Modelling the temporal interplay of molecular signalling and gene expression using dynamical nested effects models

Benedict Anchang [*], **Mohammad J. Sadeh** [*], **Juby Jacob** [*], **Marcel O. Vlad** [*] [†] [‡], **Peter Oefner** [*], **and Rainer Spang** [*]

[*]Institute of Functional Genomics, University of Regensburg, Josef-Engert-Str 9 , 93053 Regensburg, Germany,[†]Institute of Mathematical Statistics and Applied Mathematics, Casa Academiei Romane, Calea 13 Septembrie 13, Bucharest, 050711, Romania, and [‡]Department of Chemistry, Stanford University, Stanford CA 94305-5080

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Cellular decision making in differentiation, proliferation or cell death is mediated by molecular signalling processes, which control the regulation and expression of genes. Vice versa, the expression of genes can trigger the activity of signalling pathways. We introduce and describe a statistical method called Dynamical Nested Effects Model (D-NEM) for analyzing the temporal interplay of cell signalling and gene expression. D-NEMs are Bayesian models of signal propagation in a network. They decompose observed time delays of multiple step signalling processes into single steps. Time delays are assumed to be exponentially distributed. Rate constants of signal propagation are model parameters, whose joint posterior distribution is assessed via Gibbs sampling. They hold information on the interplay of different forms of biological signal propagation: Molecular signalling in the cytoplasm acts at high rates, direct signal propagation via transcription and translation at intermediate rates, while secondary effects operate at low rates. D-NEMs allow the dissection of biological processes into signalling and expression events, and the analysis of cellular signal flow. We evaluate our method in simulation experiments and demonstrate its practical use in an application to embryonic stem cell development in mice.

Modelling Perturbation Data | Network Reconstruction | Nested Effects Models

Abbreviations: D-NEM, Dynamical Nested Effects Model; NEM, Nested Effects Model

Introduction

Intracellular signalling processes control the activity of transcription factors and the expression of genes. Changes in gene expression can activate further signalling processes, leading to secondary effects, which themselves give rise to tertiary effects and so on. The result is an intricate interplay of cell signalling and gene regulation. While protein modification in the cytoplasm can propagate signals in seconds, transcription and translation processes last hours, and secondary effects often become visible only after days. Our goal is to model the temporal interplay of signalling and expression in complex biological processes involving several signalling pathways and spanning multiple rounds of cell signalling, gene regulation, and gene expression.

Numerous statistical methods have been suggested for the analysis and reconstruction of regulatory networks. Among the most widely used are relevance networks [1], graphical Gaussian models [2, 3], methods from information theory [4], Bayesian networks [5] including dynamic Bayesian networks [6], and methods based on ordinary differential equations [7, 8]. All these methods are based on pure observational data, where the network was not perturbed experimentally. Simulation [9, 10] and experimental studies [11, 9] show that perturbation experiments greatly improve performance in network reconstruction, in line with the common practice in genetics to study molecular mechanisms by observing how cells react to perturbations. Rung et al. [12] built a directed disruption graph by connecting two genes where perturbation of the first gene resulted in expression changes in the other gene. However, disruption networks do not separate direct from indirect effects. Wagner [13] uses transitive reductions to find parsimonious subgraphs explaining a disruption network. The framework of Bayesian networks was also extended to account for perturbation data [14, 15]. Yeang et al. search for topologies that are consistent with observed downstream effects of interventions [16]. Although this algorithm is not confined to the transcriptional level of regulation, it requires that most signaling genes show effects when perturbing others.

The method described here builds on Nested Effects Models (NEMs), which have been first proposed by Markowetz et al. [15] for the analysis of non-transcriptional signalling networks. NEM infer the graph of upstream/downstream relations for a set of signalling genes from downstream effects of their knock-downs. Since non-transcriptional signalling is too fast to be analyzed by delays of downstream effects, time series are not used in this approach. This changes when analyzing slow-going biological processes like cell differentiation.

Following Markowetz et al. [15], we call the perturbed genes *S-Genes* for signalling genes and denote them by $\mathbf{S} = S_1, \ldots, S_n$. The genes that change expression after perturbation are called *E-genes* and we denote them by $\mathbf{E} = E_1, \ldots, E_N$. We further denote the set of E-genes displaying expression changes in response to the perturbation of S_i by $\mathcal{D}_i$. In a nutshell: NEMs infer that S_1 acts upstream of S_2:

$$S_1 \longrightarrow S_2 \quad \text{if and only if} \quad \mathcal{D}_2 \subset \mathcal{D}_1$$

All downstream effects of a perturbation in S_2 can also be triggered by perturbing S_1. This suggests that the perturbation of S_1 causes a perturbation of S_2 and acts upstream of S_2. The graph of upstream/downstream relations is estimated from the nested structure of downstream effects. Due to noise in the data, we do not expect strict super-/subset relations. Instead, NEMs recover rough nesting.

