Biological feature validation of estimated gene interaction networks from microarray data: a case study on MYC in lymphomas

Ulrich Mansmann and Vindi Jurinovic

Submitted: 31st August 2010; Received (in revised form): 7th February 2011

Abstract

Gene expression is a dynamic process where thousands of components interact dynamically in a complex way. A major goal in systems biology/medicine is to reconstruct the network of components from microarray data. Here, we address two key aspects of network reconstruction: (i) ergodicity supports the interpretation of the measured data as time averages and (ii) confounding is an important aspect of network reconstruction. To elucidate these aspects, we explore a data set of 214 lymphoma patients with translocated or normal MYC gene. MYC (c-Myc) translocations to immunoglobulin heavy-chain (IGH) or light-chain (IGK, IGL) loci lead to c-Myc overexpression and are widely believed to be the crucial initiating oncogenic events. There is a rich body of knowledge on the biological implications of the different translocations. In the context of these data, the article reflects the relationship between the biological knowledge and the results of formal statistical estimates of gene interaction networks. The article identifies key steps to provide a trustworthy biological feature validation: (i) analysing a medium-sized network as a subnet of a more extensive environment to avoid bias by confounding, (ii) the use of external data to demonstrate the stability and reproducibility of the derived structures, (iii) a systematic literature review on the relevant issue, (iv) use of structured knowledge from databases to support the derived findings and (v) a strategy for biological experiments derived from the findings in steps (i–iv).

Keywords: interaction networks; gene expression; comparing networks; biological knowledge

INTRODUCTION

The inference of gene interaction networks from microarray data has gained a high interest. The inferred networks are considered a first step towards the elucidation of biologically meaningful relationships within a complex system.

Prototypical examples for mechanisms of interaction between two genes (G1, G2) are: (i) both genes are regulated by the same transcription factor, which binds to common promoter elements (see example in ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’ section); (ii) gene G1 encodes a transcription factor which directly regulates gene G2; (iii) gene G1 encodes a transcription factor which directly regulates an intermediate gene which encodes a transcription factor that subsequently regulates gene G2; and (iv) gene G1 encodes a kinase protein which phosphorylates the transcription factor encoded by an intermediate gene, and modulates its ability to activate transcription of gene G2.

The statistical community developed a series of Bayesian and non-Bayesian inference procedures for gene interaction networks (see for example: [1–7]).

Corresponding author. Ulrich Mansmann, Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universitaet, Muenchen. Tel: +49 89 7095 4491; Fax: +49 89 7094 7491; E-mail: mansmann@ibe.med.uni-muenchen.de

Ulrich Mansmann is a Senior Biostatistician working in clinical oncology with focus on prognosis and translational research. He is a trained mathematical physicist and for 15 years active in clinical research.

Vindi Jurinovic is a PhD student and trained mathematician. Her main interest is computational molecular medicine with focus on interaction networks and prognostic gene signatures.
Their outcomes are in general drawings of graphs, where nodes represent the genes and the edges the direct ‘interaction’ of relevance. Often, these inference procedures define interaction in terms of conditional stochastic dependence, partial correlation or related concepts combined with mechanisms to impose sparseness on the connectivity between nodes. In principle, they infer direct interaction as described by scenario (i), (ii) and (iv). An edge between $G_1$ and $G_2$ may be inferred in case of scenario (iii) if the intermediate genes are not part of the network.

These algorithms, as well as the biological processes they try to describe, are complex and it is not obvious how close a generated graph is to the biological structure represented by the data. Statistical validation strategies are used to clarify this question. They rely on statistical simulation techniques combined with concepts such as true positive rate (TPR; rate of true edges which are detected by the algorithm), false detection rate (FDR; rate of detected edges which do not exist in the graph), and the algorithm $q$-power (probability to detect at least a $q \times 100$ percentage of edges in the graph). Simulations either use formal models which are detached from the biological structures behind the data or perturbation techniques for existing data [8]. We call this validation internal because it is fully based on statistical models and on quality measures describing the fit of the estimate to the relevant model.

When an inference procedure is applied to biological data and a network is inferred, this network should sufficiently represent the real biological phenomenon of interest. The real biological phenomenon as described by the four scenarios above can be evident in gene expression profile data for systematic experimental perturbation. A successful substitution of these laborious gene inactivation experiments by expression profiles from observational studies is desirable but not easy to achieve. Biologists are uncertain of the status of such networks proposed in the literature: under what conditions are they valid and what characteristics can be considered trustworthy [9]? A successfully internally validated algorithm does not guarantee its usefulness for biological applications. In this article, we discuss strategies to extend the common statistical validation programme towards issues motivated by the biological background.

An instructive example on how algorithmic results are biologically validated is provided by Basso et al. [10]. The authors derive a hierarchical scale-free network from expression profiles of human B cells. They perform its validation against data from the literature, bioinformatics databases as well as by specific biochemical experiments. The internal validation of their algorithm as well as a series of biological validation steps is used to conclude: such high precision is a particularly desirable feature of a reverse engineering algorithm, suggesting that the method could be applied to identify candidate interactions without the need for an extensive set of biological validations, which are time- and resource-consuming.

Gene expression microarray data try to capture essential aspects of the cell’s transcriptional activities. But, the data have to be interpreted with care.

First, they represent steady-state measurements of genes. The data do not capture aspects of the dynamic responses of genes. For example, steady-state data are not informative on the specific strong coupling of two genes by a strict time delay $\delta$: $X_1(t) = X_2(t + \delta)$. The time averages over the dynamics of both genes are (partially) correlated but without information on the interaction’s specific character.

