Genetic Network Models: A Comparative Study

Eugene P. van Someren[a], L.F.A. Wessels[a,b] and M.J.T. Reinders[a]

[a] Information and Communication Theory Group, [b] Control Laboratory,
Faculty of Information Technology and Systems, Delft University of Technology,
Mekelweg 4, Delft, The Netherlands

ABSTRACT

Currently, the need arises for tools capable of unraveling the functionality of genes based on the analysis of microarray measurements. Modeling genetic interactions by means of genetic network models provides a methodology to infer functional relationships between genes. Although a wide variety of different models have been introduced so far, it remains, in general, unclear what the strengths and weaknesses of each of these approaches are and where these models overlap and differ. This paper compares different genetic modeling approaches that attempt to extract the gene regulation matrix from expression data. A taxonomy of continuous genetic network models is proposed and the following important characteristics are suggested and employed to compare the models: (1) inferential power; (2) predictive power; (3) robustness; (4) consistency; (5) stability and (6) computational cost. Where possible, synthetic time series data are employed to investigate some of these properties. The comparison shows that although genetic network modeling might provide valuable information regarding genetic interactions, current models show disappointing results on simple artificial problems. For now, the simplest models are favored because they generalize better, but more complex models will probably prevail once their bias is more thoroughly understood and their variance is better controlled.

Keywords: Genetic Networks, Modeling, Comparison, Taxonomy, Evaluation Criteria

1. INTRODUCTION

Current developments in micro-array technology have enabled a shift in the way gene interactions can be considered, namely from the reductionists' serial view to the parallel or combinatorial approach. The combinatorial approach assumes that gene activity is the result of a *combined action* of genes rather than that it is influenced by a single gene. Time-course gene expression measurements provide valuable insight in how our genes act 'in concert' to achieve certain phenotypic characteristics. Genetic network modeling provides a methodology to employ time-course data to construct a model that describes the observed phenotypic behavior based on the result of interactions between the genes and their interactions with the environment. In this philosophy, the functionality of a gene is described by the global effect it has in the cell, as caused by its interactions with other genes. Therefore, in order to unravel the functionality of genes, methods are needed that can infer relationships among genes from gene expression data.
A wide variety of approaches that can infer relationships from data and that have proven their use in other domains have recently been proposed for the inference of genetic networks (see Sec. 2). However, the inference of such networks from gene expression data is governed by aspects that are specific for gene expression data, i.e. 1) the number of genes is very large compared to the number of measured time-points, 2) the data contains a substantial amount of measurement noise, 3) the goal of genetic network modeling is to extract the underlying genetic interactions rather than to accurately predict gene expression levels and finally 4) no ground truth is known with respect to the outcome of genetic network models. These disadvantageous aspects can be balanced by the proper use of existing biological knowledge. For example, genetic networks are assumed to be redundant[1] and stable and genes are believed to be influenced on average by no more than eight to ten other genes.[2] These specific aspects imply that genetic network models 1) should be able to handle under-constrained data, 2) must be robust against noise, 3) provide interpretable results and 4) should enforce constraints on the connectivity, stability and redundancy of the inferred networks. As a result, it is unclear what the strengths and weaknesses of the proposed models are when applied

Both van Someren and Wessels contributed equally to this work.

Further author information:
E-mail: E.P.vanSomeren@its.tudelft.nl
WWW: www-ict.its.tudelft.nl\ ~IMDS

to gene expression data and in which aspects they are similar or different. This work proposes a taxonomy for the existing repertoire of genetic network models (Section 2). In Section 3 several criteria are considered which are important for succesful genetic network modeling and the models are compared with respect to these criteria. Where applicable the performance of each of the models on the proposed criteria are further investigated empirically in Section 4. The performance of each model will depend highly on the "intrinsic complexity" of the dataset. This 'intrinsic complexity' is governed by the complexity of the underlying genetic network, the number of measured time-points and the amount of measurement noise. In order to provide insight in the influence of these aspects on the performance of the models and to determine which models are best suited for which situations, the evaluation criteria are evaluated under various conditions. In previous work,[3] we have compared currently known models on datasets generated by models of different complexity. In this paper, the influence of connectivity, measured time-points and measurement noise is evaluated. This evaluation results in some preliminary reflections (Section 5) on current genetic network models and considerations for future work.

2. MODEL DESCRIPTIONS

This comparison focuses on continuous models [4–14] , i.e. models that represent gene activity with a continuous value, rather than on discrete models. Continuous models are favored, because they are more realistic, i.e. intermediate gene activity levels are known to be important for some aspects of existing interactions and because gene expression measurements are also continuous-valued they do not require explicit discretization of the measured input data prior to the inference process. Discretization is undesirable, because it might enforce erroneous relationships which are sensitive to slight variations in the used threshold value. In addition, the use of discretized values does not provide an apparent improvement in an automated inference process. Boolean networks,[15] Bayesian networks[16] and Qualitative models[17] are therefore excluded. Studying the continuous models reveals a taxonomy of continuous genetic network models that is based on the following three categories: (1) methods based on pair-wise comparisons, (2) methods that model rough networks of gene-interactions and (3) more complex models that also describe intermediate products, such as proteins.

2.1. Pair-wise Methods

These methods construct relationships between genes based solely on pair-wise comparisons. Therefore, they do *not* take into account interactions where the resulting expression-level of one gene is governed by the *combined* action of multiple other genes. Because these methods have no actual model that describes *exactly how* genes are activated by external inputs and other genes, no predictions of gene expression can be made. We now elaborate on two of these methods that were recently introduced.

Arkin and Ross '97: Correlation Metric Construction (CMC) The CMC method[4] first computes the magnitude and position at which the maximal (time-shifted) cross-correlation occurs. This provides a measure of similarity and temporal ordering, respectively. Then a distance matrix, D, is constructed, by comparing, for each pair of genes, their similarities to other genes. The significant eigenvalues of the constructed distance matrix provides an indication of the intrinsic dimensionality of the system. Single linkage hierarchical clustering is employed to find a singly linked tree that connects associated genes. This tree (association diagram) is augmented with directional and time-lag information, revealing temporal ordering.

