Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2009, Article ID 195272, 8 pages
doi:10.1155/2009/195272

Research Article

A Bayesian Network View on Nested Effects Models

Cordula Zeller,[1] Holger Fröhlich,[2] and Achim Tresch[3]

[1] *Department of Mathematics, Johannes Gutenberg University, 55099 Mainz, Germany*
[2] *Division of Molecular Genome Analysis, German Cancer Research Center, 69120 Heidelberg, Germany*
[3] *Gene Center, Ludwig Maximilians University, 81377 Munich, Germany*

Correspondence should be addressed to Achim Tresch, tresch@lmb.uni-muenchen.de

Received 27 June 2008; Revised 23 September 2008; Accepted 24 October 2008

Recommended by Dirk Repsilber

Nested effects models (NEMs) are a class of probabilistic models that were designed to reconstruct a hidden signalling structure from a large set of observable effects caused by active interventions into the signalling pathway. We give a more flexible formulation of NEMs in the language of Bayesian networks. Our framework constitutes a natural generalization of the original NEM model, since it explicitly states the assumptions that are tacitly underlying the original version. Our approach gives rise to new learning methods for NEMs, which have been implemented in the **R**/*Bioconductor* package nem. We validate these methods in a simulation study and apply them to a synthetic lethality dataset in yeast.

1. Introduction

Nested effects models (NEMs) are a class of probabilistic models. They aim to reconstruct a hidden signalling structure (e.g., a gene regulatory system) by the analysis of high-dimensional phenotypes (e.g., gene expression profiles) which are consequences of well-defined perturbations of the system (e.g., RNA interference). NEMs have been introduced by Markowetz et al. [1], and they have been extended by Fröhlich et al. [2] and Tresch and Markowetz [3], see also the review of Markowetz and Spang [4]. There is an open-source software package "*nem*" available on the platform **R**/*Bioconductor* [5, 13], which implements a collection of methods for learning NEMs from experimental data. The utility of NEMs has been shown in several biological applications (*Drosophila melanogaster* [1], *Saccharomyces cerevisiae* [6], estrogen receptor pathway, [7]). The model in its original formulation suffers from some ad hoc restrictions which seemingly are only imposed for the sake of computability. The present paper gives an NEM formulation in the context of Bayesian networks (BNs). Doing so, we provide a motivation for these restrictions by explicitly stating prior assumptions that are inherent to the original formulation. This leads to a natural and meaningful generalization of the NEM model.

The paper is organized as follows. Section 2 briefly recalls the original formulation of NEMs. Section 3 defines NEMs as a special instance of Bayesian networks. In Section 4, we show that this definition is equivalent to the original one if we impose suitable structural constraints. Section 5 exploits the BN framework to shed light onto the learning problem for NEMs. We propose a new approach to parameter learning, and we introduce structure priors that lead to the classical NEM as a limit case. In Section 6, a simulation study compares the performance of our approach to other implementations. Section 7 provides an application of NEMs to synthetic lethality data. In Section 8, we conclude with an outlook on further issues in NEM learning.

2. The Classical Formulation of Nested Effects Models

For the sake of self-containedness, we briefly recall the idea and the original definition of NEMs, as given in [3]. NEMs are models that primarily intend to establish causal relations between a set of binary variables, the signals $\mathcal{S}$. The signals are not observed directly rather than through their consequences on another set of binary variables, the effects $\mathcal{E}$. A variable assuming the value 1, respectively, 0 is called *active*, respectively, *inactive*. NEMs deterministically

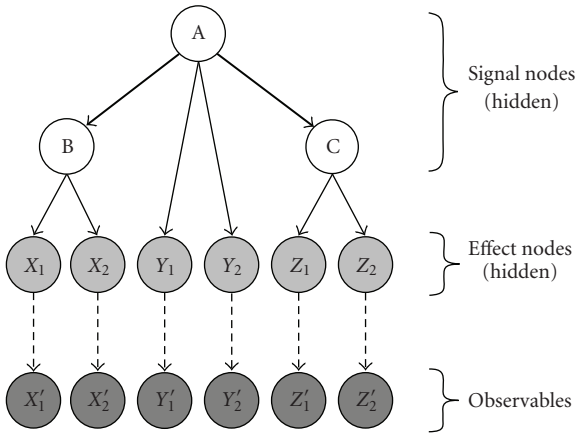

FIGURE 1: Example of a Nested effects model in its Bayesian network formulation. The bold arrows determine the graph Γ, the solid thin arrows encode Θ. Dashed arrows connect the effects to their reporters.

predict the states of the effects, given the states of the signals. Furthermore, they provide a probabilistic model for relating the predicted state of an effect to its measurements. NEMs consist of a directed graph $\mathcal{T}$ the nodes of which are the variables $\mathcal{S} \cup \mathcal{E}$. Edges represent dependencies between their adjacent nodes. An arrow pointing from a to b means that b is active whenever a is active. To be more precise, the graph $\mathcal{T}$ can be decomposed into a graph Γ, which encodes the information flow between the signals, and a graph Θ which relates each effect to exactly one signal, see Figure 1. The effects that are active as a consequence of a signal s are those effects that can be reached from s via at most one step in Γ, followed by one step in Θ. Let $\delta_{s,e}$ denote the predicted state of e when signal s is activated, and let $\Delta = (\delta_{s,e})$ be the matrix of all predicted effects.

