BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i305–i312
doi:10.1093/bioinformatics/btm178

Nested effects models for high-dimensional phenotyping screens

Florian Markowetz[1], Dennis Kostka[2], Olga G. Troyanskaya[1,*] and Rainer Spang[3]

[1]Lewis-Sigler Institute for Integrative Genomics and Department of Computer Science, Princeton University, Princeton, NJ, 08544, USA, [2]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin and [3]Institute for Functional Genomics, Computational Diagnostics Group, University of Regensburg, Josef Engertstr. 9, 93503 Regensburg, Germany

ABSTRACT

Motivation: In high-dimensional phenotyping screens, a large number of cellular features is observed after perturbing genes by knockouts or RNA interference. Comprehensive analysis of perturbation effects is one of the most powerful techniques for attributing functions to genes, but not much work has been done so far to adapt statistical and computational methodology to the specific needs of large-scale and high-dimensional phenotyping screens.

Results: We introduce and compare probabilistic methods to efficiently infer a genetic hierarchy from the nested structure of observed perturbation effects. These hierarchies elucidate the structures of signaling pathways and regulatory networks. Our methods achieve two goals: (1) they reveal clusters of genes with highly similar phenotypic profiles, and (2) they order (clusters of) genes according to subset relationships between phenotypes. We evaluate our algorithms in the controlled setting of simulation studies and show their practical use in two experimental scenarios: (1) a data set investigating the response to microbial challenge in *Drosophila melanogaster*, and (2) a compendium of expression profiles of *Saccharomyces cerevisiae* knockout strains. We show that our methods identify biologically justified genetic hierarchies of perturbation effects.

Availability: The software used in our analysis is freely available in the R package 'nem' from www.bioconductor.org

Contact: ogt@cs.princeton.edu

1 INTRODUCTION

Functional genomics has a long tradition of inferring the inner working of a cell through analysis of its response to various perturbations. Observing cellular features after knocking out or silencing a gene reveals which genes are essential for an organism (Boutros *et al*., 2004) or for a particular pathway (Gesellchen *et al*., 2005). In computational biology, the importance of perturbations for network reconstruction has been recognized in many different inference frameworks (Markowetz and Spang, 2003; Pe'er *et al*., 2001; Sachs *et al*., 2005; Van Driessche *et al*., 2005; Wagner, 2001; Werhli *et al*., 2006; Yeang *et al*., 2004).

There are several perturbation techniques suitable for large-scale analysis in different organisms, including RNA interference (Fire *et al*., 1998) and gene knockouts (Hughes *et al*., 2000). Perturbation effects can be measured either by single reporters like viability (Boutros *et al*., 2004) or by high-dimensional readouts like gene expression profiles (Boutros *et al*., 2002; Hughes *et al*., 2000; Van Driessche *et al*., 2005), metabolite concentrations (Raamsdonk *et al*., 2001), sensitivity to cytotoxic or cytostatic agents (Brown *et al*., 2006) or morphological features of the cell (Ohya *et al*., 2005). High-dimensional phenotypic profiles promise a comprehensive view on the function of genes in a cell, but only limited work has been done so far to adapt statistical and computational methodologies to the specific needs of large-scale and high-dimensional phenotyping screens.

A key obstacle to inferring genetic networks from high-dimensional perturbation screens is that phenotypic profiles generally offer only indirect information on how genes interact. Cell morphology or sensitivity to stresses are global features of the cell, which are hard to relate directly to the genes contributing to them. Gene expression phenotypes also offer an indirect view of pathway structure due to the high number of post-transcriptional regulatory events like protein modifications.

Previous work. Most previous work focused on clustering phenotypic profiles to find groups of genes that show similar effects when perturbed. The rationale is that genes with similar perturbation effects are expected to be functionally related. The most prominent method used is average linkage hierarchical clustering (Ohya *et al*., 2005; Piano *et al*., 2002). A complementary approach is ranking genes according to similarity with a query gene; e.g. the 'phenoBlast' algorithm (Gunsalus *et al*., 2004) implements lexicographic sorting. In a supervised setting, first steps have been taken to classify genes into functional groups based on phenotypic profiles (Ohya *et al*., 2005).

Both the supervised and unsupervised methods discussed above are based on a notion of similarity between phenotypic profiles. We see two limitations in such similarity-based approaches: in general, similarity measures weight observed and unobserved effects in the same way. However, in large-scale phenotyping experiments it may be more likely to miss an effect because of compensatory efforts of the cell than to see a spurious effect. So far, only phenoBlast takes this imbalance into account.

An even more important issue is that similarity-based methods may miss important features of the data, which do not relate to the similarity of profiles within a cluster, but to the relationships of effects for different clusters. For example, existing methods do not take into account subset relationships in observed perturbation effects, which can be indicative of specific cellular behaviors such as regulatory mechanisms.

*To whom correspondence should be addressed.