In the Bayesian framework of Markowetz et al. [15], networks are scored by posterior probabilities. By enumerating all network topologies, the maximum posterior network is chosen. The exhaustive search limits the method to small networks of up to 8 S-genes. Later, greedy search heuristics [17, 18] and divide-and-conquer approaches [17, 19] were

Reserved for Publication Footnotes

proposed for the analysis of large networks with hundreds of S-genes. The latter divide the graph into smaller units, use exhaustive enumeration for each subgraph, and then re-assemble the complete network. The division into subgraphs can either be into all pairs or triples of nodes [19] or data-dependent into coherent modules [17]. For a review and software see papers by Froehlich et al. [20, 21].

If S_1 is upstream of S_2 and S_2 is upstream of S_3, consistency requires that S_1 is also upstream of S_3. In fact, all proposed methods except [18] confine the model space to transitively closed graphs. While the consistency argument is valid for upstream/downstream relations, it does not hold for signal flows. Assume we have a linear cascade of S-genes where the signal flows from S_1 via S_2 to S_3. Whether there is an alternative signal flow from S_1 directly to S_3 does not become clear from the nested structure. However, evidence of the alternative signal flow comes from time delays of downstream effects. Assume that the time spent to propagate the signal from S_1 to S_2 plus the time spent to propagate the signal from S_2 to S_3 is larger than the time to propagate it from S_1 to S_3 directly, then there must exist an alternative short cut pathway from S_1 to S_3. Thus, temporal expression measurements yield additional insight into the cellular signal flow.

Fig. 1 illustrates the idea of D-NEMs in an elementary example. Shown is the hierarchical structure of a network and discrete time series data for three E-genes. A one indicates that a signal has reached the E-gene, while a zero indicates that the expression of this gene has not yet changed. Note, that the graph topology is consistent with the nested structure of ones in the final time point t_5, shown in red.

Signals starting in S_1 reach E_2 one time unit after they have arrived at E_1 suggesting that signal propagation from S_1 to S_2 takes one unit of time. The same argument using the data from perturbation of S_2 suggests that it takes two time units to propagate from S_2 to S_3. Consequently, going from S_1 to S_3 via S_2 takes 3 time units. However, the time delay from perturbation of S_1 to observing effects in E_3 is only 1 time unit (marked in blue). This suggests the existence of a direct signal flow from S_1 to S_3. Evidence comes from the two blue ones. In case they were zeros, the time delay between S_1 and S_3 would have been the sum of times spent when going via S_2. In this case, there would be no evidence for a short cut pathway and we would decide on the more parsimonious graph. A real world analysis is more difficult than the toy example. Signal propagation is a stochastic process, measurements are prone to noise, and we do not know which E-genes are controlled by which S-genes. These sources of uncertainty are addressed by D-NEMs.

We assume exponentially distributed time delays for individual signal propagation steps. The rate constants of the exponential distributions differ from case to case and are the main parameters of the model. All edges of a transitively closed network are associated with an individual rate constant, whose posterior distribution is inferred using Gibbs sampling. As explained before, molecular signalling in the cytoplasm occurs at high rates, direct signal propagation via transcription and translation at intermediate rates, and secondary effects at low rates. The joint posterior of the rate constants will be used to analyze the interplay of signalling networks and gene expression in complex biological processes. Moreover, it is used to further unravel molecular signal flow in cells.

Model

The input of a D-NEM consists of (a) a set of microarray time series that measure the response of cells to molecular perturbations, and (b) a transitively closed directed graph on vertex set $\mathbf{S}$ representing a hypothetical hierarchical structure of upstream/downstream relations. The output consists of (a) the joint posterior distribution of rate constants describing the dynamics of signal propagation, and (b) a not necessarily transitive subgraph of the input graph that describes signal flow rather than hierarchical structure. Let $D(i,k,l,s)$ denote the expression measurement of E_k in time point t_s of the l'th replication of a time series recorded after perturbation of S_i. Following Markowetz et al. [19], we assume that the data is binary, where zero encodes the wild type expression level of a gene, while one encodes that the expression of this E-gene changed due to perturbation and subsequent signal propagation.

Parameters. We assume that the time spent for propagating a signal from node S_i to node S_j is exponentially distributed with a rate constant k_{ij}. Note that the expected time spent in this step of signal transduction is $1/k_{ij}$. Fast processes are associated with high rate constants, while slow processes are associated with small rate constants. Exponential distributions are widely used to model temporal processes in complex systems [22, 23].

