Chapter 4

Inferring signal transduction pathways

The last chapter dealt with models of primary effects. We assumed that perturbing one pathway component leads to detectable changes at other pathway components. In this chapter I introduce a method designed for indirect observations of pathway activity by secondary effects at downstream genes (section 4.1). I present an algorithm to infer non-transcriptional pathway features based on differential gene expression in silencing assays. The main contribution is a score linking models to data (section 4.2). I demonstrate its power in the controlled setting of simulation studies (section 4.3) and explain its practical use in the context of an RNAi data set investigating the response to microbial challenge in Drosophila melanogaster *(section 4.4).*

4.1 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. While gene regulation leaves direct traces on expression profiles, non-transcriptional signaling does not. Thus, on microarray data gene regulatory networks can be modelled by methods described in chapters 2 and 3, while non-transcriptional pathways can not. However, reflections of signaling activity can be perceived in expression levels of other genes. We explain this in a simplified pathway model and in a real world example in *Drosophila*.

A hypothetical pathway Fig. 4.1 shows a hypothetical biochemical pathway adapted from Wagner [140]. It consists of two transcription factors, a protein kinase and a protein phosphatase and the genes encoding these proteins. The figure shows the three biological levels of interest: genome, transcriptome and proteome. The thick arrows show information flow through the pathway. The transcription factor expressed by gene *1* binds to the promoter region of gene *2* and activates it. Gene *2* encodes a protein kinase, which phosphorylates a protein phosphatase (expressed by gene *3*). This event activates the protein phosphatase, which now dephosphorylates the transcription factor produced by gene *4*. It binds to gene *5* and induces expression.

The three biological levels of DNA, mRNA and protein are condensed into a graph model on five nodes. Gene expression data only shows the mRNA level. A model inferred from expression data will only have two edges, connecting gene *1* to gene *2* and then gene *2* to gene *5*. Since genes *3* and *4* only contribute on the protein level, a model based on correlations on the mRNA level will ignore them. This holds true for all models descibed in chapter 2.

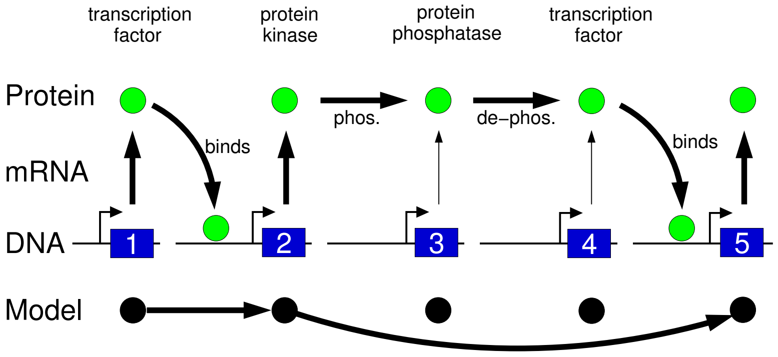

Figure 4.1: *A hypothetical biochemical pathway adapted from Wagner [140]. It shows four levels of interest: three biological and one of modeling. Inference from gene expression data alone only gives a very limited model of the pathway. The contributions of genes* 3 *and* 4 *are overlooked.*

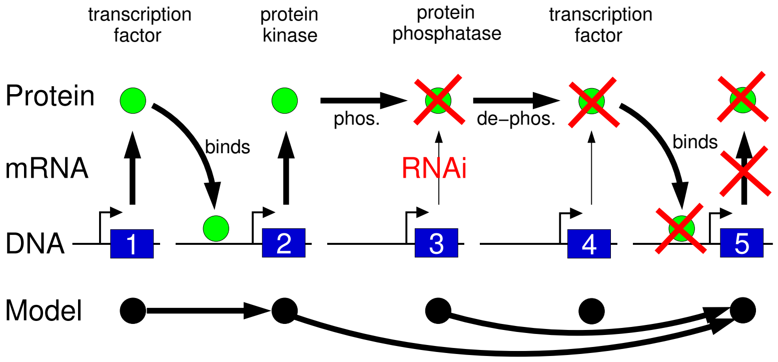

Figure 4.2: *The situation changes if we can use interventional data for model building. Silencing gene* 3 *by RNAi will cut information flow in the pathway and result in an expression change at gene* 5*. This is visible on the mRNA level and can be integrated in the model. Thus, the expanded model shows an edge from gene* 3 *to gene* 5*.*

Interventions at genes in the pathway shed light on the pathway topology. This is exemplified by an RNAi intervention at gene *3* in Fig. 4.2. Silencing gene *3* will cut information flow in the pathway and result in an expression change at gene *5*. This is reflected in the model by extending it to include an edge from gene *3* to gene *5*. Note that we have no observation of direct effects of the intervention at gene *4* in mRNA data. The only information we have are secondary effects at the transcriptional end of the pathway. This chapter will introduce novel methodology to order genes in regulatory hierarchies from secondary effects. The procedure is motivated by the logic underlying a study in *Drosophila* conducted by Michael Boutros and coworkers.

An example in Drosophila Boutros *et al.* [12] 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. Sixty minutes after 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 signaling 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*. Gene *tak* codes for protein TAK1 in Fig. 1.2, *key* for IKKγ, and *rel* is the transcription factor Relish, already discussed in the introduction in chapter 1.

Boutros *et al.* [12] 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. The interpretation is a Relish-independent response to LPS, which is also triggered by IMD and TAK but then branches off the Imd pathway. 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 is to derive an algorithm for systematic inference from indirect observations.

