Predicting Single Genes Related to Immune-Relevant Processes

C. Spieth*, F. Streichert*, N. Speer*, C. Sinzger†, K. Eberhard†, and A. Zell*
*Centre for Bioinformatics,
University of Tübingen, 72076 Tübingen, Germany
†Institute of Medical Virology,
University Hospital Tübingen, 72074 Tübingen, Germany

***Abstract*— In this paper we address the problem of predicting gene activities by finding gene regulatory dependencies in experimental DNA microarray data. Only few approaches to infer the dependencies of complete gene interconnectivity networks can be found in the literature. Due to the limited number of available data, the inferring problem is under-determined and ambiguous. Therefore, we introduce a new algorithm to infer relationships only between selected genes and the unknown gene network. This method is able to predict gene activation by mathematical modeling of the network and its simulation. The parameters of the mathematical model are determined by optimization with evolutionary algorithms. In this paper we will show that our approach is able to correctly predict gene responses in immune related regulatory processes and correctly identify some of the true genomic relationships of these genes.**

I. Introduction

Recently developed DNA microarray technology allows measurement of gene expression levels for a whole genome at the same time. Experiments using this technique provide new insights into activities of genes under different biochemical and physiological environment conditions and can therefore be used to extract relationship information of interacting genes. A gene interconnectivity network (GIN) defines the complex structure of dependencies of mRNA produced by one expressed gene influencing regulatory mechanisms of other genes. The amount of expression data grows rapidly because this technique allows for high-throughput experiments. And although increasing numbers of microarray data sets become available, mathematical methods are still infeasible to determine genome-wide regulatory networks from a small number of chips.

In this paper we propose a methodology for reverse engineering the dependencies of selected genes only, which are of special interest. The relationships between the specified gene and the other genes within the network, i.e. within the data set, are modeled mathematically. Due to the complexity of the inference problem some researchers suggested evolutionary algorithms (EA) for this purpose. We introduce an extended EA framework for evolving the mathematical models to eventually infer the parameters of these models. And in contrast to previous publications, we are not trying to reconstruct the complete network but only inferring relationships between a priori selected genes and the unknown gene network system to predict their regulatory dependencies and activation dynamics.

Section II of this publication presents an overview over related work and lists associated publications. A description of the proposed method is given in section III and applications for inferring immune-relevant genes are shown in section V. Finally, conclusions and an outlook are covered by section VI.

II. Related Work

Inferring the underlying relationships between genes is subject to current research and has recently become one of the major topics in bioinformatics and in systems biology due to the increased computing power available. There already have been some approaches in the field of system biology to solve the combinatorial problem of the inference process. A good overview of related work can be found in [6].

The earliest models to simulate regulatory systems found in the literature are Boolean or Random Boolean Networks (RBN) [13]. In Boolean Networks gene expression levels can be in one of two states: either 1 (on) or 0 (off). The quantitative level of expression is not considered. Two examples for inferring Boolean Networks are given by Akutsu *et al.* [1] and the REVEAL algorithm [18] by Liang *et al.* These models have the advantage that they can be solved with only small computational effort. But they suffer from the disadvantage of being tied to discrete system states. In contrast, qualitative network models allow for multiple levels of gene regulation. An example for this kind of approach is given by Thieffry and Thomas in [29]. But these models use only qualitative dependencies and therefore only a small part of the information hidden in the time series data. Quantitative models based on linear models for gene interconnectivity networks like the weighted matrix model by Weaver *et al.* [32] or the singular value decomposition method by Yeung *et al.* [33] consider the continuous level of gene expression. Other approaches to infer regulatory systems from time series data by using Artificial Neural Networks [14] or Bayesian Networks [11] have been recently published, but face some drawbacks as well. Bayesian networks, for example, do not allow for cyclic networks. More general examples for mathematical non-linear models like S-systems to infer regulatory mechanisms have been examined by Maki *et al.* [19] or Kiguchi *et al.* [15]. And in stochastic models, e.g. Bayesian networks, the dependencies between the components of a system are modeled by probabilistic transition

0-7803-9387-2/05/$20.00 ©2005 IEEE

values. There are many publications on this kind of model, examples can be found in [10], [7].

So far, only parameterized models have been considered in this section. Other approaches, for example non-parameterized models, to infer regulatory systems from time series data using artificial neural networks [14] or bayesian networks [11] have been recently published, but face some drawbacks as well. Bayesian networks, for example, do not allow for cyclic networks, which are known to exist in biological systems. Another kind of non-parameterized model are arbitrary differential equations, which can also be used to model regulatory structures as Ando *et al.* showed with genetic programming (GP) in [2]. In this publication, GP was used to set up a system of suitable differential equations, whose time series were then compared to the experimental data.