We do not observe the time spent for signal propagation between S-genes directly. Instead, we observe the time delay between a perturbation of an S-gene and the occurrence of downstream effects in E-genes. Following Markowetz et al. [19] we introduce parameters $\mathbf{\Theta} = (\theta_1, \ldots, \theta_N)$ to link E- to S-genes. If $\theta_k = i$, then E_k is linked to S_i. Moreover, we assume that every E-gene is linked to a single S-gene. The set of E-genes attached to the same S-gene is a regulatory module under the common regulatory control of the S-gene. The module of E-genes attached to S_i is denoted by $\mathcal{E}_i$. Finally, we introduce additional rate constants $k_{i\mathcal{E}}$ that represent the time delay between activation of S_i and regulation of its target module $\mathcal{E}_i$. A single common rate is used for all E-genes in the module.

Following ideas in Tresch and Markowetz [18], we add an additional node denoted by $+$, which is not connected to any of the S-genes. However, E-genes can be linked to this node, if they do not fit in any of the $\mathcal{E}_i$. The $+$-node implicitly selects E-genes. Genes linked to $+$ are excluded from the model.

We denote the complete set of rate constants including rates between S-genes and rates between S- and E-genes by $\mathbf{K}$. A priori, we do not know which E-genes fall into which modules. The joint posterior distribution of $\mathbf{\Theta}$ and $\mathbf{K}$ will be inferred from the data.

While the θ_k are discrete parameters by nature, rate constants are usually modelled as continuous parameters. However, for the sake of computational efficiency, we confine the rates to a discrete set of values denoted by $(\kappa_0, \ldots \kappa_T)$. If the data includes time points $(t_1, \ldots, t_T)$, we choose $(\kappa_0, 1/t_1, \ldots, 1/t_T)$, where κ_0 is set to a high value (i.e. 1,000) that represents the very fast signal transduction through post translational protein modification. Overall, we have a set of all discrete parameters $(\mathbf{K}, \mathbf{\Theta})$.

Prior distributions. Assuming independent prior distributions for $\mathbf{K}$ and $\mathbf{\Theta}$, Bayes's theorem yields $P(\Theta, K|D) = P(D|K,\Theta)P(K)P(\Theta)/P(D)$. The prior distribution $P(\mathbf{\Theta})$ can be chosen to incorporate prior knowledge on the interactions of S- with E-genes. Such information might be derived from ChIP on Chip data or regulatory motif analysis. The prior provides an interface, through which the model can be linked to different biological data types in integrative modelling approaches. Here we use the prior for calibrating E-gene selection. We set $p(\theta_k = +)$ to Δ, while distributing the remaining weight of $1 - \Delta$ uniformly on the values $1, \ldots, n$.

Similarly, the prior distribution $P(K)$ yields an interface for incorporating biological knowledge. If one knows that S_1 and S_2 fall into the same molecular signalling pathway, one can set $P(k_{12} = \kappa_0)$ to one, because signalling will operate on a high rate. In this paper we exploit the fact that transcription takes hours and set $P(k_{i\mathcal{E}} = \kappa_0)$ to zero, while assuming a uniform prior for the remaining values.

Likelihood. Let us first consider a fixed linear path g in Φ, which connects the S-gene S_i with the E-gene E_k:

$$S_i \xrightarrow{k_1} S_{j_1} \cdots \xrightarrow{k_{q-1}} S_{j_{q-1}} \xrightarrow{k_q} E_k,$$

where for simplicity of notation we reduce the double indices of rate constants to single indices and write $k_1, k_2, \ldots, k_q$ to denote the rate constants. We are interested in the time needed for propagating a signal from S_i down the path to E_k. More precisely, we want to calculate the probability, that the signal has reached E_k before some fixed time point t^*. If Z_g is the sum of q independent, exponentially distributed random variables with rate constants $k_1, \ldots, k_q$, then this probability equals $P(Z_g < t^*)$. The density function of Z_g is given by the convolution of independent exponential distributions

$$\Psi(t)_g = \int_0^\infty \cdots \int_0^\infty \delta\left(t - \sum_{u=1}^{q} \tau_u\right) \prod_{u=1}^{q} \psi_u(\tau_u)\, d\tau_1 \ldots d\tau_q,$$

where $\psi_u(\tau) = k_u \exp(-k_u \tau)$ is the density of an exponential with rate k_u. Laplace transformation yields a closed form for the cumulative distribution function of Z_g

$$F_g(t) = \sum_{b=1}^{q} \prod_{a \neq b} \left\{ \frac{k_a}{k_a - k_b} \right\} [1 - \exp(-t k_b)] \qquad [1]$$

Note that the right hand side is not defined if two or more of the k_u are identical. If two or more of the constants are identical, there are different expressions for the probability density, which contain exponential functions modulated by polynomials in time. However, as right and left limits exist and are identical, we can evaluate the probability by adding tiny distinct jitter values to the k_u.

