be (nearly) compensated by an appropriate change in the other parameter value. This prevents the parameters from being uniquely identifiable even if the model output is very sensitive to changes in the individual parameters.

After fitting parameters p to a certain data set, it is desirable to obtain some measure of the quality of the estimates which is related to uncertainty analysis. In principle, the aim is to obtain the probability distribution of the estimates or an adequate characterization of it, e.g. by computing different percentiles of the distribution. However, in most cases, this distribution is not known and it is therefore necessary to obtain approximations of it.

In Marsili-Libelli *et al.* [37], two approximations methods were considered: (i) a local linearisation of the output function, leading to the Fisher information matrix, and (ii) a quadratic expansion of the estimation error functional involving the Hessian matrix. The confidence ellipsoids obtained with the Hessian or the Fisher method coincide only when the estimation converges to the true parameters. Otherwise, they yield clearly different results, indicating an inaccurate estimation.

Assuming that the measurement noise is uncorrelated and normally distributed with zero mean and constant variance, C is also the inverse of the FIM (*cf.* equation (4)). In this case, it represents the error covariance matrix of the minimum variance unbiased estimator according to the Cramer-Rao theorem [34] and the approximate (1-α) confidence ellipsoids can be expressed as:

$$\{p : (p-\hat{p})^T C^{-1} (p-\hat{p}) \leq n_p F^{1-\alpha}_{n_p, N-n_p}\}$$

However, the confidence intervals obtained with the Hessian or the Fisher method are statistically optimistic due to the use of a linear approximation of the non-linear model in the neighborhood of the best parameter estimates [59]. Alternative, more robust techniques such as the jacknife and bootstrap methods produce parameter variances that are more realistic. As a drawback, one should mention that these methods are rather computing intensive. Another way to obtain the true confidence region of the parameters in non-linear models is by a systematic exploration of the objective functional for an extensive number of parameter combinations. This is a computing intensive task as well, because the number of evaluations increases as a power function of the number of parameters.

Finally, it should be noted that *a posteriori* identifiability analysis using the Fisher information matrix approach has been applied to optimal experimentel design (which is in principal an integral part of system identification, *cf.* [19] for an example).

1.2.3 Parameter estimation

The parameter estimation or approximation problem can be stated as follows (Jan van Schuppen, personal communication). *Consider an observed time series and a class of dynamic systems with an identifiable parametrization. Determine a dynamic system in the selected subclass such that the time series generated by that system is close to the observed time series in terms of a selected criterion. In case there is no time series but an estimate of the impulse response function, then the comparison is made between the impulse response function associated with the time series and the impulse reponse function as sociated with the dynamic system.*

Considering a system as in equations (1) and (2), with possibly (non)linear constraints as in

$$c(x(t,p), u(t), p) \geq 0$$

The optimization problem is given by the task to minimiz some measure, $V(p)$, for the discrepancy $e(p)$. The most used measure for the discrepancy is the Euclidean norm or the sum of the squares weighted with the error in the measurement [2]:

$$V_{MLE}(p) = \sum_{i=1}^{N} \frac{(g_i(x(t_i,p),u(t_i),p) - y_i)^2}{\sigma_i^2} = e^T(p)We(p)$$

with g_i the vector of observables and y_i the vector of measurement errors. Under the assumption that the experimental errors are independent and normally distributed with standard deviation σ_i, the least squares estimate $\hat{p}$ of the parameters is the value of p that minimizes the sum of squares:

$$\hat{p} = \arg\min V_{MLE}(p)$$

In recent years, there has been an increase in the number of methods approaching this problem in the biological literature. In [10] an overview is given of recent developments in parameter estimation and structure identification of biochemical and genomic systems. The most prominent search methods for parameter estimation from time series data can be grouped into gradient-based methods, stochastic search algorithms and others, reviewed in [2, 10]. Due to the frequent ill-conditioning and multi-modality of many of these problems, traditional local methods usually fail (unless initialized with very good guesses of the parameter vector). In order to surmount these difficulties, global optimization (GO) methods have been suggested as robust alternatives. These heuristic algorithms cannot garrantee to give the best answer to a given problem.

Currently, deterministic GO methods can not solve problems of realistic size within this class in reasonable computation times. In contrast, certain types of stochastic GO methods have shown promising results, although the computational cost remains large. Rodriguez-Fernandez *et al.* [47, 45] have presented hybrid stochastic-deterministic GO methods. An established way to evaluate the performance of heuristic algorithms is to try them on a sufficient number of realistic test cases. Gennemark and Wedelin have, indeed, recognized the importance of systematic evaluation of ODE identification algorithms and collected more than 40 benchmark problems [21].

A serious bottleneck of modeling biochemical networks has been the lack of sufficient data for parameter estimation. Even for low-dimensional systems the number of parameters is typically in the dozens and grows faster than linearly with the number of variables involved. However, with the advent of large-scale data collection methods the bottleneck has moved more to the computational side. While intensive, and most likely parallelized, computational effort will be unavoidable for pathways of realistic proportions much time can be saved with other approaches, see *e.g.* Voit and Almeida [61]. Gennemark and Wedelin have presented a stable and efficient algorithm based upon decoupling the systems equations [20]. This requires that time course information is known for all state variables. On top of the parameter estimation algorithm they have added a model selection algorithm that picks different tentative model structures (based on specific reaction types and an error function). A metabolic system and a genetic network were presented as examples. Voit *et al.* [62] used S-systems for glycolysis in *Lactococcus lactis*.Their strategy was to use the metabolite time courses as input functions. They are in effect 'off-line' data that enter the system as time-dependent forcing-functions. As a variation the raw data may first be smoothed with a filter or spline [36, 52, 55]. An alternative approach, avoiding the numerical integration of ODEs, is to replace the differentials by estimated slopes using simple linear interpolation, the three-point method [1] (both sensitive to noise) or an artificial neural network (webtool Webmetabol in [61]; [60]). Finally, Chou

et al. applied *alternating regression* to S-systems models, combined with methods for decoupling systems of differential equations [9].

2 Results and Discussion

Central in our approach towards modeling red blood cell differentiation is the concept of the Gene(-tic) Regulatory Network as defined in the section 1.2.1. We focus here on a small pool of genes known to be critically involved in an intricate but well-less investigated regulatory circuit. Our core network consists of the transcription factors Gata-1, Gata-2, Pu.1, EKLF and FOG-1. In addition 1 'target' gene has been included as well: β-globin, which, together with α-globin, forms the constituents of the red cells' most important protein, *i.e.* hemoglobin. Table 1 gives an overview of the regulatory interactions included in our models. Figure 1 schematically depicts the major interactions between the core network's components.

