BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 8 2005, pages 1626–1634
doi:10.1093/bioinformatics/bti226

Systems biology

Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection

Reinhard Guthke[1,*], Ulrich Möller[1], Martin Hoffmann[1], Frank Thies[1] and Susanne Töpfer[2]

[1]Hans Knoell Institute for Natural Products Research, D-07745 Jena, Beutenbergstrasse 11a, Germany and [2]BioControl Jena GmbH, D-07745 Jena, Wildenbruchstrasse 15, Germany

Received on September 3, 2004; revised on November 9, 2004; accepted on December 13, 2004
Advance Access publication December 21, 2004

ABSTRACT

Motivation: The immune response to bacterial infection represents a complex network of dynamic gene and protein interactions. We present an optimized reverse engineering strategy aimed at a reconstruction of this kind of interaction networks. The proposed approach is based on both microarray data and available biological knowledge.

Results: The main kinetics of the immune response were identified by fuzzy clustering of gene expression profiles (time series). The number of clusters was optimized using various evaluation criteria. For each cluster a representative gene with a high fuzzy-membership was chosen in accordance with available physiological knowledge. Then hypothetical network structures were identified by seeking systems of ordinary differential equations, whose simulated kinetics could fit the gene expression profiles of the cluster-representative genes. For the construction of hypothetical network structures singular value decomposition (SVD) based methods and a newly introduced heuristic Network Generation Method here were compared. It turned out that the proposed novel method could find sparser networks and gave better fits to the experimental data.

Contact: Reinhard.Guthke@hki-jena.de

1 INTRODUCTION

Discovering and understanding the complex molecular interactions that make up living organisms is one of the most interesting and challenging problems of modern molecular biology, systems biology and bioinformatics. A commonly accepted top-down approach to unravel the structure of these systems is to reverse engineer gene regulatory networks from experimental time series data (D'haeseleer *et al*., 2000; de Jong, 2002; Csete and Doyle, 2002). Usually, the measured data record spontaneously running processes, like cell division and cell differentiation, or reactions to external stimuli, like responses to bacterial infection (Boldrick *et al*., 2002). The observed changes in gene expression over time are either due to direct effects of the stimulus on specific genes or result from secondary gene–gene interactions. The aim of reverse engineering is then to detect the most likely interactions by identifying sets of relevant model parameters that are required to obtain an appropriate correspondence between measured data and model output. Often the amount and the quality of the experimental data at hand is insufficient for an unequivocal assignment of the model parameters. A widely used approach to resolve this indeterminacy is to favor simple mechanisms (Occam's razor) by requiring the set of model parameters to be minimal. For genetic networks this assumption is further justified by the observation that genetic networks are sparsely connected (Yeung *et al*., 2002 and references therein).

In standard gene expression profiling there are many more variables (N genes) than measurements (M time points). As a consequence, the gene interaction matrix ($N \times N$ entries) of linear models cannot be uniquely determined by the measurement matrix ($N \times M$ entries). Several approaches have been proposed to cope with this problem:

(1) Interpolation and subsequent resampling of the experimental time courses (e.g. D'haeseleer *et al*., 1999) being able to generate almost any number of semi-empirical measurement data (enlargement of M).

(2) Singular value decomposition (SVD) based methods (Holter *et al*., 2001; Yeung *et al*., 2002) that calculate a solution to the interaction matrix by imposing additional mathematical constraints.

(3) Methods for finding sparse interaction matrices by combinatorial search strategies (Chen *et al*., 1999; van Someren *et al*., 2001).

(4) Clustering of gene expression time series (reduction of N) and use of cluster-representatives for subsequent modeling (D'haeseleer *et al*., 2000; Wahde and Hertz, 2000; Mjolsness *et al*., 2000).

The first approach has major drawbacks since it cements microarray measurement errors and introduces some arbitrariness through the choice of interpolation method especially for undersampled data. In the present paper, clustering and a combinatorial search strategy were chosen as the primary approaches to reduce the indeterminacy of the interaction matrix.

Clustering as a means for reducing the number of variables can be justified by the presence of regulatory motifs (D'haeseleer *et al*., 2000). From a system theoretic point of view coarse graining of expression profiles means eliminating redundant information (in terms of indistinguish-ability). However, it has to be done with the highest possible accuracy in order to preserve and extract the existing data structure.

We introduce a novel approach of data-driven reverse engineering that generates probable gene regulation network models based on a combination of optimized clustering and optimized network

*To whom correspondence should be addressed.

reconstruction. While the optimization of clustering concentrates on effective cost function minimization and robust cluster validation, the optimization of network reconstruction is directed to a simultaneous minimization of both the number of interaction parameters and the model error. Both steps, optimized clustering and subsequent optimized network generation, are compared with alternative methods.

The newly proposed approach is demonstrated using data on the immune response of human blood cells to bacterial infection recorded by Boldrick *et al*. (2002). It is compared to established SVD based methods (Yeung *et al*., 2002).

Summarizing, the present reverse engineering approach consists of four steps: (1) data pre-processing, (2) optimized fuzzy clustering and cluster validation, (3) selection of cluster-representative genes by the degree of cluster membership and available biological knowledge, and (4) generation of probable dynamic network models by fitting the simulated kinetics to the experimental expression profiles at hand with a minimum number of model parameters.

2 METHODS

2.1 Data pre-processing

Gene expression data of peripheral blood mononuclear cells (PBMCs) infected by *Escherichia coli* were obtained from the internet (http://genome-www.stanford.edu/ hostresponse/; Boldrick *et al*., 2002). The data represent logarithmized ratios (log-ratios) of the expression intensities of 18 432 genes or ESTs at five time points $t(t = 0.0, 0.5, 1.0, 2.0, 4.0$ h) before and after an infection with heat-killed pathogenic *E.coli*. The log-ratios at $t = 0$ (unperturbed state) were subtracted from the respective time series, i.e. only differences with respect to the pre-infection state were considered. The resulting log-ratios range from -10.4 to 8.7 ($\log_2$-values). A total of 1336 genes was selected by requiring an upregulation or downregulation of at least a factor $8(=2^3)$. For cluster analysis the time profiles were scaled by their respective absolute temporal extreme values to focus on qualitative behavior. Missing data were imputed by using a method based on a k-nearest neighbor algorithm (Troyanskaya *et al*., 2001). The value of k, finally selected from a set of tested values, led to robust clusters and the smallest differences with respect to additionally removed and re-imputed values. For modeling the unscaled log-ratios only data with no missing values were used.