Second, networks are often presented as units of interacting components detached from a wider environment which—for an analysis at hand—is assumed to have no direct influence on the structure of the inferred network. The exclusion of confounding inside the network by components outside the network seems to be a general working assumption in analyses which use the proposed inference procedures.

Third, with steady-state and confounding as implications for the analysis, the meaning of an inferred network may not be easy to formulate. Therefore, it may be helpful to base interpretations on the comparison of networks under two conditions and to look at differential interactions as a starting point for a valid and trustworthy interpretation of the inferred networks. While a single graph represents complex information, the comparison of two graphs can focus on the parts of the network which underlie structural changes between two conditions. Thus, the comparison has the potential to remove extensive parts of a large network with no functional relevance for the phenomenon of interest and focuses the strategy for the validation of biological features described by the networks.

We study network structures by looking at the interaction between the MYC gene and the Wnt pathway. There is an extensive literature about aberrant activation of the Wnt signalling pathway,
its association with numerous human cancers, and its correlation with the overexpression or amplification of the c-MYC oncogene. The MYC translocation involving immunoglobulin (IG) loci is associated with MYC up-regulation. It is of interest how this may influence the Wnt signaling network and consequently the control of cell–cell communication. Recent data reveal that multiple extracellular, cytoplasmic and nuclear regulators intricately modulate Wnt signalling levels. This also makes the Wnt pathway a perfect candidate to study confounding.

We will follow these issues formally in ‘Methodological review’ section and study them by analysing MYC translocation as well as the Wnt pathway in a lymphoma data set [11]. ‘Methodological review’ section reviews the relevant statistical methodology. Section ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’ presents a first example studying interaction of type 1. ‘Biological feature validation in estimated gene interaction networks’ section formulates principles of biological feature validation. ‘Differential gene interaction around the MYC gene in lymphoma patients section’ formulates interesting facts about the MYC gene and presents the results from the lymphoma data set. The discussion of our ideas is given in ‘Discussion’ section.

METHODOLOGICAL REVIEW

The basic model

We adopt the modelling approach from Margolin and Califano [12]. The dynamics of a gene’s abundance at time \( t \), \( X_i(t) \), is formulated as a function of \( \mathbf{N}_{-i}(t) \), the abundance of all other measured genes in the network at time \( t \). For simplicity, the time index is removed and \( X_i(t) \), the dynamics of the abundance of gene \( i \) at time \( t \) is modelled by

\[
\frac{\text{d}X_i}{\text{d}t} = f_i(\mathbf{N}_{-i}) \text{ or in vector notation } \mathbf{\dot{X}} = f(\mathbf{X}).
\]

If the system is measured at steady-state, then \( 0 = f(\mathbf{X}) \). This equation can be further broken down into a part that describes the influence of factors that affect the rates of transcriptions (\( f_{\text{trans},i} \)) and degradation (\( f_{\text{degr},i} \)) of \( X_i \). If we assume that \( f_{\text{degr},i} \) is a function of \( X_i \) alone, it follows

\[
0 = f_{\text{trans},i}(\mathbf{N}_{-i}) - f_{\text{degr},i}(\mathbf{N}_{-i}) \text{ or } X_i = f_{\text{degr},i}^{-1} \circ f_{\text{trans},i}(\mathbf{N}_{-i}) = g(\mathbf{N}_{-i}).
\]

We assume that the function \( g_i \) is linear and use the following vector notation for the entire system with the matrix representation \( M_g \):

\[
\mathbf{N} = g(\mathbf{N}) = M_g \cdot \mathbf{N}.
\]

This idea is studied by Meinshausen and Bühlmann [13] and uses regularized linear regression techniques to estimate the sparse matrix \( M_g \).

Assuming a small random perturbation of \( X_i \) given the values of \( \mathbf{N}_{-i} \) denoted by \( \mathbf{\varepsilon} \sim N(0, \Sigma_\varepsilon) \) iid, with vector notation \( \mathbf{\varepsilon} \sim N(0, \Sigma_\varepsilon) \) it holds

\[
\mathbf{N} = M_g \cdot \mathbf{N} + \mathbf{\varepsilon} \text{ or } (1 - M_g) \cdot \mathbf{N} = \mathbf{\varepsilon} \text{ or } \text{Var}[1 - M_g] \cdot \mathbf{N} = \Sigma_\varepsilon.
\]

From here, it is possible to describe the precision matrix \( \Omega_{\mathbf{N}} \) (inverse of the covariance matrix \( \Sigma_{\mathbf{N}} \)) of \( \mathbf{N} \):

\[
\Omega_{\mathbf{N}} = \Sigma_{\mathbf{N}}^{-1} = [1 - M_g] \cdot \Sigma_\varepsilon^{-1} \cdot [1 - M_g]^T.
\]

Therefore, the estimation of the covariance or precision matrix of \( \mathbf{N} \) allows the inference of the stationary linear law which defines the dynamics of gene transcription.

This idea motivated a series of methodological developments on the estimation of covariance or precision matrices [1–3] for high-dimensional linear Gaussian systems.

