ABSTRACT

Motivation: With the exponential growth of expression and protein–protein interaction (PPI) data, the frontier of research in systems biology shifts more and more to the integrated analysis of these large datasets. Of particular interest is the identification of functional modules in PPI networks, sharing common cellular function beyond the scope of classical pathways, by means of detecting differentially expressed regions in PPI networks. This requires on the one hand an adequate scoring of the nodes in the network to be identified and on the other hand the availability of an effective algorithm to find the maximally scoring network regions. Various heuristic approaches have been proposed in the literature.

Results: Here we present the first exact solution for this problem, which is based on integer-linear programming and its connection to the well-known prize-collecting Steiner tree problem from Operations Research. Despite the NP-hardness of the underlying combinatorial problem, our method typically computes provably optimal subnetworks in large PPI networks in a few minutes. An essential ingredient of our approach is a scoring function defined on network nodes. We propose a new additive score with two desirable properties: (i) it is scalable by a statistically interpretable parameter and (ii) it allows a smooth integration of data from various sources.

We apply our method to a well-established lymphoma microarray dataset in combination with associated survival data and the large interaction network of HPRD to identify functional modules by computing optimal-scoring subnetworks. In particular, we find a functional interaction module associated with proliferation over-expressed in the aggressive ABC subtype as well as modules derived from non-malignant by-stander cells.

Availability: Our software is available freely for non-commercial purposes at http://www.planet-lisa.net.

Contact: tobias.mueller@biozentrum.uni-wuerzburg.de

1 INTRODUCTION

Construction and analysis of large biological networks have become major research topics in systems biology (Aittokallio and Schwikowski, 2006). Various aspects have been analyzed including the inference of cellular networks from gene expression (Friedman, 2004), network alignments (Flannick et al., 2006; Kelley et al., 2003; Sharan and Ideker, 2006) and other related strategies as reviewed by Srinivasan et al. (2007). At the same time, well-established microarray technologies provide a wealth of information on gene expression in various tissues and under diverse experimental conditions. Integrating protein–protein interaction (PPI) and gene-expression data generates a meaningful biological context in terms of functional association for differentially expressed genes.

Frequently, large scale expression profiling studies investigate many experimental conditions simultaneously, thereby generating multiple $P$-values. Especially in tumor biology expression profiling has become a well-established tool for the classification of different tumors and tumor subtypes. Furthermore, in the clinical context, various patient-associated data are available that—in conjunction with expression data—provide valuable information of the influence of specific genes on disease-specific pathophysiology. In particular the analysis of survival data allows to establish gene expression signatures to make predictions about the prognosis and to assess the disease relevance of certain genes. However, the cellular function of an individual gene cannot be understood on the level of isolated components alone, but needs to be studied in the context of its interplay with other gene products. The combined analysis of expression profiles and PPI data thus allows the detection of previously unknown dysregulated modules in interaction networks not recognizable by the analysis of a priori defined pathways.

Ideker et al. (2002) have proposed to identify interaction modules in this setting by devising firstly an adequate scoring function on networks and secondly an algorithm to find the high-scoring subnetworks. The underlying combinatorial problem has been proven to be $NP$-hard for additive score functions defined on the nodes of the network. The authors proposed a heuristic strategy based on simulated annealing and developed a score to measure the significance of a subnetwork that includes the integration of multivariate $P$-values. This score has been extended by Rajagopalan and Agarwal (2005) to incorporate an adjustment parameter in order to obtain smaller subgraphs in conjunction with a greedy search algorithm. This approach however, excludes the possibility to combine multiple $P$-values. Variants of greedy search strategies have also been used by Nacu et al. (2007) and Sohlér et al. (2004). Subsequently Cabusora et al. (2005) proposed an edge score by adapting the scoring concept of Ideker et al. (2002).
An alternative edge scoring based on correlation of gene expression has been proposed by Guo et al. (2007). All the former methods are heuristic approaches that cannot guarantee to identify the maximally scoring subgraph. Some of these often computationally demanding approaches tend to deliver large high-scoring networks, which may be difficult to interpret.

Here we present a novel approach (i) that is characterized by a modular scoring function, based on signal-noise decomposition implemented as a mixture model, (ii) permits the smooth integration of multivariate P-values derived from various sources, (iii) delivers provably optimal and suboptimal solutions to the maximal-scoring subgraph problem by integer-linear programming (ILP) in reasonable running time and (iv) allows to control the resultant subnetwork size by an adjustment parameter, which is statistically interpretable as false-discovery rate (FDR).