2.2 Clustering and cluster validation

The clustering results subsequently used for network modeling were obtained from the fuzzy C-means (FCM) algorithm (Bezdek and Pal, 1992). FCM was selected as the method of choice after a pre-investigation that comprised several clustering approaches (see Discussion section). The number of clusters was estimated by the vote of 42 cluster validity indices: (1) 18 generalizations of Dunn's index (Bezdek and Pal, 1998), (2) the same 18 generalizations applied to the Davis–Bouldin index (Bolshakova and Azuaje, 2003), (3) the mean cluster silhouette width (Kaufman and Rousseeuw, 1990) and (4) indices proposed by Goutte *et al*. (1999); Ray and Turi (1999); Fadili *et al*. (2001); Kim *et al*. (2001) and Pakhira *et al*. (2004). These indices capture different aspects of a clustering structure.

The FCM algorithm converges to a local optimum. The obtained result is likely to be random because the initial partition can only be chosen heuristically or randomly (Peña *et al*., 1999). Therefore, the estimated number of clusters may also be random, and no algorithmic output quantifies the significance of this estimate. Möller *et al*. (2002) have presented an approach that copes with both problems. An improved version of this approach was used here. Commonly, a validity index vector, $V = (v_2, \ldots, v_C, \ldots, v_{C_{\max}})$, is calculated based on a set, $\Pi = \{\pi_C\}$, of locally optimal candidate partitions, π_C, each with a different number of clusters, C. In the present study, however, the number of clusters was estimated from an array of convergent index curves, $\mathbf{V} = \{V_1, \ldots, V_S\}$, where each curve V_i was calculated from an independent set, Π_i. The convergence of the curves V_i, that allows for a unique estimation in the number of clusters, was achieved step by step: each partition $\pi_{C,i}$ was improved during T runs of the FCM algorithm using random run initialization, and retaining the best result, i.e. the partition $\pi_{C,i}[T_{\text{best}}], T_{\text{best}} \in \{1, \ldots, T\}$, with the smallest value of the FCM objective function. T is a measure of the optimization effort, whereby, here, multiple local optimization is used to approach the global minimum, and one run of the FCM algorithm is the scale unit of the optimization effort.

2.3 Selection of cluster-representative genes

For each cluster one representative gene was selected. The following selection criteria were used: The representative gene

- is assigned to one cluster with a high fuzzy membership degree n (MSD),
- is annotated with a known immunological function, and
- is represented by an expression profile with no missing values.

Subsequently, the expression profiles of the selected genes were used for modeling.

2.4 Dynamic modeling

The dynamics of hypothetic gene regulatory networks was modeled by systems of linear differential equations. Their general solution is a linear combination of exponentially damped (stable) or excited (unstable) oscillations. Apart from its inherent simplicity an advantage of this approach is that linear algebraic methods can be used to fit its free parameters to experimental data. The general mathematical form reads

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \sum_{j=1}^{C} w_{i,j} \cdot x_j(t) + b_i \cdot u(t) \qquad (1)$$

in which $x_i(t)$ is the expression of gene $i = 1, \ldots, C$ at time t, $w_{i,j}$ denotes a gene–gene interaction matrix and b_i represents an external (infection) stimulus response vector. $u(t)$ is the Heaviside step function: $u(t < 0) = 0$ and $u(t \geq 0) = 1$, i.e. the influence of bacterial infection is taken to be constant over time (for 4 h). In addition, the system is assumed to be at equilibrium prior to stimulation, i.e. $\mathrm{d}x_i(t < 0)/\mathrm{d}t = x_i(t < 0) = 0$.

Genetic networks are known to be sparsely connected (Yeung *et al*., 2002 and references therein). The aim of dynamic modeling and network reconstruction is thus to find a minimal set of relevant (i.e. non-zero) model parameters ($w_{i,j}$ and b_i) required to achieve an adequate fit to the expression data at hand.

2.5 Dynamic modeling using SVD

Substantiating Equation (1) for the measuring time points $t_1, \ldots, t_M$ results in a system of linear algebraic equations. The time derivatives have to be estimated from the experimental data points (presently by linear interpolation). Usually, the number of measurements M is smaller than the number of measured genes C rendering the system under-determined [infinite number of solutions for $(\boldsymbol{W})_{i,j} = w_{i,j}$]. Solving the matrix equation by SVD (Holter *et al*., 2001; Yeung *et al*., 2002) involves selection of the matrix $\boldsymbol{W}$ whose rows have the smallest Euclidean (L_2) norm. In addition, the SVD matrix decomposition provides a means of finding a solution, for which the rows of $\boldsymbol{W}$ have the smallest city block (L_1) norm (Yeung *et al*., 2002). Both methods (L_2- and L_1-norm minimization) can be regarded as regularization techniques aimed at finding a minimal set of non-zero model parameters.

2.6 Dynamic modeling using a search strategy

In this paper, a new Network Generation Method for the estimation of the interaction matrix $\boldsymbol{W}$ and the stimulus response vector $\boldsymbol{b}$ according to Equation (1) is proposed. This modeling approach is characterized by an explicit optimization of the model structure.

The method developed employs a heuristic search strategy that separates the structure identification from the parameter identification problem

by examining and comparing models with different connectivity. For each screened model structure the following procedure is performed: (1) The model parameters are fitted to the gene expression data using standard optimization techniques. (2) The resulting model is simulated to obtain the model output. (3) The mean square error (mse) between the model output and the data is determined and is subsequently used to assess the model structure.

Even for small network models it is impractical to consider all possible model structures. Therefore, the Network Generation Method employs a strategy that restricts the search space by directing the search towards simple and plausible model structures and by exploiting prior knowledge concerning the connectivity between genes. The developed approach significantly simplifies the structure search by decomposing the overall identification problem into a number of C separate identification steps, i.e. the submodels for the C gene expression time series are identified separately.

In general, a search strategy consists of three components: an initial model structure, a direction of search and a stopping criterion (van Someren *et al*., 2001). The following search strategy is applied:

Initial submodel structure. The submodel estimation starts with a simple initial submodel that represents a first order lag element. The submodel of gene i possesses two non-zero parameters; the parameter $w_{i,i}$ realizes the selfregulation effect and the parameter b_i describes the influence of the external stimulus on the expression of gene i.

Direction of search. Two directions of search are allowed: forward selection and backward elimination. The method comprises three phases:

(1) In the first phase, a forward selection of the most likely interactions is performed. Thus, the model complexity is increased by adding new gene–gene interactions or stimulus response components. Starting from the initial submodel with two parameters, in the first iteration, all possible submodel structures with three parameters are examined. The best solution with respect to the model fit is retained and further expanded in the next iteration. This so called greedy hill-climbing proceeding is continued until a stopping criterion is met.

(2) The model growing of the first phase bases on the assumption that the best intermediate solution is a part of the best final solution. Since this assumption does not have to be true, unimportant interactions may be included. Therefore, the second phase realizes a backward elimination of gene–gene interactions and stimulus response components. In order to decrease the model complexity all possible solutions that result from the removal of one interaction are considered. Again, the best solution is retained and tested for possible further removals until a stopping criterion is met.

