BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 21 2005, pages 4026–4032
doi:10.1093/bioinformatics/bti662

Systems biology

Non-transcriptional pathway features reconstructed from secondary effects of RNA interference

Florian Markowetz*, Jacques Bloch and Rainer Spang

Department of Computational Molecular Biology, Computational Diagnostics Group, Max Planck Institute for Molecular Genetics, Ihnestrasse 63–73, 14195 Berlin, Germany

Received on July 14, 2005; revised on September 2, 2005; accepted on September 3, 2005
Advance Access publication September 13, 2005

ABSTRACT

Motivation: Cellular signaling pathways, which are not modulated on a transcriptional level, cannot be directly deduced from expression profiling experiments. The situation changes, when external interventions such as RNA interference or gene knock-outs come into play. Even if the expression of the signaling genes is not changed, secondary effects in downstream genes shed light on the pathway, and allow partial reconstruction of its topology.

Results: We introduce an algorithm to infer non-transcriptional pathway features based on differential gene expression in silencing assays. We demonstrate the power of our algorithm in the controlled setting of simulation studies, and explain its practical use in the context of an RNA interference dataset investigating the response to microbial challenge in *Drosophila melanogaster*.

Contact: florian.markowetz@molgen.mpg.de

1 INTRODUCTION

Cellular signaling pathways regulate essential processes in living cells. In many cases, alterations of these molecular mechanisms lead to serious diseases including cancer. Understanding the organization of signaling pathways is hence a leading problem in modern biology. Microarray studies using cell assays with external interventions into the signaling process allow for the systematic analysis of these pathways (Spradling *et al*., 1999, Hughes *et al*., 2000). Gene-expression profiling is a well-established high-throughput technology, but until recently external interventions have been labor intensive and time consuming. With the technology of RNA interference (RNAi), this situation has changed. RNAi (Fire *et al*., 1998) is a novel method of post-transcriptional gene silencing. It has drastically reduced the time required for testing downstream effects of gene silencing (Nature insight, 2004; Boutros *et al*., 2004). In several studies, RNAi screening has been applied to such functional genomic analysis (Gönczy *et al*., 2000; Fraser *et al*., 2000).

Non-transcriptional modules in signaling pathways. A cell's response to an external stimulus is complex. The stimulus is propagated via signal transduction to activate transcription factors, which bind to promoters thus activating or repressing the transcription and translation of genes, which in turn can activate secondary signaling pathways, and so on. We distinguish between the transcriptional level of signal transduction known as gene regulation and the non-transcriptional level, which is mostly mediated by post-translational modifications. Although gene regulation leaves direct traces on expression profiles, non-transcriptional signaling does not. However, reflections of signaling activity can be perceived in expression levels of other genes. We explain this by a real world example.

An example in Drosophila. Boutros *et al*. (2002) investigate the response to microbial challenge in *Drosophila melanogaster*. They treat *Drosophila* cells with lipopolysaccharides (LPS), the principal cell wall components of Gram-negative bacteria. After 60 min of applying LPS, a number of genes show a strong reaction. Which genes and gene products were involved in propagating the signal in the cell? To answer this question a number of candidate pathway genes are silenced by RNAi. The effects on the LPS-induced genes are measured by microarrays. The observations are: with only one exception, the signaling genes show no change in expression when other signaling genes are silenced. They stay 'flat' on the microarrays. Differential expression is only observed in genes downstream of the signaling pathway: silencing *tak* reduces expression of all LPS-inducible transcripts, silencing *rel* or *mkk4*/*hep* reduces expression of disjoint subsets of induced transcripts, silencing *key* results in profiles similar to silencing *rel*.

Boutros *et al*. (2002) explain this observation by a fork in the signaling pathway with *tak* above the fork, *mkk4*/*hep* in one branch and both *key* and *rel* in the other branch. Note that this pathway topology was found in an indirect way: no information is coming from the expression levels of the signaling genes. Silencing candidate genes interrupts the information flow in the pathway, the topology is then revealed by the nested structure of affected gene sets downstream the pathway of interest. The computational challenge we address in this paper is to derive an algorithm for systematic inference from indirect observations.

Previous work. Previous methods for learning from interventions construct a model to explain primary effects of silencing genes on other genes in the pathway (Wagner, 2001, 2004; Tegner *et al*., 2003; Ideker *et al*., 2000; Akutsu *et al*., 1998). With one exception (Tegner *et al*., 2003), they are deterministic and cannot handle noise in the data. All of them aim for transcriptional networks and are unable to capture non-transcriptional modulation.

Various probabilistic methods have been developed to reconstruct regulatory networks from microarray data (Wille *et al*., 2004; Friedman, 2004; Segal *et al*., 2003; Imoto *et al*., 2002; Wessels *et al*., 2001; Friedman *et al*., 2000). Some incorporate external interventions explicitly (Di Bernardo *et al*., 2005;

*To whom correspondence should be addressed.

Markowetz *et al*., 2005; Gardner *et al*., 2003; Pe'er *et al*., 2001; Cooper and Yoo, 1999). All these methods model the joint distribution of gene-expression levels as a graphical model. This requires the expression levels of modeled genes to change from one array to another. Interactions are modeled on the transcriptional level, the non-transcriptional level is blinded out.