Graphical models and Bayesian networks

Consequently, the steady-state of a transcription network is described by a multivariate joint probability measure (JPM) which is often represented by a probabilistic graphical model (PGM). A PGM represents a JPM via vertices, representing random variables and edges representing the stochastic dependency between them. In case of multivariate Gaussian models, an informative graph can be derived from the precision matrix \( \Omega_{\mathbf{N}} \), by drawing an edge between two nodes \( X_i \) and \( X_j \) if the entry \((i,j)\) or \((j,i)\) in \( \Omega_{\mathbf{N}} \) is different from zero.

Bayesian networks (BN) are proposed by Pearl [14] to implement PGMs. BNs have been widely employed in the analysis of cellular networks [7]. They represent JPMs as a directed acyclic graph (DAG), which encodes the Markov assumption that a node is conditionally independent from its non-descendants, given its parents.

When ignoring the directions of a DAG, we get the skeleton of a DAG. Estimation of a DAG from data is difficult and computationally non-trivial: the number of possible DAGs grows super-exponentially.
with the number of nodes. Nevertheless, there are quite successful search-and-score methods for problems where the number of nodes is small or moderate. An interesting alternative to existing estimation approaches is the PC-algorithm from Spirtes et al. [4]. It starts from a complete, undirected graph and deletes recursively edges based on conditional independence decisions. This yields an undirected graph which represents a subset of non-zero off-diagonal elements of $\Omega_R$. The reason is that the entry $(i,j)$ of $\Omega_R$ represents conditional dependence given the set of all other nodes, while an edge derived from the PC algorithm represents conditional dependence given all subsets of the other nodes.

In the following, we adopt the sparse multivariate Gaussian setting and infer the skeleton of the BN by the PC algorithm.

**Confounding**

The analysis described relies on the assumption that a microarray data set measures values of all interaction variables in the network. The network is assumed to be completely observed. The biological interpretation of this assumption is that gene regulation is controlled entirely at the transcriptional level. But there is a vast network of hidden variables while the observed microarray data only probe a small portion of this network. Margolin and Califano [12] give a thorough discussion on this issue.

In the framework of linear models and multivariate normally distributed observations, the effect of hidden variables on the estimation of the dependency structure (encoded by the precision matrix $\Omega$) can be quantified. The precision matrix $\Omega$ can be partitioned in parts $\Omega_O$ and $\Omega_H$ which describe the conditional correlation structure within the observed part (O) and within the hidden part (H) of the system. The submatrices $\Omega_{OH}$ and $\Omega_{HO} (= \Omega_{OH}^T)$ describe the conditional correlation structure between the observed and unobserved components of the system:

$$\Omega = \begin{pmatrix} \Omega_O & \Omega_{OH} \\ \Omega_{HO} & \Omega_H \end{pmatrix}.$$

The precision matrix of the marginal distribution of the observed components O is given by

$$\Omega_{O}^{\text{marg}} = \Omega_O + \Omega_{OH} \cdot \Omega_H^{-1} \cdot \Omega_{HO}$$

which quantifies the amount of confounding by hidden variables on the observed variables with respect to the conditional dependence structure.

We study possible confounding by imbedding the set of nodes into a larger node set and determining the graph of interest as a subgraph of the larger network. Effects of confounding are visualized as the differences between the graph derived on the set of interesting nodes and the subgraph related to the set of interesting nodes in a larger network.

**Comparing graphs**

While a single graph represents complex information, the comparison of two graphs may focus on the parts of the network which underlie structural changes between two conditions. Possibly, the comparison can remove extensive parts of a large network with no functional relevance for the phenomenon of interest.

Several authors propose different approaches to compare graphs. Gill et al. [15] provided a strategy for conducting a differential analysis of connectivity that represents the strength of genetic association or interaction between two genes. They provide a connectivity score and propose formal statistical tests: (i) whether the overall modular structures of the two networks are different, (ii) whether the connectivity of a particular set of ‘interesting genes’ has changed between the two networks and (iii) whether the connectivity of a given single gene has changed between the two networks.

Kosta and Spang [16] addressed the problem of detecting sets of genes which are differentially coexpressed between two phenotypically different groups of expression profiles. They propose a score for measuring differential coexpression based on an ANOVA model and describe a heuristic algorithm for finding high-scoring gene sets. While Gill et al. analyse a given structure and provide strategies for inference, Kostka and Spang search for gene sets with differential co-expression in a hypothesis generating sense. Choi et al. [17] study two phenotypical subtypes and calculate the Pearson correlation coefficient for every pair of two genes within each subdataset. If the genes $i$ and $j$ are connected within both subtypes simultaneously, they are connected by a conserved link. If a subtype link is not a conserved link, it can be called a subtype-specific link. A statistical assessment of the number of conserved links is the basis of their approach. Gillis and Pavlidis [18] analyse changes in coexpression across multiple ordered groups (e.g. over time) and extensively test its validity and usefulness. Their method uses the Haar-wavelets in order to efficiently...
represent changes in coexpression at multiple time scales.

We propose two further strategies. First, we base our comparison of graphs $G_1$ and $G_2$ (or their adjacency matrices) over the same node set between two biological conditions on the normalized structural Hamming distance (nSHD). The SHD between two graphs is the number of edge insertions and deletions in order to transform one graph into the other or the total number of different edges: $SHD = \# \text{edges in } G_1 + \# \text{edges in } G_2 - 2x\#\text{common edges}$. Since the SHD largely depends on connectivity indices of the nodes in the given set, as well as the score for the entire graph.