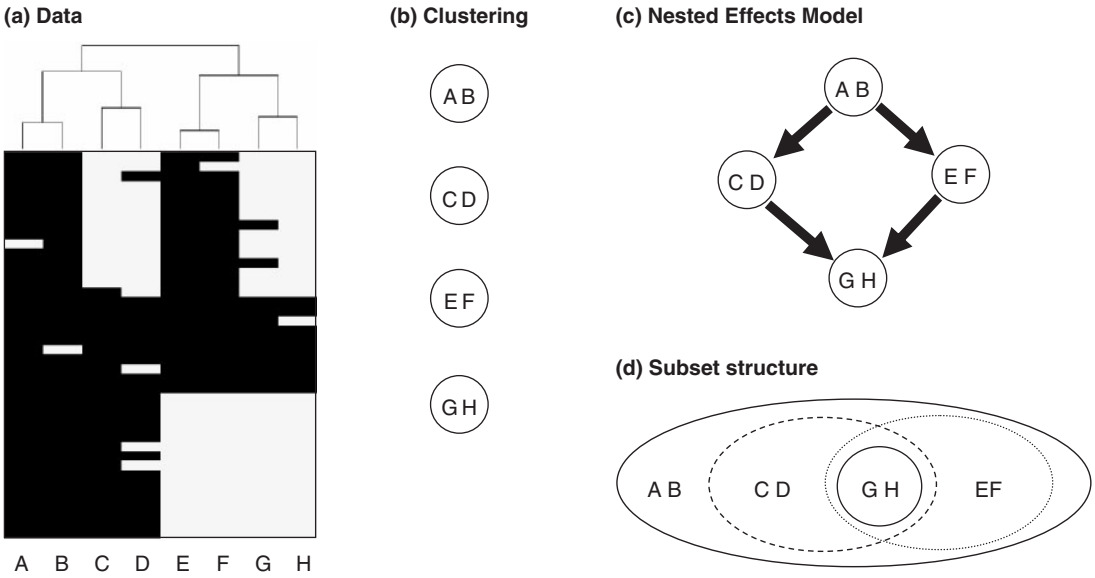

Fig. 1. An introduction to Nested Effects Models. Plot (**a**) shows a toy dataset consisting of phenotypic profiles for eight perturbed genes ($A, \ldots, H$). Each profile is binary with *black* coding for an observed effect and *white* for an effect not observed. The eight profiles are hierarchically clustered, showing that they fall into four pairs of genes with almost identical phenotypic profiles: (A, B), (C, D), (E, F) and (G, H), as shown in plot (**b**). An important feature of the data missed by clustering is the subset structure visible between the profiles in the data set: the effects observed when perturbing genes A or B are a superset to the effects observed for all other genes. The effects of perturbing G or H are a subset to all other genes' effects. The pairs (C, D) and (E, F) have different but overlapping effect sets. The directed acyclic graph (DAG) shown in plot (**c**) represents these subset relations, which are shown in plot (**d**). Compared to the clustering result in plot (b) the NEM additionally elucidates relationships between the clusters and thus describes the dominant features of the data set better.

Clustering defines groups of genes with similar phenotypic profiles, but may miss the hierarchy in the observed perturbation effects, as is exemplified in Figure 1. Perturbing some genes may have an influence on a global process, while perturbing others affects subprocesses of it. Imagine, e.g. a signaling pathway activating several transcription factors (TFs). Blocking the entire pathway will affect all targets of the TFs, while perturbing a single downstream TF will only affect its direct targets, which are a subset of the phenotype obtained by blocking the complete pathway. Boutros *et al*. (2002) show that by this reasoning non-transcriptional features of signaling pathways can be recovered from gene-expression profiles. However, no previous computational method is applicable to infer models from biological subset relations on data sets screening whole pathways.

Nested effects models. We will call a model encoding the (noisy) subset relations between the effects observed after perturbing the target genes a *Nested Effects Model* (NEM). It can be seen as a generalization of similarity-based clustering, which orders (clusters of) genes according to subset relationships between the sets of phenotypes. In this article, we develop a Bayesian method to infer NEM from large-scale data sets.

Our method builds on preliminary work by Markowetz *et al*., (2005), which is specifically designed for inference from indirect information and also takes the imbalance between spurious and missed effects into account. Previously, this method was limited to small-scale scenarios of up to six genes, where model search can be done by exhaustive enumeration. Scaling up model search to larger numbers of perturbed genes is a non-trivial problem due to the constraints imposed on the model by having only indirect information of the underlying genetic network. Here, we approach the problem of inferring a hierarchy on the set of *all* perturbed genes by constructing it from smaller *sub-models* containing only pairs or triples of genes. Such '*divide-and-conquer*'-like approaches are regularly used in high-dimensional statistical inference, e.g. for estimating large phylogenetic trees (Strimmer and von Haeseler, 1996) or learning Gaussian graphical models for regulatory networks (Wille *et al*., 2004). Our resulting method is the first one to make inference of NEMs feasible on a pathway-wide scale.

The next section introduces our novel methodology in detail. In Section 3, we demonstrate the applicability of our methods in a controlled simulation study, and in Section 4 we describe results for two experimental data sets. We show that the subset relations retrieved actually reflect the regulatory functions of the genes involved.

2 ALGORITHM

Data. We assume that data is given in the form of a binary matrix D with columns corresponding to perturbation experiments on one of n genes (replicates are possible) and rows to one of m possible effects $E_1, \ldots, E_m$. A phenotypic profile P_x of gene x consists of a binary vector of length m with a $P_x(E_i) = 1$ denoting that effect E_i occurred after perturbing gene x, and $P_x(E_i) = 0$ denoting that it did not.