In the general case a signal can be propagated from S_i to E_k via multiple alternative paths. In this case the fastest path determines the time delay for downstream effects to be seen. We enumerate all linear paths connecting S_i to E_k. For each path we construct a random variable Z_u as described above. The probability that the signal has arrived at E_k before time t^* via at least one of the paths is given by

$$P_{S_i \to E_k}(t^*) = 1 - \prod_u (1 - F_u(t^*)) \qquad [2]$$

Observations from E-genes linked to the +-node generate neutral likelihood values of 0.5 independent of all other parameters.

Equations (1) and (2) describe the stochastic nature of signal propagation in the cell. Before calculating the likelihood, we need to consider a second source of stochasticity, namely measurement error. Following Markowetz et al. [19], we denote the probabilities for false positive and false negative signals by α and β respectively. Assuming conditional independence, the likelihood factorizes into

$$\begin{aligned} P(D|K,\Theta) &= \prod_{D=1} P_{S_i \to E_k}(t_s)(1-\beta) + (1 - P_{S_i \to E_k}(t_s))\alpha \\ &\times \prod_{D=0} P_{S_i \to E_k}(t_s)\beta + (1 - P_{S_i \to E_k}(t_s))(1-\alpha), \end{aligned}$$

where the first product is over all data points, for which we observe a downstream effect, and the second product over those for which we do not.

Gibbs sampling. With N E-genes, n S-genes and L edges in the input graph, the model comprises $N + n + L$ discrete parameters. For simplicity of notation, we reduce the double indices of rate constants to single indices such that the joint posterior is written

$$P(k_1, \ldots, k_{L+n}, \theta_1, \ldots \theta_N | D).$$

We initialize the parameters with random values from their domains. Then we iteratively cycle through all rate constants updating them by sampling from the conditional posterior distributions

$$p(k_i | \mathbf{K} - \{k_i\}, \mathbf{\Theta}, D).$$

With only discrete parameters, updating is straight forward: We calculate all values

$$p(k_i = \kappa_j)\, p(D | \mathbf{K} - k_i, \mathbf{\Theta}, k_i = \kappa_j),$$

normalize them to sum up to one, and draw a new value for k_i from this distribution. The iteration is completed by similarly updating all θ_k. We typically run 10,000 iterations in which we sample from the joint posterior distribution of parameters, discard the first 1,000 iterations as burn in time, and summarize the remaining samples for inference of signal propagation. Choosing positive values for the tuning parameters α and β protects the conditional posterior distributions from singularity, and ensures good mixing properties of the Gibbs sampler.

Inference of signal flow. Under the natural assumption that perturbation effects propagate down the signalling network to all descendants of a perturbed gene, the nested structure of downstream effects resolves the network only up to its transitivity class. Network topologies with identical transitive closures produce the same nesting of downstream effects and, hence, can not be distinguished. As explained above, temporal data hold the potential of further resolving these transitivity classes. We enumerate all triplets of S-genes involving transitive edges like that shown in Fig. 1. We analyze the joint posterior distribution of the three rate constants to decide whether the transitive edge represents an alternative signal flow from S_1 directly to S_3 or whether signal exclusively flows from S_1 via S_2 to S_3. For each sample from the joint posterior of rates we calculate the difference $d = 1/k_{12} + 1/k_{23} - 1/k_{13}$ Without the transitive shortcut edge the distribution of d is expected to be centered around zero. If however a direct signal flow exists, d is expected to be positive. We discard a transitive edge e from Φ, if $P[d > 0] < p_0$ for all triplets of S-genes, where e is the transitive edge. The probability is calculated as the relative frequency of positive values along the trajectory of the Gibbs sampler.

Inference of the network topology. Due to the long running times of the Gibbs sampler it is not possible to reconstruct the network topology from scratch as was done for standard NEMs in [19, 20, 18]. Nevertheless, we use our method to discriminate between small numbers of candidate topologies using posterior odds for model comparison. Let us assume we have two hypothetical network topologies Φ_1 and Φ_2. The ratio of their posterior probabilities equals

$$\frac{P(\Phi_1)}{P(\Phi_2)} \times \frac{\int\int P(\Theta_1, K_1 | \Phi_1) P(D | \Theta_1, K_1, \Phi_1) d\Theta_1\, dK_1}{\int\int P(\Theta_2, K_2 | \Phi_2) P(D | \Theta_2, K_2 \Phi_2) d\Theta_2\, dK_2}$$

with Θ_i and K_i representing the parameters in model Φ_i. In principle, the integrals in the Bayes factor can be approximated by averages along the Gibbs sampling trajectories. In practice, this is not feasible due to the numerical representation of the tiny likelihood values. Instead we calculate the relative deviances along the trajectories of the Gibbs samplers

$$D_k(\Phi_1, \Phi_2) = \log(P(\Phi_1|D, \Theta_1^j, K_1^j)) - \log(P(\Phi_2|D), \Theta_2^j, K_2^j)$$

where Θ_i^j and K_i^j are the current settings of parameters in the Gibbs sampling trajectory for Model Φ_i after completing the j'th round of updating. Positive values for D_k suport model Φ_1, whereas negative values indicate that model Φ_2 is better supported by the data.