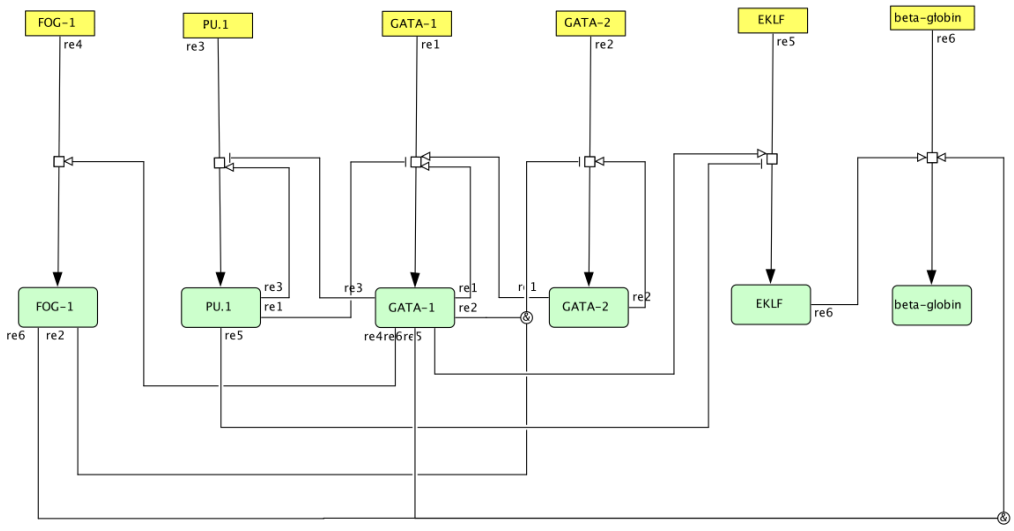

Figure 1: Diagram of the main regulatory interactions between our core GRN genes (yellow boxes) and proteins (green boxes). Activation and repression are depicted by terminal arrows and '–'-signs, resp. The '&'s indicate that proteins act in combination to affect their target.

While working with a limited number of interacting compounds in a core GRN, the underlying biochemical processes are still overly complex. A number of assumptions are required to alleviate this problem. The main processes involved in the gene expression cascade are transcription, mRNA processing and translocation, translation, and mRNA and protein degradation. The processing and transport steps were not included in our models since they are assumed to happen on a much shorter time-scale than the rest. Since transcription, is the major site of gene regulation, the combinatorics of activator and repressor binding to the different promotor regions were detailed the most in the model. The approach by Roeder *et al.* [48] was adopted (*cf.* Appendix). This is based on the assumption that transcription consists of two main steps. The first step (*initiation*) represents the binding of a specific transcription factor (activator) in complex with the DNA-dependant RNA polymerase (RNAP) to the promotor region of a specific gene. This process is considered to be reversible and in chemical equilibrium. It can therefore be characterized by only one equilibrium constant. The next step (*elongation*) is the actual mRNA production step and is assumed to be an irreversible first order process. In the case that the transcription factor that binds the promotor is a

Table 1: Regulatory Interactions in the GATA-1 GRN

Gene	Activators	Repressors
GATA-1	GATA-1, GATA-2, PU.1	(GATA-1:PU.1)
PU.1	GATA-1, PU.1	(GATA-1:PU.1)
GATA-2	GATA-2	(GATA-1:FOG-1)
FOG-1	GATA-1	
EKLF	GATA-1	(GATA-1:PU.1)
β-globin	(GATA-1:FOG-1), EKLF	

repressor of transcription, then the resulting complex is assumed to be unproductive. mRNA and protein degradation, as well as translation are represented by simple first order processes, despite the fact they consist of many processes that are (in some cases) intricately regulated. Post-translational modification and protein sorting are also unaccounted for. Briefly, the dynamics of gene-expression are assumed to be primarily determined by the transcriptional regulation. Given the limitations of the available experimental data and similar work reported (*cf.* Introduction) this is an acceptable premiss. We will now present a comprehensive listing of the molecular reactions used to build a first ('master') model. For more background we refer to Ferreira *et al.* [17]

The following notation was used for the first model:

M_1 : GATA-1 mRNA ; P_1 : GATA-1 protein
M_2 : PU.1 mRNA ; P_2 : PU.1 protein
M_3 : GATA-2 mRNA ; P_3 : GATA-2 protein
M_4 : FOG-1 mRNA ; P_4 : FOG-1 protein
M_5 : EKLF mRNA ; P_5 : EKLF protein
M_6 : β-globin mRNA ; P_6 : β-globin protein

D_i : gene i promotor region ; R : ribosome

K_i : gene i promotor association constant
Q_i : ribosome association constant
r_i : gene i transcription rate constant
l_i : mRNA i translation rate constant
n_i : number of molecules of protein i bound to promotor region i
η_i : number of ribosomes bound to mRNA i
k_{mi} : degradation rate constant mRNA i
k_{pi} : degradation rate constant protein i
τ_i : transcriptional delay of gene i
η_i : translational delay of mRNA i

The dimensions are [min] for time, [μM] for concentrations, [μM^{-1}] for association constants, and [min^{-1}] for rate constants.

GATA-1

Transcription activation and repression:

$$D_1 + n_1P_1 \overset{K_1}{\rightleftharpoons} D_1^{n_1P_1} \;;\; D_1^{n_1P_1} \overset{r_1}{\rightarrow} D_1^{n_1P_1} + M_1$$
$$D_1 + n_2P_2 \overset{K_2}{\rightleftharpoons} D_1^{n_2P_2} \;;\; D_1^{n_2P_2} \overset{r_2}{\rightarrow} D_1^{n_2P_2} + M_1$$
$$D_1 + n_3P_3 \overset{K_3}{\rightleftharpoons} D_1^{n_3P_3} \;;\; D_1^{n_3P_3} \overset{r_3}{\rightarrow} D_1^{n_3P_3} + M_1$$
$$D_1 + n_{12}P_1 + n_{13}P_2 \overset{K_{12}}{\rightleftharpoons} D_1^{n_{12}P_1 n_{13}P_2}$$

Translation:

$$M_1 + \eta_1 R \overset{Q_1}{\rightleftharpoons} M^{\eta_1 R} \;;\; M_1^{\eta_1 R} \overset{l_1}{\rightarrow} M_1^{\eta_1 R} + \eta_1 P_1$$

Degradation:

$$M_1 \overset{k_{m1}}{\rightarrow} \emptyset$$
$$P_1 \overset{k_{p1}}{\rightarrow} \emptyset$$

PU.1

Transcription activation and repression:

$$D_2 + n_4P_1 \overset{K_4}{\rightleftharpoons} D_2^{n_4P_1} \;;\; D_2^{n_4P_1} \overset{r_4}{\rightarrow} D_2^{n_4P_1} + M_2$$
$$D_2 + n_5P_2 \overset{K_5}{\rightleftharpoons} D_2^{n_5P_2} \;;\; D_2^{n_5P_2} \overset{r_5}{\rightarrow} D_2^{n_5P_2} + M_2$$
$$D_2 + n_{14}P_1 + n_{15}P_2 \overset{K_{14}}{\rightleftharpoons} D_2^{n_{14}P_1 n_{15}P_2}$$

Translation:

$$M_2 + \eta_2 R \overset{Q_2}{\rightleftharpoons} M_2^{\eta_2 R} \;;\; M_2^{\eta_2 R} \overset{l_2}{\rightarrow} M_2^{\eta_2 R} + \eta_2 P_2$$

Degradation:

$$M_2 \overset{k_{m2}}{\rightarrow} \emptyset$$
$$P_2 \overset{k_{p2}}{\rightarrow} \emptyset$$

GATA-2

Transcription activation and repression:

$$D_3 + n_6P_3 \overset{K_6}{\rightleftharpoons} D_3^{n_6P_3} \;;\; D_3^{n_6P_3} \overset{r_6}{\rightarrow} D_3^{n_6P_3} + M_3$$
$$D_3 + n_{16}P_1 + n_{17}P_4 \overset{K_{16}}{\rightleftharpoons} D_3^{n_{16}P_1 n_{17}P_4}$$

Translation:

$$M_3 + \eta_3 R \overset{Q_3}{\rightleftharpoons} M_3^{\eta_3 R} \;;\; M_3^{\eta_3 R} \overset{l_3}{\rightarrow} M_3^{\eta_3 R} + \eta_3 P_3$$