III. Mathematical And Computational Models

On an abstract level, the behavior of a cell is represented by a directed graph with N nodes representing N genes. Each gene g_i produces a certain amount of mRNA x_i when expressed and changes the concentration of the mRNA level over time: $\vec{x}(t+1) = h(\vec{x}(t))\,, \vec{x}(t) = (x_1, \cdots, x_n)$. Here, function h represents the changes of the vector of expression levels from one state to the next.

A. Model

To model and to simulate regulatory networks we decided to use S-systems since they are well-documented and examined and are flexible. Their drawback is that they have $2(N^2+N)$ parameters for a network with N genes. S-systems are a type of power-law formalism that has been suggested by Savageau [23] and can be described by a set of nonlinear differential equations:

$$\frac{dx_i(t)}{dt} = \alpha_i \prod_{j=1}^{N} x_j(t)^{\mathcal{G}_{i,j}} - \beta_i \prod_{j=1}^{N} x_j(t)^{\mathcal{H}_{i,j}} \qquad (1)$$

where $\mathcal{G}_{i,j}$ and $\mathcal{H}_{i,j}$ are kinetic exponents, α_i and β_i are positive rate constants and N is the number of equations in the system. The equations of (1) can be seen as divided into two components: an excitatory and an inhibitory component. The kinetic exponents $\mathcal{G}_{i,j}$ and $\mathcal{H}_{i,j}$ determine the structure of the regulatory network. In the case $\mathcal{G}_{i,j} > 0$, gene g_j induces the synthesis of gene g_i. If $\mathcal{G}_{i,j} < 0$, gene g_j inhibits the synthesis of gene g_i. Analogously, a positive (negative) value of $\mathcal{H}_{i,j}$ indicates that gene g_j induces (suppresses) the degradation of the mRNA level of gene g_i. We solve the equation system by integration using a fourth-order Runge-Kutta algorithm and the parameters of the S-System $\vec{\alpha}$, $\vec{\beta}$, $\mathcal{G}$, and $\mathcal{H}$ are optimized with evolutionary algorithms.

B. Optimization

Evolutionary algorithms are stochastic optimization techniques that mimic the natural evolution of mutation and selection as proposed by Charles Darwin. They have proved to be a powerful tool for solving complex optimization problems and in particular combinatorial problems. Three main types of evolutionary algorithms have been proposed in the last decades: Genetic Algorithms (GA), mainly developed by J.H. Holland [9], Evolution Strategies (ES), developed by I. Rechenberg [22] and H.-P. Schwefel [24], and Genetic Programming (GP) by J.R. Koza [16]. Each of these uses different solution representations and different operators working on them. They are, however, inspired by the same principles of natural evolution.

In our method we used a method, which separates the inference problem into two subproblems. The first task is to find the topology or structure of the network with a genetic algorithm. In the second task the parameters of a mathematical model are optimized for the given topology with an evolution strategy. The second problem can be seen as a local search phase of a memetic algorithm (MA).

1) Global Genetic Algorithm: In our implementation the genetic algorithm evolves populations of structures of possible networks. These structures are encoded as bitsets where each bit represents the existence or absence of an interaction between genes and therefore of non-zero parameters in the mathematical model. The evaluation of the fitness of each individual within the GA population uses a local search described below. Due to the binary representation of the structures, they gradually become sparse in the optimization process and thus the number of parameters to be optimized decreases dramatically.

2) Local Evolution Strategy: For evaluation of each structure suggested by the global optimizer an evolution strategy is used, which is suited for optimizing problems based on real values. The ES optimizes the parameters of the mathematical model used for representation of the regulatory network.

For assessing the quality of the locally obtained results we used the following equation for calculation of the fitness values for the ES optimization process:

$$f = \sum_{i=1}^{N} \sum_{k=1}^{T} \left(\frac{\hat{x}_i(t_k) - x_i(t_k)}{x_i(t_k)} \right)^2 \qquad (2)$$

where N is the total number of genes in the regulatory system, T is the number of sampling points taken from the time series and $\hat{x}$ and x distinguish between estimated data and experimental data. The overall problem is to minimize the fitness value f.