Chen, Filkov and Skiena '99: Activation/Inhibition Networks The method proposed by Chen *et al.*[5] expresses regulation based on whether peaks in one signal precede peaks in another signal. After thresholding and clustering, resulting in a set of prototype signals, each prototype is represented as a series of peaks. For each pair of prototypes three scores are computed, representing a possible activating, inhibiting or unmatching relationship. The regulation matrix is inferred by taking for each pair of genes the highest of these three scores.

2.2. Rough Network Models

These networks directly model effects that result from the combination of different input genes, by means of a weighted sum of their expressed levels. The term 'rough' refers to the fact that the influence of all intermediate products are summarized in the linear gene-to-gene relationship. These weights form the gene regulation matrix (hereafter GRM) which is common among all rough network models and provide information about the relationships between genes. For all gene regulation matrices, zero weights indicate the absence of interaction and a positive or

Method	Equation	$g(z)$	$R_i =$	$\lambda_i =$
Mjolsness00a[6]	(1)	$\frac{1}{1+e^{-z}}$	R_i	λ_i
Spirov97a[7]	(1)	$\frac{1}{1+e^{-z}}$	R_i	λ_i
Wahde99a[8]	(1)	$\frac{1}{1+e^{-z}}$	R_i	R_i
Weaver99a[9]	(2)	$\frac{1}{1+e^{-z}}$	R_i	0
Someren00a[10]	(2)	z	1	0
Someren00b[11]	(2)	z	1	0
D'Haeseleer99a[12]	(2)	z	1	0

Table 1. Deviation of each network model with respect to the general equations (1 & 2).

Method	Pre-Processing	Structuring	Inference
Mjolsness00a	-	EM-Clustering	SA
Spirov97a	-	-	GA + SA + GD
Wahde99a	-	-	GA
Weaver99a	Normalization	-	De-squashing + Lin. Regression
Someren00a	-	Sparse Weight Searching	Lin. Regression
Someren00b	Normalization + Thresholding	Hierarchical Clust.	Lin. Regression
D'Haeseleer99a	Interpolation	-	Lin. Regression

Table 2. Methodology to infer parameters. SA = Simulated Annealing, GA = Genetic Algorithm, GD = Gradient Descent

negative weight corresponds to stimulation or repression. The absolute value of a weight corresponds to the strength of the interaction. All genetic network models can be represented in the following generalized differential equation:

$$\frac{dx_i(t)}{dt} = R_i \cdot g\left(\sum_{j=1}^{J} W_{i,j}x_j(t) + \sum_{k=1}^{K} V_{i,k}u_k(t) + B_i\right) - \lambda_i x_i(t) \tag{1}$$

or generalized difference equation of similar form:

$$x_i[t+1] = R_i \cdot g\left(\sum_{j=1}^{J} W_{i,j}x_j[t] + \sum_{k=1}^{K} V_{i,k}u_k[t] + B_i\right) - \lambda_i x_i[t] \tag{2}$$

with the following (biological) interpretations:

$g(\cdot)$: monotonic regulation-expression (activation) function
$x_i(t),\ x_i[t]$: gene expression of gene i at time instance t.
R_i: rate constant of gene i.
$W_{i,j}$: strength of control of gene j on gene i
$u_k(t),\ u_k[t]$: k-th external input at time instance t.
$V_{i,k}$: influence of the k-th external input on gene i.
B_i: Basal expression of gene i.
λ_i: degradation constant of the i-th gene expression product.

All rough models included in this comparison are a specific instance of one of these generalized equations. An overview of the network models is given in Table 1. Table 2 summarizes the algorithms employed to infer the model parameters from measured mRNA levels.

2.3. Complex Networks

These networks are more complex in that they not only model interactions between genes based on measured mRNA levels, but also explicitly model their intermediate products, such as proteins and metabolites. Consequently,

their parameters cannot be inferred from data-sets that contain only measurements of mRNA levels, since explicit knowledge about the expression levels of the intermediates are necessary (see Chen[13]).

Chen, He and Church '99 Both mRNA as well as protein products are modeled, but it is assumed that the protein level of gene i, p_i, is translated only from mRNA of gene i, r_i.

$$\frac{dr_i(t)}{dt} = \sum_{j=1}^{J} W_{i,j} \cdot p_j(t) - \lambda_i \cdot r_i(t) \tag{3}$$

$$\frac{dp_i(t)}{dt} = L_i \cdot r_i(t) - \gamma_i \cdot p_i(t) \tag{4}$$

, which results without proteins in:

$$\frac{d^2r(t)}{dt^2} = (-W\Gamma W^{-1} - \Lambda)\frac{dr(t)}{dt} + (-W\Gamma W^{-1}\Lambda + WL)r(t) \tag{5}$$

, where $W = [W_{i,j}]$, $\Gamma = [\gamma_i]$, $\Lambda = [\lambda_i]$, $L = [L_i]$.

Savageau '99 The genetic network model proposed by Savageau[14] models mRNA, protein and metabolite expression levels. The changes in the expression level of a given chemical species (mRNA, protein or metabolite) are modeled as follows:

$$\frac{dx_i}{dt} = \alpha_i \prod_{j \in \mathcal{S}_I} x_j^{W_{i,j}} - \beta_i \prod_{k \in \mathcal{S}_D} x_k^{W_{i,k}} \tag{6}$$

The first term constitutes an activation term ($\mathcal{S}_I$ being all activating species) and the latter a degradation term ($\mathcal{S}_D$ being species involved in degradation). The 'product-power law' formulation results in expression levels that behave non-linearly, while still providing interpretable parameters. This model contains a large number of parameters in comparison to the rough and pair-wise models which makes the inference process particularly difficult.