Degradation:

$$M_3 \overset{k_{m3}}{\rightarrow} \emptyset$$
$$P_3 \overset{k_{p3}}{\rightarrow} \emptyset$$

FOG-1

Transcription activation:

$$D_4 + n_7 P_1 \overset{K_7}{\rightleftharpoons} D_4^{n_7 P_1} \;;\; D_4^{n_7 P_1} \overset{r_7}{\rightarrow} D_4^{n_7 P_1} + M_4$$

Translation:

$$M_4 + \eta_4 R \overset{Q_4}{\rightleftharpoons} M_4^{\eta_4 R} \;;\; M_4^{\eta_4 R} \overset{l_4}{\rightarrow} M_4^{\eta_4 R} + \eta_4 P_4$$

Degradation:

$$M_4 \overset{k_{m4}}{\rightarrow} \emptyset$$
$$P_4 \overset{k_{p4}}{\rightarrow} \emptyset$$

EKLF

Transcription activation and repression:

$$D_5 + n_8 P_1 \overset{K_8}{\rightleftharpoons} D_5^{n_8 P_1} \;;\; D_5^{n_8 P_1} \overset{r_8}{\rightarrow} D_5^{n_8 P_1} + M_5$$
$$D_5 + n_{18} P_1 + n_{19} P_2 \overset{K_{18}}{\rightleftharpoons} D_5^{n_{18} P_1 n_{19} P_2}$$

Translation:

$$M_5 + \eta_5 R \overset{Q_5}{\rightleftharpoons} M_5^{\eta_5 R} \;;\; M_5^{\eta_5 R} \overset{l_5}{\rightarrow} M_5^{\eta_5 R} + \eta_5 P_5$$

Degradation:

$$M_5 \overset{k_{m5}}{\rightarrow} \emptyset$$
$$P_5 \overset{k_{p5}}{\rightarrow} \emptyset$$

β-globin

Transcription activation:

$$D_6 + n_9P_1 + n_{10}P_4 \overset{K_9}{\rightleftharpoons} D_6^{n_9P_1n_{10}P_4} \;;\; D_6^{n_9P_1n_{10}P_4} \overset{r_9}{\rightarrow} D_6^{n_9P_1n_{10}P_4} + M_6$$

$$D_6 + n_{11}P_5 \overset{K_{11}}{\rightleftharpoons} D_6^{n_{11}P_5} \;;\; D_6^{n_{11}P_5} \overset{r_{11}}{\rightarrow} D_6^{n_{11}P_5} + M_6$$

Translation:

$$M_6 + \eta_6 R \overset{Q_6}{\rightleftharpoons} M_6^{\eta_6 R} \;;\; M_6^{\eta_6 R} \overset{l_6}{\rightarrow} M_6^{\eta_6 R} + \eta_6 P_6$$

Degradation:

$$M_6 \overset{k_{m6}}{\rightarrow} \emptyset$$

$$P_6 \overset{k_{p6}}{\rightarrow} \emptyset$$

Below, the differential equations are presented that derive from the mass-balance equations: the different mRNA's are each produced by transcription of the respective genes, whereas removed by degradation towards a so-called 'sink' (product is not represented by a state variable of the system). These mRNA's serve as templates for the continuous production of protein in the respective translation processes. The protein is also continuously degraded (this is another 'sink' of the system). The transcriptional terms on the right-hand-sides are rational expressions that represent the fraction of active ('open') promotor:RNAP:TF complex (*cf.* Appendix). The m_i and p_i minuscules are the variables corresponding to the mRNA and protein majuscules.The translational terms are first order in the mRNA concentration assuming that the (free) ribosome concentration r is a constant.

$$
\begin{aligned}
dm_1(t)/dt &= -k_{m1}m_1(t) + \frac{f_{1,n}(x(t))}{f_{1,d}(x(t))} \\
f_{1,n}(x) &= r_1K_1p_1(t-\tau_1)^{\eta_1} + r_2K_2p_2(t-\tau_1)^{\eta_2} + r_3K_3p_3(t-\tau_1)^{\eta_3} \\
f_{1,d}(x) &= 1 + K_1p_1(t-\tau_1)^{\eta_1} + K_2p_2(t-\tau_1)^{\eta_2} + K_3p_3(t-\tau_1)^{\eta_3} + K_{12}p_1(t-\tau_1)^{\eta_{12}}p_2(t-\tau_1)^{\eta_{13}} \\
dp_1(t)/dt &= -k_{p1}p_1(t) + \left(\frac{Q_1l_1}{\eta_1}r^{\eta_1}\right)m_1(t-\theta_1)
\end{aligned}
$$

$$
\begin{aligned}
dm_2(t)/dt &= -k_{m2}m_2(t) + f_2(x(t)) \\
f_2(x) &= \frac{r_4K_4p_1(t-\tau_2)^{\eta_4} + r_5K_5p_2(t-\tau_2)^{\eta_5}}{1 + K_4p_1(t-\tau_2)^{\eta_4} + K_5p_2(t-\tau_2)^{\eta_5} + K_{14}p_1(t-\tau_2)^{\eta_{14}}p_2(t-\tau_2)^{\eta_{15}}} \\
dp_2(t)/dt &= -k_{p2}p_2(t) + \left(\frac{Q_2l_2}{\eta_2}r^{\eta_2}\right)m_2(t-\theta_2)
\end{aligned}
$$

$$
\begin{aligned}
dm_3(t)/dt &= -k_{m3}m_3(t) + f_3(x(t)) \\
f_3(x) &= \frac{r_6K_6p_3(t-\tau_3)^{\eta_6}}{1 + K_6p_3(t-\tau_3)^{\eta_6} + K_{16}p_1(t-\tau_3)^{\eta_{16}}p_4(t-\tau_3)^{\eta_{17}}} \\
dp_3(t)/dt &= -k_{p3}p_3(t) + \left(\frac{Q_3l_3}{\eta_3}r^{\eta_3}\right)m_3(t-\theta_3)
\end{aligned}
$$

$$
\begin{aligned}
dm_4(t)/dt &= -k_{m4}m_4(t) + f_4(x(t)) \\
f_4(x) &= \frac{r_7K_7p_1(t-\tau_4)^{\eta_7}}{1 + K_7p_1(t-\tau_4)^{\eta_7}} \\
dp_4(t)/dt &= -k_{p4}p_4(t) + \left(\frac{Q_4l_4}{\eta_4}r^{\eta_4}\right)m_4(t-\theta_4)
\end{aligned}
$$

$$
\begin{aligned}
dm_5(t)/dt &= -k_{m5}m_5(t) + f_5(x(t)) \\
f_5(x) &= \frac{r_8K_8p_1(t-\tau_5)^{\eta_8}}{1 + K_8p_1(t-\tau_5)^{\eta_8} + K_{18}p_1(t-\tau_5)^{\eta_{18}}p_2(t-\tau_5)^{\eta_{19}}} \\
dp_5(t)/dt &= -k_{p5}p_5(t) + \left(\frac{Q_5l_5}{\eta_5}r^{\eta_5}\right)m_5(t-\theta_5)
\end{aligned}
$$

$$
\begin{aligned}
dm_6(t)/dt &= -k_{m6}m_6(t) + f_6(x(t)) \\
f_6(x) &= \frac{r_9K_9p_1(t-\tau_6)^{\eta_9}p_4(t-\tau_6)^{\eta_{10}} + r_{11}K_{11}p_5(t-\tau_6)^{\eta_{11}}}{1 + K_9p_1(t-\tau_6)^{\eta_9}p_4(t-\tau_6)^{\eta_{10}} + K_11p_5(t-\tau_6)^{\eta_{11}}} \\
dp_6(t)/dt &= -k_{p6}p_6(t) + \left(\frac{Q_6l_6}{\eta_6}r^{\eta_6}\right)m_6(t-\theta_6)
\end{aligned}
$$