For the probabilistic part of the model, let $d_{s,e}$ be the data observed at effect e when signal s is activated (which, by the way, need not be binary and may comprise replicate measurements), and let $D = (d_{s,e})$ be the matrix of all measurements. The stochastic model that relates the predictions Δ to the experimental data D is given by a set of "local" probabilities $\mathcal{L} = \{p(d_{s,e} \mid e = \delta_{s,e}), s \in \mathcal{S}, e \in \mathcal{E}\}$. There are several ways of specifying $\mathcal{L}$, depending on the kind of data and the estimation approach one wants to pursue (see [1–3]). An NEM is completely parameterized by $\mathcal{T}$ and $\mathcal{L}$, and, assuming data independence, its likelihood is given by

$$p(D \mid \mathcal{T}, \mathcal{L}) = \prod_{s \in \mathcal{S}, e \in \mathcal{E}} p(d_{s,e} \mid e = \delta_{s,e}). \tag{1}$$

3. The Bayesian Network Formulation of Nested Effects Models

A Bayesian network describes the joint probability distribution of a finite family of random variables (the nodes) by a directed acyclic graph $\mathcal{T}$ and by a family of local probability distributions, which we assume to be parameterized by a set of parameters $\mathcal{L}$ (for details, see, e.g., [8]). We want to cast the situation of Section 2 in the language of Bayesian networks. Assuming the acyclicity of the graph Γ of the previous section, this is fairly easy. A discussion on how to proceed when Γ contains cycles is given in Section 4. We have to model a deterministic signalling hierarchy, in which some components ($\mathcal{E}$) can be probed by measurements, and some components ($\mathcal{S}$) are perturbed in order to measure the reaction of the system as a whole. All these components $\mathcal{H} = \mathcal{S} \cup \mathcal{E}$ will be *hidden* nodes in the sense that no observations will be available for $\mathcal{H}$, and we let the topology between these nodes be identical to that in the classical model. In order to account for the data, we introduce an additional layer of observable variables (*observables*, $\mathcal{O}$) in an obvious way: each effect node $e \in \mathcal{E}$ has an edge pointing to a unique (its) observable node $e' \in \mathcal{O}$ (see Figure 1). Hence, $\mathcal{O} = \{e' \mid e \in \mathcal{E}\}$, and we call e' *the observation of* e.

Let $pa(x)$ be the set of parents of a node x, that is, the set of nodes that are direct predecessors of x. For notational convenience, we add a zero node z, $p(z = 0) = 1$, which has no parents, and which is a parent of all hidden nodes (but not of the observables). Note that by construction, $pa(x)$ is not empty unless x is the zero node. For the hidden nodes, let the local probabilities describe a deterministic relationship,

$$p(x = 1 \mid pa(x)) = \begin{cases} 1, & \text{if any parent of } x \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} = \max(pa(x)) \quad \text{for } x \in \mathcal{H}. \tag{2}$$

We slightly abuse notation by writing $\max(pa(x))$ for the maximum value that is assumed by a node in $pa(x)$. Obviously, all hidden nodes are set to 0 or 1 deterministically, given their parents. The local probabilities $p(e' \mid e)$, $e \in \mathcal{E}$, remain arbitrary for the moment. Assume that we have made an intervention into the system by activating a set of nodes $\mathcal{I} \subset \mathcal{S}$. This amounts to cutting all edges that lead to the nodes in $\mathcal{I}$ and setting their states to value 1. When an intervention $\mathcal{I}$ is performed, let $\delta_{\mathcal{I},h} \in \{0,1\}$ be the value of $h \in \mathcal{H}$. This value is uniquely determined by $\mathcal{I}$, as the next lemma shows.

Lemma 3.1. *$\delta_{\mathcal{I},h} = 1$ if and only if h can be reached from one of the nodes in $\mathcal{I}$ by a directed path in $\mathcal{T}$ (i.e., there exists a sequence of directed edges in $\mathcal{T}$, possibly of length zero, that links an $s \in \mathcal{I}$ to h). When performing an intervention $\mathcal{I}$, we, therefore, have*

$$p(h = 1) = \delta_{\mathcal{I},h} \quad \text{for } h \in \mathcal{H}. \tag{3}$$

Proof. The proof is straightforward though somewhat technical and may be skipped for first reading. Let $\mathcal{H} = \{h_1, \ldots, h_n\}$ be an ordering of the nodes compatible with $\mathcal{T}$, which means $pa(h_j) \subseteq \{h_1, \ldots, h_{j-1}\}$, $j = 1, \ldots, n$. Such an ordering exists because the graph connecting the states is acyclic. The proof is by induction on the order, the case $p(h_1 = 1) = \delta_{\mathcal{I},h_1}$ being trivial. If $h_j \in \mathcal{I}$, there is nothing to prove. Hence, we may assume $pa(h_j) \neq \varnothing$ in the graph which arises from $\mathcal{T}$ by cutting all edges that lead to a node in $\mathcal{I}$. Since $p(h_j = 1) = \max(pa(h_j))$, it follows that $\delta_{\mathcal{I},h_j} = 1$ if

and only if $h_k = 1$ for some $h_k \in pa(h_j)$. This holds exactly if $\delta_{\mathbb{1},h_k} = 1$ for some $k \in pa(h_j)$ (in particular, $k < j$). By induction, this is the case if and only if there exists an $h_i \in \mathbb{1}$ and a directed path from h_i to h_k, which can then be extended to a path from h_i to h_j.

Let $D_{\mathbb{1}} = (e' = d_{e',\mathbb{1}}; e \in \mathcal{E})$ be an observation of the effects generated during intervention $\mathbb{1}$. Marginalization over the hidden nodes yields

$$P_{BN}(D_{\mathbb{1}}) = \sum_{(b_h)\in\{0,1\}^{\mathcal{H}}} P(D \mid h = b_h; h \in \mathcal{H}) \cdot P(h = b_h; h \in \mathcal{H}). \tag{4}$$

Since by (3) there is only one possible configuration for the hidden nodes, namely, $s = \delta_{\mathbb{1},s}$, $s \in \mathcal{S}$, (4) simplifies to

$$P_{BN}(D_{\mathbb{1}}) = P(D_{\mathbb{1}} \mid h = \delta_{\mathbb{1},h}; h \in \mathcal{H}) = P(D_{\mathbb{1}} \mid e = \delta_{\mathbb{1},e}; e \in \mathcal{E}) \tag{5}$$

$$= \prod_{e\in\mathcal{E}} p(e' = d_{e',\mathbb{1}} \mid e = \delta_{\mathbb{1},e}). \tag{6}$$