Subset relations between phenotypic profiles. Instead of similarity, we will consider subset relations between phenotypic profiles. We say that gene x is *upstream* of gene y

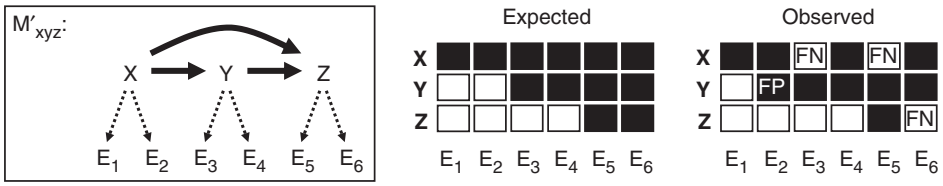

Fig. 2. A complete model. The left part of the figure shows a complete model M'_{xyz} consisting of a transitively closed graph between genes and assignments of genes to specific effects (the dashed arrows). Given the complete model, we can formulate a prediction of what effects to expect: perturbing x should cause all effects, while perturbing y should only cause E_3–E_6, and perturbing z only E_5 and E_6 (middle plot). In reality, our observations will be noisy: there can be false positive (FP) and false negative (FN) effect observations (right plot).

(and write $x \to y$) if the set of effects in P_y is a subset of the set of effects in P_x:

$$x \to y \quad \Leftrightarrow \quad \{i \,:\, P_y(E_i) = 1\} \subseteq \{i \,:\, P_x(E_i) = 1\}. \quad (1)$$

A subset relation is reflexive and transitive, and thus defines a quasi-order on phenotypic profiles. We depict the quasi-order in a directed graph in which nodes correspond to gene perturbations and edges indicate subset relations according to Equation (1). The reflexive self-loops at nodes are usually omitted. Transitivity is the key feature of our model: whenever there is a path from one node to another, we also have a directed edge between these two nodes in the graph.

2.1 Bayesian inference for NEM models

Posterior probability A Bayesian score to evaluate how well a candidate NEM fits to the observed data can be obtained in two steps (Markowetz *et al.*, 2005). First, assume that it is known which effect is specific for which perturbed gene. We call this the *complete model*, and an example is given in Figure 2. A complete model $M' = (M, \Theta)$ consists of a transitively closed graph, M, and parameters $\Theta = \{\theta_1, \ldots, \theta_m\}$. The nodes of M correspond to perturbed genes, and the parameters Θ describe the allocation of specific effects to perturbed genes (i.e. the dashed arrows in the left plot of Fig. 2). The complete model defines which effects we expect to observe (see the middle plot of Fig. 2). We can directly compute the complete likelihood of the actually observed data D under the model (M, Θ) by:

$$P(D|M, \Theta) = \prod_{i=1}^{m} \prod_{k=1}^{l} P(e_{ik}|M, \theta_i), \quad (2)$$

where, the first product is over all effects $E_1, \ldots, E_m$ and the second over all replicates of gene perturbation experiments. The probability $P(e_{ik}|M, \theta_i)$ depends on two parameters: a FP rate of seeing a spurious effect, α (type-I error rate), and a FN rate of missing an effect, β (type-II error rate).

However, in real data, it is not known which effect is specific for which intervention, i.e. Θ is unknown. Thus, in a second step, we average over Θ to gain the likelihood of the data, which is proportional to the posterior probability of the NEM and can be written as:

$$P(D|M) \propto \prod_{i=1}^{m} \sum_{j=1}^{n} \prod_{k=1}^{l} P(e_{ik}|M, \theta_i = j), \quad (3)$$

where the two products are the same as in Equation (2), and the sum is due to marginalization over Θ.

Size of model space. NEMs are defined in terms of quasi-orders, i.e. transitively closed graphs. The number of quasi-orders is known for up to 16 nodes (Sloane, 2007, seq. A000798). For $n = 7$, we already have almost 10^7 possible quasi-orders and for $n = 8$ the number is $> 6 \cdot 10^9$. Thus, exhaustive enumeration is infeasible even for medium-sized studies. For large-scale screens, we need search heuristics to explore model space. Our approach to this problem is to concentrate on small sub-models involving only pairs or triples of nodes.

2.2 Inference of pairwise relations

The smallest possible sub-model consists of pairs of genes. We infer pairwise relations by choosing between four models for each gene pair (x, y): either $x \to y$ (“upstream”, effects of x are a superset of the effects of y), or $x \leftarrow y$ (“downstream”, effects of x are a subset of the effects of y), or $x \leftrightarrow y$ (the effects of x and y are undistinguishable) or $x \cdot\cdot\, y$ (x and y are unrelated). For every pair (x, y), we compute the Bayesian score detailed above and select the maximum *aposteriori* (MAP) model $M_{xy} \in \{x \leftarrow y, x \to y, x \leftrightarrow y, x \cdot\cdot\, y\}$.

The greatest advantage of this procedure is the increase in speed. The number of models we have to score for n genes is $\binom{n}{2} \cdot 4$, which grows quadratically in the number of perturbed genes and remains feasible even for hundreds of genes. Additionally, building up the final graph is easy, since it is defined by the set of all pairwise MAP models.

These advantages come at a cost. The most serious problem is that pairwise learning treats all edges independently of each other. But in a transitive graph, there must be a shortcut $x \to y$ whenever there exists a longer path from x to y. To see how easily mistakes can be introduced in pairwise inference, consider the example in Figure 2. In the observed data (rightmost plot), the profiles of x and z seem non-overlapping (because of the FNs at E_5 and E_6), so the edge $x \to z$ could be missed. One can also think of scenarios, where noise in the data induces spurious edges in pairwise inference. To address these problems, we concentrate on triples of nodes in the next section.