In the form of a system this becomes

$$
\begin{aligned}
x(t) &= \begin{pmatrix} m_1(t) & p_1(t) & \dots & m_6(t) & p_6(t) \end{pmatrix}, \\
dx(t)/dt &= f(x(t)), \quad x(t_0) = x_0.
\end{aligned}
$$

Whereas this system cannot even considered to be a detailed mechanistic model, it still is too elaborate for our purposes. First, this is a system of coupled delay differential equations which are rather tedious to handle. Therefore, the time delays were set to zero. Given the much longer time-scale at which the system dynamics are expected to play a role, this assumption seems to be justified. Secondly, the high number of parameters (84) as compared to the number of data points (6×11) prompted us to additionally set the η's and θ's to one. This is in agreement with the fact that for the GATA-1 network multiple transcription factors of the same type binding to promotors has not been reported (*cf.* [5]. Moreover, there are a number of parameters that are clearly unidentifiable. These parameters were lumped resulting in a parameter total of 42. Then, the state variables were renamed such that variables x_1 to x_6 represent mRNA concentrations, whereas variables x_7 to x_{12} represent protein concentrations. The same order as above was thereby maintained. Finally, the state variables were scaled (dividing by an arbitrary constant equal to the expected maximum value of the respective variables).

The scaling rules are:

$$
\begin{aligned}
x_1 &\equiv m_1 \,/\, x_{1,max} \\
x_2 &\equiv m_2 \,/\, x_{2,max} \\
x_3 &\equiv m_3 \,/\, x_{3,max} \\
x_4 &\equiv m_4 \,/\, x_{4,max} \\
x_5 &\equiv m_5 \,/\, x_{5,max} \\
x_6 &\equiv m_6 \,/\, x_{6,max} \\
x_7 &\equiv p_1 \,/\, x_{7,max} \\
x_8 &\equiv p_2 \,/\, x_{8,max} \\
x_9 &\equiv p_3 \,/\, x_{9,max} \\
x_{10} &\equiv p_4 \,/\, x_{10,max} \\
x_{11} &\equiv p_5 \,/\, x_{11,max} \\
x_{12} &\equiv p_6 \,/\, x_{12,max}
\end{aligned}
$$

making the scaled variables dimensionless.

The following parameters are defined:

$$
\begin{aligned}
p_1 &\equiv r_1 K_1(x_{7,max}/x_{1,max}),\\
p_2 &\equiv K_1(x_{7,max}),\\
p_3 &\equiv r_2 K_2(x_{8,max}/x_{1,max}),\\
p_4 &\equiv K_2(x_{8,max}),\\
p_5 &\equiv r_3 K_3(x_{9,max}/x_{1,max}),\\
p_6 &\equiv K_3(x_{9,max}),\\
p_7 &\equiv r_4 K_4(x_{7,max}/x_{2,max}),\\
p_8 &\equiv K_4(x_{7,max}),\\
p_9 &\equiv r_5 K_5(x_{8,max}/x_{2,max}),\\
p_{10} &\equiv K_5(x_{8,max}),\\
p_{11} &\equiv r_6 K_6(x_{9,max}/x_{3,max}),\\
p_{12} &\equiv K_6(x_{9,max}),\\
p_{13} &\equiv r_7 K_7(x_{7,max}/x_{4,max}),\\
p_{14} &\equiv K_7(x_{7,max}),\\
p_{15} &\equiv r_8 K_8(x_{7,max}/x_{5,max}),\\
p_{16} &\equiv K_8(x_{7,max}),\\
p_{17} &\equiv r_9 K_9(x_{7,max}x_{10,max}/x_{6,max}),\\
p_{18} &\equiv K_9(x_{7,max}x_{10,max}),\\
p_{19} &\equiv r_{11} K_{11}(x_{11,max}/x_{6,max}),\\
p_{20} &\equiv K_{11}(x_{11,max}),\\
p_{21} &\equiv K_{12}(x_{7,max}x_{8,max}),\\
p_{22} &\equiv K_{14}(x_{7,max}x_{8,max}),\\
p_{23} &\equiv K_{16}(x_{7,max}x_{10,max}),\\
p_{24} &\equiv K_{18}(x_{7,max}x_{8,max}),
\end{aligned}
$$

$$
\begin{aligned}
p_{25} &\equiv k_{m1},\\
p_{26} &\equiv k_{m2},\\
p_{27} &\equiv k_{m3},\\
p_{28} &\equiv k_{m4},\\
p_{29} &\equiv k_{m5},\\
p_{30} &\equiv k_{m6},
\end{aligned}
$$

$$
\begin{aligned}
p_{31} &\equiv k_{p1},\\
p_{32} &\equiv k_{p2},\\
p_{33} &\equiv k_{p3},\\
p_{34} &\equiv k_{p4},\\
p_{35} &\equiv k_{p5},\\
p_{36} &\equiv k_{p6},
\end{aligned}
$$

$$
\begin{aligned}
p_{37} &\equiv l_1 Q_1 R(x_{1,max}/x_{7,max}),\\
p_{38} &\equiv l_2 Q_2 R(x_{2,max}/x_{8,max}),\\
p_{39} &\equiv l_3 Q_3 R(x_{3,max}/x_{9,max}),\\
p_{40} &\equiv l_4 Q_4 R(x_{4,max}/x_{10,max}),\\
p_{41} &\equiv l_5 Q_5 R(x_{5,max}/x_{11,max}),\\
p_{42} &\equiv l_6 Q_6 R(x_{6,max}/x_{12,max})
\end{aligned}
$$

Parameters $p_1, p_3, p_5, p_7, p_9, p_{11}, p_{13}, p_{15}$, and p_{19} have dimension $[\mu\mathrm{M}^{-1}min^{-1}]$; $p_2, p_4, p_6, p_8, p_{10}$, p_{12}, p_{14}, p_{16}, and p_{20} are dimensionless; $p_{17}, p_{25} - p_{42}$ have dimension $[min^{-1}]$; parameters p_{18}, $p_{21} - p_{24}$ have dimension $[\mu\mathrm{M}]$.