The connectivity index for a set of nodes is the sum of the connectivity score for a given node. The score is defined as $\frac{\text{degree}}{C_0}$. It is also possible to calculate the connectivity index as follows:

1. The matrix representation of $f_{\text{trans}}(M_{\text{trans}})$ is simple, since $\text{IGH@}$ and MYC transcription depend on the common transcription factor $T$ which is indirectly regulated by the MYC expression.
2. The matrix representation of $f_{\text{deg}}(M_{\text{deg}})$ is a diagonal matrix containing the degradation factors $d_1, d_2$ and $d_3$. 

$$M_{\text{trans}} = \begin{pmatrix} 0 & 0 & a \\ b & 0 & 0 \\ c & 0 & 0 \end{pmatrix} ; \quad M_{\text{deg}} = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}.$$ 

$$M_{g} = \begin{pmatrix} 0 & 0 & a \cdot d_1^{-1} \\ b \cdot d_2^{-1} & 0 & 0 \\ c \cdot d_3^{-1} & 0 & 0 \end{pmatrix}.$$ 

ANALYSING INTERACTION OF MYC AND THE IG GROUP IN WILD-TYPE LYMPHOMA CELLS AND CELLS WITH TRANSLOCATED MYC GENE

The MYC gene produces a transcription factor c-Myc that controls cellular proliferation, programmed cell death and differentiation. In B-cell lymphomas, common translocations of the MYC gene at 8q24 include $t$(8;14) (q24;q32) (translocation partner $\text{IGH@}$), and less often $t$(2;8) (p12;q24) ($\text{IGK@}$) and $t$(8;22) (q24;q11) ($\text{IGL@}$) [20]. Via these translocations, the MYC gene is juxtaposed with the $\text{IGH@}$ gene on the derivative chromosome 14, or the immunoglobulin light chain genes are juxtaposed with MYC on the derivative chromosome 8. The IG group therefore consists of three loci $\text{IGH@}$ (14q32.33), $\text{IGK@}$ (2p12) and $\text{IGL@}$ (22q11).

We used data from Hummel et al. [11] and selected samples from 214 patients: 59 samples with a MYC-IG translocation and 155 patients with MYC wild-type. For each of these patients, the study provides gene expression profiling using Affymetrix U133A GeneChips with RNA from mature aggressive B-cell lymphomas. Affymetrix HGU133A raw data were normalized (RMA) and statistical calculations were done using Bioconductor and R software.

Since the $t$(8;14) (q24;q32) translocation constitutes $\sim$80% of MYC-IG translocations [20], our translocated sample can be considered to be mainly of this translocation type. There is a direct interaction (prototype 1 in introduction) between MYC and $\text{IGH@}$ in MYC-IG samples: both genes share parts of the same promoter. $\text{IGH@}$ is constantly activated by its promoter in normal cells, which causes MYC overexpression in translocated cells.

This statement can be translated into the language of a dynamic process. The vector $X(t)$ is given by $[X_1(t), X_2(t), X_3(t)] = [T(t), \text{IGH@}(t), \text{MYC}(t)]$ where $T$ represents the hidden (not measured) transcription factor. The derivation of the matrix $M_g$ is straightforward:

$$M_{\text{trans}} = \begin{pmatrix} 0 & 0 & a \\ b & 0 & 0 \\ c & 0 & 0 \end{pmatrix} ; \quad M_{\text{deg}} = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}.$$ 

$$M_{g} = \begin{pmatrix} 0 & 0 & a \cdot d_1^{-1} \\ b \cdot d_2^{-1} & 0 & 0 \\ c \cdot d_3^{-1} & 0 & 0 \end{pmatrix}.$$
Assuming independent random perturbations for each component (with constant standard deviations \( \sigma \) for each component as well as noting the non-zero elements of \( M_s \) by \( \alpha, \beta \) and \( \gamma \)) allows to derive the precision matrix of the three component system \( \Omega_N \):

\[
\Omega_N = \sigma^{-2} \begin{pmatrix}
1 + \alpha^2 & -\beta & -\gamma - \alpha \\
-\beta & 1 + \beta^2 & \beta \cdot \gamma \\
-\gamma - \alpha & \beta \cdot \gamma & 1 + \gamma^2
\end{pmatrix}.
\]

The marginal precision matrix of the observable part of the two gene system is

\[
\Omega_{\text{obs}} = \sigma^{-2} \begin{pmatrix}
1 + \beta^2 + \gamma^2 & \beta \left( x_{12} + x_{13}^2 \right) \\
\beta \left( x_{21} + x_{23}^2 \right) & 1 + \gamma^2 + \left( x_{31} + x_{32}^2 \right)
\end{pmatrix}.
\]

The described setting between MYC and IGH@ suggests negative values for parameter \( \alpha \) and positive values for \( \beta \) and \( \gamma \). There are three (non-linear) equations for four unknown parameters. For fixed \( \alpha \), we remove \( \sigma \) by studying the ratio between a diagonal element and the off diagonal element and derive a functional relationship between \( \beta \) and \( \gamma \). The parameter \( \sigma \) can be estimated after the determination of \( \beta \) and \( \gamma \) from one of the diagonal elements. The numerical solutions presented in Figure 1 ignore statistical uncertainty in the estimated precision matrix for the paired observation (MYC, IGH@) \( \Omega_{\text{Myc/IGH@}} = (\pi_{ij}) \) where \( \pi_{11} = 3.399 \), \( \pi_{12} = \pi_{21} = 7.325 \) and \( \pi_{22} = 69.774 \). This formal description of the dynamic process can be the starting point for specific biological experiments.