This formula is very intuitive. It says that if an intervention $\mathbb{1}$ has been performed, one has to determine the unique current state of each effect node. This, in turn, determines the (conditional) probability distribution of the corresponding observable node, for which one has to calculate the probability of observing the data. The product over all effects then gives the desired result. □

4. Specialization to the Original NEM Formulation

In fact, (6) can be written as

$$P_{BN}(D_{\mathbb{1}}) = \prod_{e\in\mathcal{E}\mid\delta_{\mathbb{1},e}=1} p(e' = d_{e',\mathbb{1}} \mid e = 1) \cdot \prod_{e\in\mathcal{E}\mid\delta_{\mathbb{1},e}=0} p(e' = d_{e',\mathbb{1}} \mid e = 0) = \prod_{e\in\mathcal{E}\mid\delta_{\mathbb{1},e}=1} \frac{p(e' = d_{e',\mathbb{1}} \mid e = 1)}{p(e' = d_{e',\mathbb{1}} \mid e = 0)} \cdot \prod_{e\in\mathcal{E}} p(e' = d_{e',\mathbb{1}} \mid e = 0). \tag{7}$$

Let $r_{e,\mathbb{1}} = \log(p(e' = d_{e',\mathbb{1}} \mid e = 1)/p(e' = d_{e',\mathbb{1}} \mid e = 0))$, $e \in \mathcal{E}$, and $t_{\mathbb{1}} = \log\prod_{e\in\mathcal{E}} p(e' = d_{e',\mathbb{1}} \mid e = 0)$. Following the NEM formulation of [3], we consider all replicate measurements of an intervention $\mathbb{1}$ as generated from its own Bayesian network, and we try to learn the ratio $r_{e,\mathbb{1}}$ separately for each intervention $\mathbb{1}$. Therefore, we include $\mathbb{1}$ into the subscript. Taking logs in (7), it follows that

$$\log P_{BN}(D_{\mathbb{1}}) = \sum_{e\in\mathcal{E}\mid\delta_{\mathbb{1},e}=1} r_{e,\mathbb{1}} + t_{\mathbb{1}} = \sum_{e\in\mathcal{E}} \delta_{\mathbb{1},e} \cdot r_{e,\mathbb{1}} + t_{\mathbb{1}}. \tag{8}$$

Suppose that we have performed a series $\mathbb{1}_1, \ldots, \mathbb{1}_N \subseteq \mathcal{S}$ of interventions, and we have generated observations $D_1, \ldots, D_N$, respectively. Assuming observational independence, we get

$$\begin{aligned} \log P_{BN}(D_1, \ldots, D_N) &= \sum_{j=1}^{N} \log P(D_j) \\ &= \sum_{j=1}^{N} \sum_{e\in\mathcal{E}} \delta_{\mathbb{1}_j,e} \cdot r_{e,\mathbb{1}_j} + \sum_{j=1}^{N} t_{\mathbb{1}_j} \\ &= \sum_{j=1}^{N} (\Delta R)_{j,j} + \sum_{j=1}^{N} t_{\mathbb{1}_j} \\ &= \operatorname{tr}(\Delta R) + \sum_{j=1}^{N} t_{\mathbb{1}_j}, \end{aligned} \tag{9}$$

with the matrices $\Delta = (\delta_{\mathbb{1}_j,e})_{j,e}$ and $R = (r_{e,\mathbb{1}_j})_{e,j}$. The importance of (9) lies in the fact that it completely separates the estimation steps for $\mathcal{L}$ and $\mathcal{T}$. The information about the topology $\mathcal{T}$ of the Bayesian network enters the formula merely in the shape of Δ, and the local probability distributions alone define R. Hence, prior to learning the topology, one needs to learn the local probabilities only for once. Then, finding a Bayesian network that fits the data well means finding a topology which maximizes $\operatorname{tr}(\Delta R)$.

In the original formulation of NEMs, it is assumed that the set of interventions equals the set of all single-node interventions, $\mathbb{1}_s = \{s\}$, $s \in \mathcal{S}$. As pointed out in Section 2, the topology of the BN can be captured by two graphs Γ and Θ, which we identify with their corresponding adjacency matrices Γ and Θ by abuse of notation. The $\mathcal{S} \times \mathcal{S}$ adjacency matrix $\Gamma = (\Gamma_{s,t})_{s,t\in\mathcal{S}}$ describes the connections among signals, and the $\mathcal{S} \times \mathcal{E}$ adjacency matrix $\Theta = (\Theta_{s,e})_{s\in\mathcal{S},e\in\mathcal{E}}$ encodes the connection between signals and effects. For convenience, let the diagonal elements of Γ equal 1. Denote by $\overline{\Gamma}$ the adjacency matrix of the transitive closure of Γ. Check that by Lemma 3.1, $\Delta = \overline{\Gamma}\Theta$. Therefore, we seek

$$\underset{(\Gamma,\Theta);\ \Gamma\text{acyclic}}{\arg\max} \operatorname{tr}(\overline{\Gamma}\Theta R), \tag{10}$$

which for transitively closed graphs $\Gamma = \overline{\Gamma}$ is exactly the formulation in [3]. It has the advantage that given $\overline{\Gamma}$, the optimal Θ can be calculated exactly and very fast, which dramatically reduces the search space and simplifies the search for a good graph $\overline{\Gamma}$. The BN formulation of NEMs implies via (10) that two graphs Γ_1, Γ_2 are indistinguishable (likelihood equivalent, they fit all data equally well) if they have the same transitive closure. It is a subject of discussion whether the transitive closure of the underlying graph is a desirable property of such a model (think of causal chains which are observed in a stable state) or not (think of the dampening of a signal when passed from one node to another, or of a snapshot of the system where the signalling happens with large time lags), see [9].