C. Separation

To model single genes we used a separation technique, which is well known in systems theory. To identify the underlying non-linear system, we separate the equations such that only the one equation for x_i of the system (1) is optimized. In the original version, the concentrations of all genes were depending on the model found so far in the optimization process. The current mRNA concentration levels that actually influence a particular gene have a significant impact on the modeling of this gene. An error in the network model propagates to all other concentration levels. Thus, modeling large systems is a very difficult task and it is not guaranteed to find the true system in case of a large number of system components. However, separation reduces the mentioned issue by separating the complete inference process into a number of smaller sub-problems. For this, the mRNA concentration levels of all the other genes x_j , $j \neq i$ are directly taken from the experimental data and included in every time step of the integration process while optimizing the model parameters. Thus, the correct concentration levels are used in every time step.

Separation reduces the complexity of the preceding N-to-N system into N separated sub-systems, where each is only having N-to-1 interactions to be inferred as illustrated in the following figure.

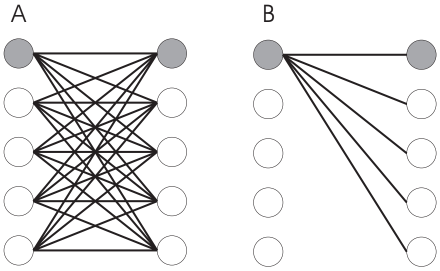

Fig. 1. Separation reduces the complexity of the N-to-N system (A) into N separated sub-systems, where each is only having N-to-1 interactions (B).

However, this separated model is only guaranteed to be valid for a single gene, and the dependencies of the other components of the complete network structure are not found during the inference, because they are not modeled. An enhancement to our method to solve the complete inference problem, may be the iterative combination of inferred genes to a larger system that eventually represents the complete system. This idea is currently implemented and will be subject of a future publication.

D. Interpolation

In the case of immune specific processes biologists distinguish between four phases: immediate early response, early response, intermediate response, and late response. The processes that were studied in the experiments stretch over a relatively long period of time (24h). Further on, microarray experiments are expensive. Therefore, the experiments have to be planned to cover all relevant time steps within the biological process with a minimum number of DNA chips. The sampling points should be distributed in a way that all mentioned phases are represented. But with such an experimental design, the resulting sampling points are not necessarily distributed equidistantly over time. Due to this issue, an interpolation scheme has to be used to calculate mRNA concentrations at intermediate time points. At present, we use a cubic spline interpolation method to interpolate necessary data points lying between sample points. This interpolating scheme has proven to be very successful in engineering problems and is guaranteed to yield the correct values at each grid point that was used to build the interpolation scheme [21].

E. Prediction

The last sampling point of the data sets was not included in the inference data set to examine the prediction capabilities of our algorithm (leave-one-out). After the inference of the network, we simulated the model to obtain a predicted value for the gene of interest and compared them to the true values of the experiments. We repeated the inference experiment for 25 times to obtain statistically sound results to be able to evaluate the predictive power of the method.

IV. Biological System

Our approach was tested on real microarray data, which were obtained by biological experiments within the 'Inflammatory and Infection' subnetwork of the German National Genome Research Network (NGFN). The details of the biological experiments will be subject of a future biological publication. Therefore, we will not describe them in details here. The data was obtained from a time course DNA microarray experiment in which human macrophages were infected by HCM viruses to study virus-specific immune response. Expression levels of $N = 22,215$ different probe sets were monitored at six time points (1h, 2h, 4h, 8h, 16h, and 24h after infection) both in uninfected and infected cells. The two expression levels of each probe set were compared using standard statistical algorithms to obtain a quantification of the differential gene expression (infected versus uninfected). The experiment was repeated 4 times under the same conditions to gain enough data sets to statistically filter the gene lists and thus to decrease the number of variables.

To reduce the size of the data sets, the first preprocessing step was to filter probe sets that did not appear to participate in the biological process of infection either because their expression signal was below a threshold, indicating experimental noise, or there was so little variation over time that these probe sets were likely not to be involved in the underlying infectious regulations. For our experiments, we decided to use a differential fold change value of ≥ 3.0 between the uninfected and the infected cells for the threshold. Further on, we used the standard statistical package SAM [30] as a second step to further reduce the complexity of the inference problem. The statistical methods merged the repeated experiments and thus raised the level of confidence that only genes were selected, which are involved in the immune response of the cell. A False-Discovery rate of $FDR = 0.04$ was chosen for the SAM analysis. After filtering and statistical preprocessing, 268 genes remained in the resulting data set.