Results

Simulations. We evaluate our method in the context of simulated data from the network shown in Fig. 2A. The topology of this network is identical to the one that we will use in an application to stem cell development in the next section. The numbers annotating the edges of the network are time delays of signal propagation along this edge. For simplicty, edges between S-genes and E-genes are not shown, nor are transitive edges. Time delays for signal propagation between S- and E-genes are set to one for all such edges, while time delays for the transitive edges are equal to the sum of the delays in the paths that they cut short. Hence, the simulated data does not generate evidence for transitive edges as described in the previous section. For all possible E-gene positions the expected data pattern across time points and perturbation experiments is calculated and artificial E-gene data is simulated by adding independent binary noise to these patterns using $\alpha = 0.1, \beta = 0.05$. We simulate data for 20 E-genes per S-gene and one measurement per time point, resulting in a data array of 840 binary values. The D-NEM procedure is run on this data using 10,000 iterations, from which the first 1,000 are discarded as burn in time. In the likelihood we set $(\alpha = 0.2, \beta = 0.1)$. Note that these values are different from those used in data generation.

Fig. 2B shows the histogram of time delays (reciprocal rate constants) along the Gibbs sampling trajectory for the transitive edge between S_5 and S_6. It is equivalent to the top most color intensity profile of the heat map in Fig. 2C. The histogram reflects the marginal posterior probability of this parameter. Note that the signal flow is

$$S_5 \xrightarrow{1} S_4 \xrightarrow{1} S_2 \xrightarrow{1} S_6$$

and that the expected time delay for signal propagation between S_5 and S_6 equals 3 units of time. Accordingly, the histogram displays a distribution with mode 3. The heat map in Fig. 2C summarizes the posterior probabilities for all edges. Small time lags of zero or one are recovered with high precision from the D-NEM, while more dispersed posterior probabilities are observed for larger time delays. Moreover we observe a slight overestimation of the high time delays. Nevertheless, posterior modes are generally close to real values. To show that the results are not limited to the set of rates used for this simulation, we have run nine additional simulations based on randomly chosen time delays. Fig. 2D shows a scatter plot of the time delays underlying the simulations vs. the posterior modes.

In order to evaluate the ability of our model to detect transitive edges we have run a second simulation experiment, in which we set the time delays for all transitive edges to 1. Note that for all transitive edges a time delay of 1 is smaller than the sum of time delays along the paths they cut short, thus simulating evidence for short cuts in signal propagation. Fig. 2E compares the relative frequency of positive values of the transitivity score d along the Gibbs trajectory in the two simulations. As expected the second simulation produced a higher posterior probability $P[d > 0]$ than the first one. With a cutoff of $p_0 = 0.5$, we correctly discard all transitive edges in the first simulation while retaining them in the second one.

Application to murine stem cell development. We apply the D-NEM approach to a data set on the molecular mechanisms of self-renewal in murine embryonic stem cells. Ivanova et al [24] used RNA interference techniques to downregulate six gene products associated with self-renewal regulatory function, namely *Nanog*, *Oct4*, *Sox2*, *Esrrb*, *Tbx3* and *Tcl1*. They combined perturbation of these gene products with time series of microarray gene expression measurements. Mouse embryonic stem cells (ESC) were grown in the presence of the leukemia inhibitory factor LIF thus retaining their undifferentiated self-renewing state (positive controls). Cell differentiation associated changes in gene expression were detected by inducing differentiation of stem cells through removing LIF and adding retinoic acid (RA) (negative controls). Finally, RNAi based silencing of the six regulatory genes was used in (LIF+, RA-) cell cultures to investigate, whether silencing of these genes partially activates cell differentiation mechanisms. Time series including (6-7) time points in one day intervals were taken for the positive control culture (LIF+, RA-), the negative control culture (LIF-, RA+), and the six RNAi perturbed assays. In the context of the D-NEM framework the six regulatory gene products *Nanog*, *Oct4*, *Sox2*, *Esrrb*, *Tbx3* and *Tcl1* are S-genes, while all genes showing significant expression changes in response to LIF depletion are used as E-genes. Downstream effects of interest are those, where the expression of an E-gene is pushed from its level in self-renewing cells to its level in differentiated cells. Our goal is to model the temporal occurrence of these effects across all time series simultaneously.