Substitution then leads to

$$
\begin{aligned}
dx_1(t)/dt &= -p_{25}x_1(t) + \frac{p_1 x_7(t) + p_3 x_8(t) + p_5 x_5(t)}{1 + p_2 x_7(t) + p_4 x_8(t) + p_6 x_9(t) + p_{21} x_7(t) x_8(t)}\\
dx_2(t)/dt &= -p_{26}x_2(t) + \frac{p_7 x_7(t) + p_9 x_8(t)}{1 + p_8 x_7(t) + p_{10} x_8(t) + p_{22} x_7(t) x_8(t)}\\
dx_3(t)/dt &= -p_{27}x_3(t) + \frac{p_{11} x_9(t)}{1 + p_{12} x_9(t) + p_{23} x_7(t) x_{10}(t)}\\
dx_4(t)/dt &= -p_{28}x_4(t) + \frac{p_{13} x_7(t)}{1 + p_{14} x_7(t)}\\
dx_5(t)/dt &= -p_{29}x_5(t) + \frac{p_{15} x_7(t)}{1 + p_{16} x_7(t) + p_{24} x_7(t) x_8(t)}\\
dx_6(t)/dt &= -p_{30}x_6(t) + \frac{p_{17} x_7(t) x_{10}(t) + p_{19} x_{11}(t)}{1 + p_{18} x_7(t) x_{10}(t) + p_{20} x_{11}(t)}\\
dx_7(t)/dt &= -p_{31}x_7(t) + p_{37}x_1(t)\\
dx_8(t)/dt &= -p_{32}x_8(t) + p_{38}x_2(t)\\
dx_9(t)/dt &= -p_{33}x_9(t) + p_{39}x_3(t)\\
dx_{10}(t)/dt &= -p_{34}x_{10}(t) + p_{40}x_4(t)\\
dx_{11}(t)/dt &= -p_{35}x_{11}(t) + p_{41}x_5(t)\\
dx_{12}(t)/dt &= -p_{36}x_{12}(t) + p_{42}x_6(t)
\end{aligned}
$$

In the form of a system this becomes

$$\begin{aligned} x(t) &= (\ x_1(t)\ \ x_2(t)\ \ \ldots\ \ x_{12}(t)\ ), \\ dx(t)/dt &= f(x(t)),\ \ x(t_0) = x_0. \end{aligned}$$

In this form the model is finally amenable to parameter estimation. First, MATLAB (The MathWorks^{TM}) was used to implement the model. We have used the System Identification Toolbox^{TM} software which is suited to construct mathematical models of dynamic systems from measured input-output data. The so-called non-linear grey-box module can be used to create a model file specifying the right-hand sides of the state and the output equations typically arrived at through physical first principle modeling. The system should be in the following form:

$$\begin{aligned} dx(t)/dt &= F(t,x(t),u(t),par1,par2,\ldots,parN) \\ y(t) &= H(t,x(t),u(t),par1,par2,...parN)+e(t) \\ x(0) &= x_0 \end{aligned}$$

with F and H arbitrary linear or nonlinear functions with N_x and N_y components, respectively. N_x is the number of states and N_y is the number of outputs. $e(t)$ is the observational error, hence the system itself is error-free.

A MATLAB scrip file (m-file) was created that describes the system structure as a MATLAB-function. A separate m-file was created to specify all parameters. These are then referred to in a second m-file that re-defines the system as an IDNLGREY-object. In this format, besides other data, the names of the variables and parameters are specified, as well as the values of the initial conditions and parameter ranges. For parameter estimation then the PEM-function was used which minimizes the trace of the weighted prediction error matrix. Without the weighting this corresponds to a traditional least-sum-of-squared-errors criterion. A line search method is, by default, automatically selected for (from (classic/adaptive) Gauss-Newton, Levenberg-Marquardt) by PEM to find a (local) minimum. To test the method we used synthetic data produced by assigning physiologically feasible values to the parameters that, upon simulation (using the ODE45 differential equation solver), generate realistic expression profiles. Random noise with zero mean was superimposed upon these data (using the RANDN random number generator). Subsequently, the data were packaged into the correct format for the PEM function using the IDDATA function. After running PEM, various output is available by calling specific MATLAB functions: first of all the plots of the experimental data together with the predicted values (*cf.* Figure 2 for an example). Furthermore, the final value of the Cost Function is returned, as well as the parameter error estimates, covariance matrix and, finally, plots of the (auto-)correlation function of the residuals (*cf.* Figure 3).

We used separate runs of PEM to estimate each parameter individually starting from a value widely (1000 times) off the real value. This lead to a list of locally unidentifiable parameters (more precisely $p_2, p_4, p_6, p_8, p_{10}, p_{12}, p_{16}, p_{36}$,and p_{42}). Closer investigation showed that, besides p_{36} and p_{42}, these parameters are part of terms that during the whole time course are too small to affect the dynamic behaviour of the system. Calculating the so-called time-dependent response coefficients (sensitivity coefficients) for all variables towards all parameters (defined as $R^{x_i}_{p_i}(t) \equiv \frac{\partial x_i(t,p)/x_i}{\partial p_j/p_j}$ for some parameter vector $p = p_0$ and for $t > t_0$, *cf.* [28]) confirmed that these parameters are 'locally' unidentifiable because they do not affect the system's behaviour significantly.

To investigate how well PEM performs when simultaneously estimating parameters, different parameter combinations were tested. This revealed that up to 10-15 parameters could be accurately

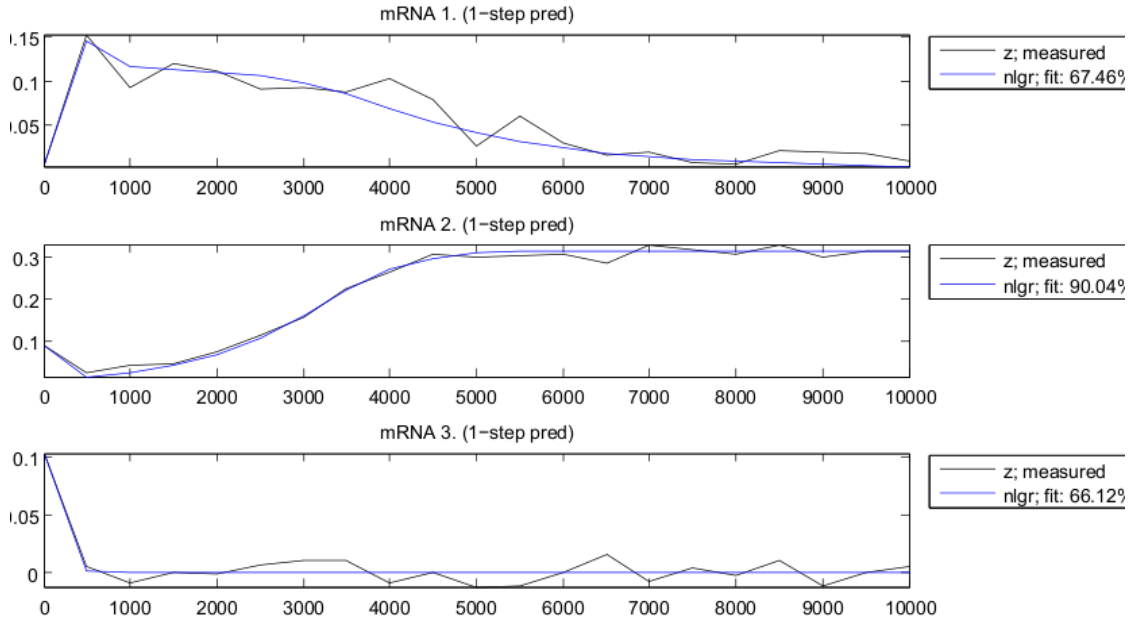

Figure 2: Exemplary plots of scaled mRNA concentrations (of GATA-1, PU.1, and GATA-2) as a function of time (in minutes). The artificial experimental data points are connected by black lines (labeled 'z' in the boxes). The corresponding predictions are connected by blue lines (labeled 'nlgr' in the boxes). Note the 'good fit' despite of the relatively high noise levels