2.3 Inference of triple relations

Inference from triples of genes instead of pairs is a natural way to extend our inference method beyond the independence

assumption between edges. To build a graph on n nodes, we propose the following two steps:

(1) **Scoring all triples:** for each triple (x, y, z), we score all 29 possible quasi-orders and select the MAP model. The number of models to be scored is $\binom{n}{3} \cdot 29$, which grows as $O(n^3)$ and is still feasible even for dozens of genes.

(2) **Edge-wise model averaging:** to combine these models into one final graph, we employ model averaging. Every edge can be part of $n-2$ different triple models. Counting how often it actually is chosen assesses the models' confidence in edge existence. For each edge, we compute

$$f(x \to y) = \frac{1}{n-2} \sum_{z \notin \{x,y\}} 1[\text{“}x \to y\text{”} \in M_{xyz}], \qquad (4)$$

where, $1[\cdot]$ is an indicator function for the existence of an edge $x \to y$ in a model M_{xyz}. The final graph is then constructed from edges whose frequency $f(x \to y)$ exceeds a certain threshold (we chose 0.5 in our applications).

Even though all triplet models are transitively closed, edge-wise model averaging and thresholding are not guaranteed to yield a transitively closed graph. However, in our experience the results are closer to a quasi-order than with the pairwise approach and empirically show an increase in precision. This observation holds over a wide range of noise levels and data set sizes, as we will show in Section 3.

2.4 Representation of inferred quasi-orders

The last two sections introduced two approaches to infer large quasi-orders. This section describes our use of standard graph algorithms to find clusters of undistinguishable profiles and to distinguish direct from indirect relationships in the graph.

Merge undistinguishable profiles into cluster. First, we identify the strongly connected components (SCCs) in the quasi-order (Cormen *et al*., 2002). SCCs are defined as subsets of nodes in which all pairs of nodes are mutually reachable by paths in the graph. In our setting—where edges encode subset relations—this corresponds to pairs of genes with undistinguishable profiles. We merge SCCs into single nodes and retrieve a transitively closed DAG with clusters of undistinguishable profiles as nodes.

Remove shortcuts. To distinguish direct from indirect relationships, we use a method for transitive reduction (Wagner, 2001) on the DAG to remove direct edges ('shortcuts') between nodes that are also connected by a longer path. The method iteratively compares the adjacency list of nodes with the adjacency lists of the nodes' children to cut edges which appear in both. The final result is the DAG with the smallest number of edges satisfying all subset relations. An overview of this process is summarized in Figure 3.

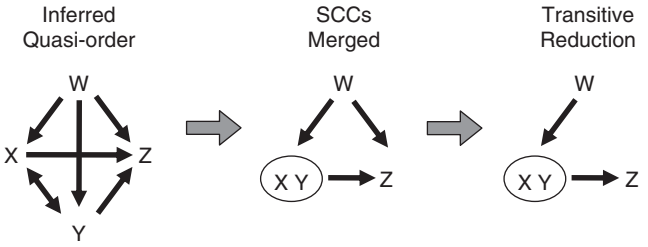

Fig. 3. Summary of the proposed algorithm. The left plot shows an example quasi-order inferred from data using either the pairs- or triples-based approach. In a first step, the SCC consisting of nodes X and Y is merged into a single node. In a second step, the shortcut $W \to Z$ is removed. In big graphs, these two steps tremendously improve readability of results.

3 EVALUATION ON SIMULATED DATA

We performed simulations to assess the performance of NEM inference for varying noise levels and data set sizes. The setup and choice of parameters is inspired by the simulation study conducted in Markowetz *et al*. (2005) to evaluate the performance of exhaustive enumeration.

(1) We randomly generate a graph with $n \in \{4, 8, 32\}$ nodes. The number of edges in the random graph are $\{4, 11, 55\}$, which ensures that the transitively closed graph contains in average half of all possible edges. The transitively closed random graph constitutes the core model M_{true}. We distribute $2n$ effect reporters randomly over the core model to generate a complete model.

(2) From the complete model, we generate data assuming error probabilities α_{data} and β_{data}. While β_{data} is fixed to 0.05, α_{data} is varied from 0.1 to 0.5. We sample 1–5 times to gain data sets with increasing numbers of replicates.

(3) From each data set, we infer an NEM model M_{NEM} by the pairwise and triple approach (and by exhaustive enumeration for $n = 4$) with parameters $\alpha_{\text{score}} = 0.1$ and $\beta_{\text{score}} = 0.3$. Note that these (arbitrarily chosen) values are different from $(\alpha_{\text{data}}, \beta_{\text{data}})$ used for data generation.

(4) We compute the positive predictive value of M_{NEM} with respect to M_{true} as the fraction of true positive edges out of all edges in M_{NEM}:

$$\text{positive prediative value } (M_{\text{NEM}}) = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad (5)$$

where TP are the true positive edges, and FP are the FP edges. The positive predictive value is 1 whenever all edges of M_{NEM} are also part of M_{true}. This measure rewards models that retrieve only correct edges, without regard to the accuracy of negative predictions, which are less helpful in guiding laboratory experiments (Myers *et al*., 2006).