3. MODEL EVALUATION

This section introduces several characteristics that are central to successful genetic network modeling and are therefore appropriate benchmarks with which genetic modeling approaches can be evaluated. These characteristics are (1) inferential power, (2) predictive power, (3) robustness, (4) consistency, (5) stability and (6) computational cost. In this section each characteristic is defined and motivated. Both pair-wise and rough network models are also subjected to a preliminary evaluation with respect to each of the characteristics. In Sec. 4 each of these models are compared empirically (where applicable) based on these characteristics. In order to simplify the discussion of the

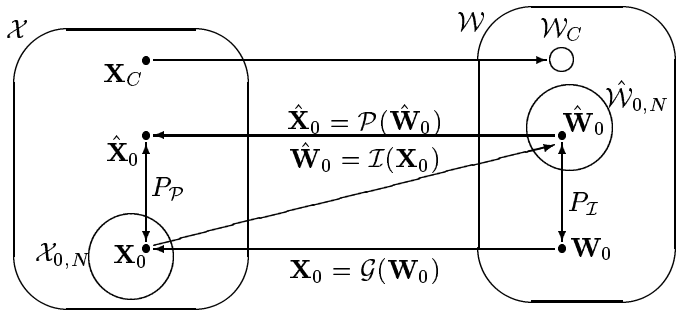

Figure 1. Graphical representation of the framework that relates the expression space ($\mathcal{X}$) to the gene regulation matrix space ($\mathcal{W}$).

different characteristics, a simple framework and nomenclature (depicted in Fig. 1) is introduced. For a particular class of genetic network models, let $\mathcal{W}$ represent the space of gene regulation matrices, with $\mathbf{W}$ an instance of a regulation matrix. Let $\mathcal{X}$ denote the space of expression data sets associated with $\mathcal{W}$, with $\mathbf{X}$ a particular data set. For experimental purposes, a particular matrix, $\mathbf{W}_0$ is chosen to represent a 'real' network.* Based on $\mathbf{W}_0$

*In reality $\mathbf{W}_0$ is not known, but needs to be inferred by the genetic network algorithms. However, in synthetic experiments, we are free to choose $\mathbf{W}_0$, allowing us the opportunity to verify the results produced by the algorithms.

(and initial conditions) a dataset, $\mathbf{X}_0$, is generated ($\mathbf{X}_0 = \mathcal{P}(\mathbf{W}_0)$). During the *inference* step (employing one of the models described in Section 2) an estimate of $\mathbf{W}_0$ is obtained based solely on $\mathbf{X}_0$, and this is represented as $\hat{\mathbf{W}}_0$ with $\hat{\mathbf{W}}_0 = \mathcal{I}(\mathbf{X}_0)$. $\hat{\mathbf{W}}_0$ can, in turn, be employed to make a *prediction* of the expression data, given the same initial conditions. This produces an approximation of $\mathbf{X}_0$, which is denoted by $\hat{\mathbf{X}}_0$, with $\hat{\mathbf{X}}_0 = \mathcal{P}(\hat{\mathbf{W}}_0)$.

3.1. Inferential power, $P_{\mathcal{I}}$

Since genetic network models are primarily employed to infer regulatory interaction patterns from expression data, it is quite important that this process delivers accurate estimates of the gene regulation matrix. Inferential power is a measure of this capability.

Definition: Formally, inferential power is measured as the similarity between the actual ($\mathbf{W}_0$) and inferred ($\hat{\mathbf{W}}_0$) gene regulation matrices: $P_{\mathcal{I}}(\mathbf{W}_0, \hat{\mathbf{W}}_0) = 0.5(1 + \rho(\mathbf{W}_0, \hat{\mathbf{W}}_0))$, with $\rho(\cdot)$ the Pearson product moment correlation.

Preliminary evaluation: Pair-wise approaches infer a regulation matrix based on pair-wise comparisons. Consequently greedy, yet sparse, sub-optimal solutions are obtained. When a gene is influenced by more than one other gene, even erroneous solutions may be obtained. Network models do not suffer from this disadvantage, but the regulation matrix is inferred *indirectly*, i.e. by minimizing the prediction error (difference between $\mathbf{X}_0$ and $\hat{\mathbf{X}}_0$) instead of maximizing $\rho(\mathbf{W}_0, \hat{\mathbf{W}}_0)$. This is done under the assumption that small prediction errors result in accurate regulation matrices. More definitive statements about the inferential power of individual models can only be made based on empirical evidence (Sec. 4).

3.2. Prediction power, $P_{\mathcal{P}}$

Prediction power is reflected in the prediction accuracy, i.e. how closely $\hat{\mathbf{X}}_0$ approximates $\mathbf{X}_0$. For network models, regulation matrices are inferred by minimizing prediction error. It is therefore important to gain insight into the relationship between $P_{\mathcal{P}}$ and $P_{\mathcal{I}}$ for the different models in the comparison.

Definition: For a given expression data set, $\mathbf{X}_0$, and the predicted approximation thereof, $\hat{\mathbf{X}}_0$, prediction power is expressed as $P_{\mathcal{P}} = 1/(1+E_{\text{MSE}})$, with the mean squared error given by $E_{\text{MSE}}(\mathbf{X}_0, \hat{\mathbf{X}}_0) = \frac{1}{TN}\sum_{i,t}(X_0(i,t) - \hat{X}_0(i,t))^2$.

Preliminary evaluation: It is to be expected that more complex models, such as the network models containing non-linearities and larger parameter sets have *potentially* greater predictive power than simpler network models such as the linear models proposed by D'Haeseleer[12] and van Someren.[11] However, more complex models require more training data (or regularization constraints) and non-deterministic inference algorithms (such as a GA[8] or SA[6]) that do not guarantee convergence to a global minimum, possibly resulting in a sub-optimal value for $P_{\mathcal{P}}$. To shed more light on these issues, an empirical investigation was performed (Sec. 4).

3.3. Robustness

Noise is an ever present phenomenon that always needs to be dealt with. Gene expression measurements are particularly noisy and it is therefore important to know to what degree an accurate gene regulation matrix will be extracted in the presence of noise.