It should be mentioned that the graph topology in our BN formulation of NEMs is necessarily acyclic, whereas the original formulation admits arbitrary graphs. This is only an apparent restriction. Due to the transitivity assumption, effects that connect to a cycle of signals will always react in the

same way. This behaviour can also be obtained by arranging the nodes of the cycle in a chain and connecting the effects to the last node of the chain. This even leaves the possibility for connecting other effects to only a subset of the signals in the cycle by attaching them to a node higher up in the chain. As a consequence, admitting cycles does not extend the model class of NEMs in the Bayesian setting.

Although the original NEM model is algebraically and computationally appealing, it has some drawbacks. Learning the ratio $r_{e,\mathbb{1}} = \log(p(e' = d_{e',\mathbb{1}} \mid e = 1)/p(e' = d_{e',\mathbb{1}} \mid e = 0))$ separately for each intervention $\mathbb{1}$ entails various problems as follows.

(1) Given an observation $d_{e'}$ at observable e' together with the state of its parent e, the quantity $p(e' = d_{e'} \mid e)$ should not depend on the intervention $\mathbb{1}$ during which the data were obtained, by the defining property of Bayesian networks. However, we learn the ratio $r_{e,\mathbb{1}}$ separately for each intervention, that is, we learn separate local parameters $\mathcal{L}$, which is counterintuitive.

(2) Reference measurements $p(e' = d_{e',\mathbb{1}} \mid e = 0)$ are used to calculate the ratio $r_{e,\mathbb{1}}$, raising the need for a "null" experiment corresponding to an unperturbed observation $\mathbb{1}_0 = \varnothing$ of the system, which might not be available. The null experiment enters the estimation of each ratio $r_{e,\mathbb{1}}$. This introduces an unnecessary asymmetry in the importance of intervention $\mathbb{1}_0$ relative to the other interventions.

(3) The procedure uses the data inefficiently since for a given topology, the quantities of interest $p(e' = d_{e'} \mid e = 1)$, respectively, $p(e' = d_{e'} \mid e = 0)$ could be learned from *all* interventions that imply $e = 1$, respectively, $e = 0$, providing a broader basis for the estimation.

The method proposed in the last item is much more time-consuming, since the occurring probabilities have to be estimated individually for each topology. However, such a model promises to better capture the real situation, so we develop the theory into this direction.

5. NEM Learning in the Bayesian Network Setting

Bear in mind that a Bayesian network is parameterized by its topology $\mathcal{T}$ and its local probability distributions, which we assume to be given by a set of local parameters $\mathcal{L}$. The ultimate goal is to maximize $P(\mathcal{T} \mid D)$. In the presence of prior knowledge, (we assume independent priors for the topology and the local parameters), we can write

$$\begin{aligned} P(\mathcal{T}, \mathcal{L} \mid D) &= \frac{P(D \mid \mathcal{T}, \mathcal{L})P(\mathcal{T}, \mathcal{L})}{P(D)} \\ &\propto P(D \mid \mathcal{T}, \mathcal{L})P(\mathcal{T})P(\mathcal{L}), \end{aligned} \tag{11}$$

from which it follows that

$$\begin{aligned} P(\mathcal{T} \mid D) &= \int P(\mathcal{T}, \mathcal{L} \mid D)\mathrm{d}\mathcal{L} \\ &\propto P(\mathcal{T}) \int P(D \mid \mathcal{T}, \mathcal{L})P(\mathcal{L})\mathrm{d}\mathcal{L}. \end{aligned} \tag{12}$$

If it is possible to solve the integral in (12) analytically, it can then be used by standard optimization algorithms for the approximation of $\arg\max_{\mathcal{T}} P(\mathcal{T} \mid D)$. This full Bayesian approach will be pursued in Section 5.1. If the expression in (12) is computationally intractable or slow, we resort to a simultaneous maximum a posteriori estimation of $\mathcal{T}$ and $\mathcal{L}$, that is,

$$\begin{aligned} (\widehat{\mathcal{T}}, \widehat{\mathcal{L}}) &= \arg\max_{\mathcal{T},\mathcal{L}} P(\mathcal{T}, \mathcal{L} \mid D) \\ &= \arg\max_{\mathcal{T}} \left(\arg\max_{\mathcal{L}} P(D \mid \mathcal{T}, \mathcal{L})P(\mathcal{L}) \right) P(\mathcal{T}). \end{aligned} \tag{13}$$

The hope is that the maximization $\widehat{\mathcal{L}}(\mathcal{T}) = \arg\max_{\mathcal{L}} P(D \mid \mathcal{T}, \mathcal{L})P(\mathcal{L})$ in (13) can be calculated analytically or at least very efficiently, see [3]. Then, maximization over $\mathcal{T}$ is again done using standard optimization algorithms. Section 5.2 is devoted to this approach.

5.1. Bayesian Learning of the Local Parameters. Let the topology $\mathcal{T}$ and the interventions $\mathbb{1}_j$ be given. Let N_{eik} denote the number of times the observable e was reported to take the value k, while its true value was i, and let N_{ei} be the number of measurements taken from e when its true value is i:

$$\begin{aligned} N_{eik} &= \left| \{ j \mid \delta_{\mathbb{1}_j,e} = i, d_{e',\mathbb{1}_j} = k \} \right|, \\ N_{ei} &= \left| \{ j \mid \delta_{\mathbb{1}_j,e} = i \} \right|. \end{aligned} \tag{14}$$

Binary Observables. The full Bayesian approach in a multinomial setting was introduced by Cooper and Herskovits [10].

The priors are assumed to follow beta distributions:

$$\beta_0 \sim \mathrm{Beta}(\alpha_0, \beta_0), \qquad \beta_1 \sim \mathrm{Beta}(\alpha_1, \beta_1). \tag{15}$$

Here, $\alpha_0, \alpha_1, \beta_0$, and β_1 are shape parameters, which, for the sake of simplicity, are set to the same value for every effect e. This assumption can be easily dropped and different priors may be used for each effect.