The dataset consists of 64 Affymetrix microarrays consisting of eight time series with eight time points at one day intervals each. One time series for self-renewing stem cells (LIF+, RA-), one time series for cells passing through early differentiation (LIF-, RA+), and six time series for LIF stimulated ESCs with one of the six regulators *Nanog*, *Oct4*, *Sox2*, *Esrrb*, *Tbx3* and *Tcl1* silenced by RNAi. In a comparison of the (LIF+, RA-) to the (LIF+, RA-) cell cultures 135 genes showed a more than 2 fold up or down regulation across all time points. These were used as E-genes in our analysis. The two time series without RNAi were also used to discretize the time series of pertubation experiments following the discretization method of Markowetz et al.[15], thereby setting an E-gene state to 1 in an RNAi experiment, if its expression value is far from the positve controls, and 0 otherwise. Genes that did not show any 1 after discretization across all experiments were removed, leaving us with 122 E-genes for further analysis.

D-NEMs assume that once a perturbation effect has reached an E-gene it persists until the end of the time series. In other words, a one at time point t indicates that a downstream effect has reached the E-gene prior to t and not that it is still observable at this time. Hence, a typical discretized time series starts with zeros, eventually switches to ones and then stays one until the end of the series (e.g. 0001111), we refer to these patterns as admissible patterns. In general, the discretized data was roughly following admissible patterns. Nevertheless, exceptions were observed most likely due to measurement noise or cellular compensation of primary downstream effects. We replaced the time series for each gene by the closest admissible pattern, based on edit dis-

tances. In the case where several admissible patterns had the same edit distance to the time series, we chose the pattern holding the most ones. This curated data was used in further analysis.

Since long computation times for Gibbs sampling prohibit the reconstruction of the nework's topology from scratch using D-NEMs, we used the triplet search approach for the standard nested effect approach [19] applied to the final time point to determine a topology for the network. Note, that the final time point of an admissible pattern accumulates information along the time series, because it reports a one whenever a downstream signal has reached the E-gene at any time. The binary data of the last time point across all S-gene perturbations is shown in Fig. 3A, while Fig. 3B shows the reconstructed network. A nested structure is visible. For example, the top 4 rows in Fig. 3A show a staircase like pattern of nested sets consistent with the linear cascade $Nanog \longrightarrow Sox2 \longrightarrow Oct4 \longrightarrow Tcl1$. The topology is based exclusively on the nesting of downstream effects. Time delays of signal propagation can now be used for fine tuning the topology: Originally, the triplet search suggested a bidirectional arrow between *Oct4* and *Tcl1* ($Oct4 \longleftrightarrow Tcl1$) indicating that the same sets of E-genes show downstream effects in *Oct4* and *Tcl1* silencing assays. Moreover, most of these E-genes also show effects, when silencing other E-genes suggesting that both *Oct4* and *Tcl1* are at the downstream end of the network. Using the time delay might resolve this interaction further. In order to resolve the network further, we used deviance based model comparison. We compare the two network hypotheses Φ_1 and Φ_2, which place *Oct4* up- or downstream of *Tcl1*. The posterior distribution of the deviance $D(\Phi_1, \Phi_2)$ has mass only on positive values far from zero, which is strongly supporting that *Oct4* is upstream of *Tcl1*.

Next, we exploit the D-NEM Gibbs sampler trajectories associated with the network topology from Fig. 3B for inference of time delays and regulatory control of E-genes. The posterior heat map for all edges is shown in Fig. 3C. The posteriors are spread out wider than in the simulation results shown in Fig. 2B. Nevertheless, the posterior mass either concentrates around zero indicating no time delay for this step of signal propagation, on intermediate values, or exclusively on large values explaining secondary and tertiary effects. Note that Fig. 3B shows transitive short cut edges inferred from the analysis of the posterior distributions of d-values shown in Fig. 3D. Note that with only one exception, all transitive short cut edges are strongly supported by obserbved time delays, emphasizing the high degree of connectivity of this developmental network.

Discussion

Signalling through post-translational modifications is not directly observable on microarrays. The phosphorylation of a signalling molecule (an S-gene) does not change its expression and appears to be under the radar of the microarray. Inspite of this, the phosphorylation can activate the signalling molecule and this stimulus is then propagated via signal transduction to activate transcription factors, which bind to promoters activating or repressing the transcription of genes, leaving indirect traces on a microarray. Even if the expression of the S-gene is not changed, reflections of signaling dynamics are perceived in expression levels of other genes, the E-genes. Unlike dynamical Bayesian networks, which model how the expression of one gene influence the expression of another, D-NEMs model the indirect downstream effects of pertubations described above. Inference on the dynamics of signal propagation between S-genes becomes possible even if the expression of S-genes is flat across the time series.

Through Gibbs sampling we do not only obtain estimates for rate constants of signal propagation and E-gene positions but also natural measures of the uncertainty associated with these estimates. This is important, because in many applications including the example given in this manuscript the data at hand will not hold enough information to fully recover the dynamics of the entire network. Some aspects will have strong support in the data while others will not. The joint posterior of parameters allows us to identify badly resolved features and can help to design further experiments, leading to a better model of the network.