For evaluating the performance of the method we decided to infer eleven genes in the data set, which are known to be relevant to immune specific response activities of an infected cell and which have been suggested by the participating biologists because they are of special interest. Due to the limited space in this publication, we will only show three of the predicted genes here:

One of the relevant genes is the interferon-induced protein **G1P2** (UniGene Id *Hs.432233*). References can be found in several biological articles, for example [20]. Gene Ontology (GO) [3] classifies this gene in the category 'Biological Process' into the class '6955 - immune response'. The second gene to be inferred by the proposed algorithm was the Human p53 cellular tumor antigen mRNA **P53** (UniGene Id *Hs.426890*) and the related tumor suppressor phosphoprotein **TP53**. It was subject to several publications as for example in [31]. According to GO the gene takes part in the cell cycle and in the apoptosis pathway and is classified in the category 'Biological Process' into the main class '6915 - Apoptosis'. And the third gene was the **RASA3** RAS p21 protein activator 3 (UniGene Id *Hs.119274*). GO classifies this gene in '7242 - intracellular signaling cascade' and a reference on the gene is given in [5].

The parameters of the inferred models have been evaluated in a postprocessing step in collaboration with medical researchers and with the help of the TRANSPATH [4] and the KEGG [12] database. For each gene that was modeled using the proposed inference strategy, we examined the strongest relationships indicated by the significantly largest parameter values in the mathematical model. These dependencies were then evaluated by searching for corresponding relationships in the databases and in the literature. The biological interpretation sections for each gene list some of the found and confirmed regulatory effects.

V. Results

The GA evolved a population of 250 possible structures with a tournament selection with a tournament group size of $t_{group} = 8$, 3-Point-crossover recombination with $p_c = 1.0$ and a mutation probability $p_m = 0.1$. The local optimization was started using a (μ,λ)-ES with $\mu = 10$ parents and $\lambda = 50$ offsprings together with a Covariance Matrix Adaptation (CMA) mutation operator without recombination. The probabilities of crossover and mutation were chosen as $p_c = 0.0$ and $p_m = 1.0$. These parameters were determined in preliminary experiments and have shown to yield the best results. Overall, the MA evolved the individuals for $250,000$ generations. And due to the separation of the system, the maximum number of variables was $2N + 2 = 538$ instead of $2N + 2N^2 = 72,360$ variables in case of the complete inference problem.

To evaluate the quality of the solutions and to show that the proposed memetic algorithm is learning and thus finding solutions with better fitness values in each generation, we compared it to a totally random sampling of model parameters. These results are also given in the corresponding fitness graphs.

A. Interferon Alpha-inducible Protein G1P2

The expression pattern of the first gene is highly dynamic as can be seen in figure 2, where the differences between uninfected and infected is plotted with a straight line. The maximum change of the gene expression levels is 13.98 for the compared cell states.

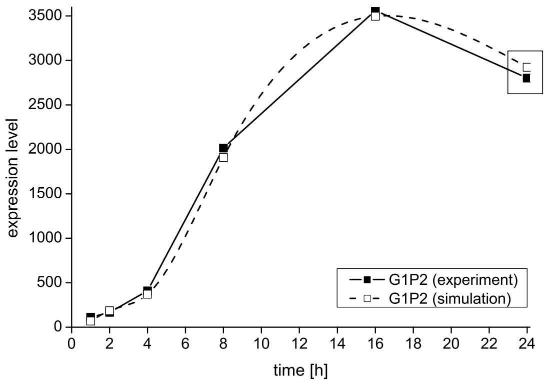

Fig. 2. Time dynamics of gene **G1P2**. Given are the time course of the experiment data and the simulated course.

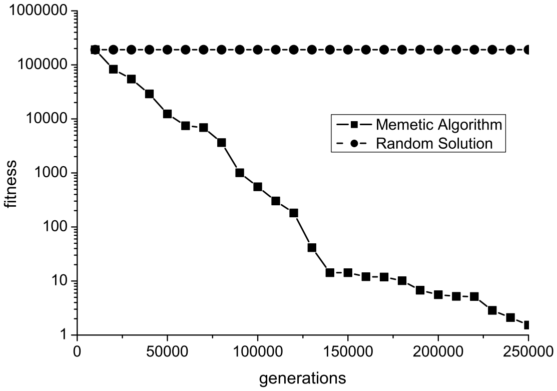

Fig. 3. Fitness courses for gene **G1P2** obtained by the inference process using a separated memetic algorithm compared to a totally random set of model parameters.

Figure 2 shows the inferred time course of gene **G1P2** found by the optimizing process. The lines represent the time dynamics of the experimental and the simulated gene expression. The fitness course for gene **G1P2** of the inference process is shown in figure 3 and suggests a very good compliance with the experimental data as the values decrease continuously over time. We implemented the optimization objective as a problem to minimize the fitness values and the algorithm eventually finds a solution with a fitness of 1.53057.