In this special setting with binomial nodes with one parent, the well-known formula of Cooper and Herskovitz can be simplified to

$$\begin{aligned} &P(D_1, \ldots, D_N \mid \mathcal{T}) \\ &\quad = \prod_{j=1}^{N} \prod_{e \in \mathcal{E}} \prod_{i \in \{0,1\}} \frac{\Gamma(N_{ei0} + \alpha_i)\Gamma(N_{ei1} + \beta_i)\Gamma(\alpha_i + \beta_i)}{\Gamma(N_{ei} + \alpha_i + \beta_i)\Gamma(\alpha_i)\Gamma(\beta_i)} \\ &\quad \propto \prod_{j=1}^{N} \prod_{e \in \mathcal{E}} \prod_{i \in \{0,1\}} \frac{\Gamma(N_{ei0} + \alpha_i)\Gamma(N_{ei1} + \beta_i)}{\Gamma(N_{ei} + \alpha_i + \beta_i)}. \end{aligned} \tag{16}$$

Continuous Observables. Let us assume $p(e' \mid e = k)$ to be normally distributed with mean a_{ek} and variance σ^2_{ek}, $e \in \mathcal{E}$, $k \in \{0, 1\}$. We refer to the work of Neapolitan [8] for the calculation of this section. Let the prior for the precision $r_{ek} = 1/\sigma^2_{ek}$ follow a Gamma distribution,

$$\rho(r_{ek}) = \mathrm{Gamma}\left(r_{ek}; \frac{\alpha}{2}, \frac{\beta}{2} \right). \tag{17}$$

Given the precision r_{ek}, let the conditional prior for the mean a_{ek} be

$$\rho(a_{ek} \mid r_{ek}) = \mathcal{N}\left(a_{ek}; \mu, \frac{1}{\nu r_{ek}}\right). \tag{18}$$

So the Data of observable e' given its parent's state $\delta_{\mathbb{1}_j,e} = k$ is

$$\rho(d_{e',\mathbb{1}_j} \mid a_{ek}, r_{ek}) = \mathcal{N}\left(d_{e',\mathbb{1}_j}; a_{ek}, \frac{1}{r_{ek}}\right), \quad \delta_{\mathbb{1}_j,e} = k. \tag{19}$$

Then,

$$\begin{aligned} P(D_1,\ldots,D_N \mid \mathcal{T}) &= \prod_{e\in\mathcal{E}} \prod_{k\in\{0,1\}} \left(\frac{1}{2\pi}\right)^{N_{ek}/2} \left(\frac{\nu}{\nu+N_{ek}}\right)^{1/2} 2^{N_{ek}/2} \frac{\Gamma((\alpha+N_{ek})/2)}{\Gamma(\alpha/2)} \\ &\quad \cdot \frac{|\beta|^{\alpha/2}}{\left|\beta + s_{ek} + (\nu N_{ek}/(\nu+N_{ek}))\left(\overline{x}_{ek}-\mu\right)^2\right|^{(\alpha+N_{ek})/2}} \\ &\propto \prod_{e\in\mathcal{E}} \prod_{k\in\{0,1\}} \left(\frac{\nu}{\nu+N_{ek}}\right)^{1/2} \\ &\quad \times \frac{\Gamma((\alpha+N_{ek})/2)}{\left|\beta + s_{ek} + (\nu N_{ek}/(\nu+N_{ek}))\left(\overline{x}_{ek}-\mu\right)^2\right|^{(\alpha+N_{ek})/2}}. \end{aligned} \tag{20}$$

The data enters this equation via

$$\overline{x}_{ek} = \frac{1}{N_{ek}} \sum_{j\mid\delta_{\mathbb{1}_j,e}=k} d_{e',\mathbb{1}_j}, \qquad s_{ek} = \sum_{j\mid\delta_{\mathbb{1}_j,e}=k} \left(d_{e',\mathbb{1}_j} - \overline{x}_{ek}\right)^2. \tag{21}$$

5.2. Maximum Likelihood Learning of the Local Parameters. Let the topology $\mathcal{T}$ and the interventions $\mathbb{1}_j$ be given. For learning the parameters of the local distributions $p(e' \mid e)$, we perform maximum likelihood estimation in two different settings. The observables are assumed to follow either a binomial distribution or a Gaussian distribution.

Binary Observables. For an effect $e \in \mathcal{E}$, let its observable e' be a binary random variable with values in $\{0,1\}$, and let $p(e' = 1 \mid e = x) = \beta_{e,x}$, $x \in \{0,1\}$. The model is then completely parameterized by the topology $\mathcal{T}$ and $\mathcal{L} = \{\beta_{e,x} \mid e \in \mathcal{E},\ x \in \{0,1\}\}$.

Note that

$$\begin{aligned} P(D_1,\ldots,D_N \mid \mathcal{T},\mathcal{L}) &= \prod_{j=1}^{N} \prod_{e\in\mathcal{E}} p(e' = d_{e',\mathbb{1}_j} \mid e = \delta_{\mathbb{1}_j,e}) \\ &= \prod_{e\in\mathcal{E}} \prod_{x\in\{0,1\}} \prod_{j\mid\delta_{\mathbb{1}_j,e}=x} p(e' = d_{e',\mathbb{1}_j} \mid e = x) \\ &= \prod_{e\in\mathcal{E}} \prod_{x\in\{0,1\}} B(k = N_{ex1}; n = N_{ex}, p = \beta_{e,x}), \end{aligned} \tag{22}$$

with $B(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$. The parameter set $\widehat{\mathcal{L}}$ that maximizes expression (22) is

$$\widehat{\beta}_{e,x} = \frac{N_{ex1}}{N_{ex}}, \quad e \in \mathcal{E},\ x \in \{0,1\} \tag{23}$$

(the ratios with a denominator of zero are irrelevant for the evaluation of (22) and are set to zero).