Some approaches use hidden variables to capture non-transcriptional effects (Nachman *et al*., 2004; Rangel *et al*., 2001, 2004). None of them makes use of interventional data. To keep model selection feasible they have to introduce a number of simplifying assumptions: either the hidden nodes do not regulate each other, or the hidden structure is not identifiable. In both cases, the models do not allow inference of non-transcriptional pathways.

Another class of algorithms searches for topologies which are consistent with observed downstream effects of interventions (Yeang *et al*., 2004). Although these algorithms are not confined to the transcriptional level of regulation, they require that most signaling genes show effects when perturbing others.

In summary, none of the methods designed to infer transcriptional networks can be applied to reconstruct non-transcriptional pathways. The major problem is that these algorithms require direct observations of expression changes of signaling genes, which are not fully available in datasets such as those of Boutros *et al*. (2002).

Our general objective is similar to epistasis analysis with global transcriptional phenotypes (Driessche *et al*., 2005). Nevertheless, there are several important difference. First, we model whole pathways and not only single gene–gene interactions. Second, we treat an expression profile not as one global phenotype but as a collection of single-gene phenotypes.

Overview of our approach. In this paper, we present a computational framework for the systematic reconstruction of pathway features from expression profiles relating to external interventions. Our approach is based on the nested structure of affected downstream genes, which are themselves not a part of the model. Here we give a short overview of our method before presenting it in all details in Section 2.

We distinguish two kinds of genes: the candidate pathway genes, which are silenced by RNAi, and the genes, which show effects of such interventions in expression profiles. We call the first ones *S*-genes (*S* for 'silenced' or 'signaling') and the second ones *E*-genes (*E* for 'effects'). Since large parts of signaling pathways are non-transcriptional, there will be little or no overlap between *S*-genes and *E*-genes. Elucidating relationships between *S*-genes is the focus of our analysis, the *E*-genes are only needed as reporters for signal flow in the pathway. *E*-genes can be considered as transcriptional phenotypes. *S*-genes have to be chosen depending on the specific question and pathway of interest. *E*-genes are identified by comparing measurements of the stimulated and non-stimulated pathway: genes with a high-expression change are taken as *E*-genes.

Our approach models how interventions interrupt the information flow through the pathway. Thus, *S*-genes are silenced, while the pathway is stimulated to see which *E*-genes are still reached by the signal. Optimally, the gene-expression experiments are replicated several times. This results in a dataset representing every signaling gene by one or more microarrays. These requirements are the same as in epistasis analysis (Avery and Wasserman, 1992), but they are not satisfied in all datasets monitoring intervention effects (Hughes *et al*., 2000).

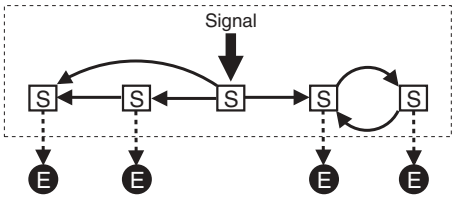

Fig. 1. A schematic summary of our model. The dashed box indicates one hypothesis: it contains a directed graph T on genes contributing to a signaling pathway (*S*-genes). A signal enters the pathway at one (or possibly more than one) specified position. Interventions at *S*-genes interrupt signal flow through the pathway. *S*-genes regulate *E*-genes on the second level. Together the *S*- and *E*-genes form an extended topology T'. We observe noisy measurements of expression states of *E*-genes. The objective is to reconstruct relationships between *S*-genes from observations of *E*-genes in silencing experiments.

The main contribution of this paper is a scoring function, which measures how well hypotheses about pathway topology are supported by experimental data. Input to our algorithm is a list of hypotheses about the candidate pathway genes. A hypothesis is characterized by (1) a directed graph with *S*-genes as nodes and (2) the possibly many entry points of signal into the pathway. This setting is summarized in Figure 1. Our model is based on the expected response of an intervention given a candidate topology of *S*-genes and the position of the intervention in the topology. Pathways with different topology can show the same downstream response to interventions. We identify all pathways, which make the same predictions of intervention effects on downstream genes, by one so-called silencing scheme. Sorting silencing schemes by our score shows how well candidate pathways agree with experimental data. Output of the algorithm is a strongly reduced list of candidate pathways. The algorithm is a filter, which helps to direct further research.

Applications beyond RNAi. Our motivation to develop this algorithm results from the novel challenges the RNAi technology poses to bioinformatics. At present, RNAi appears to be the most efficient technology for producing large-scale gene-intervention data. However, our framework is flexible and any type of external interventions can be used, which reduces information flow in the pathway. This includes traditional knock-out experiments and specific protein inhibiting drugs. An important requirement for any perturbation technique used is high specificity. Off-target effects impair our method since intervention effects can no longer be uniquely predicted.

In the next section we develop our model in detail. Then we test it in simulation studies (Section 3.1) and demonstrate its use on real data (Section 3.2).

2 METHODS

First, we describe our model for signaling pathways with transcriptional phenotypes. Predictions from pathway hypotheses are summarized in a silencing scheme. In the main part of the section, we develop a Bayesian method to estimate a silencing scheme from data.