The presented algorithm is, to our knowledge, the first approach that really tackles and solves the original problem raised by Ideker et al. (2002) to optimality. We strongly believe that the optimal and suboptimal solutions produced by our method provide a considerable benefit over heuristic solutions in that they allow for a sound evaluation and adaptation of the underlying model. Given only a heuristic solution it is impossible to decide whether poor results are due to inappropriate parameter settings or due to the optimality gap. Based on extensive simulations we evaluate our exact approach in comparison to the heuristic method proposed by Ideker et al. (2002). Finally, analyzing a comprehensive microarray and survival dataset of lymphoma patients we detect functional modules, extending the results of Rosenwald et al. (2002).

The remainder of this article is organized as follows: after describing the data and methods we use (Section 2), we introduce our approach and its application to lymphoma interactome data in Section 3. Section 4 presents the subnetworks we found and a validation of our approach in comparison to the method by Ideker et al. We conclude with Section 5, where we discuss our findings.

2 METHODS

2.1 Microarray and survival data

We used the published gene-expression data set from diffuse large B-cell lymphomas (DLBCL) (Rosenwald et al., 2002). In particular, gene expression data from 112 tumors with the germinal center B-like phenotype (GCB DLBCL) and from 82 tumors with the activated B-like phenotype (ABC DLBCL) were included in this study. Gene expression analysis was performed on the Lymphochip including 12 196 cDNA probe sets (Rosenwald et al., 2002). In addition, survival information was available from 190 patients (Rosenwald et al., 2002).

Statistical analyses were performed using the software package R (R Development Core Team, 2006) and Bioconductor (Gentleman et al., 2004) and the routines implemented in limma (Smyth, 2004). For normalization of gene expression within arrays we used the loess method, normalization between arrays was performed by using the scale method to adjust the log ratios to the same median absolute deviation (MAD) across arrays as detailed in Bleakley et al. (2007). For subsequent analyses the expression values for different spots of the same gene have been aggregated by taking the median. Significance of differential expression between the two subtypes ABC and GCB was calculated using robust statistics based on linear models and a moderated t-test (Smyth, 2004). Survival analysis was performed by Cox regression as implemented in the R-package survival (Andersen and Gill, 1982; Therneau et al., 1990).

2.2 Network

The dataset of literature-curated human PPI has been obtained from HPRD (Mishra et al., 2006; Peri et al., 2003). Using R and the network structures and algorithms in the Bioconductor packages graph and RBGL (Cary et al., 2005), we derived an interactome network consisting of 36 504 interactions between 9392 different proteins. With the subset of genes shared between the interactome dataset and the chip a Lymphochip-specific interactome network was derived as the vertex-induced subgraph. The resulting network comprises 2561 different gene products and 8538 interactions, with a large connected component of 2034 proteins (79.4%) and 8399 interactions (98.4%). The remaining proteins were either non-interacting singletons (472) or tiny clusters of sizes between two and six (23). Our analysis focuses exclusively on the giant connected component: visualization and further network analysis was performed with Cytoscape (Cline et al., 2007; Shannon et al., 2003).

2.3 Optimization algorithm

Our algorithm is based on the software dbase (district heating) from Ljubic et al. (2006). We have extended the C++ code in order to generate suboptimal solutions and have created several Python scripts to control the transformation to a Steiner tree problem, the use of dbase and the retransformation to a PPI subnetwork. The dbase code uses the commercial CPLEX callable library version 9.030 by ILOG, Inc. (Sunnyvale, CA) (Sunnyvale, CA). All experiments were run on a 64 bit 2.2 GHz Opteron Intel with 8 GB of main memory. Our software and the datasets used in this study are publicly available for academic and research purposes within the baniz (heaviest induced subgraph) package of the open source library LoSA (http://www.planet-lisa.net).

3 SCORING FUNCTION AND ALGORITHM

This section introduces our new integrated exact approach to support the identification of functional modules in PPI networks. Section 3.2 focuses on the order statistics-based method to determine score values for the network nodes. We illustrate our approach by analyzing a network obtained by combining the data from an expression profiling study of lymphoma patients (Rosenwald et al., 2002) with the comprehensive interactome data from HPRD (Peri et al., 2003). We derive P-values from the analysis of differential expression between two tumor subtypes (ABC and GCB) as well from the analysis of survival data by Cox regression for each node in the interaction network.