Models for primary effects cannot be applied here In chapter 3, we discussed models to explain primary effects of silencing genes on other genes in the pathway. Some are deterministic and graph based, some are probabilistic and able to handle noise in the data. All of them aim for transcriptional networks and are unable to capture non-transcriptional modulation. Some approaches use hidden variables to capture non-transcriptional effects [89, 104, 105] without making 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. In graphical models with hidden variables non-transcriptional effects are considered nuisance, not the main target of pathway reconstruction. In summary, none of the methods designed to infer transcriptional networks can be applied to reconstruct non-transcriptional pathways from microarray data. The major problem is: these algorithms require direct observations of expression changes of signaling genes, which are not fully available in datasets like that of [12]. There exist only two methodologies comparable to ours in being able to identify non-transcriptional pathway features from microarray data: *physical network models* and *epistasis analysis*.

Physical network models Yeang *et al.* [149] introduce a maximum likelihood based approach to combine three different yeast datasets: protein–DNA, protein–protein, and single gene knock-out data. The first two data sources indicate direct interactions, while the knock-out data only contains indirect functional information. The algorithm searches for topologies which are consistent with observed downstream effects of interventions. While it is not confined to the transcriptional level of regulation, it also requires that most signaling genes show effects when perturbing others. It is not designed for a dataset like that of Boutros *et al.* [12] described above.

Epistasis analysis Our general objective is similar to epistasis analysis with global transcriptional phenotypes. Regulatory hierarchies can be identified by comparing single-knockout phenotypes to double-knockout phenotypes. Driessche *et al.* [31] use gene expression time-courses as phenotypes and reconstruct a regulatory system in the development of *Dictyostelium discoideum*, a soil-living amoeba. Yet, there are several important differences. 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. This will be made clear in the following overview.

How to learn from secondary effects We present a computational framework for the systematic reconstruction of pathway features from expression profiles relating to external interventions. The approach is based on the nested structure of affected downstream genes, which are themselves not part of the model. Here we give a short overview of the method before presenting it in all details in section 4.2. The model distinguishes 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"). Because 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.

The basic idea is to model 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

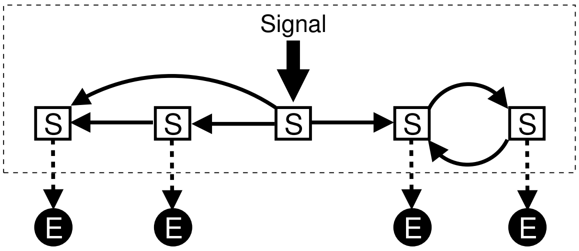

Figure 4.3: *A schematic summary of our model. The dashed box indicates one hypothesis: it contains a directed graph T on genes contributing to a signaling pathway (S-genes). A signal enters the pathway at one (or possibly more than one) specified position. Interventions at S-genes interrupt signal flow 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 changes of E-genes. The objective is to reconstruct relationships between S-genes from observations of E-genes in silencing experiments.*

replicated several times. This results in a data set representing every signaling gene by one or more microarrays. These requirements are the same as in epistasis analysis [6], but they are not satisfied in all datasets monitoring intervention effects. In the Rosetta yeast compendium [61], for example, there is no external stimulus by which the interruption of signal flow through a pathway of interest could be measured.

The main contribution of this chapter is a scoring function, which measures how well hypotheses about pathway topology are supported by experimental data. *Input* to the 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 Fig. 4.3. The 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. All pathways, which make the same predictions of intervention effects on downstream genes, are identified by one so called *silencing scheme*. Sorting silencing schemes by our scoring function 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.

4.2 Gene silencing with transcriptional phenotypes

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.

4.2.1 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 $\mathbf{S} = \{S_1, \ldots, S_p\}$. 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. While 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 predicted from the pathway hypothesis. Mathematically, a silencing scheme is the transitive closure of pathway T implying 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. An example is given in Fig. 4.4. Note that the E-genes do not appear in Φ, which only describes interactions between S-genes. The E-genes come into play when inferring silencing schemes. Reduced signaling strength of S-genes due to interventions in the pathway cannot be observed directly on a microarray, but secondary effects are visible 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 $\mathbf{S}$. In particular, the E-genes do not interact with each other. 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. 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

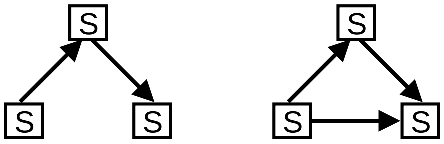

Figure 4.4: *Transitive closure. The right topology is the transitive closure of the left topology. When adding an entry point for signal, both are valid pathway hypotheses. Both are represented by a silencing scheme, which has the same topology as the right graph.*

can be experimentally observed as differential expression on microarrays. Due to the observational noise or stochastic effects in signal transduction, we expect a number of false positive and false negative observations.

4.2.2 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 several microarrays per silenced gene are included in the dataset. Microarrays are indexed by $k = 1, \dots, l$. 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. The result is a binary matrix $M = (e_{ik})$, where $e_{ik} = 1$ if E-gene E_i shows an effect in experiment k. Thus, our data only consists in coarse qualitative information. We do not consider whether an E-gene was up- or down-regulated 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 The positions of E-genes are included as model parameters $\Theta = \{\theta_i\}_{i=1}^m$ with $\theta_i \in \{1, \dots, n\}$ and $\theta_i = j$ if E_i is attached to S_j. Let us first consider a fixed extension T' of T, that is, 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 T' predicts effects at all E-genes, which are attached to an S-gene in the influence region. Expected effects can be compared to observed effects in the data to choose the topology, which fits the data best. Due to measurement noise no topology T' is expected to be in complete agreement with all observations. Deviations from predicted effects are allowed by introducing global error probabilities α and β for false positive and negative calls, respectively.