We continue the example with a typical statistical analysis of the interaction of four genes (MYC, IGH@, IGK@ and IGL@) under two biological conditions (wild-type, MYC translocation). The precision matrices of the multivariate Gaussian distribution for the four genes can easily be estimated under both conditions. It is also possible to calculate the \( P \)-values for the relevant partial correlations. The results are given in Table 1. The upper triangle represents the entries of the precision matrices; the lower triangle reports the \( P \)-values for partial correlation.

The wild-type samples show no statistical evidence for any direct interaction between the four selected genes (after Bonferroni adjustment for 6-fold testing). The samples with MYC-translocation show convincing direct interaction between MYC and IGH@ (as expected from biology) as well as convincing direct interaction between IGK@ and IGL@.

A more elaborate analysis using the PC algorithm compares the direct interaction graphs of the four genes between the wild-type and MYC-translocated samples (restricted analysis). Additionally, we embed the small four genes network into a larger system by adding the genes of the Wnt and Cell cycle pathway (extended analysis). The results are given in Figure 2.

In the restricted analysis, the PC algorithm indicates a direct interaction between MYC and IGH@ as well as IGK@ and IGL@ under both conditions. For wild-type samples, it additionally gives a direct interaction between IGH@ and IGL@.

The extended analysis shows a direct interaction between MYC and IGH@ only in the translocated samples. The direct interaction between IGK@ and IGL@ is present under both conditions.

This section started with a biological statement for a specific type of MYC-translocated cells. The statement was first translated in the language of dynamic processes which we introduced in 'Methodological review’ section. The resulting model was fitted to observed data. The estimated model parameters can be experimentally validated and conclusions about the quality of the proposed model can be drawn. Second, the qualitative estimates of a small network were compared between wild-type and

\[\text{Figure 1: Estimates for the parameters of the dynamic MYC-IGH@ model.}\]
Myc-translocated samples. The comparison was performed in a restricted as well as an extended setting. The biological implications derived from the analysis were only compliant with existing biological knowledge in the extended setting. Is the detected direct interaction between IGK@ and IGL@ of biological relevance? We could replicate the structure of the graph in the extended analysis for a series on breast cancer data sets (E-GEOD-6532, E-GEOD-4922, E-GEOD-11121, E-GEOD-7390, E-GEOD-12093, E-GEOD-2603, E-GEOD-5462, E-GEOD-9936, E-GEOD-5847, E-MTAB-7, E-GEOD-1561, E-TABM-43, E-MEXP-440, E-GEOD-11965, E-GEOD-6772, E-GEOD-9574, E-GEOD-4917, E-GEOD-3494, E-GEOD-2990). Papers in which the IGK and IGL locus are simultaneously discussed are sparse and do not elucidate the observed 'direct' interaction biologically ([21] and related articles).

### Table 1: Precision matrices (upper triangle) and P-values (lower triangle) for direct (linear) influence within the small MYC-IG gene network in MYC-translocated and wild-type samples

<table>
<thead>
<tr>
<th>MYC-IG translocated</th>
<th>Normal karyotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IGH@</td>
</tr>
<tr>
<td>IGH@</td>
<td>7.960</td>
</tr>
<tr>
<td>IGL@</td>
<td>0.7883</td>
</tr>
<tr>
<td>IGK@</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>MYC</td>
<td>0.7137</td>
</tr>
</tbody>
</table>

**Figure 2:** Direct interaction in the MYC-IG-network as inferred by the PC algorithm: left side under restricted analysis, right side under consideration of cell cycle as well as Wnt pathway genes. Black (red) edges: Direct interaction present for the translocated as well as the normal samples. Broken - right panel (green) edges: Direct interaction present only for the translocated samples. Dotted - left panel (purple) edges: Direct interaction present only for the normal samples. A colour version of this figure is available at http://bib.oxfordjournals.org.

### Biological feature validation in estimated gene interaction networks

Statistical validation of a network algorithm takes place in the framework of theoretical models. It quantifies model based true detection rate (TDR) or false detection rates (FDR) of edges as well as the positive (negative) predictive value of a given (missing) edge in the estimated graph. Biological feature validation relates the estimated structure to biological knowledge of the context which generated the data analysed. Thus, biological feature validation denoted the process which demonstrates to what extent a network estimated from gene expression data will lead to valid biological knowledge.

A borderland between both strategies is the handling of confounding as indicated in ‘Methodological review’ and ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’ sections: analysing a network between nodes of interest as a subnet within a larger network of interacting nodes. The specific choice of the larger environment of nodes may be based on biological reasoning or on a variety of larger neighbourhoods which allow to study the stability of the results derived in the subnet. To our knowledge, forms of extended analyses (as exemplified at the end of ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’ section) for small or medium sized networks (up to a few hundred genes) have not been applied in the literature.

It is our goal to draft a validation master plan (VMP) which sketches the biological feature validation programme executed for a study. The VMP has three parts: (i) using reports in the scientific literature, (ii) using available structured and curated biological knowledge and (iii) using results of specifically designed new experiments.
Using reports in the scientific literature

Using reports in the scientific literature links the specific statistical output to the results of already existing and published scientific work. This strategy stresses the relevance of the derived result in terms of producing subject matter evidence and is often used in papers presenting methodological innovations in bioinformatics. There, a new method is applied to a published subject matter example and the authors mostly argue that their new method allows a deeper insight into the issue than the former approach did. Furthermore, suitable available microarray data sets may be reanalysed under the specific focus of the researchers’ problem to reproduce his/her findings. Ideas as presented in ref. [8] may be helpful to study the stability of the results.