2.1 Signaling pathway model

Core topology on S-genes. The set of *E*-genes is denoted by $\mathbf{E} = \{E_1, \ldots, E_m\}$, and the set of *S*-genes by $\mathbf{S} = \{S_1, \ldots, S_n\}$. As a pathway model, we assume a directed graph T on vertex set $\mathbf{S}$. The structure of T is not further

restricted: there may be cycles and it may decompose into several subgraphs. The external stimulus acts on one or more of the *S*-genes as specified by the hypothesis. *S*-genes can take values 1 and 0 according to whether signaling is interrupted or not. State 0 corresponds to a node, which is reached by the information flow through the pathway. This is the natural state when the pathway is stimulated. State 1 describes a node, which is no longer reached by the signal, because the flow of information is cut by an intervention at some node upstream in the pathway. An *S*-gene in state 1 is in the same state as if the pathway had not been stimulated. Although the pathway is stimulated, experimental interventions break the information flow in the pathway. An intervention at a particular *S*-gene first puts this *S*-gene's state to 1. The silencing effect is then propagated along the directed edges of T.

From pathways to silencing schemes. We call the subset of *S*-genes, which are in state 1 when *S*-gene S is silenced, the 'influence region of S'. The set of all influence regions is called a 'silencing scheme Φ'. It summarizes the effects of interventions we predict from the pathway hypothesis. Mathematically, a silencing scheme is the transitive closure of pathway T defining a partial order on $\mathbf{S}$. Drawn as a graph, Φ contains an edge between two nodes whenever they are connected by a directed path in T. Different pathway models can result in the same silencing scheme. Note that the *E*-genes do not appear in Φ, which only describes interactions between *S*-genes. The *E*-genes come into play when we want to infer silencing schemes: reduced signaling strength of *S*-genes owing to interventions in the pathway cannot be observed directly on a microarray, but we can see secondary effects on *E*-genes.

Secondary effects on E-genes. The extended topology on $\mathbf{S} \cup \mathbf{E}$ is called T'. We assume that each *E*-gene has a single parent in $\mathbf{S}$. We interpret the set of *E*-genes attached to one *S*-gene as a regulatory module, which is under the common control of the *S*-gene. To account for the frequent case where more than one *S*-gene regulates an *E*-gene, we will use model averaging. The reaction of *E*-genes to interventions in the pathway depends on where the parent *S*-gene is located in the silencing scheme. *E*-genes are set to state 1 if their parent *S*-gene is in the influence region of an intervention; else they are in state 0. The state of *E*-genes can be experimentally observed as differential expression on microarrays. Owing to the observational noise or stochastic effects in signal transduction, we expect a number of false positive and false negative observations.

Filtering hypotheses. The input to the algorithm is a list $H_1, \ldots, H_N$ of pathway hypotheses. Each hypothesis makes predictions of effects at *E*-genes downstream of the pathway. In the next section we develop a Bayesian method to score silencing schemes given noisy observations of *E*-genes. In general, different topologies can have identical scores; hence the algorithm does not uniquely reconstruct the pathway, but returns a strongly reduced list of optimally scoring pathways of length $M \ll N$.

2.2 Likelihood of a silencing scheme

Data. In each experiment, one *S*-gene is silenced by RNAi and effects on *E*-genes are measured by microarrays. Each *S*-gene needs to be silenced at least once, but ideally the silencing assays are repeated and we have several microarrays per silenced gene. We index the microarrays by $k = 1, \ldots, l$. The expression data are assumed to be discretized to 1 and 0—indicating whether interruption of signal flow was observed at a particular gene or not. As a result we get a binary matrix $D = (e_{ik})$, where $e_{ik} = 1$ if *E*-gene E_i shows an effect in experiment k. Thus, our data only consist of coarse qualitative information. We do not consider whether an *E*-gene was upregulated or downregulated or how strong an effect was. Each single observation e_{ik} relates the intervention done in experiment k to the state of E_i. In the following, the index 'i' always refers to an *E*-gene, the index 'j' to an *S*-gene, and the index 'k' to an experiment.

Likelihood. We introduce the position of the *E*-genes as model parameters $\Theta = \{\theta_i\}_{i=1}^m$ with $\theta_i \in \{1, \ldots, n\}$ and $\theta_i = j$ if E_i is attached to S_j. Let us first consider a fixed extension T' of T, i.e. the parameters Θ are assumed to be known. For each *E*-gene, T' encodes to which *S*-gene it is connected. In a silencing experiment we predict effects at all *E*-genes, which are attached to an *S*-gene in the influence region. Expected effects can be compared with observed effects in the data to choose the topology, which fits the data best. Owing to measurement noise we cannot expect to find a topology T' in complete agreement with all observations. We allow deviation from predicted effects by introducing global error probabilities α and β for false positive and negative calls, respectively.

We model the expression levels of *E*-genes on the various microarrays as binary random variables E_{ik}. The distribution of E_{ik} is determined by the silencing scheme Φ and the error probabilities α and β. For all *E*-genes and targets of intervention, the conditional probability of *E*-gene state e_{ik} given silencing scheme Φ can then be written in tabular form as

$$P(e_{ik}|\Phi, \theta_i = j) = \begin{array}{c|cc} & \underline{e_{ik}=1} & \underline{e_{ik}=0} \\ & & \\ & \alpha & 1-\alpha \quad \text{if } \Phi \text{ predicts } \textit{no effect} \\ & 1-\beta & \beta \quad \text{if } \Phi \text{ predicts } \textit{effect} \end{array}$$

This means that if E_i is not in the influence region of the *S*-gene silenced in experiment k, the probability of observing $E_{ik} = 1$ is α (probability of false alarm, type-I error); the probability to miss an effect and observe $E_{ik} = 0$ even though E_i lies in the influence region is β (type-II error). The likelihood $P(D|\Phi, \Theta)$ of the data is then a product of terms from the table for every observation.