(3) The third phase aims to obtain an improved model fit by adapting the type of dynamic dependency between the interacting genes. The general model structure Equation (1) involves first order dynamics for all submodels. In order to overcome this limitation, the presented Network Generation Method allows to identify submodels that consist of R differential equations and that, consequently, represent lag elements of order R. The search strategy tests different dynamic orders up to a pre-defined maximum dynamic order $R_{\max}$ and selects the best fitting one. Although, the dynamic behavior of the higher order submodels included changes significantly, their allowed parameterization is strongly restricted to transfer functions with R equal poles and no zeros. Higher order submodels are well suited to identify regulatory interactions that are characterized by significant time delays. They preserve the connectivity of the network model and have the form

$$\begin{aligned}\frac{\mathrm{d}x_{i,1}}{\mathrm{d}t} &= \sum_{j\in D_i} w_{i,j}\cdot x_j(t) + w_{i,i}\cdot x_{i,1}(t) + b_i\cdot u(t)\\ \frac{\mathrm{d}x_{i,r}}{\mathrm{d}t} &= x_{i,r-1}(t) + w_{i,i}\cdot x_{i,r}(t), \quad r = 2,\ldots,R-1\\ \frac{\mathrm{d}x_i}{\mathrm{d}t} &= x_{i,R-1}(t) + w_{i,i}\cdot x_i(t)\end{aligned} \tag{2}$$

Here, the genes j with $j \in D_i$ have been found to significantly influence the expression of gene i.

Stopping criterion. In the forward selection mode, interactions are added if the following conditions are met: (1) The increased model complexity leads to a considerably improved model fit. (2) The number of parameters of the expanded submodel is smaller than the number of data points in the corresponding time series. (3) The number of interactions of the expanded submodel stays below a pre-defined limit. (4) In order to avoid overfitting, the mse of the submodel to be expanded is still larger than a pre-defined maximum allowed submodel error $E_{\max}$.

In the backward elimination mode, interactions are removed if the following conditions are fulfilled: (1) The decreased model complexity only leads to a marginally worsened model fit. (2) The sparser model structure remains biologically plausible. Submodel structures with only one non-zero parameter $w_{i,i}$ (the self-regulation parameter) are meaningless with respect to interactions and are generally excluded by the applied search strategy.

Model parameter identification for a given submodel structure is a repeatedly executed operation. In this approach, the parameter identification is performed by a constrained nonlinear optimization algorithm that minimizes the mean square error between the model fit and the pre-processed expression data. The self-regulation parameters, $w_{i,i}$, are constrained by the condition $w_{i,i} < 0$, i.e. the generated submodels are locally stable. Suitable initial parameters for the iterative non-linear optimization are obtained using a linear optimization method. The required time derivatives are calculated based on Hermite interpolation between the data points. These time derivatives are exclusively used in order to find initial parameter values for the iterative nonlinear optimization procedure.

3 RESULTS

3.1 Clustering and cluster validation

Figure 1 shows how the clustering and cluster-validation procedure provided guidance for the visual determination of the number of clusters. According to Section 2.2, each panel of Figure 1 presents the array of validity index curves, $\mathbf{V} = \{V_1, \ldots, V_S\}$, after T runs of clustering. After $T = 5$ runs, the validity index curves exhibited random courses. At this stage partitions were obtained that contained a redundant cluster, and one of the unique expression patterns, present in the data, remained unrecognized. With increasing optimization effort the curves became more similar until (for $T = 100$) they exhibited a consistent pattern with respect to their indicative extrema (Fig. 1) and only then an unequivocal interpretation was possible. The computation effort, T, that was necessary for an unequivocal interpretation depended on (1) the cluster validity index (Fig. 1, $T = 50$; one index yielded estimates of 6 or 7, the other index a unique estimate of 6), (2) the number of clusters, (3) the dataset (result of the pre-investigation, not shown) and (4) other parameters, e.g., the strength of the FCM termination criteria and the fuzzy exponent m. The subsequent results, obtained for $m = 1.5$, proved to be robust, i.e. choosing $m = 2.0$ yielded similar results.

The clear majority vote of the 42 validity indices suggested that the data set has a coarse structure of two clusters, and a finer structure of six clusters. Thirty two indices had their global optimum at $C = 2$. Twenty eight indices exhibited a clear extremum at $C = 6$, being the global optimum for 6 and the first local optimum for 22 of these indices. Because the 6-cluster partition appears to be biologically more meaningful than the 2-cluster partition, it was used in the subsequent modeling study (Table 1, Fig. 2).

3.2 Selection of cluster-representative genes

Table 1 shows the selected genes. The required selection criteria met perfectly with very high MSD (>0.95) and without missing data.