In summary, the present paper extends the framework of nested effects models to time series data. To our knowledge, D-NEMs are currently the only method to model the dynamics of a cell's response to perturbation from unspecific downstream events. They compliment dynamic Bayesian networks and might play an important role for understanding slow-going biological processes and their disruption in human disease.

ACKNOWLEDGMENTS. This research was supported by the Bavarian Genome Network BayGene and the Reform-M program of the Regensburg school of medicine. In addition, M.O. Vlad was supported by the National Science Foundation and CEEX grant M1-C2-3004/2006-Response of the Romanian Ministry or Research and Education.

1. Stuart, J and Segal,E and Koller,D and Kim, K (2003) A gene-coexpression network for global discovery of conserved genetic modules Science 302:249-55.
2. Wille,A and Zimmermann, P and Vranov, E and Furholz, A and Laule, O and Bleuler, B and Hennig, L and Prelic, A and Rohr, P V and Thiele, L and Zitzler, E and Gruissem, L and Buhlmann, P (2004) Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana Genome Biol 5:R92.
3. Schaefer, J and Strimmer, K (2005) An empirical Bayes approach to inferring large-scale gene association networks Bioinformatics 21:754-64.
4. Basso, K and Margolin, A A and Stolovitzky, G and Klein, U and Dalla-Favera, R and Califano, A (2005) Reverse engineering of regulatory networks in human B cells Nat Genet 37:382-390.
5. Friedman, N and Linial, M and Nachman, I and Pe'er, D (2000) Using Bayesian networks to analyze expression data Journal of Computational Biology 7:601-620.
6. Husmeier, D (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks Bioinformatics 19:2271-2282.
7. Quach, M and Brunel, N and d'Alch-Buc, N (2007) Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference Bioinformatics 23:3209-3216.
8. Klipp, E and Liebermeister, W (2006) Mathematical modeling of intracellular signaling pathways BMC Neurosci 7:S10.
9. Werhli, A and Grzegorczyk, M and Husmeier, D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks Bioinformatics 22:2523-2531.
10. Markowetz, F and Spang, R (2003) Evaluating the effect of perturbations in reconstructing network topologies Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
11. Sachs, K and Perez, O and Peer, D and Lauffenburger, D A and Nolan, G P (2005) Causal proteinsignaling networks derived from multiparameter single-cell data Science 308:523-9.
12. Rung, J and Schlitt, T and Brazma, A and Freivalds, K and Vilo, J (2002) Building and analysing genomewide gene disruption networks Bioinformatics 18:202-210.
13. Wagner, A (2002) Estimating Coarse Gene Network Structure from Large-Scale Gene Perturbation Data Genome Res. 12:309-315.
14. Pe'er, D and Regev, A and Elidan, G and Friedman, N (2001) Inferring subnetworks from perturbed expression profiles Bioinformatics 17:S215-S224.
15. Markowetz, F and Bloch, J and Spang, R (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interferenc eBioinformatics 21:4026-32.
16. Yeang, C H and Ideker, T and Jaakkola, T (2004) Physical network models Journal of Computational Biology 11:243-262.

17. Froehlich, H and Fellmann, M and Sueltmann, H and Poustka, A and Beissbarth, T (2007) Large scale statistical inference of signaling pathways from RNAi and microarray data BMC Bioinformatics 8:386.

18. Tresch, A and Markowetz, F (2008) Structure learning in Nested Effects Models Stat Appl Genet Mol Biol 7 :Article9.

19. Markowetz, F and Kostka, D and Troyanskaya, OG and Spang, R (2007) Nested effects models for high-dimensional phenotyping screens Bioinformatics 13:i305-12.

20. Froehlich, H and Fellmann, M and Sueltmann, H and Poustka, A and Beissbarth, T (2008) Estimating Large Scale Signaling Networks through Nested Effect Models with Intervention Effects from Microarray Data Bioinformatics 10.

21. Froehlich, H and Beissbarth, T and Tresch, A and Kostka, D and Jacob, J and Spang, R and Markowetz, F (2008) Analyzing Gene Perturbation Screens With Nested Effects Models in R and Bioconductor Bioinformatics 2.

22. Vlad, M O and Moran, F and Tsuchiya, M and Cavalli-Sforza L L and Oefner, P J and Ross, J (2002) Neutrality condition and response law for nonlinear reaction-diffusion equations, with application to population genetics Phys. Rev. E. 65, 061110: 1-17.

23. Vlad, M O and Arkin, A and Ross, J (2004) Response experiments for nonlinear systems with application to reaction kinetics and genetics PNAS 101:7223-7228.

24. Ivanova, N and Dobrin, R and Lu, R and Kotenko, I and Levorse, J and Decoste, C and Schafer, X and Lun, Y and Lemischka, I (2006) Dissecting self-renewal in stem cells with RNA interference Nature 442:533-538.