It has to be stressed that the size of data sets we use is very small compared to the data set sizes in other simulation studies (e.g. Basso *et al*., 2005; Hartemink, 2005), where performance on networks of 20 genes is evaluated on hundreds or thousands of measurements. Performing five replicate measurements for each gene perturbation is the practical upper limit in almost all real-world studies. Our evaluation focuses on an amount

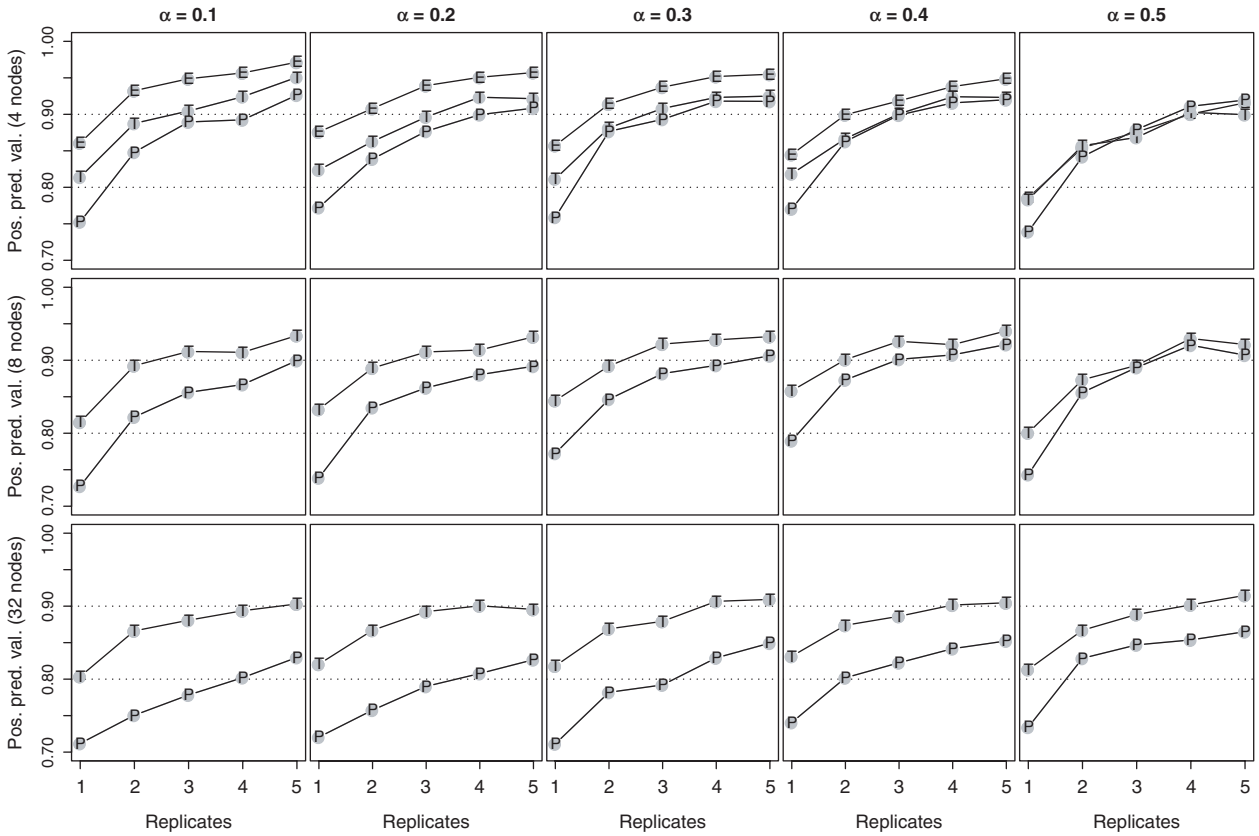

Fig. 4. Results of simulations. Rows correspond to the number of perturbed genes in the simulated data sets (4, 8, 32), while columns represent different levels of noise ($\alpha_{\text{data}} = 0.1, \ldots, 0.5$). In each plot, the y-axis measures the positive predictive value (the number of correct edges in the inferred graph), while the x-axis ranges over different numbers of replicates. The lines in the plot correspond to the different inference methods: 'E' for exhaustive enumeration, 'P' for pairwise inference and 'T' for inference from triples. Exhaustive enumeration is not possible for more than six genes, thus the lower two rows only compare pairs with triples. The simulation results show that inference from triples beats pairwise inference. All methods stay robust to changes in model size and levels of noise.

of data that can actually be achieved in real experimental studies.

Simulation results. The mean results of 250 simulation runs are shown in Figure 4. These plots show:

(1) Our methods are precise: the fraction of correct edges reaches 90% and above for all noise levels and model sizes.

(2) Our methods are robust: the fraction of correct edges stays stable with increasing noise and only slowly decreases with increasing model size.

Inference from triples always beats inference from pairs and is close to exhaustive enumeration on the small data sets. Inference from pairs is the quickest method, but suffers from the independence assumption imposed on edge existence. Inference from triples is slower than inference from pairs, but proves to be more reliable and is still feasible for graphs of the size of complete pathways. Overall these results show that NEMs can be constructed efficiently and accurately over a wide range of model sizes and noise levels.

4 RESULTS ON EXPERIMENTAL DATA

We show the practical use of the methodology developed in Section 2 in two experimental scenarios: (1) a data set investigating the response to microbial challenge in *Drosophila melanogaster*, and (2) a compendium of expression profiles of *Saccharomyces cerevisiae* knockout strains. We show that our method identifies biologically justified genetic hierarchies of perturbation effects.

4.1 Immune response in *Drosophila*

As a first proof-of-principle example on real data, we apply our method to data from an RNAi study on innate immune response in *Drosophila* (Boutros *et al.*, 2002). The experiment probes how transcriptional response to lipopolysaccharides (LPS) is regulated by signal transduction pathways in the cell.