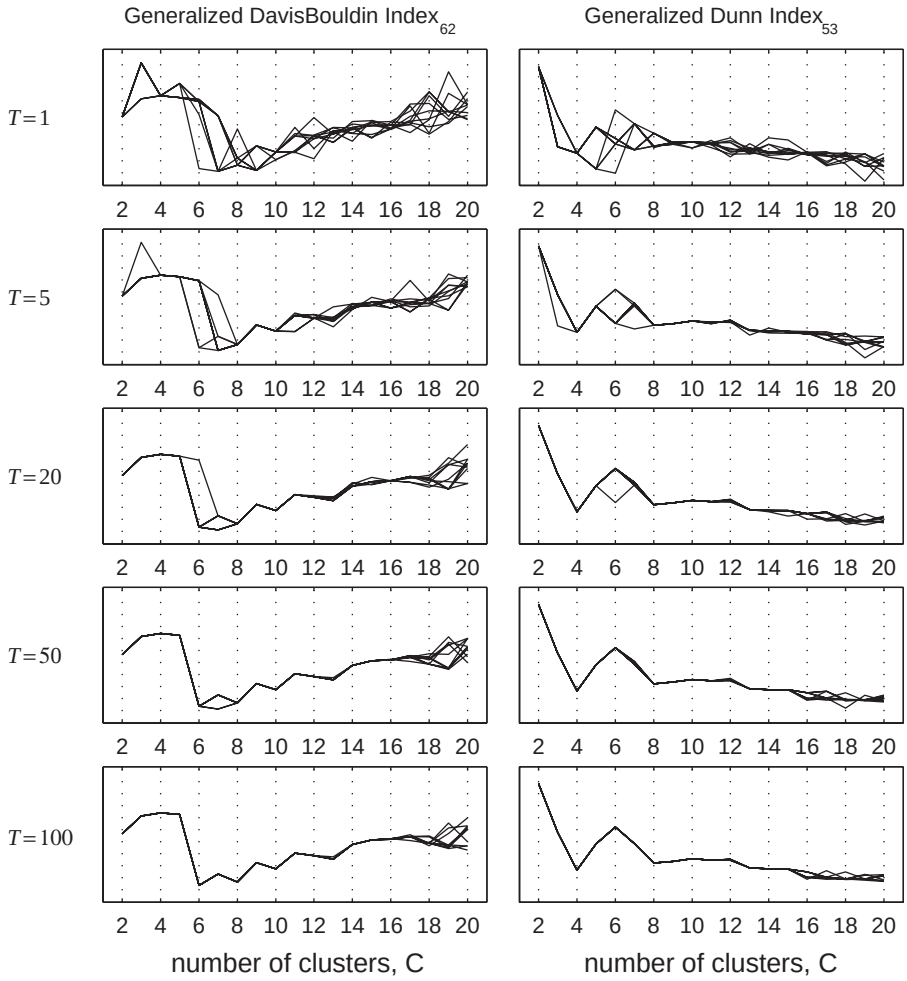

Fig. 1. Array of cluster validity index vectors, $\mathbf{V} = \{V_1, \ldots, V_{10}\}$, recorded after T runs of the FCM algorithm, as a function of the number of clusters, C. The ten curves are superimposed in each box. Left: Generalized Davis–Bouldin index (DBI) with the Hausdorff metric for measuring the distance between clusters, and the average interpoint distance for measuring the cluster diameter (DBI scale: 0.8–1.1). Right: Generalized Dunn index (DI) with the average-to-centroid distance between clusters and the points-to-centroid distance for the cluster diameter. (DI scale: 0.4–1.3). A minimum of the DBI and a maximum of the DI are estimates in the number of clusters. More than one clear extremum indicates structure at different levels of resolution.

Table 1. Number N of genes belonging to cluster c ($c = 1, \ldots, C$) with the MSD >50% as well as the selected representative genes (MSD Symbol and Function)

c	*N*	MSD	Symbol	Function
1	494	0.992	*IL1A*	Interleukin 1, alpha
2	269	0.958	*CD59*	Antigen
3	97	0.989	*NFKBIE*	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon
4	67	0.999	*STAT1*	Signal transducer and activator of transcription 1
5	137	0.995	*STAT5A*	Signal transducer and activator of transcription 5A
6	188	1.000	*HLA-DMA*	Major histocompatibility complex II, DM alpha

The cluster means and standard deviations are shown in Figure 2.

IL1A, *NFKBIE*, *STAT1*, *STAT5A* and *HLA-DMA* are known to be involved in immune response after infection.

3.3 Dynamic modeling

Figure 3 shows the gene expression kinetics obtained from SVD-based dynamic modeling according to Equation (1). The simulated kinetics for the L_2- and L_1-fits are graphically indistinguishable. For the L_2-approach all of the 42 possible parameters, i.e. 36 gene–gene interaction coefficients $w_{i,j}$ and 6 stimulus associated coefficients b_i, are present (fully connected network). The L_1-fit according to Yeung *et al*. (2002) reduces the number of non-vanishing model parameters n to 31 (mse = 1.512).

The optimized model structures obtained from the proposed Network Generation Method configured with a maximum allowed submodel error of $E_{\max} = 1$ and a maximum dynamic order of $R_{\max} = 1$ and $R_{\max} = 3$ are shown in Figures 4 and 5, respectively. The number of non-zero parameters n was reduced to 14 and 15,

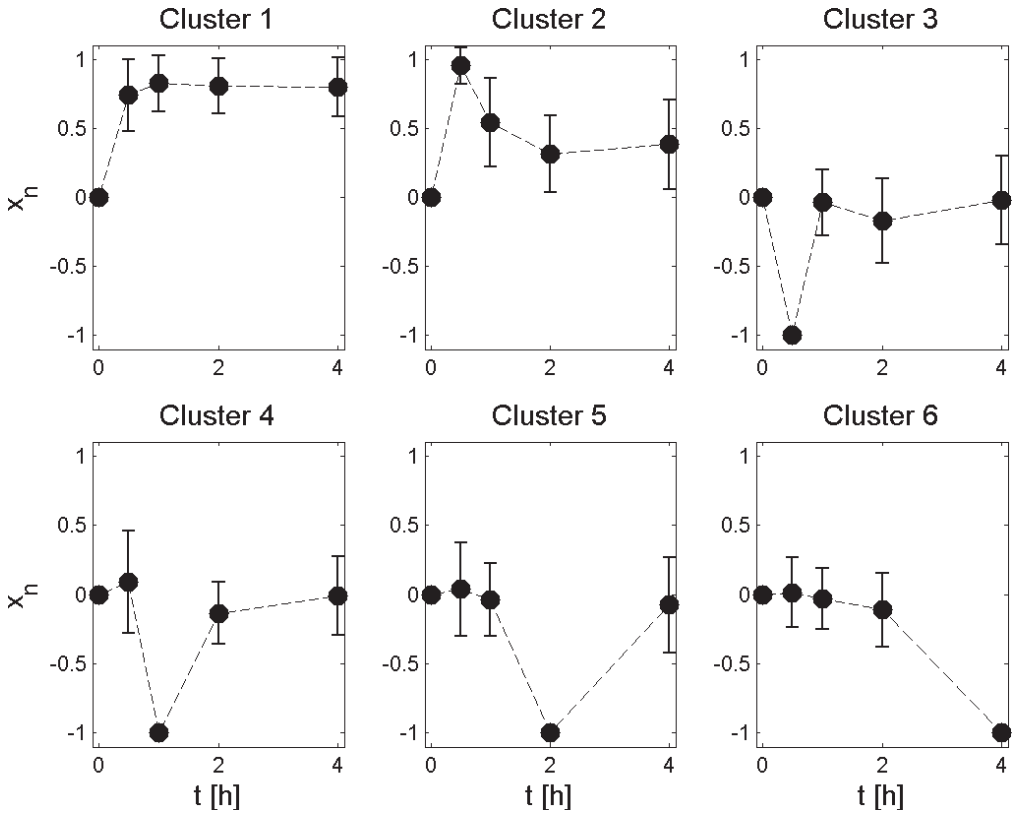

Fig. 2. Result of the FCM clustering with six clusters: mean normalized gene expression profiles with standard deviation averaged over the N genes for the respective cluster (Table 1).

respectively. Model (3) describes the structure of Figure 4 in detail. (The variables $x_1, \ldots, x_6$ are the log-ratios of *IL1*, *CD59*, *NFKBIE*, *STAT1*, *STAT5A* and *HLA-DMA*, respectively.)