The last data point at $t = 24h$ (indicated by the frame in figure 2) is the predicted gene behavior that results from our simulation. As one can see, the predicted value of the gene response (dashed line) and the true behavior of the gene (straight line) are very similar. The experiments yield an expression value of 2801.6 for the last sampling point $t = 24h$. And the simulation of the network that was found in the optimization process results in a very close value of 2922.7. Furthermore, due to the high quantitative level of expression, the difference between these values is almost negligible, especially in case of noisy biological experiments.

As mentioned in the previous section, we evaluated the inferred models by searching for correspondence of strong relationships in the model and known relationships in biology. In case of the **G1P2** model, we found relationships in the model parameters between the Interferon Regulatory Factor 3 and 4 **IRF3**, **IRF4**, respectively, and the gene product **ISG15(h)** of **G1P2** represented by a positive dependency value. This is affirmed by the work of Meraro *et al.* in [20].

IRF-3(h) → ISG15(h) (activation; transregulation), and

IRF-4(h) → ISG15(h) (activation; transregulation).

Further on, the model suggests relationships between **ICSBP** and **G1P2**, which can also be found in the database:

ICSBP(h) → ISG15(h) (activation; transregulation).

However, the inference process was not able to find the relationship between **ISGF3G(h)** and **G1P2** with statistical significance although **ISGF3G(h)** was included in the experimental data set. Because the time dynamics of the inferred model matches the biological system so well, further investigations of the model relationships are necessary to evaluate unknown alternative regulatory effects that bypass this known regulation.

B. Tumor Protein P53

The straight line in figure 4 shows the differential expression signals between uninfected and infected cells with a maximum change level of 4.7.

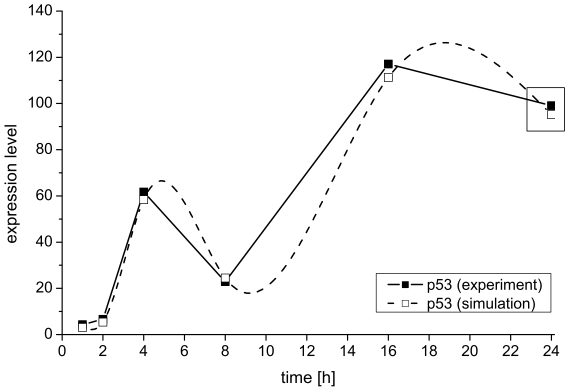

Fig. 4. Time dynamics of gene **P53**. Given are the time course of the experiment data and the simulated course.

Figure 4 shows the inferred time course of gene **P53** found by the optimizing process. The lines represent the time dynamics of the experimental and the simulated gene expression. As in the previous section, the proposed method finds a model that shows very good resemblance to the biological experiment. The expression levels of gene **P53** and the simulated gene differ only marginally. The fitness course for gene **P53** of the inference process is shown in figure 5 and as in the inference before, the values decrease to very good fitness values, which also corresponds to the good data match in figure 4.

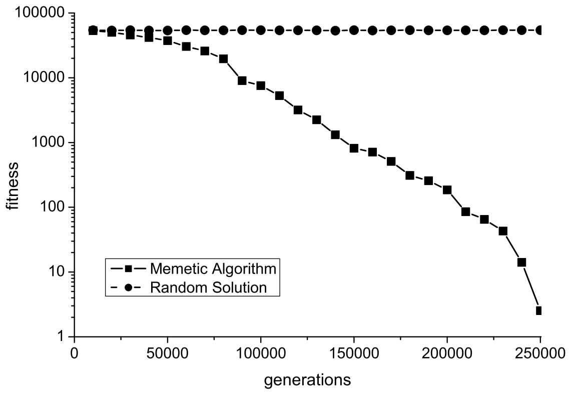

Fig. 5. Fitness courses for gene **P53** obtained by the inference process using a separated memetic algorithm compared to a totally random set of model parameters.

As in the example above, we excluded the last data point ($t = 24h$) from the inference process. Figure 4 shows the value of the experiment together with the predicted value obtained from the simulation. The predicted value of the gene response (dashed line) and the true behavior of the gene (straight line) are very similar: 99.1 for the true value and 95.2 resulting from our algorithm. This result, as the result of the previous experiment, indicate a very high predictive power of our method.

One of the strongest interactions of system components in the inferred model was the dependency between **P53** and the paired box gene 8, **PAX8(h)**, which can also be found in TRANSFAC

PAX8(h) → p53(h) (transregulation).