However, in reality we do not know the 'correct' extension T' of a candidate topology T. The positions of *E*-genes are unknown and they may be regulated by more than one *S*-gene. We also do not aim to infer extended topologies from the data: the model space of extended topologies is huge, and model inference is unstable. We are only interested in the silencing scheme Φ of *S*-genes. To deal with these issues, we interpret the position of edges between *S*- and *E*-genes as nuisance parameters, and average over them to obtain a marginal likelihood. This is what we describe next.

2.3 Marginal likelihood of a silencing scheme

We define a scoring function that evaluates how well a given silencing scheme Φ fits the data. For now, we assume the silencing scheme Φ and the error probabilities α and β to be fixed. But in contrast to the last section, the connection parameters Θ are unknown. By Bayes' formula we can write the posterior of silencing scheme Φ given data D as

$$P(\Phi|D) = \frac{P(D|\Phi)P(\Phi)}{P(D)}. \tag{1}$$

The normalizing constant $P(D)$ is the same for all silencing schemes; we can neglect it for model comparison. The model prior $P(\Phi)$ can be chosen to incorporate biological prior knowledge. Here, we assume it to be uniform over all possible models. What remains is the marginal likelihood $P(D|\Phi)$. It equals the likelihood $P(D|\Phi, \Theta)$ of the data averaged over the nuisance parameters Θ. To compute it, we make three assumptions:

(1) Given silencing scheme Φ and fixed positions of *E*-genes Θ, the observations in D are sampled independently and distributed identically:

$$P(D|\Phi, \Theta) = \prod_{i=1}^{m} P(D_i|\Phi, \theta_i) = \prod_{i=1}^{m} \prod_{k=1}^{l} P(e_{ik}|\Phi, \theta_i),$$

where D_i is the i-th row in data matrix D.

(2) Parameter independence. The position of one *E*-gene is independent of the positions of all the other *E*-genes:

$$P(\Theta|\Phi) = \prod_{i=1}^{m} P(\theta_i|\Phi).$$

(3) Uniform prior. The prior probability to attach an *E*-gene is uniform over all *S*-genes:

$$P(\theta_i = j|\Phi) = \frac{1}{n} \quad \text{for all } i \text{ and } j.$$

The last assumption can easily be dropped to include existing biological prior knowledge about regulatory modules. With the assumptions above, the

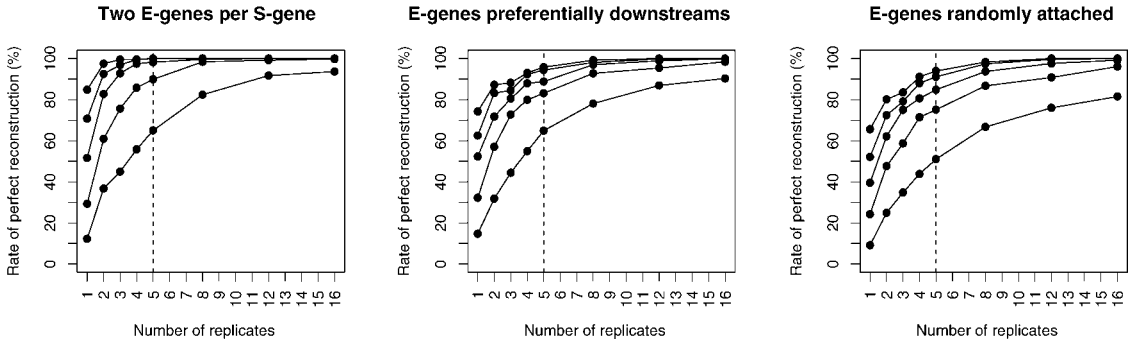

Fig. 2. Results of simulation experiments on random graphs. The number of replicates r in the data are on the x-axis, whereas the y-axis corresponds to the rate of perfect reconstructions in 1000 runs. Each plot corresponds to a different way of attaching E-genes to S-genes. The curves in each plot correspond to $\alpha_{\text{data}} = 0.1, \ldots, 0.5$ in descending order; the lower the curve, the higher the noise in data generation. The dashed vertical line indicates performance with $r = 5$ replicates—a practical upper limit for most microarray studies. The plots show excellent results for low noise levels. Even with $\alpha_{\text{data}} = 0.5$ the method does not break down, but identifies the complete true pathway in more than half of all simulation runs.

marginal likelihood can be calculated as follows. The numbers above the equality sign indicate which assumption was used in each step.

$$\begin{aligned} P(D|\Phi) &= \int P(D|\Phi,\Theta)P(\Theta|\Phi)\mathrm{d}\Theta \\ &\overset{[1,2]}{=} \prod_{i=1}^{m}\int P(D_i|\Phi,\theta_i)P(\theta_i|\Phi)\mathrm{d}\theta_i \\ &\overset{[3]}{=} \frac{1}{n^m}\prod_{i=1}^{m}\sum_{j=1}^{n}P(D_i|\Phi,\theta_i=j) \\ &\overset{[1]}{=} \frac{1}{n^m}\prod_{i=1}^{m}\sum_{j=1}^{n}\prod_{k=1}^{l}P(e_{ik}|\Phi,\theta_i=j). \end{aligned} \tag{2}$$

The marginal likelihood in Equation (2) contains the error probabilities α and β as free parameters to be chosen by the user. In Section 3.2 we will show how to estimate these parameters from data.