$$\begin{aligned}
\frac{dx_1}{dt} &= -2.99 \cdot x_1 - 14.8 \cdot u(t) \\
\frac{dx_2}{dt} &= -2.41 \cdot x_1 - 2.20 \cdot x_2 + 14.9 \cdot u(t) \\
\frac{dx_3}{dt} &= 4.43 \cdot x_1 - 2.03 \cdot x_3 - 21.5 \cdot u(t) \\
\frac{dx_4}{dt} &= 2.15 \cdot x_3 - 1.52 \cdot x_4 \\
\frac{dx_5}{dt} &= 2.59 \cdot x_4 + 2.31 \cdot u(t) \\
\frac{dx_6}{dt} &= -1.02 \cdot x_1 + 3.82 \cdot u(t)
\end{aligned} \tag{3}$$

The simulated kinetics are displayed in Figure 6. The mse was 0.6304 and 0.1710, respectively. The influence of two configuration parameters, the maximum dynamic order $R_{\max}$ and the maximum allowed submodel error $E_{\max}$ was investigated. For $R_{\max} = 2$, a similar structure to that for $R_{\max} = 3$ (Fig. 5) was obtained. However, it contained only second order lag elements for *CD59* and *STAT1* and had an error of mse $= 0.2337$. Setting $E_{\max} = 2$ and $R_{\max} = 1$ resulted in a structure that preserved the interrelations between *IL1A*, *NFKBIE* and *HLA-DMA* in comparison with those shown in Figures 4 and 5 ($n = 13$; mse $= 0.5250$).

Randomly disturbed input data were used for a bootstrapping study to assess the impact of measurement error and test the reliability of the structures generated. The analyses were repeated 1000 times using input data obtained by adding normal distributed random deviates with a standard deviation σ. With $R_{\max} = 1$, $E_{\max} = 1$ and $\sigma = 0.1$ the structure shown in Figure 4 was confirmed 961 times, i.e. in 96% of the cases, except for the negative link from *IL1A* to *CD59* which was found only 499 times. The exciting cascade from the infection via *IL1A* to *NFKBIE* as well as the inhibitory link from infection to *NFKBIE* was found to be the consensus structure for all 1000 runs with $\sigma = 0.1$, 896 runs (90%) with $\sigma = 0.5$ and 645 times (65%) with $\sigma = 1.0$.

4 DISCUSSION AND CONCLUSION

The current study proposes a systems biology approach to analyze the dynamic behavior of the immune response to bacterial infection. It demonstrates how to reconstruct the structure and dynamics of a functional module of the immune system by analyzing stimulus–response data from perturbation experiments and by using available knowledge. The reverse engineering approach presented in this paper combines clustering techniques with network inference. Similar ideas have already been published (D'haeseleer *et al.*, 2000; Mjolsness *et al.*, 2000; Wahde and Hertz, 2000). However, both methods were optimized in this work.

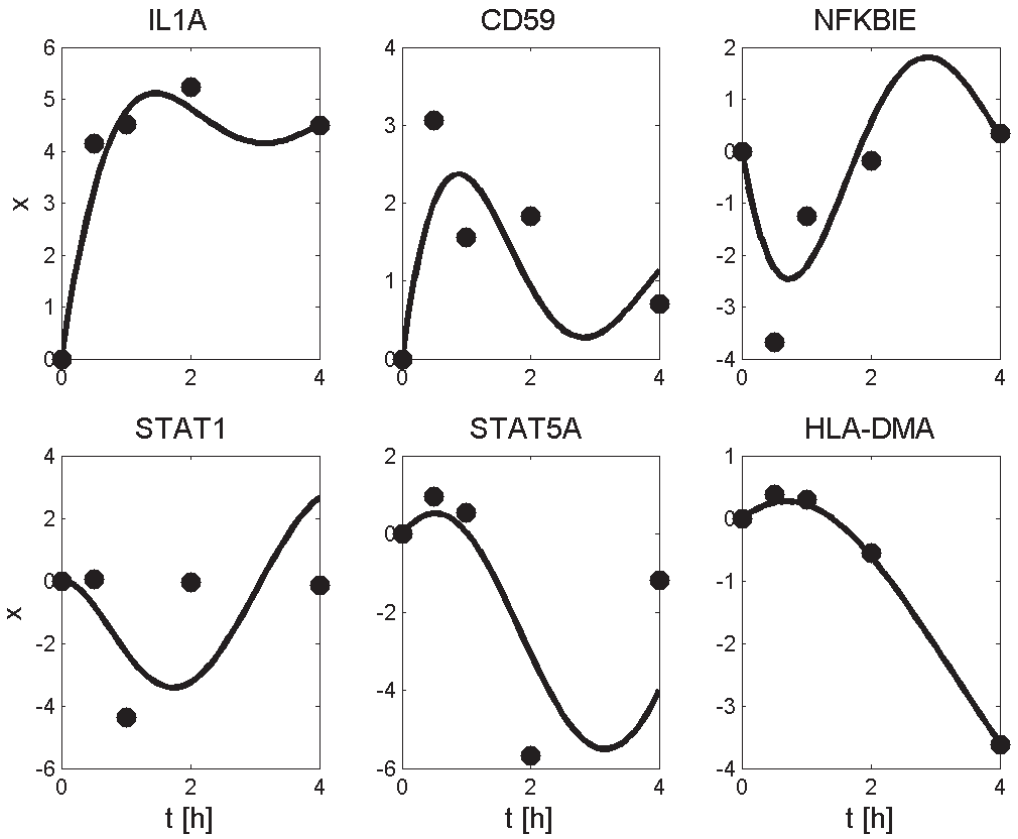

Fig. 3. Measured and simulated expression kinetics (log-ratios) for the genes selected as representatives of the 6 clusters (Table 1). The simulated kinetics (lines) were obtained from Equation (1) and SVD. mse = 1.512.

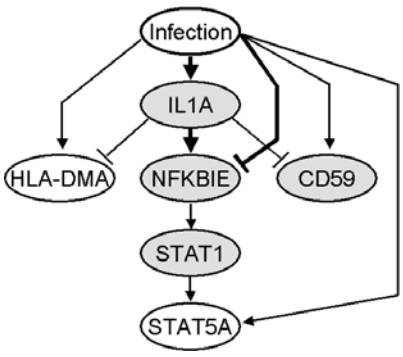

Fig. 4. Structure of the dynamic system described by Equation (3) for the gene expressions of the representatives of clusters 1–6 generated by the proposed Network Generation Method configured by $R_{\max} = 1$ and $E_{\max} = 1$. The arrows represent stimuli or activations. The T-shaped links ($\perp$) represent inhibitions. Grey boxes denote elements with non-zero (decay or self-regulation) elements $w_{i,i}$. The thick links indicate the connections confirmed by bootstrapping.