At the end of section ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’, we reported the shortage of literature on a direct IGK@ and IGL@ interaction. Consequently, we analysed a series of other microarray experiments to see if this direct interaction is present (using series with the Affymetrix U133A GeneChip and performing the same analysis as the extended analysis in ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’ section). This validates the reproducibility of the detected pattern in a larger set of biological samples. Reproducibility is a relevant target of each validation strategy.

In general, the selection of reports or data does not follow objective rules. The authors do not justify their choices or do not clarify how representative the selected example(s) is (are) to support the biological validity of the reported finding. To reproduce the direct interaction between IGK@ and IGL@ we should perform a meta-analysis or a systematic review of graphs on a representative series of biological samples. A systematic review is a literature review focused on a research question that tries to identify, appraise, select and synthesize all high-quality research evidence relevant to that question. But, this technique is still not well established in the field of Bioinformatics.

Using available structured and curated biological knowledge

The bioinformatics counterpart of the systematic review (as practised in evidence-based medicine) is the structured and curated biological knowledge. Such knowledge is updated and accessible in a series of public and commercial databases. It is available in formal structures which enable the researcher to use it in algorithmic procedures. Ontologies play an important role to represent knowledge besides databases on the DNA structure, genes, proteins, etc. [22]. Their concept is imported from computing science to describe different conceptual frameworks that guide the collection, organization and publication of biological data. An ontology is similar to a paradigm but has very strict implications for formatting and meaning in a computational context. The Gene Ontology (http://www.geneontology.org/) is a prominent example which influenced the development of knowledge-based statistical computing and inspired the development of gene set enrichment analysis (GSEA [23]). GSEA links prior biological knowledge to newly generated data and uncovers the biological features which conform to the observed data. GSEA is developed for differential gene expression analysis. But, it is also helpful to apply enrichment strategies to networks as demonstrated in Basso et al. [10]. Alternative strategies allow the combination of biological knowledge formalized in ontologies with multiple testing procedures to enhance the biological interpretation of a statistical analysis [24–26]. The transfer of such strategies from the field of differential gene expression to networks or differential interaction analysis between networks would create new formal instruments for biological feature validation strategies.

Transcription factors provide prototypical examples for mechanisms of interaction between two genes. Therefore, gene sets that contain genes that share a transcription factor binding site can be used to validate the computed interaction graph on a set of genes. This information is for example available for download under http://www.broadinstitute.org/gsea/msigdb/index.jsp and is helpful to impute the interaction Scenario 1 as described in Introduction section. Interaction Scenarios 2 and 3 link two genes by relating the coded transcription factor of one gene with the transcription factor binding site of the second gene or building chains of corresponding gene pairs.

Using results of specifically designed new experiments

This strategy is the real biological validation. Basso et al. [10] validate new MYC targets by ChiP assay to show direct binding of c-MYC to their promoter
regions in vivo. Chip-on-chip technology can be used to identify binding sites for DNA binding proteins on a genome-wide basis. This technology allows the identification of interaction scenarios of prototype 2 because for the transcription factor as a protein of interest, one can determine its transcription factor binding sites throughout the genome. The more elaborated interactions of prototype 3 or 4 need specifically designed gene inactivation (silencing) experiments to be detected. Different experiments are needed to explore valid parameter estimates proposed by the simple MYC-IGH@ interaction as described in ‘Analysing interaction of MYC and the IG group in wild-type lymphoma cells and cells with translocated MYC gene’ section. They may follow ideas as presented in ref. [27]. The design of the corresponding experimental programme relates on the findings derived from the data in the several analysis and validation steps.

Differential gene interaction around the MYC gene in lymphoma patients

The Wnt/β-catenin signalling pathway regulates cell proliferation, differentiation and stem cell renewal and is linked to cancer [28]. In the presence of Wnt-stimulation, β-catenin binds to transcription factors such as c-Myc and activates their expression. We want to find genes within the Wnt pathway which change their direct interaction pattern between wild-type cells and MYC-translocated cells.

Again, the data is taken from Hummel et al. [11] and consists of samples from 214 patients: 59 samples with MYC-IG translocation and 155 patients with MYC wild-type. The Wnt pathway is represented by 261 probe sets (representing 135 genes) on the Affymetrix HGU133A array.

The MYC-IG translocation causes overexpression of the MYC gene and overproduction of the transcription factor c-Myc. A straightforward differential gene expression analysis [29] detects the genes with differential transcription as shown in Figure 3. A detailed numerical report on the content of Figure 3 is given in the Supplementary Table A3.

The strong overexpression of the MYC gene in MYC-IG samples (represented by probe set 202431_s_at) is clearly visible as second high column in the left third of the plot. Furthermore, as expected, many genes of the Wnt pathway are up-regulated in the MYC-IG samples. There are 21 probe sets with a strong up-regulation (influence >200) representing 15 genes (CCND3, DVL1, SMAD3, MYC, NFACT3, PPP2R5C, PPP3CC, PRKACB, PRKCA, MAP3K7, TP53, RUVBL1, PLCB1, CACYPB, LEF1). A strong down-regulation (influence >200) in MYC-IG samples can be observed for 14 probe sets representing 8 genes (CCND2, CTBP2, JUN, MMP7, PSEN1, RAC2, TBL1X, FZD1).