The human immunodeficiency virus type I enhancer binding protein 1 **HIVEP1** (KEGG) or **PRDII-BF1(h)** (TRANSFAC) expresses the Gatekeeper of Apoptosis Activating Proteins 1 **GAAP-1(h)** and regulates **P53** as described in [17]. This chain of regulation can also be found in the parameters of our model, where this relationship is represented by a direct link between **PRDII-BF1(h)** and **P53**. This shortcut is due to the absence of **GAAP-1(h)** in the experimental data set such that this intermediate product is not covered by our model.

Regulation in TRANSFAC:

PRDII-BF1(h) → GAAP-1(h) (expression)

GAAP-1(h) → p53(h) (activation; transregulation).

Regulation in the inferred Model:

PRDII-BF1(h) → p53(h) (activation; transregulation).

Furthermore, the model suggests several downstream relationships. Two of them are the following dependencies: first, the model parameter suggest a relation between the apoptosis-related cysteine protease **CASP1** and **P53**, which is confirmed by the work of Gupta *et al.* in [8]:

p53(h) → caspase-1(h) (activation; transregulation).

And secondly, Shin *et al.* examined the transcriptional stimulation of the transforming growth factor alpha **TGFA** by **P53** in [25], which can also be found in our model, where

p53(h) → TGFalpha(h) (activation; transregulation).

Because the tumor protein **P53** is a well studied gene due to its participation in cancer disease, several known interactions of this gene with other system components can be found in the literature and thus several additional matches between the databases and the model can be found. Due to space restrictions, only the mentioned three most significant relationships are listed here.

C. RAS p21 Protein Activator RASA3

Figure 6 gives the time dynamics of **RASA3** with a maximum change of 5.1 between the uninfected and the infected state (straight line).

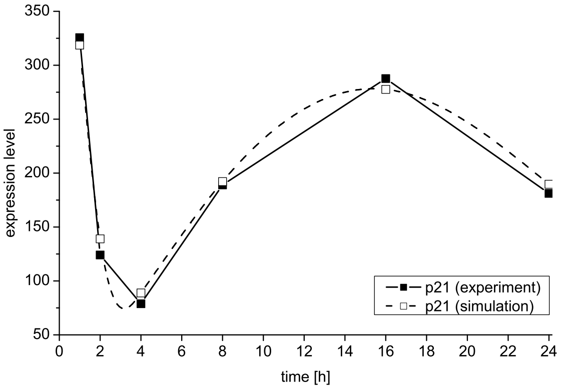

Fig. 6. Time dynamics of gene **RASA3**. Given are the time course of the experiment data and the simulated course.

The predicted dynamics of gene **RASA3** resulting from the simulation are given figure 6 with a straight line. As in the previous examples, the simulated activity level of gene **RASA3** shows a very good correlation with the true expression value of the experiment. The actual value for the last sampling point ($t = 24h$) in the simulation was 181.2 compared to a true value of 189.6 in the experiment. This again suggests a high predictive power of the algorithm.

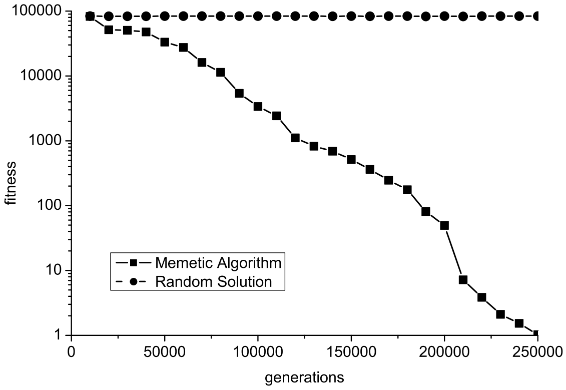

Fig. 7. Fitness courses for gene **RASA3** obtained by the inference process using a separated memetic algorithm compared to a totally random set of model parameters.

One of the lesser known genes that was modeled with our approach is **RASA3**, which yielded a very good model fitness. This gene represents a class of genes for which their particular regulating roles in the immune process are not as well studied as for example those of the **P53** gene. And thus, only little information can be found in the literature. Further on, no interactions with other system components are listed in the databases. We found two genes of this class, which yielded very good model fitness values but which were almost not covered by any publication or database entries. However, the inferred model suggests several relationships between **RASA3** and other genes that are to date unknown. These new hypothesis of regulatory effects will be subject of further analysis by biological researchers at our facility in future work.

D. Remaining Gene Set

Due to the limited space available in this paper, we showed only three example predictions. As mentioned in the beginning, we inferred eleven different genes from which we have been able to reliably model nine with our method with very good prediction values. The remaining two genes did not fit the true values correctly: one activity prediction was completely wrong (99.2 in the experiment and 231.9 in our simulation) and the other value showed only little resemblance (78.6 in the experiment and 99.1 in the simulation). The reason for the prediction to fail in these two cases may be that these genes are participating in many different biological processes and are thus stimulated by a large number of non-immune-relevant genes, which may not be included in the experiment data set due to the filtering. Another reason for the failure of our modeling technique may be the variance of the biological experiments, since the cell lines and their environmental conditions have a strong impact on the resulting gene expression levels as well.