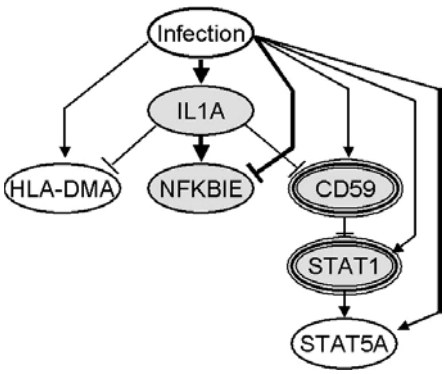

Fig. 5. Alternative structure of the dynamic system with third order time lag elements for *CD59* and *STAT1* obtained from the proposed Network Generation Methods configured by $R_{\max} = 3$ and $E_{\max} = 1$.

The proposed algorithm was also applied to gene expression time series with more time points, e.g. recordings of the *E.coli* stress response during recombinant protein expression (Schmidt-Heck *et al*., 2004). In the present study, we focused on an application with few time points which is typical and most common in infection biology research due to the high costs of microarrays.

The reverse engineering approach proposed in this paper consists of four steps: (1) data pre-processing, (2) data clustering and cluster validation, (3) selection of representative genes and (4) dynamic modeling of the kinetic behavior of the cluster-representatives. The aim of the first two steps is to reduce the number of variables (in

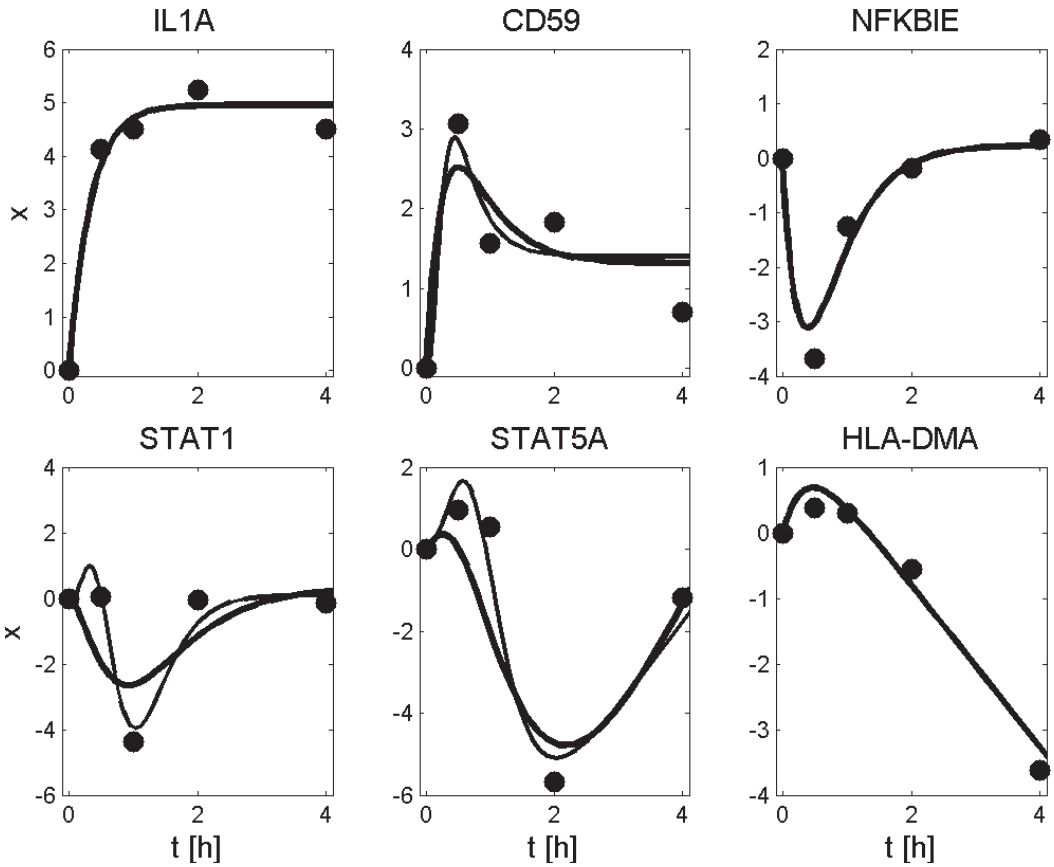

Fig. 6. Measured and simulated expression kinetics for the genes selected as representatives of the six clusters. Simulations using optimised model structures shown in Figures 4 (thick lines; mse = 0.6304) and 5 (thin lines, mse = 0.1710).

our example from 18432 to 6), i.e. to identify the main components that represent the dynamic answer to the infection stimulus.

From a statistical learning perspective, clustering is subdivided into (1) combinatorial algorithms, (2) mixture modeling and (3) mode seeking (Hastie *et al*., 2001). No single tool has emerged as the method of choice for gene expression analysis. We selected type (1), because it is most widely applied. First, several algorithms were tested on various simulated and gene expression data. Hierarchical clustering, with 32 combinations of linkage method and distance measure, provided highly inconsistent results. The 'best' cluster trees, with the largest cophenetic coefficient, led to inappropriate partitions. Prototype-based clustering together with validity indices captured known (simulated) clustering structures more adequately. Here, the best results of the fuzzy (FCM) analysis yielded stronger evidence of the clusters than hard clustering based on a local or global optimization scheme (Möller *et al*., 2002). Self-organizing maps (SOMs) depended on the map size, where a novel SOM validation (Wu and Chow, 2004) often failed to estimate the number of non-trivial simulated clusters.

The correct estimation of the number of clusters is a fundamental issue. This number directly affects the inferred network model by determining the number of network nodes. A wrong number of clusters may thus lead to an inappropriate model and misleading biological conclusions. Therefore, the FCM result served as the input for the proposed Network Generation Method. Advantages of utilizing FCM have already been presented (Guthke *et al*., 2000; Gasch and Eisen, 2002).

One may view the above prototype-based clustering (PC) as an alternative choice to the model-based clustering (MC) used, e.g. by Mjolsness *et al*. (2000). Whereas MC involves a statistical model choice problem (Yeung *et al*., 2001), PC includes a parameter choice problem. A novel solution is proposed here for the local optima problem that occurs in both the MC and PC approaches. This solution is a monitoring of the change in cluster validity measures depending on the computational effort for solving the local optima problem. Spurious random clusterings, due to a limited computational effort, can be avoided. The novel approach relieves the user of a critical part of the parameter choice problem, i.e. of a heuristic decision that is difficult to make. This can be interpreted as one option to increase the 'accuracy' of microarray data analysis (Vilo and Kivinen, 2001; Campbell, 2003). The procedure offers room to further optimize the calculations, e.g. for a particular dataset, number of clusters, and algorithmic parameters (such as the fuzzy exponent). Nevertheless, our type of multistep analysis is only one possibility. The choice of a suitable cluster validity index is a problem with a long history which has not yet been solved (cf. Pakhira *et al*., 2004). Some indices are correlated, because they quantify similar partition properties. However, if the approach of relative cluster validity has become

the method of choice, votes of different indices for the same value tend to increase the confidence (cf. Bezdek and Pal, 1998). Other techniques, including resampling (Dudoit and Fridlyand, 2002) and bootstrapping (Hastie *et al*., 2001), are worth of being considered in future studies.