Estimated position of E-genes. Given a silencing scheme Φ, we can calculate the posterior probability for the edge between S_j and E_i as

$$P(\theta_i = j|\Phi,D) = \frac{1}{Z}\prod_{k=1}^{l}P(e_{ik}|\Phi,\theta_i=j), \tag{3}$$

with a uniform prior for E-gene position and normalizing constant Z chosen such that the probabilities for E_i sum up to one over all S-genes. The E-genes attached with high probabilty to an S-gene are interpreted as a regulatory module, which is under the common control of the S-gene.

3 RESULTS

We demonstrate the potential of our algorithm in two steps. First, we investigate accuracy and sample size requirements in a controlled simulation setting. In a second step, we show that our approach is also useful in a real biological scenario by applying it to a dataset on *Drosophila* immune response.

3.1 Accuracy and sample size requirements

We performed simulations consisting of five steps:

(1) S-genes. Randomly generate a directed acyclic graph T with 20 nodes and 40 edges. This is the core topology of S-genes.

(2) E-genes. Connect 40 E-genes to core T. This forms an extended topology T'. To evaluate how the position of E-genes affects the results we try three different ways of attaching E-genes to S-genes: deterministically two E-genes per S-gene, uniformly distributed positions, or preferentially downstream positions (also random but with a higher probability for S-genes at the end of pathways).

(3) Data. Generate one random dataset D from the extended topology T'. We use eight different repetition numbers per knock-out experiment ($r \in \{1, 2, 3, 4, 5, 8, 12, 16\}$). The experiment consists of $20 \cdot r$ 'microarrays', each corresponding to one of r repeated knock-outs of one of the 20 signaling genes. For each knock-out experiment the response of all E-genes is simulated from T' using error probabilities α_{data} and β_{data}. The false negative rate is fixed to $\beta_{\text{data}} = 0.05$ and the false positive rate α_{data} is varied from 0.1 to 0.5.

(4) Hypotheses. Randomly select three existing edges in the graph T, and three pairs of non-connected nodes. Using these six edges, there are $2^6 = 64$ possible modifications of T, including the original pathway T itself: some of the selected edges in T may be missing and some new links may be added. We take the 64 pathways as input hypotheses of our algorithm.

(5) Scoring. Score the pathway hypotheses by marginal likelihood with parameters $\alpha_{\text{score}} = 0.1$ and $\beta_{\text{score}} = 0.3$. Note that these (arbitrarily chosen) values are different from $(\alpha_{\text{data}}, \beta_{\text{data}})$ used for data generation. If the best score is achieved by the original pathway T we count this as a perfect reconstruction. Even with a single incorrect edge we count the reconstruction as failed.

We report the average number of perfect reconstructions for every $(\alpha_{\text{data}}, r)$-pair over 1000 simulation runs. Results are summarized in Figure 2.

The plots show that rates of perfect reconstruction are best when each S-gene has two E-genes as reporters and worst for purely random E-gene connections. The frequency to identify the correct pathway quickly increases with the number of replicates. With five replicates and low noise levels, the rate of perfect reconstruction is >90% in all simulations. Even with a noise level of 50% we correctly identified the right hypothesis in more than half of the runs.

3.2 Application to *Drosophila* immune response

We applied our method to data from a study on innate immune response in *Drosophila* (Boutros *et al*., 2002), which was already described as an example in Section 1. Selectively removing

signaling components (*S*-genes in our terminology) blocked induction of all, or only parts, of the transcriptional response to LPS (*E*-genes in our terminology).

Data preprocessing. The dataset consists of 16 Affymetrix microarrays: 4 replicates of control experiments without LPS and without RNAi (negative controls), 4 replicates of expression profiling after stimulation with LPS but without RNAi (positive controls) and 2 replicates each of expression profiling after applying LPS and silencing 1 of the 4 candidate genes *tak*, *key*, *rel* and *mkk4*/*hep*. For preprocessing, we perform normalization on probe level using a variance stabilizing transformation (Huber *et al*., 2002), and probe set summarization using a median polish fit of an additive model (Irizarry *et al*., 2003). In this data, 68 genes show a >2-fold up regulation between control and LPS stimulation. We used them as *E*-genes in our analysis.

Discretization and error rates. Next, we transformed the continuous expression data to binary values. We set an *E*-gene's state in an RNAi experiment to `1` if its expression value is sufficiently far from the mean of the positive controls, i.e. if the intervention interrupted the information flow. If the *E*-genes expression is close to the mean of positive controls, we set its state to `0`.

Let C_{ik} be the continuous expression level of E_i in experiment k. Let μ_i^+ be the mean of positive controls for E_i, and μ_i^- the mean of negative controls. To derive binary data E_{ik}, we defined individual cutoffs for every gene E_i by

$$E_{ik} = \begin{cases} 1 & \text{if } C_{ik} < \kappa \cdot \mu_i^+ + (1-\kappa) \cdot \mu_i^-, \\ 0 & \text{else.} \end{cases}$$

We tried values of κ from 0.1 to 0.9 in steps of 0.1. To control the false negative rate, we chose $\kappa = 0.7$: it is the smallest value where all negative controls are correctly recognized. This discretization is consistent with a small value of false negative rate β. We set it to $\beta = 0.05$. The false positive rate α was estimated from the positive controls: the relative frequency of negative calls there was just below 15%. Thus we set $\alpha = 0.15$. Trying different values of α and β did not change the results qualitatively, except when very large und unrealistic error probabilities were chosen.