Definition: Let $\mathcal{X}_{0,N} \subset \mathcal{X}$ denote the set of expression data sets that are obtained when noise, $\mathbf{N}(0,\sigma)$ [†], is added to $\mathbf{X}_0$, i.e. $\mathcal{X}_{0,N} = \{\mathbf{X}_{0,\sigma} | \mathbf{X}_{0,\sigma} = \mathbf{X}_0 + \mathbf{N}(0,\sigma), \sigma \in [0, \sigma_{\text{MAX}}]\}$. (See also Fig. 1). Let $\hat{\mathcal{W}}_{0,N}$ denote the set of regulation matrices that are obtained by performing inference on each element of $\mathcal{X}_{0,\sigma}$, i.e. $\hat{\mathcal{W}}_{0,N} = \{\hat{\mathbf{W}}_{0,\sigma} | \hat{\mathbf{W}}_{0,\sigma} = \mathcal{I}(\mathbf{X}_{0,\sigma}), \mathbf{X}_{0,\sigma} \in \mathcal{X}_{0,N}\}$. A measure of robustness, $P_{\mathcal{R}}$, is defined as the minimal correlation amongst the inferred regulation matrices in $\hat{\mathcal{W}}_{0,N}$: $P_{\mathcal{R}} = min_{[\{\hat{\mathbf{W}}_{0,\sigma_1}, \hat{\mathbf{W}}_{0,\sigma_2}\} \in \hat{\mathcal{W}}_{0,N}]}[0.5(1 + \rho(\hat{\mathbf{W}}_{0,\sigma_1}, \hat{\mathbf{W}}_{0,\sigma_2}))]$.

Preliminary evaluation: None of the techniques contain features that were explicitly included to increase robustness. However, some implicit characteristics may improve robustness, For example: 1) several techniques[5,6,11] perform a clustering step prior to the inference step which could increase robustness since inference is performed on the prototype (*averaged* signals in a cluster); 2) CMC[4] also employs an averaging process (cross correlation) which could contribute to increased robustness and 3) the method proposed by Chen et al.[5] may also result in improved robustness by creating a filtered signal representation prior to the inference step. However, many models also use a kind of selection step (e.g. thresholding or clustering) such that small changes in expression profiles could cause completely different regulation matrices. Therefore, robustness can be most effectively evaluated empirically by monitoring the behavior of E_R as a function of the noise level, σ.

[†] zero-mean, Gaussian noise with a variance of σ.

3.4. Consistency

One of the most striking properties of available gene expression data is the relatively large number of genes compared to the number of measured time-points. This so called 'dimensionality problem' is an important cause of inconsistency in the inferred genetic network models, and therefore an important characteristic to investigate.

Definition: A genetic network model is said to be inconsistent if multiple parameter sets can be inferred from the same expression data. Formally, for an arbitrary expression data set, $\mathbf{X}_C \in \mathcal{X}$ (See Figure 1), the following set of gene regulation matrices is defined: $\mathcal{W}_C = \{\mathbf{W}_C | \mathbf{W}_C = \mathcal{I}(\mathbf{X}_C), P_{\mathcal{I}}(\mathcal{P}(\mathbf{W}_C), \mathbf{X}_C) > 1 - \epsilon\}$. If the cardinality of $\mathcal{W}_C$ is greater than one, the model is inconsistent and the degree of inconsistency is given by $P_C = min_{[\{\mathbf{W}_{C,i}, \mathbf{W}_{C,j}\} \in \mathcal{W}_C]}\ 0.5(1 + \rho(\mathbf{W}_{C,i}, \mathbf{W}_{C,j}))$, $i \neq j$.

Preliminary evaluation: Without a principled mechanism to select the most appropriate regulation matrix from a set of possible solutions, the inferred parameters will be meaningless. Inconsistency originates from the combination of two causes, 1) the dimensionality problem, i.e. the data represents an incomplete description of the underlying process, or 2) the predictive power of the model exceeds the intrinsic complexity of the data. The first cause can be only partly corrected by interpolation. To handle the second cause one needs to control the predictive power of the model, which can be attained by thresholding (to eliminate very noisy signals), clustering (grouping signals that cannot be distinguished by the model) and/or the introduction of appropriate constraints. Consistency is a very difficult criterion to determine, because the set of possible solutions can generally not be determined.

3.5. Stability

Due to limited energy and storage within a cell, concentrations of gene expression products, such as mRNA, remain bounded. All real genetic networks are therefore stable by definition. Consequently, inferred genetic network models should also be stable in order to be realistic.

Definition: A model, parameterized by a specific gene regulation matrix, is stable if the predicted expression levels remain bounded over all time, for any finite initial state. If infinitely large training sets are available, and E_{MSE} remains bounded, stability can be guaranteed. On finite data sets, the requirement for bounded E_{MSE} must be augmented by other stability requirements such as the existence of a Lyapunov function.

Preliminary evaluation: Since the pair-wise methods have no explicit model to generate signals, no indication can be given about their stability. Both linear network models[11,12] are stable when the magnitude of the eigenvalues of the weight matrix are all smaller than or equal to one. In contrast, the network models that have a sigmoidal transfer function are ensured to have a bounded output and are therefore by definition stable. However, the plausibility of genetic network model that generates gene expression profiles that are oscillating rapidly between minimal and maximal expression bounds are questionable. The determination of the maximal eigenvalue provides some indication whether predictions are expected to be stable or to show unstable or oscillation behavior. None of the methods described in Section 2 are equipped with explicit mechanisms to ensure that a stable network is inferred.

3.6. Computational cost

Methods that require extremely long computation times to reach a solution are obviously undesirable, unless the additional waiting time results in substantially improved results. Methods that can not compute analytical solutions, but rely on iterative solution approaches require, in general much longer computation times. Empirical evaluation was employed to obtain an indication of the time required to reach a solution for each of the approaches.

4. EXPERIMENTAL EVALUATION

In order to test the evaluation criteria experimentally, each model is tested on a set of artificially generated datasets with varying properties. The performance of each tested model will depend highly on the "intrinsic complexity" of the dataset. Presently, there is no known method to determine this "intrinsic complexity", but factors that will influence it are 1) the number of genes, 2) the redundancy[‡] and 3) the connectivity of the underlying genetic network, as well as 4) the no. of measured time-points and 5) the measurement noise. To provide insight in the influence of these factors on the performance of the models, we have employed the following data generation scheme.

For a given number of genes and a given connectivity[§], a random gene regulation matrix (GRM) is generated. To

[‡] redundancy implies that functionality is shared among genes.