Section 3.2 describes how the score values will be used as input for the maximum-weight connected subgraph (MWCS) problem and a novel algorithmic approach based on mathematical optimization. Our algorithm solves this problem to provable optimality and, furthermore, is able to enumerate sufficiently distinct suboptimal solutions.

3.1 Statistics of scoring function

3.1.1 Aggregation statistics of P-values

Having annotated each node of the interaction network with experimentally derived P-values, we are faced with the problem to aggregate these P-values into one number. A simple aggregation statistics proposed in the literature is the so-called Fisher’s method, which combines n P-values $p_i$ by $-2 \sum_{i=1}^{n} \log(p_i) = \chi^2(n)$ (Fisher, 1948). This method however does not provide the necessary flexibility to control the number of significant P-values, instead it only provides a significance measurement over the entire set of P-values. Due to the heterogeneous nature of the data a more flexible approach is
required. Therefore we use an aggregation statistic based on the distribution of the order statistics of \( P \)-values. Let \( X_1, \ldots, X_n \) be independently identically distributed (iid) then the density of the \( i \)th smallest observation \( X_{(i)} \) is given by

\[
f_{X_{(i)}}(x) = \frac{n!}{(n-i)(i-1)!} f(x)^{i-1}(1-F(x))^{n-i},
\]

where \( F(x) \) denotes the probability density function of \( X_i \) for \( i = 1, \ldots, n \) (Lindgren, 1993). Now we propose to aggregate the \( P \)-values at each node in the network by asking for its \( i \)th order statistic of the associated \( P \)-values, resulting in one \( P \)-value of \( P \)-values. Because \( P \)-values are uniformly distributed under the null hypothesis (Wasserman, 2005), we apply Equation (1) with density \( f(x) = 1 \) and density function \( F(x) = x \) and get

\[
X_{(i)} \sim \frac{n!}{(n-i)(i-1)!} x^{i-1}(1-x)^{n-i}, \quad 0 \leq x \leq 1,
\]

or, in other words, \( X_{(i)} \sim B(n, n-i+1) \) with the associated cumulative distribution function

\[
F_{X_{(i)}} = \frac{n!}{(n-i)(i-1)!} \int_0^x t^{i-1}(1-t)^{n-i} \, dt,
\]

where \( B(\cdot, \cdot) \) denotes the beta distribution. Applying Equation (2), each gene in the network can be assigned an overall \( P \)-value independently identically distributed (iid) then the density of the \( i \)th order statistic. This approach is also applicable in case of missing data: for missing \( P \)-values the \( i \)th order statistic can be used after correcting the parameter \( n \) in Equation (2) appropriately.

3.1.2 Signal-noise decomposition Based on these aggregated \( P \)-values we derive our new scoring function. Following Pounds and Morris (2003) we consider the distribution of the \( P \)-values as a mixture of a noise and a signal component. The signal is assumed to be \( B(a, 1) \) distributed whereas the noise is \( B(1, 1) \) uniformly distributed under the null hypothesis.

The \( B(a, b) \) distribution is given by

\[
f(x) = \frac{\Gamma(a+b)}{a^a b^b \Gamma(b)} x^{a-1}(1-x)^{b-1},
\]

where \( \Gamma(\cdot) \) denotes the gamma function. Thus the distribution of the derived \( P \)-values reduces to

\[
f(x | a, \lambda) = \lambda + (1-\lambda) x^{-a-1} \quad \text{for} \quad 0 < x < 1; \quad 0 < a < 1
\]

with mixture parameter \( \lambda \) and shape parameter \( a \) of the beta distribution. For given data \( x = x_1, \ldots, x_n \) the log likelihood is defined as

\[
\log L(\lambda, a; x) = \sum_{i=1}^n \log \lambda + (1-\lambda) x_i^{a-1},
\]

and consequently the maximum-likelihood estimations of the unknown parameters are given by \( \hat{\lambda}, \hat{a} = \arg\max \sum_{i=1}^n L(\lambda, a; x) \).