The expression levels of E-genes on the various microarrays are modelled 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

$$p(e_{ik}|\Phi,\theta_i = j) = \left\{ \begin{array}{ccl} e_{ik} = 1 & e_{ik} = 0 & \\ \hline \alpha & 1-\alpha & \text{if } S_j = 0 \\ 1-\beta & \beta & \text{if } S_j = 1 \end{array} \right. \tag{4.1}$$

This means: if the parent of 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 $p(M|\Phi,\Theta)$ of the data is then a product of terms from the table for every observation, that is,

$$p(M|\Phi,\Theta) \ = \ \prod_{i=1}^{m}\prod_{k=1}^{l} p(e_{ik}|\Phi,\theta_i) \ = \ \alpha^{n_{10}}\beta^{n_{01}}(1-\alpha)^{n_{00}}(1-\beta)^{n_{11}}, \tag{4.2}$$

where n_{se} is the number of times we observed E-genes in state e when their parent S-gene in Φ was in state s.

However, in reality the “correct” extension T' of a candidate topology T is unknown. 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 described next.

4.2.3 Marginal likelihood of a silencing scheme

This section defines a scoring function to link models with observations. It evaluates how well a given silencing scheme Φ fits the experimental 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 position parameters Θ are unknown. By Bayes' formula the posterior of silencing scheme Φ given data M can be written as

$$p(\Phi|M) = \frac{p(M|\Phi)p(\Phi)}{p(M)}. \tag{4.3}$$

The normalizing constant $p(M)$ is the same for all silencing schemes, it can be neglected for relative model comparison. The model prior $p(\Phi)$ can be chosen to incorporate biological prior knowledge. In the following, we assume it to be uniform over all possible models. What remains is the marginal likelihood $p(M|\Phi)$. It equals the likelihood $p(M|\Phi,\Theta)$ 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 M are sampled independently and distributed identically:
$$p(M|\Phi,\Theta) = \prod_{i=1}^{m} p(M_i|\Phi,\theta_i) = \prod_{i=1}^{m}\prod_{k=1}^{l} p(e_{ik}|\Phi,\theta_i),$$
where M_i is the ith row in data matrix M.
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 distribution. The prior probability to attach an E-gene is uniform over all S-genes:
$$P(\theta_i = j|\Phi) = \frac{1}{p} \quad \text{for all } i \text{ and } j.$$

The last assumption can easily be dropped to include existing biological prior knowledge about regulatory modules. With the assumptions above, the marginal likelihood can be calculated as follows. The numbers above the equality sign indicate which assumption was used in each step.

$$\begin{aligned}
p_{\alpha,\beta}(M|\Phi) &= \int p_{\alpha,\beta}(M|\Phi,\Theta)\; p(\Theta|\Phi)\; \mathrm{d}\Theta \\
&\overset{[1,2]}{=} \prod_{i=1}^{m} \int p_{\alpha,\beta}(M_i|\Phi,\theta_i)\; p(\theta_i|\Phi)\; \mathrm{d}\theta_i \\
&\overset{[3]}{=} \frac{1}{p^m}\prod_{i=1}^{m}\sum_{j=1}^{p} p_{\alpha,\beta}(M_i|\Phi,\theta_i = j) \\
&\overset{[1]}{=} \frac{1}{p^m}\prod_{i=1}^{m}\sum_{j=1}^{p}\prod_{k=1}^{l} p_{\alpha,\beta}(e_{ik}|\Phi,\theta_i = j). \qquad (4.4)
\end{aligned}$$

The marginal likelihood in Eq. (4.4) contains the error probabilities α and β as free parameters to be chosen by the user. This is indicated by subscripts. In section 4.4 we will show how to estimate these parameters when discretizing the data.

Estimated position of E-genes Given a silencing scheme Φ, the posterior probability for an edge between S_j and E_i is given by

$$P_{\alpha,\beta}(\theta_i = j|\Phi, M) \;=\; \frac{p(\theta_i = j \mid \Phi)}{p_{\alpha,\beta}(M_i \mid \Phi)}\prod_{k=1}^{l} p_{\alpha,\beta}(e_{ik} \mid \Phi, \theta_i = j) \qquad (4.5)$$

where the prior $p(\theta_i = j|\Phi)$ is again chosen to be uniform. In general, the prior could take any other form as long as it is the same as in the computation of marginal likelihood above. 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.

4.2.4 Averaging over error probabilities α and β

The likelihood in Eq. (4.4) is a polynomial in α and β. In a full Bayesian approach we would again average over possible values of α and β given a prior distribution. This problem can be cast in a way accessible to standard Bayesian theory, as it is also used when averaging over LPD parameters to gain the marginal likelihood in Bayesian network structure learning (see section 2.12). So far, we assumed that all E-genes share the distribution specified in Eq. (4.1) and α and β are indeed global parameters applicable to every E-gene. This simplifying assumption was introduced to keep inference feasible. Else, we would have to estimate parameters (α_i, β_i) for every E-gene E_i. When averaging over LPD parameters, we will drop the assumption of parameter sharing. Instead we augment the three assumptions above by three additional ones.