[§] For this experiment the connectivity is considered to be equal for all genes, i.e. it corresponds to the number of non-zero weights in each row of the gene regulation matrix

ensure that the GRM is stable[¶], the eigenvalues are determined and all elements in the matrix are divided by the magnitude of the largest eigenvalue. The resulting GRM is used in combination with a linear model and random initial states to generate for each predefined total number of time-points (hereafter T), ten duplicate training datasets and ten duplicate test datasets. The number of time-points, T, was varied from 3 to twice the number of genes in order to reflect under-determined as well as over-determined data. All elements in each of these datasets were corrupted with gaussian noise of zero mean and given standard deviation ($\sigma^2 \in \{0, .0001, .01, .1\}$) to mimic the effect of measurement noise. The effect of noise was so drastic that noise levels were kept low. In total 6720 datasets of fifteen genes were generated: 20 duplicates of 28 different lengths and 3 different levels of connectivity (from sparse to fully connected, connectivity = 2, 8 and 15 resp.) and 4 noise levels.

Six models (Heaseleer99a, Someren00a, Someren00b, Arkin97a, Chen99b and Wahde99a) and a reference model were implemented and were trained on each of the 3360 training datasets. The inferred GRM's were compared with the original to determine the inferential power as defined in section 3. The used computation time of the inference process was measured as an indication of the computational costs. From the inferred GRM's the magnitude of the eigenvalue was determined as an indication of the stability. The inferred GRM's were also used with the corresponding models to predict signals and to determine three different types of predictive power, i.e. 1) free-run predictive power on the training dataset, using only the initial state of the training dataset, 2) free-run predictive power on the test dataset, using only the initial state of the test dataset and 3) one-step predictive power on the test dataset, using consecutively the previous state of the test dataset to predict the next.

For each criterion a suitable reference is determined. In the case of the inferential power, the stability and the computational cost, the reference model returns a random stable matrix with a connectivity of seven. This connectivity was chosen as is in order to check the criteria for a possible bias with respect to a certain connectivity. However, no distinguishable difference is observed in the performance of the reference model for the various cases of connectivity and no bias may be assumed. The reference model for the predictive power is based on a zero-order hold, i.e. the model returns the state at the previous known time-step. For free-run predictions this means the initial state, whereas for one-step predictions the previous state is returned. The grey patches in each of the plots indicate the average of the reference model plus and minus two standard deviations. For example, for a randomly generated GRM, the inferential power is expected to average around .5. To get an indication of the relation between the pair-wise GRM and the rough GRM, the GRM's returned by both pair-wise models were used in combination with a linear model to determine their "predictive power".

Figure 2 depicts from top to bottom, the measured inferential ($P_{\mathcal{I}}$) power, free-run prediction ($P_{\mathcal{P}}$) power on training data, the free-run prediction power on test data, the one-step prediction power on the test data, the stability and the computational costs of the implemented models under the various conditions [‖] without measurement noise. From left to right, the different columns correspond to decreasing connectivity, namely fully connected (15), half connected (8) and sparsely connected (2). In each plot, the x-axis corresponds to increasing numbers of time-points in the dataset. Each line shows the performance of one of the models, where line markers distinguish the models (see also the legend).

Chen99b: The inferential power of the Chen99b model seems to have random behavior at best, however the inferred GRM's are generally stable with "predictive power" comparable to the reference model. The Chen99b model performs constant, i.e. its performance is independent of connectivity and no. time-points except that the computational costs grow logarithmically with T.

Arkin97a: The inferential power of the Arkin97a model performs slightly above random irrespective of the connectivity, but increases slightly when the number of time-points are increased. Because the inferred GRM's are unstable (in linear sense) the predictive power remains zero. The computational costs stay virtually constant for increased T.

Someren00b: As expected, the inferential power of the Someren00b model increases with T and even becomes perfect when the model is over-constrained. The model was designed for redundant GRM's, but shows no improvement for less connected GRM's in the under-determined case and reduced performance for very sparse matrices in the over-determined case. With very low connectivity the generated signals tend to be very simple, indicating a dataset with low intrinsic complexity. Although the model captures this low intrinsic complexity by using a small number of clusters, the genes need not necessarily share the same weights, hence the decrease in inferential power. The computational costs increase more strongly for the over-determined case than for the under-determined case.

[¶] In this paper, a GRM is determined to be stable, when the magnitudes of all of its eigenvalues are smaller than or equal to one. This definition equals the requirement that the GRM is a marginally stable linear system.

[‖] Each point depicts a ten-fold average; standard deviations not depicted in the interest of clarity.