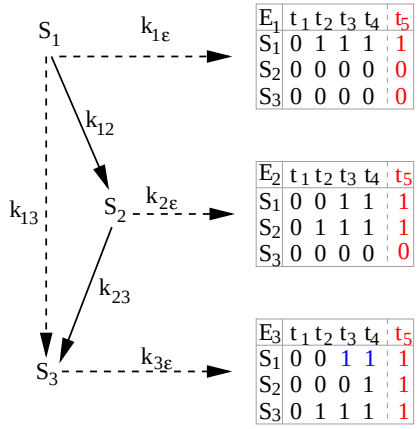

Fig. 1. Elementary example of a D-NEM together with underlying binary time series data.

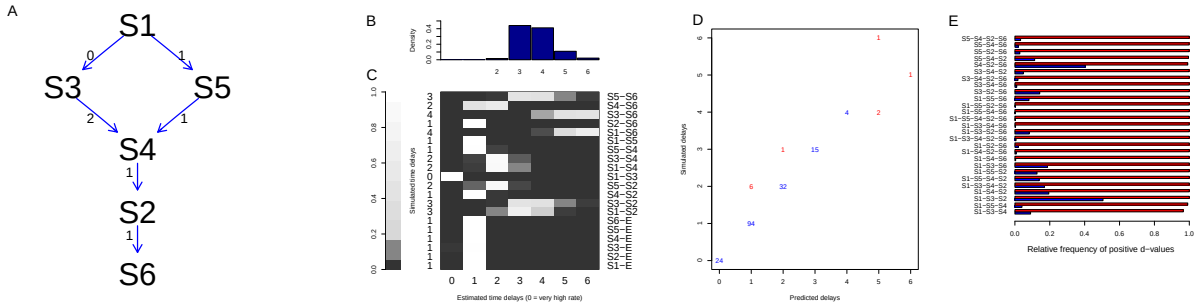

Fig. 2. **A** The topology and time delays underlying our simulation experiments. Edges between S-genes and E-genes are not shown, nor are transitive edges. Time delays for signal propagation between S- and E-genes are set to one for all such edges, while time delays for the transitive edges are equal to the sum of the delays in the paths that they cut short. **B** The marginal posterior distribution of time delays for the transitive edge between S_5 and S_6 shown as a histogram. **C** Heat map summarizing the marginal posterior distribution for all edges. Rows correspond to edges of the network including those between S- and E-genes, while columns refer to time delays. Marginal posterior probabilities are color coded. The top row corresponds to the histogram above. The simulated time delays are shown on the y-axis to the left of the heat map. **D** Estimation performace on 9 models with random time delays. The x-axis shows the time delays underlying the simulations, while the y-axis shows the estimated time delays, given by marginal posterior modes. The blue numbers on the diagonal count correct estimations, while the red numbers off the diagonal count deviating estimations. **E** Transitivity scores. The bars represent relative frequencies of positive transitivity scores along the Gibbs trajectory. The blue bars correspond to the simulation with and the blue bars to the simulation without short cut edges.

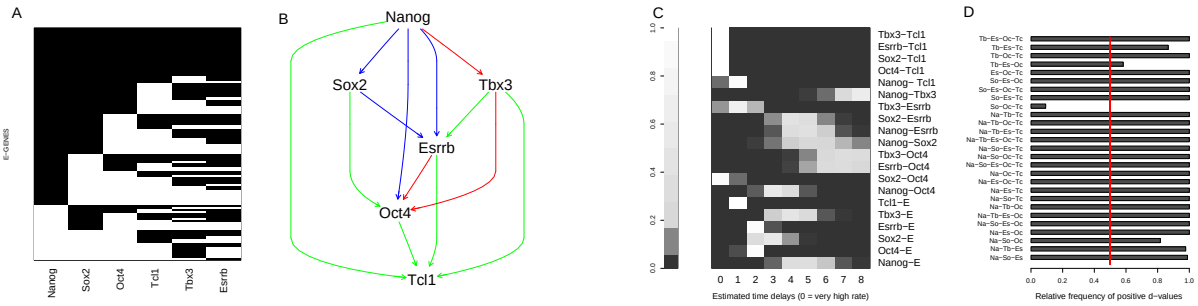

Fig. 3. Stem cell data analysis: **A** The discretized data of the last time point across E-gens (rows) and S-gene perturbations (colums), with black representing downstream effects and white no effects. **B** The network structure estimated by triplet search and subsequent time delay analysis. Note that this network has the same transitive closure as the one in Fig. 2A. Edge colors correpond to estimated time delays: fast signal propagation (green), intermediate signal propagation (blue) and slow signal propagation (red). **C** Heat map of the posterior distribution of time delays. Rows correspond to edges of the network including those between S- and E-genes, while columns refer to time delays. Marginal posterior probabilities are color coded. **D** Transitivity scores: The bars represent relative frequencies of positive transitivity scores along the Gibbs trajectory. The red line shows the cut-off value of 0.5 used in this analysis.