Continuous Observables. There is an analogous way of doing ML estimation in the case of continuous observable variables if one assumes $p(e' \mid e = x)$ to be a normal distribution with mean $\mu_{e,x}$ and variance $\sigma^2_{e,x}$, $e \in \mathcal{E}$, $x \in \{0,1\}$.

Note that

$$\begin{aligned} P(D_1,\ldots,D_N \mid \mathcal{T},\mathcal{L}) &= \prod_{j=1}^{N} \prod_{e\in\mathcal{E}} p(e' = d_{e',\mathbb{1}_j} \mid e = \delta_{\mathbb{1}_j,e}), \\ &= \prod_{e\in\mathcal{E}} \prod_{x\in\{0,1\}} \prod_{j\mid\delta_{\mathbb{1}_j,e}=x} p(e' = d_{e',\mathbb{1}_j} \mid e = x) \\ &= \prod_{e\in\mathcal{E}} \prod_{x\in\{0,1\}} \mathcal{N}\left(\{d_{e',\mathbb{1}_j} \mid \delta_{\mathbb{1}_j,e} = x\};\ \mu_{e,x}, \sigma_{e,x}\right), \end{aligned} \tag{24}$$

with

$$\mathcal{N}\left(\{x_1,\ldots,x_k\};\ \mu,\sigma\right) = \left(\frac{1}{\left(\sqrt{2\pi}\sigma\right)^k}\right) \cdot \exp\left(-\left(\sum_{j=1}^{k} \frac{(x_j-\mu)^2}{2\sigma^2}\right)\right). \tag{25}$$

The parameter set $\widehat{\mathcal{L}}$ maximizing expression (24) is

$$\begin{aligned} \widehat{\mu}_{e,x} &= \frac{1}{N_{ex}} \sum_{j\mid\delta_{\mathbb{1}_j,e}=x} d_{e',\mathbb{1}_j}, \\ \widehat{\sigma}_{e,x} &= \frac{1}{N_{ex}} \sum_{j\mid\delta_{\mathbb{1}_j,e}=x} \left(d_{e',\mathbb{1}_j} - \widehat{\mu}_{e,x}\right)^2, \quad e \in \mathcal{E},\ x \in \{0,1\} \end{aligned} \tag{26}$$

(quotients with a denominator of zero are again irrelevant for the evaluation of (24) and are set to zero). Note that in both the discrete and the continuous case, $\widehat{\mathcal{L}}$ depends on the topology $\mathcal{T}$, since the topology determines the values of $\delta_{\mathbb{1}_j,e}$, $j = 1,\ldots,N$, $e \in \mathcal{E}$.

5.3. Structure Learning. It is a major achievement of NEMs to restrict the topology of the underlying graphical structure in a sensible yet highly efficient way, thus, tremendously reducing the size of the search space. There is an arbitrary "core" network consisting of signal nodes, and there is a very sparse "marginal" network connecting the signals to the effects. It is, however, by no means necessary that the core network and the signal nodes coincide. We propose another partition of the hidden nodes into core nodes $\mathcal{C}$ and marginal nodes $\mathcal{M}$, $\mathcal{H} = \mathcal{C} \cup \mathcal{M}$, which may be distinct from the partition into signals and effects, $\mathcal{H} = \mathcal{S} \cup \mathcal{E}$. No restrictions are imposed on the subgraph generated by the

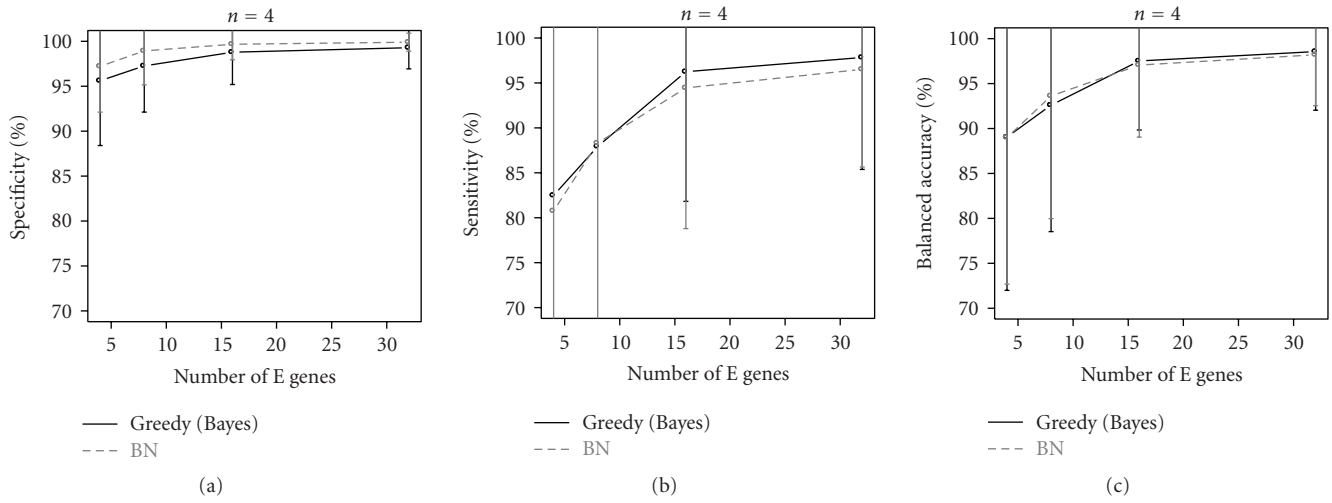

Figure 2: Results (specificity, sensitivity, and balanced accuracy) of simulation run. The continuous line (greedy (Bayes)) describes the performance of the traditional NEM method, the dashed line stands for our new approach via Bayesian networks.