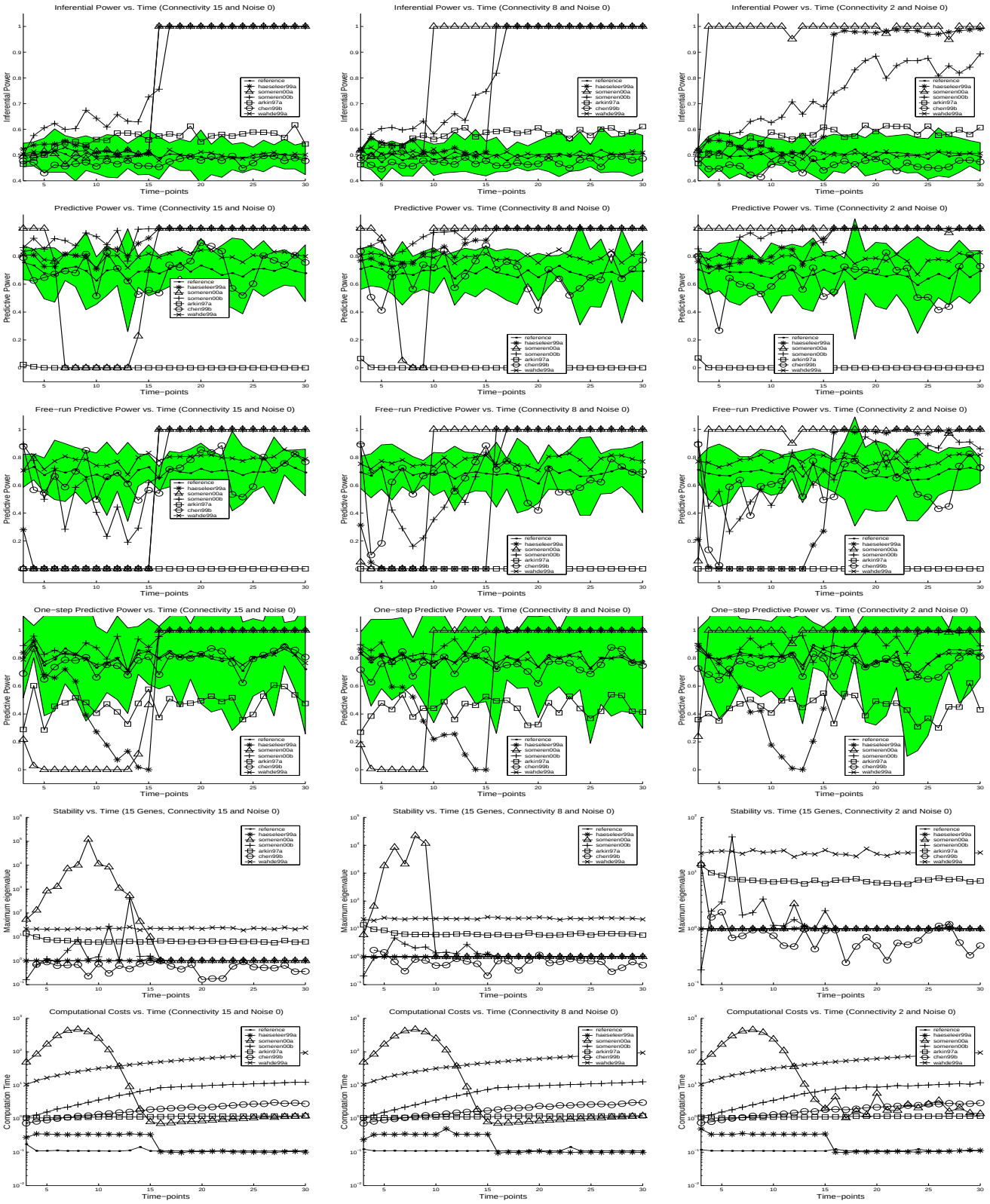

Figure 2. Noise-free experiments of seven genetic network models as a function of no. time-points in the dataset (x-axis). Rows: Six evaluation criteria, i.e inferential power and three types of predictive power, stability and computational costs. Columns: three different levels of connectivity. See text for more detailed description.

Someren00a: The Someren00a model, which is designed especially for sparse networks, shows excellent inferential performance which is directly related to the connectivity. The inferential performance is perfect for all cases where the number of time-points exceeds the connectivity by more than one! Such a performance would indicate that large genetic networks of thousands of genes for which the connectivity is expected to be below ten, can be learned using no more than twelve time-points! Of course this result is not surprising because the model that generated the data is also linear and no measurement noise was added. The Someren00a model is, however, computationally not tractable for larger networks. The predictive power also drops in the same cases as the inferential power, except for very small T on the training data, caused by overfitting**of the training data.
Heaseleer99a: The cubic interpolation of Haeseleer99a gives good results in the over-determined cases, but scores no better than random in the under-determined cases. In our experience, interpolation does not resolve the dimensionality problem.
Wahde99a: The Wahde99a model stays within the random patch for all cases. In order to obtain a result in reasonable time, the number of generations was limited to 20 which resulted in a lack of convergence. Only with extremely long computation times, the performance increases (results not shown).
Figure 3 again depicts the six criteria, but now for the case of small measurement noise ($s^2 = .01$). With this amount of noise the signals show only slight differences. Both pair-wise models show no difference with respect to the noise-free performance indicating strong robustness, which is for the Arkin97a model partly due to the averaging effect of the correlation measure, but also (and completely for the Chen99b model) due to the already poor performance which cannot get worse. In the over-constrained case, both the Someren00a as the Someren00b models drop sharply and show very similar inferential performance, i.e. a gradual increase with T. For small values of T, the performance of the Someren00a model drops drastically, as expected, but the Someren00b model seems less strongly effected in the under-determined case. A possible indication of its robustness that is probably due to the averaging effect of the cluster prototypes.
Figure 4 shows the inferential power for the two remaining noise levels, i.e. $\sigma^2 \in \{.0001, .1\}$. It illustrates the drastic decrease of inferential performance in the presence of measurement noise. The predictive power on the training data remained virtually unaffected for increased levels of noise (not shown), indicating that the models faithfully fitted the provided data, without distinguishing noise from signal. Although the decreased inferential performance indicated above indicates a poor robustness for the models in situations with decreased performance, the actual measurement of the robustness criterion, as defined in Section 3, required a different experimental setup.
Similar to the previous experiment, 10 duplicate datasets were generated using a GRM of fifteen genes for each of the three levels of connectivity and number of time-points, T. From each of these 840 noise-free datasets (each representing one data-point), nine additional datasets were created by adding gaussian measurement noise with zero mean and a variance of .1. Again all models were trained on the datasets, generating ten estimates for the GRM for each datapoint. The robustness can now be determined by calculating the minimal correlation between two solutions out of the ten estimates.
The resulting measures of robustness are shown in Figure 5. The Arkin97a model shows a robustness that stays above average for all time-points. The robustness of the Haeseleer99a model is high for $T < 7$, slightly above average for $T > 23$ and below average when T is slightly less than the number of genes. This decrease in robustness matches exactly the cases where the interpolation has the most influence. This strong robust behavior is probably due to the fact that it employs a pseudo inverse to solve for the least mean square solution. It seems that the interpolation degrades the robustness of the pseudo-inverse.
A more general interpretation of the experimental results is provided in the discussion (Section 5).

5. DISCUSSION

In this paper a taxonomy of continuous modeling approaches was proposed consisting of three classes: pair-wise, rough, and complex network models. Two generalized equations are given that describe all rough network models. In addition, a set of evaluation criteria which are central to genetic network modeling, were proposed and, were applicable, quantitative measures were given. A qualitative as well as an empirical comparison of a set of models was performed based on these evaluation criteria.
Two of the most striking results of the experimental evaluation is that current genetic network models perform disappointing when learned on datasets consisting of a relatively low number of time-points and when corrupted by a

**Overfitting corresponds to the situation when a model adapts too strongly to the observed data, losing its generalizing capabilities.