Figure 3 shows the continuous and discretized data as used in our analysis. Silencing *tak* affects almost all *E*-genes. A subset of *E*-genes is additionally affected by silencing *mkk4*/*hep*, another disjoint subset by silencing *rel* and *key*. Note that expression profiles of *rel* and *key* silencing are almost indistinguishable both in the continuous and discrete datamatrix.

Results. We took all possible pathways on four genes as input to our algorithm. The four *S*-genes can form $2^{12} = 4096$ pathways, which result in 355 different silencing schemes. The distribution of marginal likelihood over the 30 top ranked silencing schemes in Figure 4 shows a clear peak: a single silencing scheme achieves the best score. It is well separated from a group of four silencing schemes having almost the same second-best score. Only after a wide gap all other silencing schemes follow.

The topology of the best silencing scheme is shown in Figure 4b. It can be constructed from three different pathway hypotheses: one is the topology shown in Figure 4, which is transitively closed, the other two miss either the edge from *tak* to *rel* or from *tak* to *key*. The key features of the data are preserved in all of them. The signal runs through *tak* before splitting into two pathway branches, one containing *mkk4*/*hep*, the other both *key* and *rel*. There is no hint to cross-talk between the two branches of the pathway. All in all, our result fits exactly to the conclusions Boutros *et al*. (2002) drew from the data.

The order of *key* and *rel* cannot be resolved from this dataset (see the nearly identical profiles in Fig. 3). However, it is known that *rel* is the transcription factor regulating the downstream genes (Boutros *et al*., 2002). This knowledge could have been easily introduced into a model prior $P(\Phi)$ penalizing topologies not showing *rel* below *key*. We refused to do this on purpose; our results here show how well pathway features can be reconstructed just based on experimental data, without any biological prior knowledge.

4 DISCUSSION

We have described a computational method for reconstructing pathway topologies based on differential gene expression in assays using external interventions like RNAi. Unlike previous work, our method is designed to deal with indirect observations. Simulation studies have demonstrated the power of our algorithm to choose pathways well supported by data. The applicability of our method to real world data could be shown in an application to a small RNAi study in *Drosophila*.

A measure for uncertainty. In Bayesian terminology, maximizing the marginal likelihood is equivalent to calculating the mode of the posterior distribution on model space, assuming a uniform prior. When scoring all possible pathways, we have derived a complete posterior distribution on model space, which does not only estimate a single pathway model, but also accurately describes the uncertainties involved in the reconstruction process. A flat posterior distribution indicates ambiguities in reconstructing the pathway. What we find in Figure 4 is a well pronounced maximum, which shows that we found the dominant structure in the data with high certainty. Still, we can only reconstruct features of the pathway, not the full topology. This stems from inherent limits of reconstruction from indirect observations. We discuss here prediction equivalence and likelihood equivalence.

Prediction equivalence. More than one pathway hypothesis result in the same silencing scheme if they only differ in transitive edges. An example are the three topologies sharing the silencing scheme of Figure 4 as discussed above. Since our score is defined on silencing schemes and not on topologies directly, the hypotheses with the same silencing scheme are not distinguishable. Assuming parsimony, we can represent each silencing scheme by a graph with minimal number of edges. This technique is called transitive reduction (Aho *et al*., 1972; van Leeuwen, 1990; Wagner, 2001, 2004).

Likelihood equivalence. We can also construct cases, where two hypotheses with different silencing schemes produce identical data. Figure 5 shows an example with a cycle of *S*-genes and a linear cascade, where all *E*-genes are attached on the downstream end. All *E*-genes react to interventions at every *S*-gene. In this case, the data do not prefer one silencing scheme over the other; both will have the same likelihood.

Epistatic effects. The model we use in this paper is very simple. Additional constraints, which are not dealt with in our model, are imposed by epistatic effects: one gene can mask the effect of another gene. These effects can be included into our model by