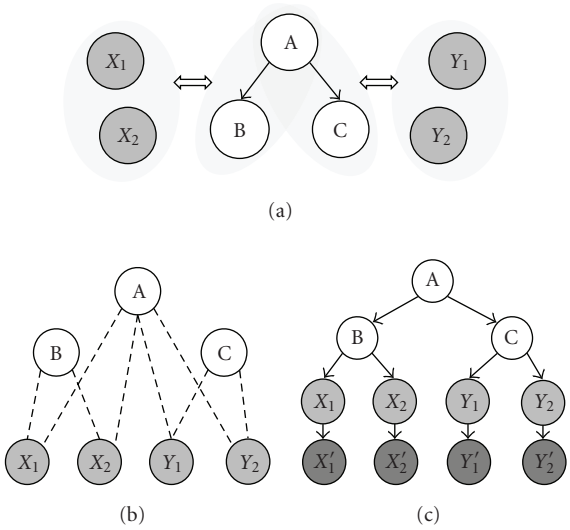

Figure 3: Schematic reconstruction of a signalling pathway through synthetic lethality data. (a) A situation in which there are two pairs of complementary pathways ($\{A, B\}, \{X_1, X_2\}$ and $\{A, C\}, \{Y_1, Y_2\}$). (b) Model of the situation as follows: the primary knockouts are considered signals $\{A, B, C\}$ (they are not observed). As those are our genes of interest, they will also form the core nodes. The secondary effects are accessible to observation and, therefore, represented by the effects X_1, X_2, Y_1, and Y_2. Each SL pair is connected by a dashed line. (c) NEMs that might be estimated from (b), using binary observables and one of the approaches in Sections 5.1 or 5.2.

core nodes (except that the graph has to be acyclic). The key semantics of NEMs is that marginal nodes are viewed as the terminal nodes of a signalling cascade. The requirement that the marginal nodes have only few or at most one incoming edge can be translated into a well-known structure prior $P(\mathcal{T})$ (see, e.g., [12]) which penalizes the number of parents of marginal nodes:

$$\log P(\mathcal{T}) = -\nu \cdot \sum_{m \in \mathcal{M}} \max\left(\left|pa(m)\right| - 1, 0\right). \tag{27}$$

For the penalty parameter $\nu = \infty$, this is the original NEM restriction. If $\nu = 0$, each marginal node can be assigned to all suitable core nodes. As a consequence, there is always a best scoring topology with an empty core graph. ν makes signalling to the marginal nodes "expensive" relative to signalling in the core graph. It is unclear how to choose ν optimally, so we stick to the choice $\nu = \infty$ for the applications. Simulation studies have shown that a simple gradient ascent algorithm does very well in optimizing the topology of the Bayesian network, compared to other methods that have been proposed [7].

6. Simulation

6.1. Network and Data Sampling. The ML and the Bayesian method for parameter learning have been implemented in the *nem* software [13], which is freely available at the **R**/*Bioconductor* software platform [5]. To test the performance of our method, we conducted simulations with randomly created acyclic networks with $n = 4$ signals. The out-degree d of each signal was sampled from the power-law distribution

$$p(d) = \frac{1}{Z} d^{-2.5}, \tag{28}$$

where Z is an appropriate normalization constant. Binary data (1 = effect, 0 = no effect) was simulated for the perturbation of each signal in the created network using 4 replicate measurements with type-I and type-II error rates α and β, which were drawn uniformly from $[0.1, 0.5]$ and $[0.01, 0.2]$ for each perturbation separately. This simulates individual measurement error characteristics for each experiment.

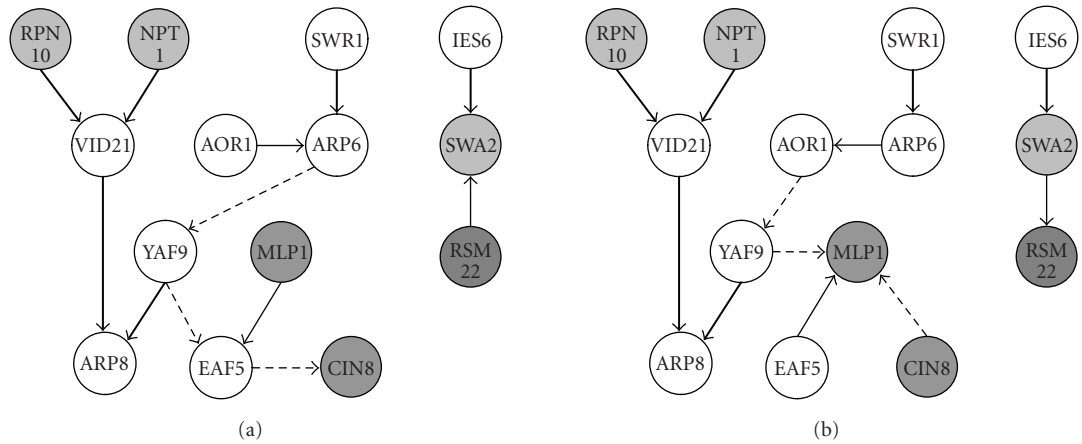

FIGURE 4: NEMs constructed from the SL data. Only core genes that have at least one edge are shown. (a) The ML estimate. (b) The Bayesian estimate (the prior choice (see (15)) was $\beta_{e0} \sim \text{Beta}(5,2)$, respectively, $\beta_{e1} \sim \text{Beta}(2,5)$). Nodes with the same shading pertain to the same clusters that were defined by Ye et al. [11]. Bold arrows appear in both reconstructions, thin arrows reverse their direction, and dashed arrows are unique to each reconstruction.

6.2. Results. We compared our Bayesian network model with the classical NEM using a greedy hill-climbing algorithm to find the best fitting connection between signals. We simulated $m = 25, 50, 100$ and 250 effect nodes, and for each number of effects, 100 random networks were created as described above. Figure 2 demonstrates that both approaches perform very similarly.