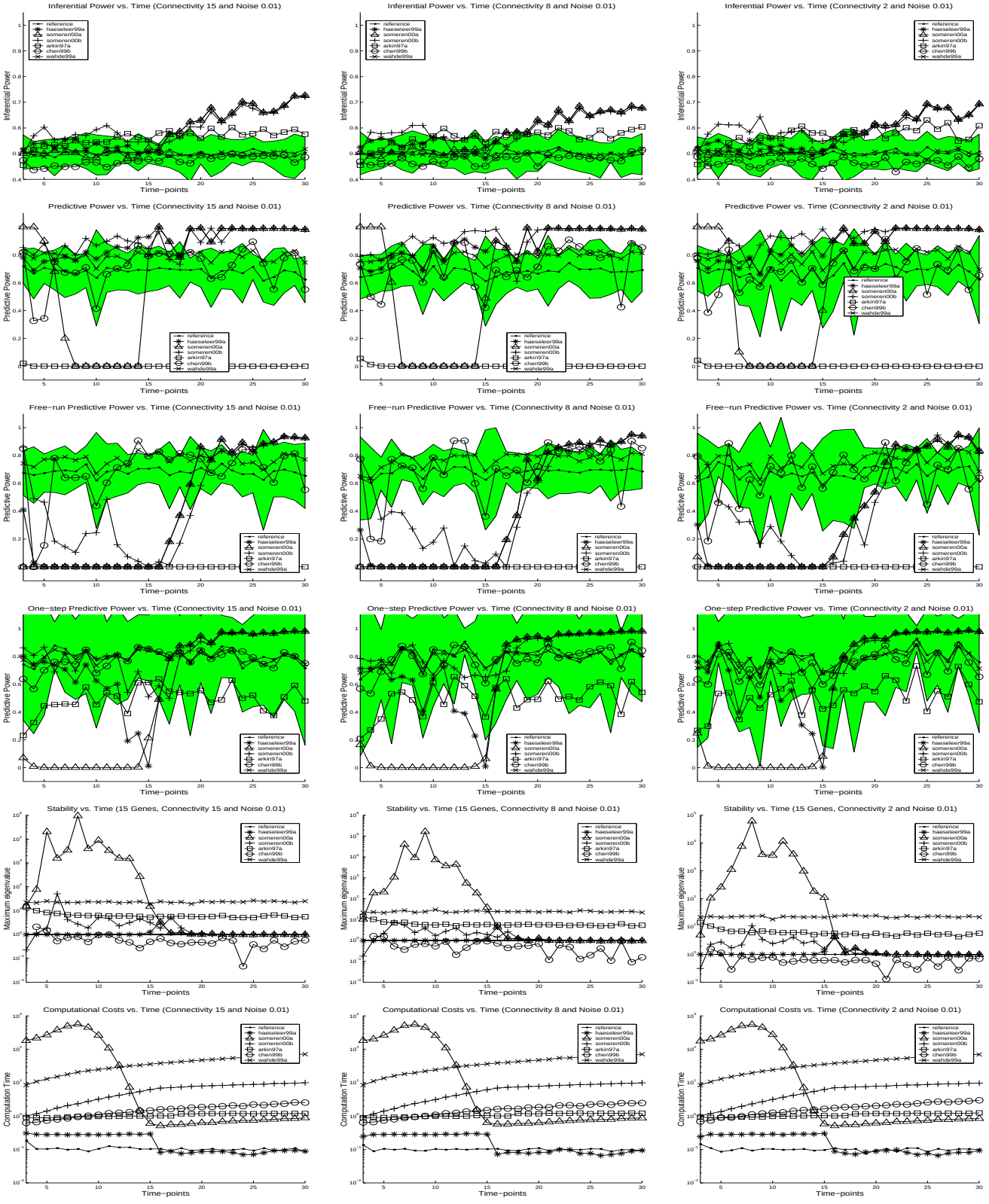

Figure 3. Measurement noise ($\sigma^2 = .01$) experiments of seven genetic network models as a function of no. time-points in the dataset (x-axis). Rows: Six evaluation criteria. Columns: Three different levels of connectivity. See text for more detailed description.

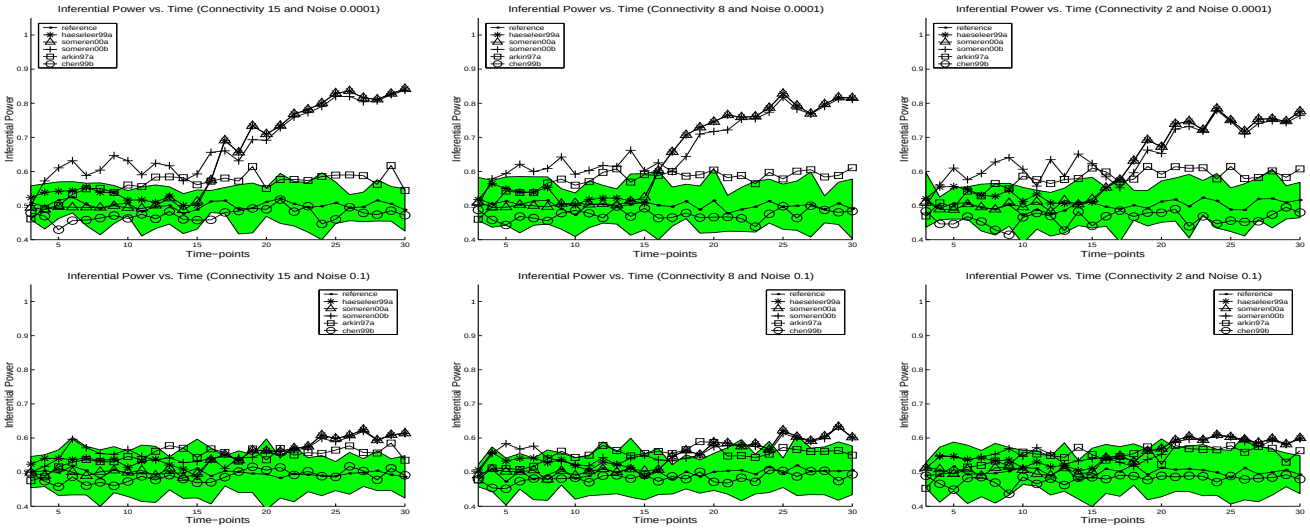

Figure 4. Inferential power of seven different models as a function of no. time-points in the dataset (x-axis). Rows: Different levels of measurement noise (top: very small ($\sigma^2 = .0001$), bottom: more realistic ($\sigma^2 = .1$)). Rows: Three different levels of connectivity. See text for more detailed description.