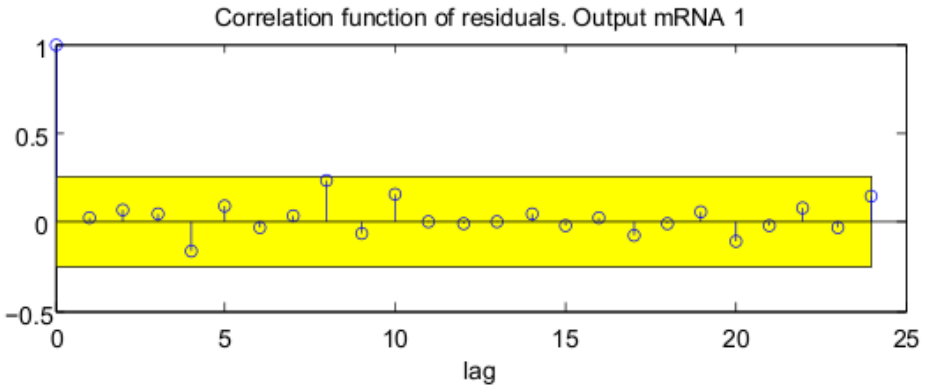

Figure 3: Exemplary plot of the autocorrelation of the residuals (of output variable GATA-1) as a function of the so-called lag parameter. A confidence interval (depicted by the yellow band) indicates whether the autocorrelations lie within the expected range.

estimated from realistic noisy data (in practice this means for example 21 time points for all 6 mRNA species, with a 10 percent standard error). Noteworthy, at this number of parameters the computational time on a ordinary desktop computer becomes a limiting factor.

Other program packages also include parameter estimation modules. For comparative testing we used MATHEMATICA^{TM} and the biochemical simulation software COPASI [26]. Mathematica contains some powerful tools for nonlinear fitting. NMINIMIZE is a function that allows for so-called *global* fitting using Differential Evolution, Random Search, Simulated Annealing and/or Nelder-Mead (a so-called direct local method *i.e.* not explicitly using derivatives) algorithms separately or combined. COPASI has an even more extensive repertoire of algorithms for global fitting, besides classic line search methods. However, in this instance the methods cannot be combined automatically. At this point our experience is that these programs yield parameters estimates that are less accurate than those obtained with MATLAB. In particular the simulated annealing and evolutionary methods took several hours to simultaneously estimate up to 15 parameters. A more detailed report on the comparison of the different packages, both for synthetic data as well as an application to real experimental data from microarray analysis, will be presented in the follow-up (part (2)) of this technical report.

Because of the clear discrepancy between model size and the limitations of the software and computing power available, we proceeded with further simplifying the model. We chose the so-called S-systems framework which has been frequently reported to be useful for modeling gene regulatory networks (*e.g.* [10]). S-systems consist of equations derived from a canonical form that gives the production of some chemical compound as the difference of a production and a degradation term that both contain factors (effector concentrations) raised to some power (apparent kinetic order) multiplied by specific rate constants. Like Linear Systems and Mass Action (MA) systems, S-systems models are special cases of Generalized Mass Action (GMA) systems which have the following generic form:

$$\dot{X}_i \;=\; \sum_{k=1}^{P_i}\left(\pm\gamma_{i,k}\prod_{j=1}^{n}X_j^{g_{i,j,k}}\right), \quad i=1,2,...,n. \quad \forall g_{ij}\in\mathbb{R}^{+}.$$

The generic form of an S-system is

$$\dot{X}_i \;=\; \alpha_i\prod_{j=1}^{n}X_j^{g_{ij}} \;-\; \beta_i\prod_{j=1}^{n}X_j^{h_{ij}}, \quad i=1,2,...,n. \quad \forall g_{ij},h_{ij}\in\mathbb{R}^{+}\cup\mathbb{R}^{-}.$$

The kinetic orders of S-systems may be non-integer and non-positive. This is in contrast with the GMA systems kinetic orders which can only be non-positive and the MA systems which can only have integer values for the their kinetic orders. Clearly the stoichiometric relations are no longer respected, however, the structure is ideal to represent the different regulatory influences (activators with positive powers and repressors with negative powers) on the synthesis/degradation of every compound in the system. In other words, it is a concise representation, in the sense that the kinetic description of each process takes a nearly minimal number of parameters. Furthermore, all the parameters have intuitive physical meaning, facilitating insight and *a priori* estimation. It also yields closed-form analytical steady-state solutions, which are also power laws. This feature greatly facilitates the analysis of steady-state behavior. Finally, it provides accurate approximations of nonlinear systems.

The following system was derived for our GATA-1 GRN. It should be noted that there are only 6 instead of 12 state variables left (the protein level is simply assumed to be proportional to the

mRNA level). Since all variables are observables and every parameter occurs in only one differential equation of the system, it is now suited for decoupling methods [20]. Furthermore the number of parameters is further reduced to 26. Like for the preceding system, the degradation terms are assumed to be first order. Nevertheless, in accordance with preliminary results, this system is expected to be sufficiently flexible to account for observed/expected expression patterns. However, most parameters occurring as exponents makes it in principle more cumbersome for system identification: stiff system are one possible outcome. Furthermore, the success of parameter estimation is reported to crucially depend on the parameter constraints [10]. Our initial trials seem to confirm these issues.

We end this section with the simplified (S-)system for the GATA-1 network:

$$\begin{aligned}
\dot{X}_1 &= \alpha_1 X_1^{g_{1,1}} X_2^{g_{1,2}} X_3^{g_{1,3}} - \beta_1 X_1 \\
\dot{X}_2 &= \alpha_2 X_1^{g_{2,1}} X_2^{g_{2,2}} - \beta_2 X_2 \\
\dot{X}_3 &= \alpha_3 X_1^{g_{3,1}} X_3^{g_{3,3}} X_4^{g_{3,4}} - \beta_3 X_3 \\
\dot{X}_4 &= \alpha_4 X_1^{g_{4,1}} - \beta_4 X_4 \\
\dot{X}_5 &= \alpha_5 X_1^{g_{5,1}} X_2^{g_{5,2}} - \beta_5 X_5 \\
\dot{X}_6 &= \alpha_6 X_1^{g_{6,1}} X_4^{g_{6,4}} X_5^{g_{6,5}} - \beta_6 X_6
\end{aligned}$$

The rate constants α_i and β_i are non-negative and the kinetic orders g_{ij} and h_{ij} are reals with typical values between -1 and +3 [62].

$$\begin{aligned}
X(t) &= \begin{pmatrix} X_1(t) & X_2(t) & \dots & X_6(t) \end{pmatrix} \\
\dot{X} &= f(X) \qquad X(t_0) = X_0
\end{aligned}$$

3 Conclusion

Stem cell differentiation is a fundamental biological process that is still poorly understood. The recent advent of systems biology has brought about a (renewed) interest to build mathematical models of biological processes. The resulting efforts are boosted by the development of new high through-put, genome-scale experiments. In this particular case we have constructed models to describe the time-dependent expression profiles for a gene regulatory network that is proposed to be central to red blood cell differentiation.

In this first technical report of this investigation a (relatively) detailed mechanistic model (with 84 parameters) has been built from which two types of models have been constructed that can be used for parameter estimation. Given the large number of model parameters that are poorly characterized, this is, indeed, an essential step towards reliable model predictions. The first model has significantly less parameters (42) than our 'master' model. It was used to set up a work-flow for parameter estimation with the programs MATLAB, MATHEMATICA, and COPASI. Synthetically generated data were used for testing. The results indicate that, despite of its more limited repertoire of search algorithms, MATLAB performs the best. Still, the number of parameters that can be simultaneously fitted is rather limited, both because of computational as well as model related issues. Moreover, some parameters are shown to be *a priori* (locally) unidentifiable. A simpler model was therefore built based on the S-systems approach. Despite of several advantages, the first tests also indicate some problems related to stiffness and numerical instabilities.