TABLE I

RESULTS OF THE EXPRESSION PREDICTION FOR EACH OF THE ELEVEN MODELED GENES.

Gene	ID	Averaged Fitness	Predicted Value
RASA3	Hs.119274	1.022	181.2 (189.6)
P53	Hs.426890	2.531	99.1 (95.2)
G1P2	Hs.432233	1.530	2922.7 (2801.6)
PHB	Hs.75323	1.341	4.02
GJA1	Hs.74471	2.726	4.02
FASTK	Hs.75087	1.943	4.02
DMD	Hs.169470	2.776	4.02
EIF1AY	Hs.155103	3.127	4.02
HSA9947	Hs.128866	3.591	4.02
PEX6	Hs.301636	3.998	231.9 (99.2)
TCF8	Hs.232068	4.194	99.1 (78.6)

Table I gives an overview of the results of the predicted expression values for each gene that was modeled. The table lists the gene symbol and the corresponding UniGene identifier followed by the averaged fitness value that was achieved for the model parameters. The last column gives the predicted expression level for the gene and the true value in brackets.

VI. DISCUSSION

The problem of inferring complete GINs is a very difficult problem due to the limited data available and the large number of unknown variables in the system. And the ability of predicting the expression levels of genes under certain conditions is an important task in bioinformatics. Thus, we proposed a method to infer only selected genes in the context of the whole experimental data. We showed that separation is dramatically reducing the complexity of the problem. Moreover, our method is able to find biological relationships between genes in immune specific systems and is thereby able to predict gene behavior to regulatory factors very accurately. We were not only able to predict 82% of the selected genes correctly and reliably in several runs. Moreover, we were able to reproduce most of the dependencies of the genes that are known to date, limited by the statistical preprocessing and the filtering of genes that afterwards turned out to be important.

The capabilities of the separation strategies suggest to divide a complete system inference into a preprocessing with separation and a second phase, where the separated genes are combined. However, preliminary computational experiments showed that good fitness values are not necessarily an indicator for a biologically correct model. Particularly, complete inference experiments on artificial data sets advised that due to limited available data the overall problem is highly underdetermined and therefore many different solutions may yield comparably good fitness values. To cope with this issue, we plan to enhance our method by introducing a new adaptive separation strategy, where separated genes are iteratively combined or split in the optimization process to build larger system or building blocks. Another way of increasing confidence in the inferred models may be the usage of techniques like island model optimization or virtual knock-out strategies as suggested by the authors in [28] and [27], respectively.

Further on, we plan to include a-priori information into the inference process, like partially known pathways or information about co-regulated genes, which can be found in literature or public databases. Additionally, other models for gene interconnectivity networks will be examined for simulation of the non-linear interaction system to overcome the problems with those gene interconnectivity networks that hardly can be modeled by S-systems. To further reduce the dimensionality of the data set, we plan to use cluster methods. Recently, we developed a new method that incorporates biological ontologies [26] and thus yields biologically more plausible clustering results.

We will continue to test our method with real microarray data in close collaboration with biological researchers at our facility in the area of signal transduction. And the computational results of this publication will be subject to detailed verification processes, especially those of the **RASA3** protein activator. Our algorithm will be tested on repeated experiment data sets to increase confidence in the model. And further on, the medical researchers at our facility will try to verify the mathematical model to gain information about the underlying process with additional experiments.

ACKNOWLEDGEMENT

This work was supported by the National Genome Research Network (NGFN) of the Federal Ministry of Education and Research in Germany under contract number 0313323.

REFERENCES

[1] Tatsuya Akutsu, Satoru Miyano, and Satoru Kuhura. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 17–28, 1999.

[2] Shin Ando, Erina Sakamoto, and Hitoshi Iba. Evolutionary modeling and inference of gene network. *Information Sciences*, 145(3-4):237–259, 2002.

[3] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[4] Claudia Choi, Mathias Krull, Alexander Kel, Olga Kel-Margoulis, Susanne Pistor, Anatolij Potapov, Nico Voss, and Edgar Wingender. TRANSPATH - a high quality database focused on signal transduction. *Comparative and Functional Genomics*, 5(2):163 – 168, 2004.