Data. The data set consists of 16 Affymetrix microarrays: four replicates of control experiments without LPS and without RNAi (negative controls) four replicates of expression profiling after stimulation with LPS but without RNAi (positive controls) and two replicates each of expression profiling after applying LPS and silencing one of the four candidate

genes *tak*, *key*, *rel* and *mkk4/hep*. Selectively removing one of these signaling components blocks induction of all, or only parts, of the transcriptional response to LPS. Boutros *et al.* (2002) show that this observation can be explained by a fork in a signaling pathway below *tak*, with a *key* and *rel* on the one side and *mkk4/hep* on the other. This result clarified the contributions of different pathways to immune response in *Drosophila* (Royet *et al.*, 2005).

Estimating effects. Data preprocessing and discretization were performed as in (Markowetz *et al.*, 2005): if by silencing a gene in the LPS stimulated cell, the expression of an LPS-inducible gene moved close to its expression in the negative controls, this was counted as an *effect* of the intervention; if a gene's expression stayed close to its expression in the positive controls, the gene was counted as being *not affected* by the intervention. From the positive and negative controls, it is possible to estimate the two error rates as $\alpha = 0.15$ and $\beta = 0.05$.

Results and stability analysis. The small number of silenced genes allows us to compare our novel pairs- and triples-based methodologies to exhaustive enumeration. Figure 5 gives an overview of the results. All three methods succeed in recovering the true pathway structure. To show that our methods are robust to changes in the model parameters, we varied α and β from 0.05 to 0.95 and assessed the precision of our methods as in the simulation studies. The results are shown in Figure 6 and indicate a wide range of parameter combinations that succeed to perfectly reconstruct the true pathway structure.

Experimental design. What makes the data set of Boutros *et al.* (2002) especially well suited for our analysis are two main features of the experimental setup: first, the study is targeted towards a specific pathway. Stimulation by LPS turns the target pathway on, and breaking the signal flow by gene silencing allows conclusions about the pathway structure. Second, the study contains two kinds of control measurements, which makes it possible to compare the expression profiles after gene silencing to expression profiles of both the unstimulated and the stimulated cell. This experimental design allows us to estimate informative effects of interventions, which is important since NEMs crucially depend on a meaningful definition of intervention effects. As more and more gene perturbation studies focusing on a specific pathway of interest become available, we believe that the typical application of NEMs will be to pathway-wide data sets of around 50 genes.

4.2 Compendium of yeast knockouts

As a second experimental data set we chose a gene expression compendium of yeast gene knockout mutants (Hughes e*t al.*, 2000). The yeast compendium consists of 300 microarray measurements taken after perturbing yeast cells by either single gene knockouts, double gene knockouts or treatments with drugs and small compounds. It is one of the most frequently used data sets for computational studies in yeast functional genomics (Pe'er *et al.*, 2001; Rung *et al.*, 2002; Wagner, 2002; Yeang *et al.*, 2004).

In contrast to the *Drosophila* data set discussed above, the yeast compendium is not targeted towards a specific pathway but gives a broad overview over a wide range of yeast knockouts. The data set includes no replicate measurements

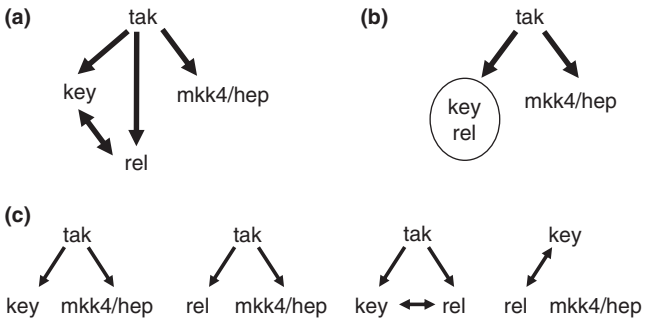

Fig. 5. Results on *Drosophila* data set. Plot (**a**) shows the inferred quasi-order, which is the same for all inference methods and agrees with the biological knowledge of this pathway. Plot (**b**) shows the result after merging two genes with undistinguishable profiles into a single node and removing shortcuts. Plot (**c**) enumerates the four triple models inferred, which all perfectly match the true structure.

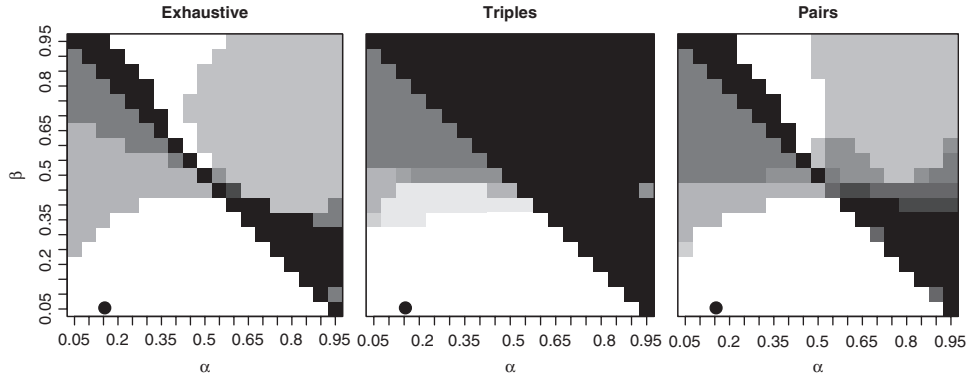

Fig. 6. Robustness against parameter choice. Each plot corresponds to one inference method. The ***x***-axis represents the parameter α, and the *y*-axis the parameter β. For each pair (α, β), we show the performance in reconstructing the true pathway structure on the *Drosophila* data. White corresponds to a positive predictive value of 1, and the darker a spot is, the more spurious edges were introduced. In all three plots, we see a wide white area of parameters for models containing only true edges. The parameter pair estimated from data is indicated by the black point and lies well within this area. Inference by triples produces very conservative results and returns an empty graph if (α, β) are both set unreasonably high.