In the next report (part (2)) we will give a more systematic comparison of the different programs we have introduced here, with respect to synthetic data. More emphasis will be on defining parameter constraints to reduce the search space. We will furthermore present a decoupling strategy for the S-systems model. Since experimental data have become available and have been processed, our findings will be used for a real case. This will be paralleled by a simple (local) identifiability analysis to gain more insight in the information content of the data and possibly to suggest new experiments. Since modeling is *par excellence* an iterative process we expect to further refine the existing models. A few widely applied information criteria (Akaike's Information Criterium (AIC), etc.) will be tested for their usefulness. Taken together, the final model will hopefully enable us to get a better idea of certain parameter values, and therefore the biochemical processes associated with them. Importantly, parameters can be 'sloppy', which highlights the power of collective fits and suggestes that modelers should focus on predictions rather than on parameters [24]. Bearing this in mind, our model analysis will also be focus on parameter sensitivity [28, 51] and robustness analysis [38].

4 Acknowledgements

DDV is funded by NBIC via project SP2.3.2 of the BioRange Program. The research advisors are Jan van Schuppen of CWI and Sjaak Philipsen and Frank Grosveld of ErasmusMC. JvS and SP are acknowledged in particular for many discussions with regard to this work. Arnaud Meironeinc is acknowledged for initiating the modeling work.

References

[1] J. S. Almeida and E. O. Voit. Neural-network-based parameter estimation in S-system models of biological networks. *Genome Inform*, 14:114–123, 2003.

[2] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom. Systems biology: parameter estimation for biochemical models. *FEBS J.*, 276:886–902, Feb 2009.

[3] S. Audoly, G. Bellu, L. D'Angi, M. P. Saccomani, and C. Cobelli. Global identifiability of nonlinear models of biological systems. *IEEE Trans Biomed Eng*, 48:55–65, Jan 2001.

[4] G. Bellu, M. P. Saccomani, S. Audoly, and L. D'Angi. DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comput Methods Programs Biomed*, 88:52–61, Oct 2007.

[5] P. Bokes, J. R. King, and M. Loose. A bistable genetic switch which does not require high co-operativity at the promoter: a two-timescale model for the PU.1-GATA-1 interaction. *Math Med Biol*, 26:117–132, Jun 2009.

[6] P. Brazhnik, A. de la Fuente, and P. Mendes. Gene networks: how to put the function in genomics. *Trends Biotechnol.*, 20:467–472, Nov 2002.

[7] M. J. Chappel, K. R. Godfrey, and S. Vajda. Global identifiability of the parameters of nonlinear systems with specified inputs: a comparison of methods. *Math. Biosci.*, 102:41–73, 1990.

[8] V. Chickarmane, C. Troein, U. A. Nuber, H. M. Sauro, and C. Peterson. Transcriptional dynamics of the embryonic stem cell switch. *PLoS Comput. Biol.*, 2:e123, Sep 2006.

[9] I. C. Chou, H. Martens, and E. O. Voit. Parameter estimation in biochemical systems models with alternating regression. *Theor Biol Med Model*, 3:25, 2006.

[10] I. C. Chou and E. O. Voit. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math Biosci*, 219:57–83, Jun 2009.

[11] C. Cobelli and J. J. DiStefano. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *Am. J. Physiol.*, 239:7–24, Jul 1980.

[12] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, 9:67–103, 2002.

[13] J. Delforge, A. Syrota, and B. M. Mazoyer. Identifiability analysis and parameter identification of an in vivo ligand-receptor model from PET data. *IEEE Trans Biomed Eng*, 37:653–661, Jul 1990.

[14] M. Dobrzynski, J. V. Rodrguez, J. A. Kaandorp, and J. G. Blom. Computational methods for diffusion-influenced biochemical reactions. *Bioinformatics*, 23:1969–1977, Aug 2007.

[15] N. D. Evans, M. J. Chapman, M. J. Chappell, and K. R. Godfrey. Identifiability of uncontrolled nonlinear rational systems. *Automatica*, 38:1799–1805, 2002.

[16] M. Farina, R. Findeisen, E. Bullinger, S. Bittanti, F. Allgöwer, and P. Wellstead. Results towards identifiability properties of biochemical reaction networks. *Proc. 45th IEEE Conference Decision and Control*, pages 2104–2109, 2006.

[17] R. Ferreira, K. Ohneda, M. Yamamoto, and S. Philipsen. GATA1 function, a paradigm for transcription factors in hematopoiesis. *Mol. Cell. Biol.*, 25:1215–1227, Feb 2005.

[18] R. Ferreira, A. Wai, R. Shimizu, N. Gillemans, R. Rottier, M. von Lindern, K. Ohneda, F. Grosveld, M. Yamamoto, and S. Philipsen. Dynamic regulation of Gata factor levels is more important than their identity. *Blood*, 109:5481–5490, Jun 2007.

[19] K. G. Gadkar, R. Gunawan, and F. J. Doyle. Iterative approach to model identification of biological networks. *BMC Bioinformatics*, 6:155, 2005.

[20] P. Gennemark and D. Wedelin. Efficient algorithms for ordinary differential equation model identification of biological systems. *IET Syst Biol*, 1:120–129, Mar 2007.

[21] P. Gennemark and D. Wedelin. Benchmarks for identification of ordinary differential equations from time series data. *Bioinformatics*, 25:780–786, Mar 2009.

[22] K. R. Godfrey and J. J. DiStefano. *Identifiability of Model Parameters in Identifiability of Parametric Models*, pages 1–20. Pergamon Press, Oxford, 1987.

[23] K. R. Godfrey and W. R. Fitch. The deterministic identifiability of nonlinear pharmacokinetic models. *J Pharmacokinet Biopharm*, 12:177–191, Apr 1984.

[24] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.*, 3:1871–1878, Oct 2007.

[25] S. Hengl, C. Kreutz, J. Timmer, and T. Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23:2612–2618, Oct 2007.

[26] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI–a COmplex PAthway SImulator. *Bioinformatics*, 22:3067–3074, Dec 2006.

[27] S. Huang, Y. P. Guo, G. May, and T. Enver. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.*, 305:695–713, May 2007.

[28] B. P. Ingalls and H. M. Sauro. Sensitivity analysis of stoichiometric networks: an extension of metabolic control analysis to non-steady state trajectories. *J. Theor. Biol.*, 222:23–36, May 2003.

[29] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, Jun 1961.

[30] J. A. Jacquez and P. Greif. Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Math. Biosci.*, 77:201–227, 1985.

[31] E. M. Landaw and J. J. DiStefano. Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. *Am. J. Physiol.*, 246:R665–677, May 1984.

[32] P. Laslo, C. J. Spooner, A. Warmflash, D. W. Lancki, H. J. Lee, R. Sciammas, B. N. Gantner, A. R. Dinner, and H. Singh. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, 126:755–766, Aug 2006.

[33] J. R. Leis and M. A. Kramer. Odessa–an ordinary differential equation solver with explicit simultaneous sensitivity analysis. *ACM Trans. Math. Software*, 14:61–67, 1988.

[34] L. Ljung. *System Identifiaction: Theory for the User*. Prentice Hall, Englewood Cliffs, NJ, 1999.

[35] L. Ljung and T. Glad. On global identifiabillity for arbitrary model parametrizations. *Automatica*, 30:265–276, 1994.

[36] S. Marino and E. O. Voit. An automated procedure for the extraction of metabolic network information from time series data. *J Bioinform Comput Biol*, 4:665–691, Jun 2006.