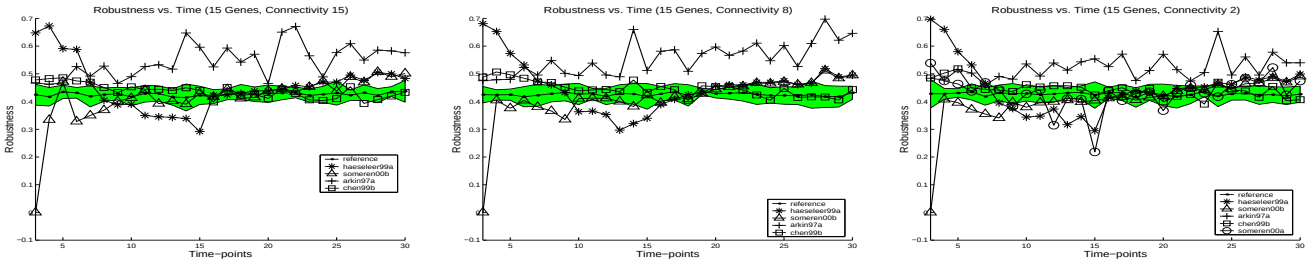

Figure 5. Robustness of seven different models as a function of no. time-points in the dataset (x-axis) and different levels of connectivity (columns) in the case of realistic measurement noise (0.1). See text for more detailed description.

considerable amount of measurement noise. These two aspects are especially typical for time-course gene expression measurements.

A wide variety of models that have been successful in other domains have been proposed for genetic network modeling. However, the relationship between bias and variance that held for other domains no longer holds for the domain of genetic network modeling. This lack of an evident relation between inferential and prediction power is a direct result of the inherent dimensionality problem of gene expression datasets. The limited number of measured time-points compared to the number of genes allows even the most simple models to perfectly fit the given dataset. The disappointing results of some genetic network models provides a strong indication that the relationship between bias and variance of genetic network models should be better understood. Therefore, genetic network modeling can only be made successful, when the variance of the models is better controlled, possibly through the use of biological knowledge. As an example, the Someren00a model that exploits the fact that genetic networks are believed to be sparse, outperforms all other models in the case of extremely under-constrained data. Although this model is of no practical use it shows that the use of biological knowledge can adequately balance the shortcomings of the datasets. Another striking result of the experimental evaluation is the drastic drop in performance of all models when even the slightest amount of measurement noise is introduced. Only the Arkin97a model showed substantial robustness against noise, but exhibits low inferential power. Because none of the implemented models performed well on noisy data, the question arises whether or not it is mathematically possible to distinguish noise from signals when only a limited amount of time-points are available? The use biological knowledge to tackle these kinds of problems seems to be the essential tool in revealing the interactions of our genes.

Acknowledgment

This work was funded by the Intelligent Molecular Diagnostic Systems program of the Delft Inter-Faculty Research Center at the TU Delft.

REFERENCES

1. P. D'Haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: From co-expression clustering to reverse engineering," *Bioinformatics* **16**(8), pp. 707–726, 2000.
2. A. Arnone and B. Davidson, "The hardwiring of development: Organization and function of genomic regulatory systems," *Development* **124**, pp. 1851–1864, 1997.
3. L. Wessels, E. van Someren, and M. Reinders, "A comparison of genetic network models," *Pacific Symposium on Biocomputing 2001* **6**, pp. 508–519, 2001.
4. A. Arkin, P. Shen, and J. Ross, "A test case of correlation metric construction of a reaction pathway from measurements," *Science* **277**, 1997.
5. T. Chen, V. Filkov, and S. Skiena, "Identifying gene regulartory networks from experimental data," *Proceedings of the third annual international conference on Computational molecular biology (RECOMB99)* , pp. 94–103, 1999.
6. E. Mjolsness, T. Mann, R. Castao, and B. Wold, "From coexpression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data," *Advances in Neural Information Processing Systems* **12**, pp. 928–934, 2000.
7. A. Spirov and et al., "The inverse problem for ode models of genetic networks: The determination of parameters of the system from experimentally observed data," *http://academic.mssm.edu/molbio/tmp/circuits/hox1circ.html* , 1997.
8. M. Wahde and J. Hertz, "Coarse-grained reverse engineering of genetic regulatory networks," *IPCAT 99* , 1999.
9. D. Weaver, C. Workman, and G. Stormo, "Modeling regulatory networks with weight matrices," *Pacific Symposium on Biocomputing '99* **4**, pp. 112–123, 1999.
10. E. van Someren, L. Wessels, and M. Reinders, "Information extraction for modeling gene expressions," *Proceedings of the 21st Symposium on Information Theory in the Benelux* , pp. 215–222, 2000.
11. E. van Someren, L. Wessels, and M. Reinders, "Linear modeling of genetic networks from experimental data," *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* , pp. 355–366, 2000.
12. P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear modeling of mrna expression levels during cns development and injury," *Pacific Symposium on Biocomputing '99* **4**, pp. 41–52, 1999.
13. T. Chen, H. He, and G. Church, "Modeling gene expression with differential equations," *Pacific Symposium on Biocomputing '99* **4**, pp. 29–40, 1999.
14. M. Savageau, "Rules for the evolution of gene circuitry," *Pacific Symposium on Biocomputing '98* **3**, pp. 54–65, 1998.
15. A. Wuensche, "Genomic regulation modeled as a network with basins of attraction," *Pacific Symposium on Biocomputing '98* **3**, pp. 89–102, 1998.
16. N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Submitted* , 1999.
17. D. Thieffry and R. Thomas, "Qualitive analysis of gene networks," *Pacific Symposium on Biocomputing '98* **3**, pp. 77–88, 1998.