and only one kind of wild-type measurements as controls. Additionally, knockouts may provoke complex reactions in the cell and result in intense compensatory efforts of the organism to an even greater extent than RNAi knockdowns. This results in a greater uncertainty in estimating specific effects of gene perturbations. However, even in this challenging situation we can show (1) that our general reasoning of building NEM models also applies to knockout data, and (2) that general features of the yeast transcriptional hierarchy can be reconstructed with NEMs.

Estimating effects. For each knockout the data set contains an expression profile of 6210 yeast genes, which shows the transcriptional response to each gene perturbation. The expression profiles are given as log ratios comparing the knockout strain to wild-type measurements. We applied a discretization procedure especially tailored to decrease the number of FP effects. For all knockouts, the log-ratio values show symmetric distributions, strongly concentrated at 0 (which corresponds to 'no change'). For each knockout, we count values more than 3 standard deviations from the mean as observing an effect and values within 3 standard deviations from the mean as observing no effect. This thresholding yields discretized phenotypic profiles, which can be used in our method.

Choosing NEM parameters. We chose a FP rate of $\alpha = 0.01$ (based on the fact that for a normal distribution the probability to fall outside of 3 SDs around the mean is $\sim 1\%$). We chose the FN rate as $\beta = 0.05$, which makes it five times as likely to miss an effect than to see a spurious one and thus accounts for efforts of the cell to compensate for gene loss.

Comparing double and single knockouts. Our motivation for using NEMs stems from the observation that perturbing key regulators may have an influence on a global process, while perturbing other genes may only affect more specific subprocesses. Thus, the difference between affecting global or specific processes should be visible as subset relations in the expression patterns. To test this assumption, we used the three gene pairs in the yeast compendium, for which we have expression profiles of both the single mutants and the double mutant: DIG1 and DIG2, FUS3 and KSS1, ISW1 and ISW2. The *Saccharomyces* Genome Database (www.yeastgenome.org) describes all three gene pairs as showing *phenotypic enhancement*, which means that the double knockout was found to show a more pronounced phenotype than any of the single knockouts. We fit NEMs to each of the three gene pairs (with one node corresponding to the double knockout and one node for each of the single knockout mutants) excluding all effect reporters that do not show an effect in any of the expression profiles. In all three cases we came to concurrent conclusions: the effects observed in the double knockout were found to be a (noisy) superset of the effects for the single mutants. The results are independent of the inference method we used and are robust over a wide range of model parameters. This result encouraged us to try our method on a larger scale and reconstruct global features of the regulatory organization of yeast.

Hierarchical structure of the yeast regulatory network. In a recent study, Yu and Gerstein (2006) show that the regulatory network of TFs in yeast can be organized as a four-layered (generalized) hierarchy, with most TFs at the bottom levels and only a few master TFs on top. This hierarchy is completely built from TF-DNA binding data and does not incorporate information from gene expression and knockout data, from which we build NEM models. In the following, we use the hierarchy of Yu and Gerstein (2006) as an independent test bed for our general assumption that the subset pattern of observed effects in expression profiles shows whether a TF has a global or specific function.

For 37 TFs in the hierarchy of Yu and Gerstein (2006), we also find expression profiles of knockout mutants in the yeast compendium of Hughes *et al*. (2000). These TFs include examples of three of the four levels in the hierarchy of Yu and Gerstein (2006). In the expression profiles of the 37 TFs, we exclude genes that are not affected in at least five knockout experiments (our results are robust to changes in this number), overall reducing the number of effect reporters to 100 . This matches the original analysis in Hughes *et al*. (2000), where only few significantly affected genes were found for most knockouts.

From this data, we inferred NEMs using both the pairs and triples approaches with the same parameters as above. We removed shortcuts by computing the transitive reduction of the NEM graphs. To be able to compare our results to those of Yu and Gerstein (2006), we then performed the algorithm they propose to organize the graph into several layers. For all pairs of TFs (x, y), we assessed how well the relationship 'x is on the same or a higher level than y' agrees between the hierarchy of Yu and Gerstein (2006) and our hierarchies inferred from knockout data. For the pairwise approach, we found 338 true positives and 93 true negatives with 132 FNs and 103 FPs. For the triple approach, we got slightly better numbers: 344 true positives and 99 true negatives with 126 FNs and 97 FPs. A chi-squared test for statistical independence between our hierarchies, and the one of Yu and Gerstein (2006) rejects the null-hypothesis at P-values of $1.5 \cdot 10^{-6}$ for the pairwise approach and $3.8 \cdot 10^{-9}$ for the triple approach. This shows that our hierarchies built from expression profiles of TF knockout mutants, and the hierarchy built from TF-DNA binding data by Yu and Gerstein (2006) correspond remarkably well.