7. Application

We apply the BN formulation of the NEM methodology to a dataset of synthetic lethality interactions in yeast. We reveal hierarchical dependencies of protein interactions. Synthetic lethality (SL) is the phenomenon that a cell survives the single gene deletion of a gene A and a gene B, but the double deletion of A and B is detrimental. In this case, A and B are called SL partners or an SL pair. It has been shown in [11] that it is not so much SL partners themselves whose gene products participate in the same protein complex or pathway, rather than genes that *share* many SL partners. The detection of genetic interactions via synthetic lethality screens and appropriate computational tools is a current area of research, see [14]. Ye and Peyser define a hypergeometric score function to test whether two genes have many SL partners in common. They apply their methodology to a large SL data set [15] for finding pairs (and, consequently, clusters) of genes whose products are likely to participate in the same pathway. We extend their approach as explained in Figure 3. SL partnership arises (not exclusively, but prevalently) among genes pertaining to two distinct pathways that complement each other in a vital cell function. If a gene A is upstream of gene B in some pathway, a deletion of gene A will affect at least as many pathways as a deletion of gene B. Hypothesizing a very simplistic world, all SL partners of B will as well be SL partners of A; but this subset relation can be detected by NEMs. Take the primary knockout genes as core nodes, and the secondary knockout genes as marginal nodes, which are active given a primary knockout whenever SL occurs. We used the dataset from [15] and chose 40 primary knockout genes having the most SL interaction partners as core genes, and included all their 194 SL partners as marginal nodes. An NEM with binary observables was estimated, both with the maximum likelihood approach and in the Bayesian setting. It should be emphasized that NEM estimation for this dataset is only possible in the new BN setting because there is no canonical "null experiment," which enables us to estimate the likelihood ratios $r_{\mathbb{1},e}$ needed in the classical setting in (7), (8), [14].

Figure 4 displays the results of the NEM reconstruction. The NEMs estimated by both methods agree well as far as the hierarchical organisation of the network is concerned. However, they do not agree well with the clusters found in [11]. We refrain from a biological interpretation of these networks, since the results are of a preliminary nature. In particular, the reconstruction does not take advantage of prior knowledge, and the postulated edges were not validated experimentally.

8. Summary and Outlook

Some aspects of the classical NEM concept appear in a different light when stated in the BN framework. Mainly, these are three folds: (1) the learning of the local parameters, for which we proposed new learning rules; (2) the structural constraints, they can be cast as priors on the NEM topology; (3) the distinction between hidden and observable nodes, which can be different from that of core nodes and marginal nodes.

We proposed some new lines of investigation, like a full Bayesian approach for the evaluation of $P(\mathcal{T}|D)$, and a smooth structure prior with continuous penalty parameter ν. It is much easier to proceed in the BN framework and implement, for example, a boolean logic for the signal

transduction, which is less simplistic than in the current model. A straightforward application of NEMs in their BN formulation to synthetic lethality data demonstrated the potential of the NEM method, with the purpose of stimulating further research in that field.

Acknowledgments

The authors like to thank Peter Bühlmann and Daniel Schöner for proposing the application of NEMs to synthetic lethality data. This work was supported by the Deutsche Forschungsgemeinschaft, the Sonderforschungsbereich SFB646. H. Fröhlich is funded by the National Genome Research Network (NGFN) of the German Federal Ministry of Education and Research (BMBF) through the platforms SMP Bioinformatics (OIGR0450) and SMP RNA (OIGR0418).

References

[1] F. Markowetz, J. Bloch, and R. Spang, "Non-transcriptional pathway features reconstructed from secondary effects of RNA interference," *Bioinformatics*, vol. 21, no. 21, pp. 4026–4032, 2005.

[2] H. Fröhlich, M. Fellmann, H. Sültmann, A. Poustka, and T. Beissbarth, "Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data," *Bioinformatics*, vol. 24, no. 22, pp. 2650–2656, 2008.

[3] A. Tresch and F. Markowetz, "Structure learning in nested effects models," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, article 9, 2008.

[4] F. Markowetz and R. Spang, "Inferring cellular networks—a review," *BMC Bioinformatics*, vol. 8, supplement 6, pp. 1–17, 2007.

[5] R. C. Gentleman, V. J. Carey, D. M. Bates, et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, article R80, pp. 1–16, 2004.

[6] F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang, "Nested effects models for high-dimensional phenotyping screens," *Bioinformatics*, vol. 23, no. 13, pp. i305–i312, 2007.

[7] H. Froehlich, M. Fellmann, H. Sueltmann, A. Poustka, and T. Beissbarth, "Large scale statistical inference of signaling pathways from RNAi and microarray data," *BMC Bioinformatics*, vol. 8, article 386, pp. 1–15, 2007.

[8] R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River, NJ, USA, 2003.

[9] J. Jacob, M. Jentsch, D. Kostka, S. Bentink, and R. Spang, "Detecting hierarchical structure in molecular characteristics of disease using transitive approximations of directed graphs," *Bioinformatics*, vol. 24, no. 7, pp. 995–1001, 2008.

[10] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

[11] P. Ye, B. D. Peyser, X. Pan, J. D. Boeke, F. A. Spencer, and J. S. Bader, "Gene function prediction from congruent synthetic lethal interactions in yeast," *Molecular Systems Biolog*, vol. 1, article 2005.0026, p. 1, 2005.

[12] S. Mukherjee and T. P. Speed, "Network inference using informative priors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 38, pp. 14313–14318, 2008.

[13] H. Fröhlich, T. Beißbarth, A. Tresch, et al., "Analyzing gene perturbation screens with nested effects models in R and bioconductor," *Bioinformatics*, vol. 24, no. 21, pp. 2549–2550, 2008.

[14] N. Le Meur and R. Gentleman, "Modeling synthetic lethality," *Genome Biology*, vol. 9, no. 9, article R135, pp. 1–10, 2008.

[15] A. H. Y. Tong, G. Lesage, G. D. Bader, et al., "Global mapping of the yeast genetic interaction network," *Science*, vol. 303, no. 5659, pp. 808–813, 2004.