Differential interaction structure in the Wnt pathway for MYC-IG samples compared to wild-type samples is shown in Figure 4. The graphs under both conditions are estimated by the PC algorithm [4].

The permutation-based test using the normalized SHD (described in 'Methodological review' section) assesses a significant difference between both graphs ($P = 0.020$). The figure does not contain the MYC gene. This indicates no strong evidence for differential interaction of MYC with other partners in the network comparing MYC-IG and wild-type samples.

The DDIS method (see ‘Methodological review’ section) identifies nine genes with differential interaction (nine probe sets representing nine genes have a $P$-value below 0.05: CAMK2G, CCND2, PPARD, RAC2, SFRP1, SKP1, TBL1X, CUL1 and FZD10). Adjustment for multiple testing (Bonferroni) results in no significant nodes. All genes detected on the unadjusted level by the DDIS method represent hubs in Figure 4.

Three of these genes (CCND2, RAC2, TBL1X) are contained in the set of strongly down regulated genes and no one is contained in the set of strongly up regulated genes.

Figure 5 illustrates the differences in estimates of the interaction structure between MYC-IG and wild-type samples when the Wnt pathway is considered as a subnet in a larger network (enlarged by genes from the cell cycle pathway, extended analysis). Figure 5 is based on the same node set as Figure 4.

Eleven genes interact with a partner gene in the restricted analysis (Figure 4) and lose their interactions in the extended analysis (Figure 5): DAAM1, TCF7, WNT1, RNUDL1, CCND3, LRP5, PRKX, WNT5B, TNND1, CER1 and CACYPB. This indicates indirect interactions where the interaction chain includes nodes of the extended setting (possible for the interaction prototypes 3 or 4).
**Figure 3:** Geneplot from globaltest for the Wnt pathway.

**Figure 4:** Differential interaction structure for the Wnt pathway. Graph 1: Interaction estimate for IG-translocated samples and Graph 2: Interaction estimate for samples with normal karyotype.
Focusing on SKP1, the following can be observed: (i) the restricted inference indicates interaction between SKP1 and WNT1 as well as SKP1 and RUVBL1 for MYC-IG samples which cannot be seen for wild-type samples; (ii) The restricted inference also shows interactions between SKP1 and CSNKIA, SKP1 and CER1, as well as SKP1 and CACYBP in wild-type samples which are not detected for MYC-IG samples; (iii) The extended inference gives interactions between SKP1 and FBXW1 as well as SKP1 and RAC1 in wild-type samples which are not shown in MYC-IG samples; and (iv) All other interactions are simultaneously present.

Figure 6 shows that a large part of differential interaction found in the restricted inference is conserved in the extended analysis. A total of 42 differential interactions are lost comparing the extended to the restricted setting. Eight differential interactions are solely detected under the extended inference.

To motivate the meaning of this finding one should remember that there are three-dimensional normal distributions for $(X,Y,Z)$ where $X$ is independent of $Y$ in the marginal distribution but $X$ is dependent on $Y$ given $Z$ [as pseudocode for the R language: $x \leftarrow \text{rnorm}(n, 0, 1)$; $y \leftarrow \text{rnorm}(n, 0, 1)$; $z \leftarrow \text{rnorm}(n, x+y, \sigma)$]. This formal model represents interaction prototype 2 where genes $X$ and $Y$ inside the subnet act independently from each other but do regulate gene $Z$ outside the subnet which belongs to the extended setting.

Figure 6 shows that the three Wnt pathway genes with differential gene expression as well as differential direct interaction structure (RAC2, TBL1X, CCND2) in the restricted analysis also represent differential interaction under the extended analysis.

Appendix Table A1 of Supplementary summarizes information on all genes with at least three interaction partners in the graph of Figure 5 (independent of the edge colour). The information is taken from the corresponding entries in the NCBI GENE database (http://www.ncbi.nlm.nih.gov/gene?term=). The interacting partners of each gene are given. It is also indicated if the interaction was observed in the wild-type (WT) or MYC-IG translocation sample (TL). Additionally the up arrows or down arrows indicate how a gene is regulated with respect to the WT measurements. The NCBI GENE database entry of a gene offers links to relevant papers as well as established interactions.

For example, the entry on the SKP1 gene states that SKP1 interacts with CUL1 to form part of the Cull1-Rbx1-Skp1-F Boxskp2 Scf ubiquitin ligase complex. This interaction is not present in our
But an interaction with the RBX1 gene is proposed in wild-type samples \[30\].

Wnt signals are transduced in at least two distinct ways; a well-established 'canonical' or Wnt/β-catenin pathway, and a non-canonical pathway which is β-catenin independent. The canonical Wnt pathway describes a series of events that occur when Wnt proteins bind to cell-surface receptors of the Frizzled family, causing the receptors to activate dishevelled family proteins and ultimately resulting in a change in the amount of β-catenin that reaches the nucleus. The non-canonical Wnt signalling pathways are less well understood. These non-canonical pathways have also been termed the Wnt/Calcium and Wnt/JNK. Briefly, activation of the Wnt/Calcium pathway involves Wnt binding to a Frizzled receptor, leading to release of intracellular calcium and the activation of enzymes such as CamKII and PKC. \[31\]. Some of the interesting genes are annotated to either type of pathway.