First we define $\eta_i = (\eta_{i0}, \eta_{i1}) = (\alpha_i, 1-\beta_i)$, then for one E-gene E with parent S holds $\eta_{is} = P(E_i = 1 | S_{\theta_i} = s)$. We make the following assumptions on the prior distribution $p(\eta|\Phi,\Theta)$ of $\eta = (\eta_i)_{i=1,\dots,m}$:

4. Global and local parameter independence. Parameters are independent for every E-gene E_i and for different states of the parent S-gene, that is,

$$p(\eta|\Phi,\Theta) \;=\; \prod_{i=1}^{m} p(\eta_i|\Phi,\theta_i) \;=\; \prod_{i=1}^{m} \prod_{s\in\{0,1\}} p(\eta_{is}|\Phi,\theta_i).$$

5. The prior $p(\eta_{is}|\Phi,\theta_i)$ is chosen as a beta distribution, which is conjugate to the multinomial distribution of the E_i [49], that is,

$$p(\eta_{is}|\Phi,\theta_i) \;=\; \eta_{is}^{a_{is}-1}(1-\eta_{is})^{b_{is}-1}.$$

6. All local priors $p(\eta_{is}|\Phi,\theta_i)$ share the same parameters, that is,

$$a_{is} = a_s \quad \text{and} \quad b_{is} = b_s \quad \text{for all } i = 1,\dots,m.$$

The last assumption limits the number of parameters. It is parameter sharing not on the level of distribution parameters but on the level of parameters of prior distributions, which are themselves independent. With these assumptions we can compute the marginal likelihood with respect to position parameters Θ and effect probabilities η by

$$\begin{aligned} p(M|\Phi) \;&=\; \iint p(M|\Phi,\Theta,\eta)\; p(\eta|\Phi,\Theta)\; p(\Theta|\Phi)\; \mathrm{d}\eta\; \mathrm{d}\theta \\ &\overset{[4]}{=}\; \prod_{i=1}^{m} \int \left(\int p(M_i|\Phi,\theta_i,\eta_i)\; p(\eta_i|\Phi,\theta_i)\; \mathrm{d}\eta_i \right) p(\theta_i|\Phi)\; \mathrm{d}\theta_i. \end{aligned} \tag{4.6}$$

We first concentrate on one fixed E_i. Then Φ and θ_i specify the parent S-gene S_{θ_i} and its state $S_{\theta_i} = s$. The data M_i split into two subsets M_i^s and M_i^{1-s}, where

$M_i^s = \{e_{ik} | S_{\theta_i} = s\}$. Each batch of data follows the same binomial distribution in Eq. (4.1). The inner integral in Eq. (4.6) splits into two integrals, one for each parent state s, which can be computed as follows:

$$\begin{aligned} \int p(M_i^s|\Phi, \theta_i, \eta_{is}) p(\eta_{is}|\Phi, \theta_i) \, \mathrm{d}\eta_{is} &= \\ &\overset{[5,6]}{=} \frac{\Gamma(a_s + b_s)}{\Gamma(a_s)\Gamma(b_s)} \int \eta_{is}^{n_{is1}+a_s-1} (1-\eta_{is})^{n_{is0}+b_s-1} \, \mathrm{d}\eta_{is} \\ &= \frac{\Gamma(a_s + b_s)}{\Gamma(a_s)\Gamma(b_s)} \cdot \frac{\Gamma(n_{is1} + a_s)\Gamma(n_{is0} + b_s)}{\Gamma(n_{is1} + n_{is0} + a_s + b_s)}, \end{aligned} \tag{4.7}$$

where the counts n_{ise} denote the number of experiments, in which we observed $E_i = e$ while the parent S-gene S_{θ_i} was in state s. Note that this computation is identical to marginalizing LPD parameters in discrete Bayesian networks (section 3.4.2). The reason is that our model can be viewed as a highly restricted Bayesian network, in which the LPDs at S-genes are deterministic and the E-genes follow a conditional binomial distribution.

The data likelihood $p(M_i|\Phi, \theta_i)$ for gene E_i is a product of terms on the right hand side of Eq. (4.7) for both S-gene states. The marginalization over E-gene positions Θ works exactely as in section 4.2.3 and results in the following full marginal likelihood:

$$p(D|\Phi) = \frac{1}{p^m} \prod_{i=1}^{m} \sum_{j=1}^{p} \prod_{s\in\{0,1\}} \frac{\Gamma(a_s + b_s)\Gamma(n_{is1} + a_s)\Gamma(n_{is0} + b_s)}{\Gamma(a_s)\Gamma(b_s)\Gamma(n_{is1} + n_{is0} + a_s + b_s)}. \tag{4.8}$$

Estimated position of E-genes Similar to Eq. (4.5), the posterior probability for an edge between S_j and E_i with marginalization over α and β is given by

$$\begin{aligned} P(\theta_i = j|\Phi, M) &= \frac{1}{Z} \prod_{k=1}^{l} p(e_{ik} \mid \Phi, \theta_i = j) \\ &= \frac{1}{Z} \prod_{s\in\{0,1\}} \frac{\Gamma(a_s + b_s)\Gamma(n_{is1} + a_s)\Gamma(n_{is0} + b_s)}{\Gamma(a_s)\Gamma(b_s)\Gamma(n_{is1} + n_{is0} + a_s + b_s)}. \end{aligned} \tag{4.9}$$

where Z is a normalizing constant ensuring that the sum over all S-genes is 1. This equation allows to estimate E-gene positions given the beta prior on the local distribution parameters of E_i.

Summary of parameters Table 4.1 gives an overview of the ingredients to the formulas developed in this section. It shows counts, distribution parameters and prior parameters for the four possible combinations of E-gene state and parent S-gene state. The counts are E-gene specific, while the parameters (α, β) and prior parameters (a_0, b_0, a_1, b_1) apply to all E-genes. Having four prior parameters to specify, while before there were only two distribution parameters, may seem as a disadvantage of marginalization. But there are two considerations to keep in mind. First, a model is much more stable against choices of prior parameters than of distribution parameters.