The nodes of the gene regulatory network were selected from a sorted list of genes ranked by the fuzzy MSD obtained from cluster analysis (Table 1). Due to the currently limited knowledge about the physiological function of genes and their translational products, this selection is somewhat arbitrary and other genes may be considered as well. For instance, *IL6* as well as *TNF*α can be used as representatives of cluster 1 instead of *IL1A*. Similarly, *STAT6* can be selected for cluster 2, the CCAAT-box-binding protein for cluster 3, the *MAP kinase 4* for cluster 5 and *CD31* for cluster 6. Cluster-representative genes were selected from the fuzzy membership ranked gene list by using the available expert knowledge. This can be supported by text mining tools (Shatkay and Feldman, 2003; Chiang *et al*., 2004), e.g. by searching for known links between the infection stimulus and the considered genes. The literature hit rate resulting from such searches can be combined with (multiplied by) the fuzzy MSD obtained from cluster analysis in order to obtain a ranking score for the cluster-representative gene.

The network models obtained from the SVD procedure are not optimal with respect to a low number of model parameters and a low mse. The modeling results, and specifically the reconstructed network connectivities, strongly depend on the actual values assumed for the time derivatives. However, due to the sparseness of the gene expression data the time derivatives cannot be determined reliably. From a system identification point of view (Ljung, 1999), the SVD method realizes a prediction error identification that leads to biased parameter estimates in the presence of measurement noise. The proposed Network Generation Method, on the other hand, realizes a model output identification and thus circumvents both drawbacks (i.e. the need for time derivates and the bias of the estimated parameters).

Nevertheless, the solution of nonlinear optimization problems is very time-consuming. The separation of the whole identification problem into distinct subproblems significantly alleviates this problem, since each submodel parameter optimization involves a few parameters only.

The solution according to Yeung *et al*. (2002) in which the rows of the interaction matrix have the smallest possible city block (L_1) norm reduced the number of non-vanishing model parameters from 42 to 31 while leaving the time courses almost unchanged. The proposed Network Generation Method, on the other hand, optimized the model structure by minimizing the number of non-vanishing model parameters as well as the mse. For the immune response problem studied here the number of parameters was reduced from 42 to 14 (Fig. 4) and 15 (Fig. 5), i.e. by 67 and 64%, respectively. The mse was reduced from 1.5 (Fig. 3) to 0.63 and 0.17 (Fig. 6), i.e. by 58 and 89%, respectively.

In the presented reverse engineering approach the inclusion of available knowledge is possible through the selection of cluster-representative genes (Table 1) and by the configuration of the algorithm. Different configurations can generate different model structures. The influence of two configuration parameters, the maximum dynamic order $R_{\max}$ and the maximum allowed submodel error $E_{\max}$ was illustrated. The links found between the infection stimulus, *IL1A*, *NFKBIE* and *HLA-DMA* were found to be stable for several parameter values $E_{\max}$ and $R_{\max}$. We used $R_{\max} = 1$ and $E_{\max} = 1$ as default configuration. $R_{\max} = 1$ means starting with the simplest model (first order lag element). $E_{\max}$ should be related to the experimental noise. $E_{\max} = 1$ means that a fold change >2 is considered to be a significant change. In general, the complexity of the model and the number of model parameters increase when $R_{\max}$ is increased and $E_{\max}$ is decreased.

Prior knowledge that concerns the existence or absence of either gene–gene interactions or the influence of environmental factors can be included in the proposed Network Generation Method by pre-specification of initial submodel structures. The pre-defined interactions or stimulus–response components are preserved by the search strategy. For instance, the interactions between infection, *IL1A* and *NFKBIE* highlighted in Figures 4 and 5 could be used as advance information for further studies (data not shown). The cascade from the infection stimulus via *IL1A* to *NFKBIE* and the fact that *NFKBIE* is primarily down-regulated by the infection was found as a consensus structure for different configurations (Figs 4 and 5) and bootstrapping and can therefore be considered to be highly probable. This finding is corroborated by biological knowledge since *NFKB* that is inhibited by *NFKBIE* is a transcription factor involved in inflammatory immune response. The present results suggest the following response mechanism. The infection stimulates the expression of *NFKB* dependent genes via pro-inflammatory cytokine effected phosphorylation and subsequent degradation of *NFKB* inhibitor proteins (IkBs) such as *NFKBIE* (*IL-1* signal transduction pathway). In addition *NFKBIE* turns out to be transcriptionally suppressed by the infection stimulus, thereby enhancing the transcription of *NFKB* dependent genes such as *IL-1*. Evidently, IL-1 in turn induces NFKBIE expression as a counter-regulation and thus limits its own over-expression and that of other *NFKB* dependent genes.

In order to ensure network model plausibility, a submodel is required to have a non-zero, negative self-regulation parameter $w_{i,i}$. Positive self-regulation parameters lead to locally unstable submodels and are excluded by the method. However, self-regulation parameters with zero value cannot be avoided, if a gene expression time series, such as for *STAT5A* (x_5) and *HLA-DMA* (x_6), has not yet reached a steady state during the measurement. Then, any parameter optimization algorithm sets the corresponding self-regulation parameter $w_{i,i}$ to $\sim$0 and, therefore, the applied search strategy removes this parameter in the backward elimination mode [i.e. $w_{5,5} = w_{6,6} = 0$ in Equation (3)]. Then, the reconstructed interactions of the respective genes are less reliable than those estimated for time series that have reached a steady state. Thus, the reconstruction of regulatory interactions concerning *STAT5A* and *HLA-DMA* is quite vague.

The simulated gene expression for *NFKBIE* and *STAT1* reach stationary values near the initial ones (log-ratios of 0.2 and 0.3, respectively), whereas those of *IL1* and *CD59* reach up-regulated stationary values (log-ratios of 4.9 and 1.4, respectively). Thus, *NFKBIE* and STAT1 are down-regulated only temporarily, whereas *IL1* and *CD59* are permanently up-regulated during the infection.

The relaxation to a unique steady state is a general property of stable linear differential equations with a constant external forcing. Multiple steady states as observed for some biological systems are caused by non-linearities. Non-linear terms can be included in the proposed modeling algorithm when they are pre-defined from prior knowledge. The automatic identification of additional non-linear model terms in general requires more independent experimental

data in order to ensure a stable convergence of the algorithm to a unique model structure. Due to the wide range of expression values logarithmized data are preferred for analysis. Modeling the non-logarithmized data instead of the log-ratios confirmed the three links between 'infection', *IL1* and *NFKBIE* shown as thick lines in Figure 4.