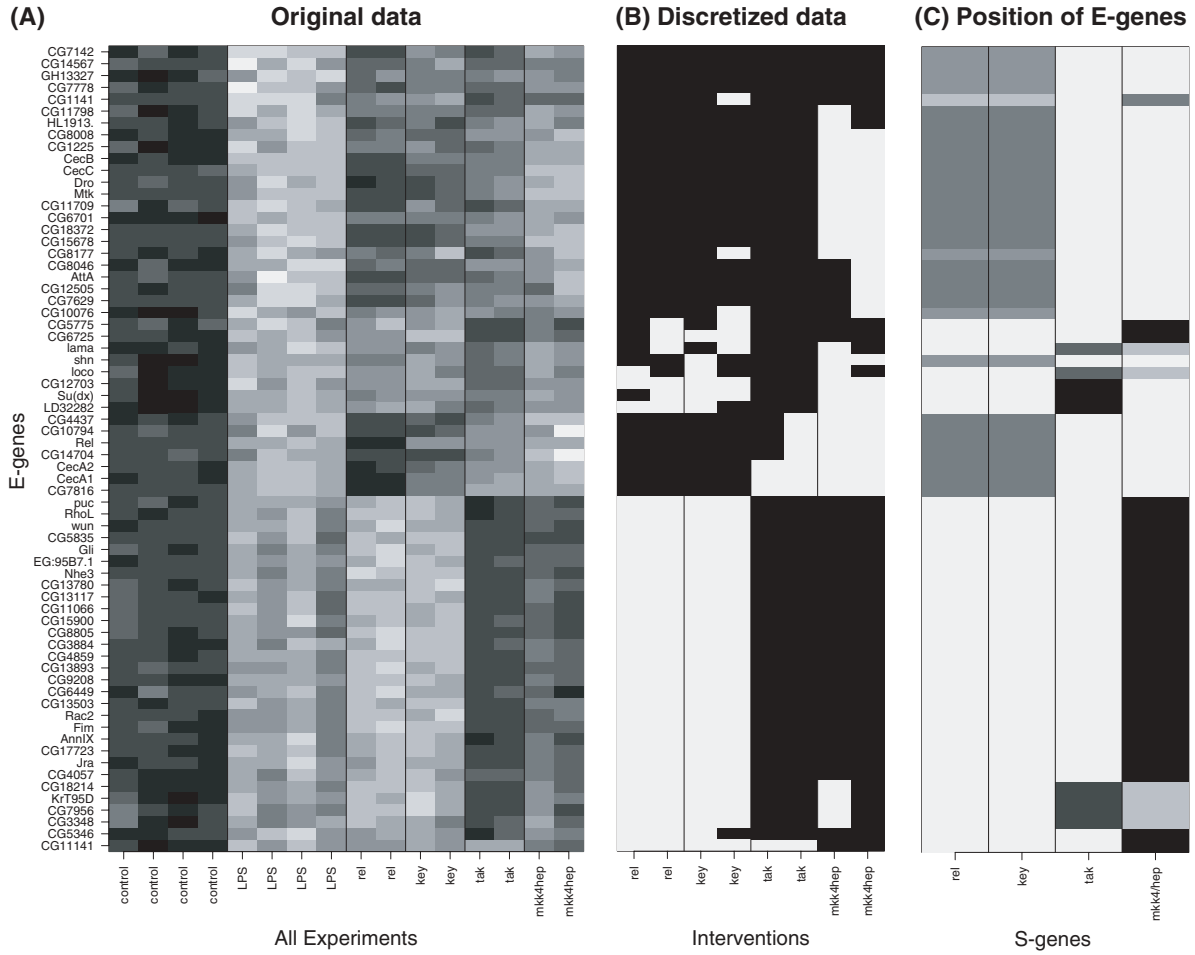

Fig. 3. Data on *Drosophila* immune response. (**A**) the normalized, genewise scaled data from (Boutros *et al*., 2002). Black stands for low expression and white for high expression. Rows are *E*-genes selected for differential expression after LPS stimulation (as seen in the first eight colums). The second eight columns correspond to silencing the four *S*-genes: silencing *rel* or *key* cuts the signal flow to the upper part of *E*-genes. Silencing *tak* reduces expression in all *E*-genes. Silencing *mkk4*/*hep* affects the lower half of *E*-genes. (**B**) the data from silencing experiments after discretization ($\kappa = 0.7$) as used in our analysis. (**C**) the expected position of *E*-genes given the silencing scheme with highest marginal likelihood of the data (Figure 4) computed from Equation (3). The lower half of *E*-genes is attributed to *mkk4*/*hep*, the upper half mostly to *key* and *rel*, which show almost the same intervention profiles in the middle figure.

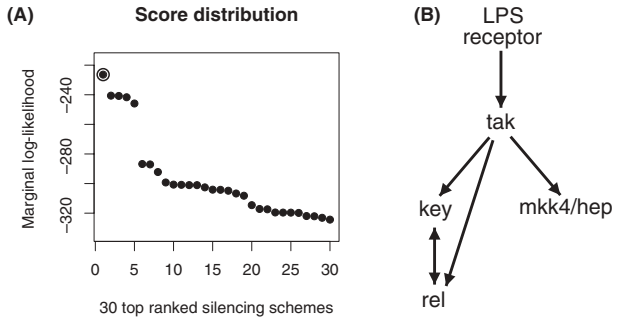

Fig. 4. Results on *Drosophila* data. (**A**) the score distribution over the 30 top scoring silencing schemes. One silencing scheme (circled) achieves the best score. It is well separated from a small group of four lagging behind with a pronounced gap to the rest. (**B**) the topology of the top-scoring silencing scheme.

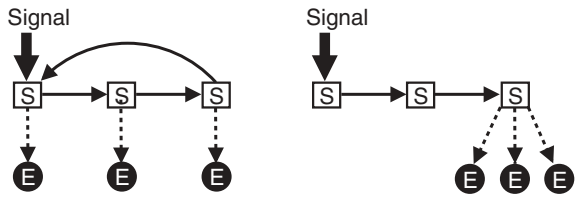

Fig. 5. Likelihood equivalence: the two plots show different topologies of *S*-genes with two distinct silencing schemes. However, both pathways will produce the same data: all *E*-genes react to interventions at every *S*-gene.

introducing a set of Boolean functions $F = \{f_S, S \in \mathbf{S}\}$. Each $f_S \in F$ determines the state of *S*-gene *S* given the states of its parents in *T*. Two simple examples of local functions f_S are AND- and OR-logics. In an AND-logic, all parent nodes must be affected by an intervention (i.e. have state `1`) to propagate the silencing effect to the child.