		E_i	
		1	0
S	0	n_{i01}	n_{i00}
	1	n_{i11}	n_{i10}

	Eq. (4.4)	
	1	0
0	α	$1-\alpha$
1	$1-\beta$	β

	Eq. (4.8)	
	1	0
0	a_o	b_0
1	a_1	b_1

Table 4.1: *The table describes the main terms of the marginal likelihoods computed in this section. It focusses on one E-gene (columns) and its parent S-gene (rows). The left table contains the counts from the data for the four possible combinations of E-gene and parent state. They are E-gene specific and used in all formulas. To compute the marginal likelihood of Eq. (4.4) error probabilities α and β need to be specified, which are the same for all E-genes. For the full marginal likelihood of Eq. (4.8) the user needs to choose prior parameters (a_0, b_0) and (a_1, b_1), which are shared by all E-genes.*

In situations with little knowledge on error rates in experiments it is safer to use the full marginal likelihood of Eq. (4.8) than the marginal likelihood of Eq. (4.4). Second, the four prior parameters fall in two categories: (a_0, b_1) give weights for observing errors, while (a_1, b_0) give weights for observing the predicted state. This motivates to use only two values for the prior parameters: one for a_0 and b_1, and another one for a_1 and b_0. Because we expect there to be more signal than noise in the data, the value of $a_0 = b_1$ should be considerably smaller than that of $a_1 = b_0$. We will see an example in the application to *Drosophila* data in section 4.4.

4.2.5 Limits of learning from secondary effects

The method we described 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 *data equivalence*.

Prediction equivalence More than one pathway hypothesis result in the same silencing scheme if they only differ in transitive edges. An example is given in Fig. 4.4. Both topologies there can be considered as pathway hypotheses, but only the right one is transitively closed and thus a silencing scheme. 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, each silencing scheme can uniquely be represented by a graph with minimal number of edges. This technique is called *transitive reduction* [1, 75, 142, 140].

Data equivalence There exist cases, where two hypotheses with different silencing schemes produce identical data. Fig. 4.5 shows an example with a cycle of S-genes and a linear cascade, where all E-genes are attached at the downstream end. In both pathways, all E-genes react to interventions at every S-gene. In this case, the data does not prefer one silencing scheme over the other.

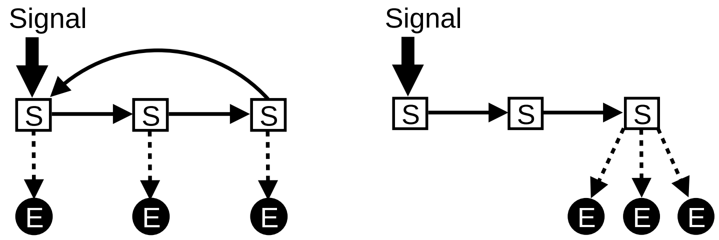

Figure 4.5: *Data equivalence: The two plots show different topologies of S-genes with two distinct silencing schemes. However, both pathways will produce the same data: All E-genes react to interventions at every S-gene.*

4.2.6 Extending the basic model

Epistatic effects The model described above is very simple. Additional constraints are imposed by epistatic effects: one gene can mask the effect of another gene. These effects can be included into the model by 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 $\mathbf{S}$. Fig. 4.6 gives an example, how local logics constrain influence regions and change silencing schemes.

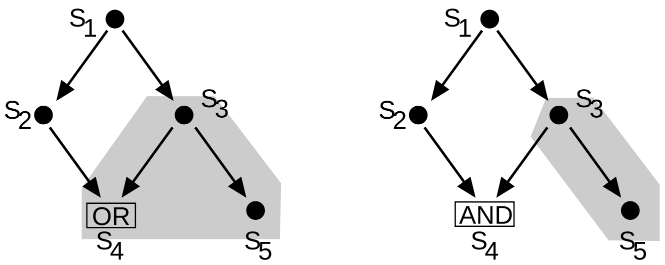

Figure 4.6: *Influence regions are constrained by local logics. The left plot shows in grey the influence region of S_3 if S_4 is reigned by an OR-logic. If the logic changes to an AND, S_4 lies no longer in the influence region of S_3, because the second parent S_2 lies outside of it.*

Multiple knockouts Since epistatic effects involve more than one gene, they cannot be deduced from single knock-out experiments. The model has to be extended to data attained by silencing more than one gene at the same time. This will not change the scoring function, but more sophisticated silencing schemes have to be developed, which encode predictions both from single-gene and multi-gene knockouts. Since the number of possible multiple knockouts increases exponentionally, tools to choose the most informative experiments are needed. Experimental design or *active learning* deals with deciding which interventions to perform to learn the structure of a model as quickly as possible and to discriminate optimally between alternative models. This is an active area of research in Machine Learning [138, 88]. For reconstruction of regulatory networks, a number of methods have been proposed in different frameworks: for Bayesian networks [103, 152], physical network models [150], Boolean networks [63], and dynamical modeling [135].