5 DISCUSSION

We introduced a Bayesian method to approach two problems central to the analysis of large-scale and high-dimensional phenotyping screens: (1) that real effects are more likely to be missed than spurious effects are to appear, and (2) to recover features of the regulatory hierarchy of the cell. Our proposed method, NEM, estimates a quasi-order on the set of perturbed genes by combining probabilistic modeling with graph-based algorithms. In a simulation study, we show that NEM can be inferred accurately by building them from smaller sub-models. On real data sets, we show that the results actually reflect the functions of the genes involved. The methodology introduced in this article significantly increases the applicability of NEM, which were so far limited to small data sets containing < 6 perturbed genes.

Key to our method is inferring a non-symmetric relation between genes instead of symmetric gene relations as it is done in similarity-based clustering. In this sense, our methodology is related to asymmetric distance measures used in graph-

based clustering to identify protein complexes (Pipenbacher *et al*., 2002).

We introduced methodology to build large models from small building blocks and decided against alternative ways of inference like local search methods as hill-climbing or simulated annealing, which are e.g. used to learn Bayesian networks (Acid and de Campos, 2003; Friedman *et al*., 1999). Applying such approaches to our setting is complicated by the transitivity requirement inherent in subset relations. Even a small change in the model—like removing or adding an edge—can make many more changes necessary to preserve transitivity. The scoring function will be quite volatile and the score landscape will not be smooth. Building a model from small submodels avoids this problem and is robust to parameter changes and noise.

While clusters in the data can also be identified by a similarity-based method, our approach is the first to unravel the hierarchy of gene function based on global and specific effects of gene perturbations. Our method is exploratory, and we believe that it provides a good starting point for a more detailed analysis. Ultimately, NEMs have to be combined with other models and data sources in an integrated approach to uncover details of gene regulation.

There are several potential extensions to our approach. Currently, the method is only developed for binary effects and treats effect reporters as independent random variables. However, defining subset relations for quantitative data and capturing dependencies between observed effects could help to improve our method.

We believe that the analysis of large-scale and high-dimensional phenotyping screens will be a powerful way to infer regulatory hierarchies. NEM allow analysis of whole pathways and reconstruction of the information flow from the observed effects of interventions. The method we proposed here is a first step into a relatively new area of research that can profit from additional computational and statistical modeling.

ACKNOWLEDGEMENTS

We thank Edo Airoldi, Matthew Hibbs and Curtis Huttenhower (all LSI Princeton) for comments and helpful discussions. Tim Beissbarth (DKFZ Heidelberg) and Claudio Lottaz (MPI-MG Berlin) helped to create the 'nem'R-package.

F.M. is supported by NIH grant R01 GM071966 and NSF grant IIS-0513552 to OGT. This research was partly supported by NIGMS Center of Excellence grant P50 GM071508 and by NSF grant DBI-0546275. R.S. was supported by the *Bayerisches Genomforschungsnetz (BayGene)*.

Conflict of Interest: none declared.

REFERENCES

Acid,S and de Campos,L.M. (2003) Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *J. Artifi. Intell. Res*., **18**, 445–490.

Basso,K. *et al*. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet*., **37**, 382–390.

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

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

Brown,J.A. *et al*. (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol. Syst. Biol*., **2**, 2006.0001.

Cormen,T.H. *et al*. (2002) *Introduction to Algorithms*. 2 edn. McGraw-Hill.

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

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

Friedman,N. *et al*. (1999) Learning Bayesian network structures from massive data sets: the sparse candidate algorithm. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Stockholm, Sweden, pp. 206–215.

Gesellchen,V. *et al*. (2005) An RNA interference screen identifies Inhibitor of Apoptosis Protein 2 as a regulator of innate immune signalling in Drosophila. *EMBO Rep*., **6**, 979–984.

Gunsalus,K.C. *et al*. (2004) RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res*., **32**, D406–D410.

Hartemink,A.J. (2005) Reverse engineering gene regulatory networks. *Nat. Biotechnol*., **23**, 554–555.

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

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

Markowetz,F. *et al*. (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, **21**, 4026–4032.

Myers,C.L. *et al*. (2006) Finding function: evaluation methods for functional genomic data. *BMC Genomics*, **7**, 187.

Ohya,Y. *et al*. (2005) High-dimensional and large-scale phenotyping of yeast mutants. *Proc. Natl Acad. Sci. USA*, **102**, 19015–19020.

Pe'er,D. *et al*. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17** (Suppl. 1), S215–S224.

Piano,F. *et al*. (2002) Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. *Curr. Biol*., **12**, 1959–1964.

Pipenbacher,P. *et al*. (2002) Proclust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics*, **18** (Suppl. 2), S182–S191.

Raamsdonk,L.M. *et al*. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol*., **19**, 45–50.

Royet,J. *et al*. (2005) Sensing and signaling during infection in drosophila. *Curr. Opin. Immunol*., **17**, 11–17.

Rung,J. *et al*. (2002) Building and analysing genome-wide gene disruption networks. *Bioinformatics*, **18** (Suppl. 2), 202S–210S.

Sachs,K. *et al*. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Sloane,N.J.A. (2007) The on-line encyclopedia of integer sequence. Available at http://www.research.att.com/njas/sequences/

Strimmer,K. and von Haeseler,A. (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol*., **13**, 964–969.

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

Wagner,A. (2002) Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Res*., **12**, 309–315.

Werhli,A.V. *et al*. (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, **22**, 2523–2531.

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

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

Yu,H. and Gerstein,M. (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl Acad. Sci. USA*, **103**, 14724–14731.