[5] P.J. Cullen, J.J. Hsuan, O. Truong, A.J. Letcher, T.R. Jackson, A.P. Dawson, and R.F. Irvine. Identification of a specific Ins(1,3,4,5)P4-binding protein as a member of the GAP1 family. *Nature*, 376(6540):527–30, 1995.

[6] Hidde de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, January 2002.

[7] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601?620, 2000.

[8] Sanjeev Gupta, Vegesna Radha, Yusuke Furukawa, and Ghanshyam Swarup. Direct transcriptional activation of human caspase-1 by tumor suppressor p53. *Journal of Biological Chemistry*, 276(14):10585–10588, 2001.

[9] John H. Holland. *Adaption in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Systems*. University Press of Michigan, 1975.

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

[11] Seiya Imoto, Tomoyuki Higuchi, Takao Goto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. In *Proceedings of the IEEE Computer Society Bioinformatics Conference*, pages 104 –113, 2003.

[12] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[13] Stuart A. Kauffman. *The Origins of Order*. Oxford University Press, 1993.

[14] E. Keedwell, A. Narayanan, and D.A. Savic. Modelling gene regulatory data using artificial neural networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, volume 1, pages 183–188, 2002.

[15] Shinichi Kikuchi, Daisuke Tominaga, Masanori Arita, Katsutoshi Takahashi, and Masaru Tomita. Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics*, 19(5):643–650, 2003.

[16] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, 1992.

[17] Christophe Lallemand, Marta Palmieri, Brigitte Blanchard, Jean-Francois Meritet, and Michael G. Tovey. GAAP-1: a transcriptional activator of p53 and IRF-1 possesses pro-apoptotic activity. *EMBO reports*, 3(2):153?158, 2002.

[18] Shoudan Liang, Stafanie Fuhrman, and Roland Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 3, pages 18–29, 1998.

[19] Yukihiro Maki, Daisuke Tominaga, Masahiro Okamoto, Shoji Watanabe, and Yukihiro Eguchi. Development of a system for the inference of large scale genetic networks. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 6, pages 446–458, 2001.

[20] David Meraro, Merav Gleit-Kielmanowicz, Hansjorg Hauser, and Ben-Zion Levi. IFN-stimulated gene 15 is synergistically activated through interactions between the myelocyte/lymphocyte-specific transcription factors, PU.1, IFN regulatory factor-8/IFN consensus sequence binding protein, and IFN regulatory factor-4: characterization of a new subtype of IFN-stimulated response element. *Journal of Immunology*, 168(12):6224–6231, 2002.

[21] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of scientific computing*. Cambridge University Press, 1992.

[22] Ingo Rechenberg. *Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, 1973.

[23] Michael A. Savageau. 20 years of S-systems. In E.O. Voit, editor, *Canonical Nonlinear Modeling. S-systems Approach to Understand Complexity*, pages 1–44, 1991.

[24] Hans-Paul Schwefel. *Numerical optimization of computer models*. John Wiley & Sons, 1981.

[25] T.H. Shin, A.J. Paterson, and J.E. Kudlow. p53 stimulates transcription from the human transforming growth factor alpha promoter: a potential growth-stimulatory role for p53. *Molecular and Cellular Biology*, 15:4694–4701, 1995.

[26] Nora Speer, Christian Spieth, and Andreas Zell. A memetic clustering algorithm for the functional partition of genes based on the gene ontology. In *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 252–259, 2004.

[27] Christian Spieth, Felix Streichert, Nora Speer, and Andreas Zell. Iteratively inferring gene regulatory networks with virtual knockout experiments. In *Proceedings of the European Workshop on Evolutionary Bioinformatics*, volume 3005 of *Lecture Notes in Computer Science*, pages 102–111, 2004.

[28] Christian Spieth, Felix Streichert, Nora Speer, and Andreas Zell. Utilizing an island model for EA to preserve solution diversity for inferring gene regulatory networks. In *Proceedings of the IEEE Congress on Evolutionary Computation*, volume 1, pages 146–151, 2004.

[29] Denis Thieffry and R. Thomas. Qualitative analysis of gene networks. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 77–87, 1998.

[30] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. In *Proceedings of the National Academy of Sciences*, volume 98, pages 5116–5121, 2001.

[31] B. Vogelstein and K.W. Kinzler. p53 function and dysfunction. *Cell*, 70(4):523–526, 2002.

[32] D.C. Weaver, C.T. Workman, and G.D. Stormo. Modeling regulatory networks with weight matrices. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 4, pages 112–123, 1999.

[33] M. K. Stephen Yeung, Jesper Tegner, and James J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. In *Proceedings of the National Academy of Science*, volume 99, pages 6163–6168, 2002.