4.3 Accuracy and sample size requirements

Section 4.2 introduced a Bayesian score to find silencing schemes explaining the data well. We will demonstrate its potential 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. This section evaluates how our algorithm responds to different levels of noise in the data, how accurate it is and how many replicates of intervention screens are needed for reliable pathway reconstruction. To answer these questions, we performed simulations consisting of five steps:

1. We randomly generated a directed acyclic graph T with 20 nodes and 40 edges. This is the core topology of S-genes.
2. Then, we connected 40 E-genes to the core T of S-genes. Together they form an extended topology T'. To evaluate how the position of E-genes affects the results, we implemented three different ways of attaching E-genes to S-genes: either two E-genes are assigned to each S-gene, or E-gene positions are distributed uniformly, or positions are chosen preferentially downstream (also random but with a higher probability for S-genes at the end of pathways).
3. From the extended topology T' we generated random datasets using eight different repetition numbers per knockout experiment ($r \in \{1, \ldots, 5, 8, 12, 16\}$). The experiment then consists of $20 \cdot r$ "microarrays", each corresponding to one of r repeated knockouts of one of the 20 signaling genes. For each knockout 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. We randomly selected 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. The 64 pathways were used as input hypotheses of our algorithm.
5. We scored the 64 pathway hypotheses by the marginal likelihood of Eq. (4.4) with parameters $\alpha_{\text{score}} = 0.1$ and $\beta_{\text{score}} = 0.3$. Note that these (arbitrarily chosen)

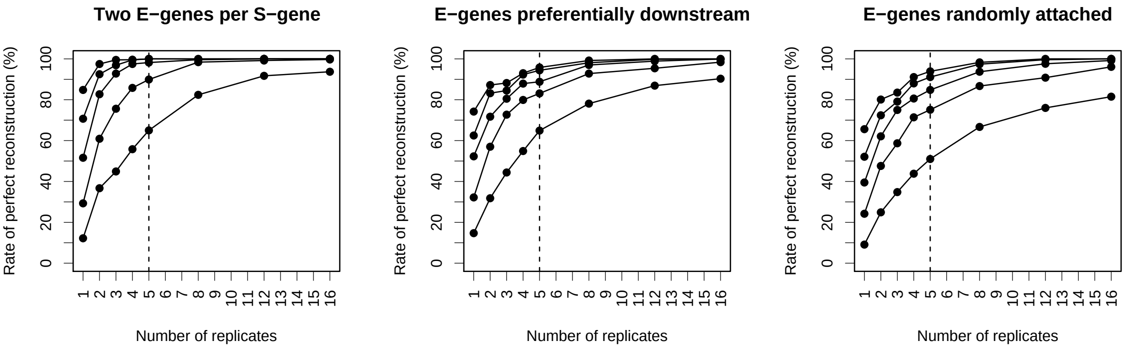

Figure 4.7: *Results of simulation experiments on random graphs. The number of replicates r in the data are on the x-axis, while the y-axis corresponds to the rate of perfect reconstructions in 1000 runs. Each plot corresponds to a different way of attaching E-genes to S-genes. The curves in each plot correspond to $\alpha_{data} = 0.1, \ldots, 0.5$ 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 $\alpha_{data} = 0.5$ the method does not break down, but identifies the complete true pathway in more than half of all simulation runs.*

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 this is counted as a perfect reconstruction. Even with a single incorrect edge the reconstruction is counted as failed.

Simulation results Fig. 4.7 depicts the average number of perfect reconstructions for every $(\alpha_{\text{data}}, r)$-pair over 1000 simulation runs. The plots show: 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 above 90% in all simulations. Even with a noise level of 50% the algorithm correctly identified the right hypothesis in more than half of the runs.

The impact of these simulation results becomes apparent when comparing it to results by graphical models of the correlation structure of expression values. Basso *et al.* [7] show that their own method, ARACNe, compares favorably against static Bayesian networks on a simulated network with 19 nodes. The smallest sample size used in the comparison is 100 observations, the biggest 2000. They show a steady increase in performance, which levels off at around 1000 observations. Hartemink [55] finds dynamical Bayesian networks to be even more accurate than ARACNe on the same simulation network with the same dataset sizes. In summary, at least 1000 observations are needed to reliably reconstruct a 19 node network by Bayesian networks or ARACNe. Our simulations show that less than 100 samples are needed to reconstruct a network of the same size when using gene silencing screens. This is one order of magnitude less. For 20 nodes, 100 observations correspond to five replicates per intervention, which give an almost consummate rate of perfect reconstruction in Fig. 4.7.

4.4 Application to Drosophila immune response

We applied our method to data from a study on innate immune response in *Drosophila* [12], which was already described as an example in the introduction. 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 one of the four candidate genes *tak*, *key*, *rel*, and *mkk4/hep*. For preprocessing, we performed normalization on probe level using a variance stabilizing transformation [60], and probe set summarization using a median polish fit of an additive model [67]. In this data, 68 genes show a more than 2-fold up-regulation between control and LPS stimulation. We used them as E-genes in the model.

Adaptive discretization Next, we transformed the continuous expression data to binary values. An E-gene's state in an RNAi experiment is set 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. Formally, this strategy is implemented as follows. 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:

$$E_{ik} = \begin{cases} 1 & \text{if } C_{ik} < \kappa \cdot \mu_i^+ + (1-\kappa) \cdot \mu_i^-, \\ 0 & \text{else.} \end{cases} \tag{4.10}$$

We tried values of κ from 0 to 1 in steps of 0.1. Fig. 4.8 shows the results. To control the false negative rate, we chose $\kappa = 0.7$: It is the smallest value where all negative controls are correctly recognized.