We also used GSEA genesets (http://www.broad institute.org/gsea/index.jsp) transcription factor targets (TFT)—gene sets (that contain genes that share a transcription factor binding site) to study if interactions given in Figure 5 are potentially from interaction prototype 1. Appendix Table A2 of Supplementary lists the pairs with common binding sites for transcription factors and the related factors. There is a total of 20 out of 51 pairs of gene interactions where the partners share common binding sites. Only one out of 20 (the pair TBL1X and ROC2) is interacting in wild-type as well as MYC-IG samples. There are 22 (9 sharing common binding sites) interacting pairs in MYC-IG samples and 28 (10 sharing common binding sites) interacting pairs in wild-type samples.

The above presentation shows that only a few specific features of the network presented in Figure 5 are discussed in the literature. Figure 5 also shows that genes switch interaction partners when a translocation is present. There are about 20 interactions
under either condition and about half of them consist between pairs which share common binding sites. The next step is the design of experiments which focus on the biological meaning of the derived graph. These experiments are under planning and their presentation is out of the range of this report.

**DISCUSSION**

There is an emerging consensus in the statistical community about the need to validate statements derived from complex statistical analysis processes. Two seminal contributions to this field are presented by Breiman [32] and Hjort [33]. Simulation studies, resampling/cross-validation (inner validation), and the use of independent data (external validation) are relevant strategies to benchmark the outcome of complex statistical analyses. All three approaches are also appropriate to see if the tools for network inference have good statistical properties: simulation strategies are helpful to quantify the detection rates for edges of a specific inference tool, resampling strategies help to explore the stability of networks under controlled variation, external data are useful to see if the network is reproducible in new material.

This report is concerned with strategies to assess the biological relevance of results derived from statistical network analyses. We focus on small- and medium-sized networks (up to several hundred nodes), which are used to disentangle the relevance of pathways in specific disease processes. The focus is not on the analysis of genome-wide gene interaction networks as presented by Basso et al. [10].

We see the extended analysis as a first step towards more reliable statements on networks. The extended analysis studies the pathway as a subnet in a larger environment. It has the potential to remove false positive edges caused by confounding. The extended analysis includes more intermediate partners which build a longer interaction chain between two genes (interaction prototypes 3 and 4). The extended analysis can also create a new edge for two genes inside the subnet if they regulate simultaneously a gene outside the subnet which belongs to the defined larger environment. Margolin and Califano [12] propose to use hidden variables to resolve this problem. Our recommendation for medium-sized networks is to use the available whole-genome information from a microarray and to choose appropriate larger environments. The properties of the algorithm used may define restrictions on the size of the larger environment.

It is also of interest to study the stability of a network in different biological conditions. We presented statistical test strategies to establish significant differences between networks over the same set of nodes. There are global tests for general statements on the entire networks over all nodes or subset of nodes. It is also possible to test if single nodes have a different interaction structure within the environment between biological conditions. Comparing networks opens a more focused view on those nodes where the interaction structure is strongly modified.

Accounting for confounding as well as differential interaction analysis are relevant steps to prepare the process of embedding the derived statistical results into the available body of biological knowledge.

The next step is to assess how the inferred network reflects biological knowledge provided in the literature. For the interesting nodes in the inferred network, we use summaries on literature or physical interaction as provided by the NCBI GENE database. We identified the need for more formal techniques which support the systematic review of published biological information. The representative selection of relevant literature is essential to assess the external validity of the main findings of the network analysis.

One way to strengthen the reproducibility of relevant aspects of the inferred network over a larger series of biologically different populations is to use available microarray data sets from repositories like ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) or Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). In our example, we could replicate the wild-type interaction pattern between MYC, IGH@, IGK@ and IGL@ in a large series of solid cancer data sets.

A further part of the biological feature validation is the use of more general structural knowledge as provided by many bioinformatics databases and ontologies. We used information on genes which share common binding sites and assessed the potential that derived interactions may be of interaction prototype 1 (both genes are regulated by the same transcription factor, which binds to common promoter elements).

In the field of differential gene expression, several methods are proposed which combine statistical reasoning with structured biological knowledge. Gene set enrichment analyses are a prominent example.
More refined approaches were developed [24, 25, 35]. For example, the focus level method [24] starts with a whole genome wide differential gene expression analysis and refines its resolution in a hierarchical way to detect the relevant gene sets of the Gene Ontology where differential gene expression is evident. It is an interesting methodological challenge to combine the comparison of network structures as proposed by refs. [15, 19] with a knowledge guided hierarchical testing procedure. This combination would provide a theoretical approach to zoom into a network and detect a set of nodes with relevant differential interaction structure.

We also propose the use of the elements presented so far to write a validation master plan for a complex network analysis before starting. The final goal of this plan should be a focused set of lab experiments to address the most relevant hypotheses derived from the network analysis.

As a counterpart to the more qualitative validation, we also presented a toy example for a quantitative biological validation. A biological hypothesis was transformed in a formal dynamic model and fitted to the data of a microarray experiment. The analysis gave quantitative functional relationships between the model parameters which can be used to design an experiment for a final assessment of the process. An interesting link between the qualitative and quantitative validation approach is presented by Chuan-Yun Li and coworkers [34].

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

better focus on ‘interesting genes’ (which change their interaction with the neighbours dramatically).

Acknowledgements
We thank Khadija ElAmrani for providing the software to visualize the comparison of graphs. We also thank three anonymous reviewers for their constructive and helpful comments.

FUNDING
German Ministry of Education and Research (BMBF) medical systems biology project HaematoSYS (BMBF-FKZ 03415452I to V.J.).

References