The proposed Network Generation Method identifies differential equation systems from measured time courses and available knowledge directly. Thus, it suggests qualitative biological relations between the considered genes. This data- and knowledge-driven modeling allows to generate models that represent alternative hypotheses for the underlying gene regulatory network. Incorporating information about the measurement error (in terms of the maximum allowed submodel error $E_{\max}$) and the available biological knowledge on immune response can help to select plausible network structures. Given alternative network structures (differing e.g. in whether STAT1 is activated by NFKBIE as shown in Figure 4 or inhibited by CD59 as shown in Fig. 5) the corresponding dynamic models can be used to design suitable perturbation experiments aimed at an optimal discrimination between these structures (Ideker *et al*., 2000). Having in mind the large number of expressed genes, the sparseness of genetic networks and the limitations of today's biological knowledge, the present study has shown that the concerted application of optimized clustering methods, data- and knowledge-driven reverse engineering and experimental planning is a viable and promising approach to enlighten the jungle of biochemical networks.

ACKNOWLEDGEMENTS

We thank Dörte Radke and Andreas Fischer for their support in the implementation of cluster validation methods. This work has been supported by the German Federal Ministry of Education and Research (BMBF, grants no. 0312704 D).

REFERENCES

Bezdek,J.C. and Pal,S.K. (1992) *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*. IEEE Press, New York.

Bezdek,J.C. and Pal,N.R. (1998) Some new indexes of cluster validity. *IEEE Trans. Syst. Man Cybern.*, **B28**, 301–315.

Boldrick,J.C., Alizadeh,A.A., Diehn,M., Dudoit,S., Liu,C.L., Belcher,C.E., Botstein,D., Staudt,L.M., Brown,P.O. and Relman,D.A. (2002) Stereotyped and specific gene expression programs in human innate immune response to bacteria. *Proc. Natl Acad. Sci., USA*, **99**, 972–977.

Bolshakova,N. and Azuaje,F. (2003) Cluster validation techniques for genome expression data. *Signal Process.*, **83**, 825–833.

Campbell,C. (2003) New analytical techniques for the interpretation of microarray data. *Bioinformatics*, **19**, 1045.

Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **4**, 29–40.

Chiang,J.-H., Yu,H.-C. and Hsu,H.-J. (2004) GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*, **20**, 120–121.

Csete,M.E. and Doyle,J.C. (2002). Reverse engineering of biological complexity. *Science*, **295**, 1664–1669.

De Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.

D'haeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

D'haeseleer,P., Wen,X., Fuhrman,S. and Somogyi,R. (1999) Linear modeling of mRNA expression levels during CNS development and injury, *Pac. Symp. Biocomput.*, **4**, 41–52.

Dudoit,S. and J. Fridlyand (2002) A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biol.*, **3**, 0036.1–0036.21.

Fadili,M.J., Ruan,S., Bloyet,D. and Mazoyer,B. (2001) On the number of clusters and the fuzziness index for unsupervised FCA application to BOLD fMRI time series. *Med. Image Anal.*, **5**, 55–67.

Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol.*, **3**, Research0059.1–0059.22

Goutte,C., Toft,P., Rostrup,E., Nielsen,F. and Hansen,L.K. (1999) On clustering fMRI time series. *NeuroImage*, **9**, 298–310.

Guthke,R., Hahn,D., Fahnert,B., Kroll,T. and Wölfl,S. (2000): Gene expression data mining by fuzzy C-means clustering and fuzzy rule generation. *Proceedings of the 11th International Biotechnology Symposium*, Berlin, Vol. 1, pp. 230–232.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning.* Springer, New York.

Holter,N.S., Maritan,A., Cieplak,M., Fedoroff,N.V. and Banavar,J.R. (2001) Dynamic modeling of gene expression data, *Proc. Natl Acad. Sci. USA*, **98**, 1693–1698.

Ideker,T.E., Thorsson,V. and Karp,R.M. (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.*, **5**, 305–316.

Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data*. Wiley, New York.

Kim,D.J., Park,Y.W. and Park,D.J. (2001) A novel validity index for determination of the optimal number of clusters. *IEICE Trans. Inf. Syst.*, **E84-D(2)**, 281–285.

Ljung,L. (1999) *System Identification—Theory for the User*. Prentice Hall, Upper Saddle River, NJ.

Mjolsness,E., Mann,T., Castano,R. and Wold,B. (2000) From coexpression to coregulation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data. In Stolla,S.A., Leen,T.K. and Muller,K.R. (eds) *Advances in Neural Information Processing Systems* 12. MIT Press, Cambridge, MA, pp. 928–934.

Möller,U., Ligges,M., Georgiewa,P., Grünling,C., Kaiser,W.A., Witte,H. and Blanz,B. (2002) How to avoid spurious cluster validation? A methodological investigation on simulated and fMRI data. *NeuroImage*, **17**, 431–446.

Pakhira,M.K., Bandyopadhyay,S. and Maulik,U. (2004) Validity index for crisp and fuzzy clusters. *Pattern Recogn.*, **37**, 487–501.

Peña,J.M., Lozano,J.A. and Larrañaga,P. (1999) An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recogn. Lett.*, **20**, 1027–1040.

Ray,S. and Turi,R.H. (1999) Determination of number of clusters in K-means clustering and application in colour image segmentation. In Pal,N.R., De,A.K. and Das,J. (eds), *Proceedings of ICAPRDT'99*, Calcutta, India, pp. 137–143.

Schmidt-Heck,W., Guthke,R., Toepfer,S., Reischer,H., Dürrschmid,K. and Bayer,K. (2004) Reverse engineering of the stress response during expression of a recombinant protein. *Proceedings of the EUNITE 2004 Conference*,Verlag Mainz, Aachen, pp. 407–412.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing values methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

van Someren,E.P., Wessels,L.F.A., Reinders,M.J.T. and Backer,E. (2001) Searching for limited connectivity in genetic network models. *Proceedings of the International Conference on Systems Biology*, Pasadena, CA, pp. 222–230.

Vilo,K. and Kivinen,K. (2001) Regulatory sequence analysis: application to the interpretation of gene expression. *Eur. Neuropsychopharmacol.*, **11**, 399–411.

Wahde,M. and Hertz,J. (2000) Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, **55**, 129–136.

Wu,S. and Chow,T.W.S. (2004) Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recogn.*, **37**, 175–188.

Yeung,K.Y., Fraley,C., Murua,A., Raftery,A.E. and Ruzzo,W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Yeung,M.K., Tegner,J. and Collins,J.J. (2002) Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl Acad. Sci., USA*, **99**, 6163–6168.