Figure 4.9 shows the continuous and discretized data as used in the 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 data matrix. The subset structure observed by Boutros *et al.* [12] is visible, but obscured by noise. Some of it can be attributed to noise inherent in biolgical systems and to measurement noise. Some of it may be due to our selection of E-genes. Including more biological knowledge on regulatory modules in *Drosophila* immune response would help to clarify the picture. The following results show that even

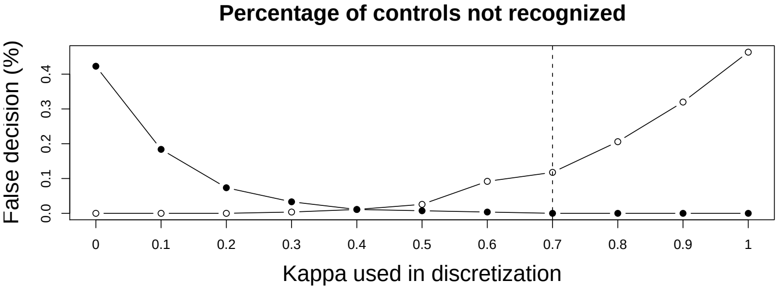

Figure 4.8: *Discretizing according to Eq. (4.10) with κ varying from* 0 *to* 1 *(x-axis). The black dots show, which percentage of negative controls was not recognized,* i.e. *set to 0 instead of 1. The circles show, which percentage of positive controls wrongly assigned to state 1. The dashed line indicates the smallest value of κ, at which all negative controls were correctly identified (the black dots hit zero).*

from noisy data the dominant biological features of the dataset can be reconstructed without having to rely on prior knowledge.

Score parameters We used the two scoring functions developed in this chapter. To compute the marginal likelihood of Eq. (4.4) we need to specify the global error rates α and β. The 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. We compare these results with the results obtained from using the full marginal likelihood of Eq. (4.8). There we have to specify four prior parameters. We set $a_0 = b_1 = 1$. Both values correspond to false observations (see Table 4.1) and should be small compared to the other two weights, if there is a clear signal in the data. We chose a_1 and b_0 to be equal and varied their value from 1 to 10.

Results Input hypotheses to the algorithm were all silencing schemes on four genes. The four S-genes can form $2^{12} = 4096$ pathways, which result in 355 different silencing schemes. Fig. 4.10 compares the result from applying both scoring functions. The distribution of marginal likelihood from Eq. (4.4) over the 30 top ranked silencing schemes in Fig. 4.10 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 ranking of silencing schemes is stable, when using different values of α and β, but the gap is sometimes less pronounced.

For the full marginal likelihood of Eq. (4.8) and low values of a_1 and b_0, we get a fully connected graph as the best model: no structure was found in the data. When the value increases, the scoring landscape looks more and more similar to the results obtained from Eq. (4.4). For $a_1 = b_0 = 5$, both scores result in the same winning

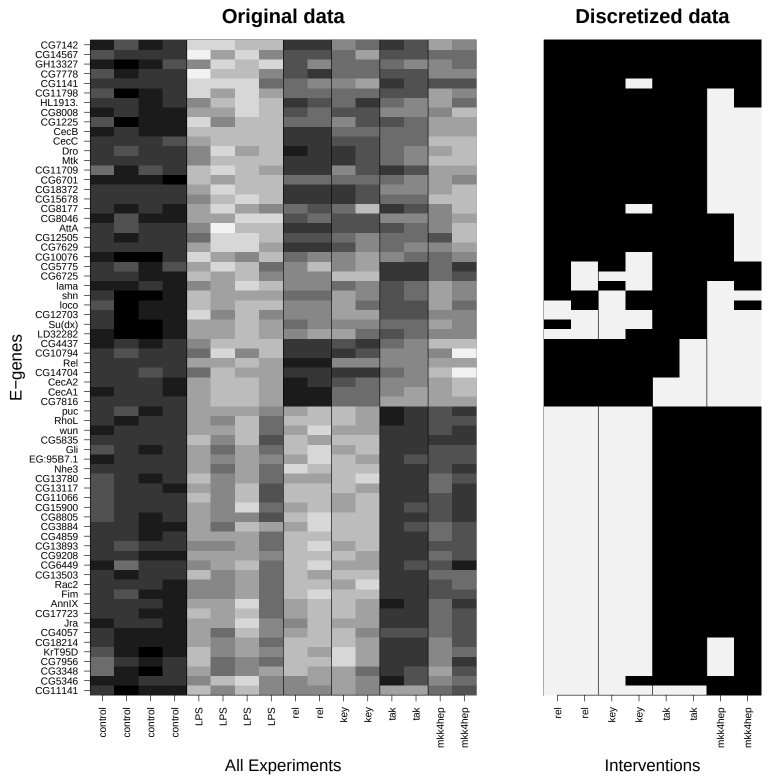

Figure 4.9: *Data on* Drosophila *immune response.* **Left:** *the normalized, gene-wise scaled data from [12]. Black stands for low expression and white for high expression. Rows are E-genes selected for differential expression after LPS stimulation (as seen in the first eight colums).* **Right:** *The data from silencing experiments after discretization ($\kappa = 0.7$) as used in our analysis. We only show the eight columns in the data matrix corresponding to RNAi experiments. The subset structure is visible, but obscured by noise.*

model. In the right plot of Fig. 4.10 we show the result for $a_1 = b_0 = 9$. It is the smallest value for which both scores agree on the five highest ranked models.

The topology of the best silencing scheme obtained from both scoring functions is shown in Fig. 4.11. It can be constructed from three different pathway hypotheses: One is the topology shown in Fig. 4.11, which is transitively closed, the other two miss either the edge from *tak* to *rel* or from *tak* to *key*. This is an example of prediction equivalence. The key features of the data are preserved in all three pathway topologies. 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 of cross-talk between the two branches of the pathway. All in all, our result fits exactly to the conclusions Boutros *et al.* [12] drew from the data.