This describes redundancy in the pathway: if two genes fulfill alternative functions, both have to be silenced to stop signal flow through the pathway. In an OR-logic, one affected parent node is enough to set the child's state to 1. This describes a set of genes jointly regulating the child node; silencing one of the parents destroys the collaboration. The topology T together with the set of functions F defines a deterministic Boolean network on $\mathbf{S}$.

Multiple knock-outs. Since epistatic effects involve more than one gene, they cannot be deduced from single knock-out experiments. One can extend our method to data attained by silencing more than one gene at the same time. This will not change our scoring function, but more sophisticated silencing schemes have to be developed, which encode predictions both from single-gene and multi-gene knock-outs. Since the number of possible multiple knock-outs increases exponentionally, we need tools to choose the most informative experiments (Yoo and Cooper, 2004; Yeang *et al*., 2005).

In summary. This is the first paper addressing pathway reconstruction from indirect observations. Our algorithm reconstructs pathway features from the nested structure of affected downstream genes. Pathway features are encoded as silencing schemes. They contain all information to predict a cell's behavior to an external intervention. In simulation studies we confirmed small sample size requirements and high accuracy. Limitations only result from the information content of indirect observations.

ACKNOWLEDGEMENTS

We thank Michael Boutros for providing the expression data and for introducing us to the world of RNAi. Special thanks go to Chen-Hsiang Yeang and Achim Tresch for many inspiring discussions on network reconstruction. We are also grateful to Stefanie Scheid and Dennis Kostka for their valuable comments. This research has been supported by BMBF grants 03U117 and 01GR0455 of the German Federal Ministry of Education and Research.

Conflicts of Interest: none declared.

REFERENCES

Aho,A.V. *et al*. (1972) The transitive reduction of a directed graph. *SIAM J. Comput*., **1**, 131–137.

Akutsu,T., Kuhara,S., Maruyama,O. and Miyano,S. (1998) Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 695–702.

Akutsu,T. (1999) Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pac. Symp. Biocomput*., 17–28.

Avery,L. and Wasserman,S. (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet*., **8**, 312–316.

Boutros,M. *et al*. (2002) Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Dev. Cell*, **3**, 711–722.

Boutros,M. *et al*. (2004) Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science*, **303**, 832–835.

Cooper,G.F. and Yoo,C. (1999) Causal discovery from a mixture of experimental and observational data. In Laskey,K. and Prade,H. (eds), *Proceedings of the Fifthteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufman, San Francisco, CA, pp. 116–125.

Di Bernardo,D. *et al*. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineering gene networks. *Nat. Biotechnol*., **23**, 377–383.

Driessche,N.V. *et al*. (2005) Epistasis analysis with global transcriptional phenotypes. *Nat. Genet*., **37**, 471–477.

Fire,A. *et al*. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.

Fraser,A.G. *et al*. (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature*, **408**, 325–330.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Friedman,N. *et al*. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol*., **7**, 601–620.

Gardner,T.S. *et al*. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

Gönczy,P. *et al*. (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature*, **408**, 331–336.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T. *et al*. (2000) *Pac. Symp. Biocomput*., **5**, 302–313.

Imoto,S. *et al*. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pac. Symp. Biocomput*., **7**, 175–186.

Irizarry,R.A. *et al*. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, e15.

Markowetz,F. and Spang,R. (2003) Evaluating the effect of perturbations in reconstructing network topologies. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Markowetz,F., Grossmann,S. and Spang,R. (2005) Probabilistic soft interventions in conditional Gaussian networks. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*.

Nachman,I. *et al*. (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, **20** (Suppl. 1), i248–i256.

Nature insight RNA interference (2004), **431**, 338–378.

Pe'er,D. *et al*. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, S215–S224.

Rangel,C., Wild,D.L., Falciani,F., Ghahramani,Z. and Gaiba,A. (2001) Modeling biological responses using gene expression profiling and linear dynamical systems. In *Proceedings of the 2nd International Conference on Systems Biology*. Madison, WI, Omnipress, pp. 248–256.

Rangel,C. *et al*. (2004) Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, **20**, 1361–1372.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.

Spradling,A.C. *et al*. (1999) The Berkeley *Drosophila* Genome Project gene disruption project: single P-element insertions mutating 25% of vital *Drosophila* genes. *Genetics*, **153**, 135–177.

Tegner,J. *et al*. (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamic modeling. *Proc. Natl Acad. Sci. USA*, **100**, 5944–5949.

van Leeuwen,J. (1990) Graph algorithms. In van Leeuwen,J. (ed.), *Handbook of Theoretical Computer Science*. Elsevier, pp. 525–632.

Wagner,A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. *Bioinformatics*, **17**, 1183–1197.

Wagner,A. (2004) Reconstructing pathways in large genetic networks from genetic perturbations. *J. Comput. Biol*., **11**, 53–60.

Wessels,L. *et al*. (2001) A comparison of genetic network models. *Pac. Symp. Biocomput*., **2001**, 508–519.

Wille,A. *et al*. (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol*., **5**, R92.

Yeang,C.H. *et al*. (2004) Physical network models. *J. Comput. Biol*., **11**, 243–262.

Yeang,C.H. *et al*. (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol*., **6**, R62.

Yoo,C. and Cooper,G.F. (2004) An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. *J Artif. Intell. Med*, **31**, 169–182.