[37] S. Marsili-Libelli, S. Guerrizio, and N. Checchi. Confidence regions of estimated parameters for ecological systems. *Ecol. Model.*, 165:127–146, 2003.

[38] P. Melke, H. Jnsson, E. Pardali, P. ten Dijke, and C. Peterson. A rate equation approach to elucidate the kinetics and robustness of the TGF-beta pathway. *Biophys. J.*, 91:4368–4380, Dec 2006.

[39] A. Munack. *Optimization of sampling. In Schügerl, K. and Rehm, H. J. and Reed, G. (Eds.), Biotechnology, a Multi-Volume Comprehensive Treatise, vol. 4*, pages 251–264. 1991.

[40] F. Němcová. *Rational systems in control and system theory* . PhD thesis, Vrije Universiteit, Amsterdam, the Netherlands, 1996.

[41] F. Ollivier. *Le problème de l'identifiabilité structurelle globale: étude théorique, méthodes effectives et bornes de complexité*. PhD thesis, Ecole Polytechnique, Paris, France, 1996.

[42] J. Pahle. Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. *Brief. Bioinformatics*, 10:53–64, Jan 2009.

[43] R. L. M. Peeters and B. Hanzon. Identifiability of homogeneous systems using the state isomorphism approach. *Automatica*, 41:513–529s, 2005.

[44] H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Math. Biosc.*, 41:21–33, 1978.

[45] M. Rodriguez-Fernandez, J. A. Egea, and J. R. Banga. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics*, 7:483, 2006.

[46] M. Rodriguez-Fernandez, P. Mendes, and J. R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems*, 83:248–265, 2006.

[47] M. Rodriguez-Fernandez, P. Mendes, and J. R. Banga. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *BioSystems*, 83:248–265, 2006.

[48] I. Roeder and I. Glauche. Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. *J. Theor. Biol.*, 241:852–865, Aug 2006.

[49] N. J. Savill, W. Chadwick, and S. E. Reece. Quantitative analysis of mechanisms that govern red blood cell age structure and dynamics during anaemia. *PLoS Comput. Biol.*, 5:e1000416, Jun 2009.

[50] T. Schlitt and A. Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8 Suppl 6:S9, 2007.

[51] J. H. Schwacke and E. O. Voit. Computation and analysis of time-dependent sensitivities in Generalized Mass Action systems. *J. Theor. Biol.*, 236:21–38, Sep 2005.

[52] C. Seatzu. A fitting based method for parameter estimation in S-systems. *Dyn. Syst. Appl.*, 9:77–98, 2000.

[53] P. Smolen, D. A. Baxter, and J. H. Byrne. Mathematical modeling of gene networks. *Neuron*, 26:567–580, Jun 2000.

[54] K. Takahashi, S. N. Arjunan, and M. Tomita. Space in systems biology of signaling pathways–towards intracellular molecular crowding in silico. *FEBS Lett.*, 579:1783–1788, Mar 2005.

[55] K. Y. Tsai and F. S. Wang. Evolutionary optimization with data collocation for reverse engineering of biological networks. *Bioinformatics*, 21:1180–1188, Apr 2005.

[56] S. Vajda, K. R. Godfrey, and H. Rabitz. Similarity transformation approach to structural identifiability of nonlinear models. *Math. Biosci.*, 93:217–248, 1989.

[57] J. M. van den Hof. *System Theory and System Identification of Compartmental Systems*. PhD thesis, Rijksuniversiteit Groningen, the Netherlands, 1996.

[58] E. P. van Someren, L. F. Wessels, E. Backer, and M. J. Reinders. Genetic network modeling. *Pharmacogenomics*, 3:507–525, Jul 2002.

[59] P. A. Vanrolleghem and D. Dochain. *Bioprocess Model Identification. In Advanced Instrumentation, Data Interpretation, and Control of Biotechnological Process.*, pages 251–318. Kluwer Academic Publishers, 1998.

[60] M. Vilela, C. C. Borges, S. Vinga, A. T. Vasconcelos, H. Santos, E. O. Voit, and J. S. Almeida. Automated smoother for the numerical decoupling of dynamics models. *BMC Bioinformatics*, 8:305, 2007.

[61] E. O. Voit and J. Almeida. Decoupling dynamical systems for pathway identification from metabolic profiles. *Bioinformatics*, 20:1670–1681, Jul 2004.

[62] E. O. Voit, J. Almeida, S. Marino, R. Lall, G. Goel, A. R. Neves, and H. Santos. Regulation of glycolysis in Lactococcus lactis: an unfinished systems biological case study. *Syst Biol (Stevenage)*, 153:286–298, Jul 2006.

[63] K. Z. Yao, B. M. Shaw, B. Kou, K. B. McAuley, and D. W. Bacon. Modeling Ethylene/Butene Copolymerization with Multi-site Catalyis: Parameter Estimability and Experimental Design. *Polym. React. Eng.*, 11:563–588, 2003.

[64] D. E. Zak, G. E. Gonye, J. S. Schwaber, and F. J. Doyle. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. *Genome Res.*, 13:2396–2405, Nov 2003.

A APPENDIX

We here illustrate how the transcriptional rate equations were derived, taking the transcription rate of GATA-1 as an example. First, the fraction of active over total promotor:RNAP:TF complexes is expressed, after which the equilibrium binding assumption is used to substitute for the different complex concentrations (represented as d_i minuscules corresponding to the DNA (promotor) majuscules). The other notations are explained above.

$$v_{transc,D1} = \frac{r_1 d_1^{n_1 P_1} + r_2 d_1^{n_2 P_2} + r_3 d_1^{n_3 P_3}}{d_1 + d_1^{n_1 P_1} + d_1^{n_2 P_2} + d_1^{n_3 P_3} + d_1^{n_{12} P_1 n_{13} P_2}}$$

$$K_1 = \frac{d_1^{n_1 p_1}}{d_1 (p_1)^{n_1}}, \quad K_2 = \frac{d_1^{n_2 p_2}}{d_1 (p_2)^{n_2}}, \quad K_3 = \frac{d_1^{n_3 p_3}}{d_1 (p_3)^{n_3}}, \quad K_{12} = \frac{d_1^{n_{12} p_1 n_{13} p_2}}{d_1 (p_1)^{n_{12}} d_1 (p_2)^{n_{13}}}$$

Substituting and dividing leads to

$$v_{transc,D1} = \frac{r_1 K_1 p_1^{\eta_1} + r_2 K_2 p_2^{\eta_2} + r_3 K_3 p_3^{\eta_3}}{1 + K_1 p_1^{\eta_1} + K_2 p_2^{\eta_2} + K_3 p_3^{\eta_3} + K_{12} p_1^{\eta_{12}} p_2^{\eta_{13}}}$$

And similarly for the other transcription rate equations ...

Contents

Centrum Wiskunde & Informatica (CWI) is the national research institute for mathematics and computer science in the Netherlands. The institute's strategy is to concentrate research on four broad, societally relevant themes: earth and life sciences, the data explosion, societal logistics and software as service.

Centrum Wiskunde & Informatica (CWI) is het nationale onderzoeksinstituut op het gebied van wiskunde en informatica. De strategie van het instituut concentreert zich op vier maatschappelijk relevante onderzoeksthema's: aard- en levenswetenschappen, de data-explosie, maatschappelijke logistiek en software als service.

Bezoekadres:
Science Park 123
Amsterdam

Postadres:
Postbus 94079, 1090 GB Amsterdam
Telefoon 020 592 93 33
Fax 020 592 41 99
info@cwi.nl
www.cwi.nl

CWI

Centrum Wiskunde & Informatica