Fig. 4.12 shows the expected position of E-genes given the optimal silencing scheme of Fig. 4.11. Both predictions agree very well and show only subtle differences. The double-headed arrow in Fig. 4.11 indicates that the order of *key* and *rel* cannot be

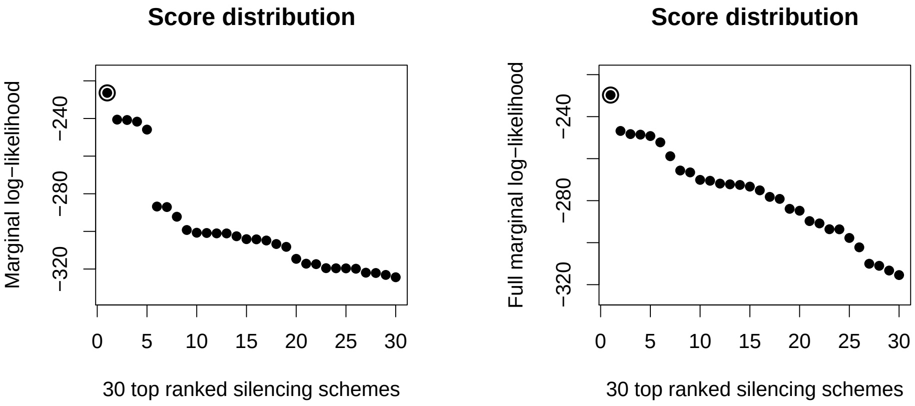

Figure 4.10: *The score distribution over the* 30 *top scoring silencing schemes. The same silencing scheme (circeled) achieves the best score in both plots. In the left plot (Eq. 4.4,* $\alpha = 0.15$*,* $\beta = 0.05$*), it is well separated from a small group of four lagging behind with a pronounced gap to the rest. In the right plot (Eq. 4.8,* $a_0 = b_1 = 1$*,* $a_1 = b_0 = 9$*), the distribution is more continuous. The five top ranking silencing schemes are the same for both scoring functions. If the value of* a_1 *and* b_0 *is further increased, the right plot converges towards the left one and shows a clear gap between the best ranking silencing schemes and the rest.*

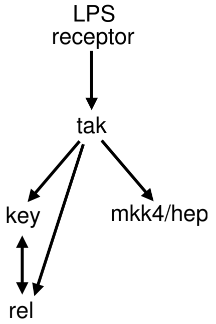

Figure 4.11: *Topology of the top-scoring silencing scheme on the* Drosophila *data. It clearly shows the fork below* tak *with* key *and* rel *on one side and* mkk4/hep *on the other. The double-headed arrow between* key *and* rel *indicates that they are undistinguishable from this data.*

resolved from this dataset, which was to be expected from the nearly identical profiles in Fig. 4.9. This is also the reason, why the posterior position of E-genes in the upper half of Fig. 4.12 is distributed equally on both S-genes. The data is undecided about the relative position of *key* and *rel*, and so is the posterior. However, it is known that *rel* is the transcription factor regulating the downstream genes (see chapter 1). 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. The results here show how well pathway features can be reconstructed just based on experimental data, without any biological prior knowledge.

A measure of 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

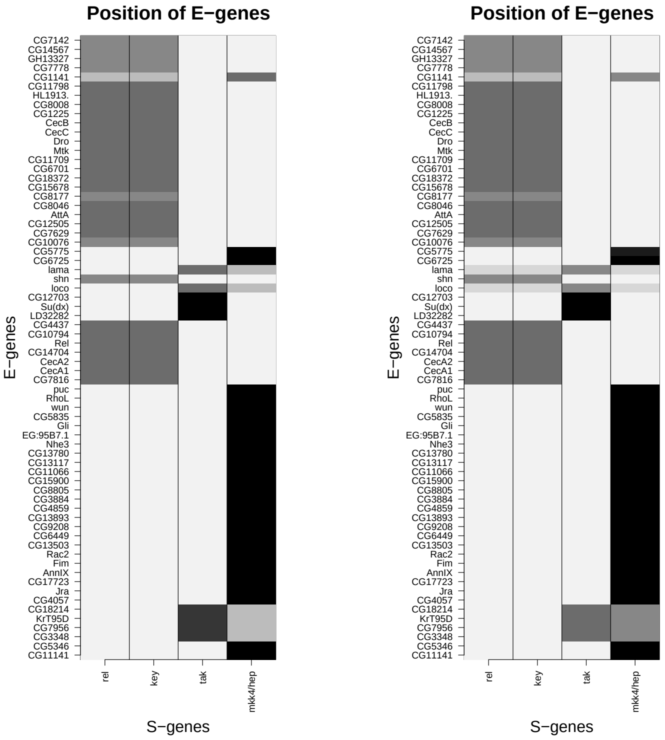

Figure 4.12: *Expected position of E-genes on the* Drosophila *data.* **Left:** *The expected position of E-genes given the silencing scheme with highest marginal likelihood of the data computed from Eq. (4.5). The lower half of E-genes is attributed to* mkk4/hep, *the upper half mostly to* key *and* rel, *which show almost the same intervention profiles (see Fig. 4.9).* **Right:** *Expected position of E-genes computed from Eq. 4.9.*

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 Fig. 4.10 shows is a well pronounced maximium for both scores. This indicates that we found the dominant structure in the data with high certainty. This conclusion is strengthened by inspecting the four silencing schemes achieving the second best score in both plots in Fig. 4.10. They all share the fork beneath *tak* and only differ from the best solution in Fig. 4.11 by missing one or two of the edges between *tak*, *key* and *rel*. All of them represent well the key features of the data.