We obtain both parameters by numerical optimization using the L-BFGS-B method (Byrd et al., 1995) as implemented in \( R \). For the lymphoma dataset analyzed here we obtained a value of 0.536 for the mixture parameter \( \lambda \) and 0.276 for the shape parameter \( a \) of the beta distribution. This relates to signal and noise proportions of 46.4\% versus 53.6\%, respectively.

Since \( P \)-values are uniformly distributed under the null hypothesis the noise component will be adequately modeled by a uniform distribution. Modeling the signal component by a beta distribution is justified by Figure 1 and a Quantile–Quantile (Q–Q) plot (data not shown). This is furthermore supported by the associated Q–Q plot of the fitted density function with the empirical distribution function, which is extremely close to a straight line (data not shown).

These analyses indicate that the signal is well-captured by the beta distribution.

Our aim is to develop an additive score, where positive values signify signal content and negative values denote background noise. Inspired by the ideas of the likelihood ratio test our approach is as follows: for the fitted parameter \( a \) at the signal component is equal to the \( B(a, 1) \) density, whereas that of the noise component is given by \( B(1, 1) \). Since \( B(1, 1) \) is equivalent to the uniform distribution the denominator is 1 for the score, which is given by

\[
S(x) = \frac{\log B(a, 1)(x)}{\log B(1, 1)(x)} = \log(a) + (a-1) \log(x).
\]

Obviously for \( a \to 1 \) the density of the signal component converges to that of the background model and consequently the score converges to 0 for all \( x \). In particular even very low \( P \)-values will be scored zero. Moreover, for a fitted parameter \( a \) and \( \lambda \):

\[
S(x) \xrightarrow{\lambda \to 1} \log(a) \quad \text{and} \quad S(x) \xrightarrow{\lambda \to 0} +\infty.
\]

This demonstrates that the score properly combines the parameters \( a \) and \( \lambda \).

Similar to classical hypothesis tests where a certain significance level is proposed, we derive a threshold value that discriminates signal from noise. As detailed in (Pounds and Morris, 2003) the mixture model allows the estimation of the FDR. From this we calculate a threshold \( P \)-value \( \tau (\text{FDR}) \), which controls the FDR for the positively scoring \( P \)-values. Thus we derive an adjusted log likelihood ratio score given as

\[
S_{\text{FDR}}(x) = \log \left( \frac{1}{x^{a-1}} \right) = (a-1) \left( \log(x) - \log(\tau (\text{FDR})) \right), \quad (3)
\]
3.2 Mathematical optimization algorithm

Combinatorially, the problem from the previous section can be cast as finding an optimal-scoring subgraph in a vertex-weighted graph:

**Problem 1.** (Maximum-Weight Connected Subgraph Problem, MWCS) Given a connected undirected, vertex-weighted graph $G = (V, E, w)$ with weights $w : V \rightarrow \mathbb{R}$, find a connected subgraph $T = (V_T, E_T)$ of $G$, $V_T \subseteq V, E_T \subseteq E$, that maximizes the score $w(T) = \sum_{v \in V_T} w(v)$.

It is easy to see that any solution for MWCS can always be trimmed to a tree of same weight, and, if all node weights are positive, an optimal solution is easily computed by determining any spanning tree. In case of both positive and negative edge weights, finding the MWCS is not so easy. In fact, in the supplement of Ideker et al. (2002), Karp has shown that MWCS is an $NP$-complete problem, and the authors use this as a justification for their heuristic search method.

We propose to solve this problem by provable optimality using techniques from mathematical programming. More precisely, we transform the problem into the well-known prize-collecting Steiner tree problem (PCST) and use a mathematical programming-based algorithm for PCST to find subgraphs of maximum weight. As our computational results in the next section show, this approach is very successful and reliable in that it finds provably optimal subnetworks in short computation time for all biologically relevant instance sizes.

The PCST problem occurs in classical applications from Operations Research such as planning district heating or telecommunications networks, where profit-generating customers and a connecting network have to be chosen in the most profitable way. Formally, it can be stated as follows:

**Problem 2.** (Prize-Collecting Steiner Tree Problem, PCST) Given a connected undirected vertex- and edge-weighted graph $G = (V, E, c, p)$ with vertex profits $p : V \rightarrow \mathbb{R}^{\geq 0}$ and edge costs $c : E \rightarrow \mathbb{R}^{\geq 0}$, find a connected subgraph $T = (V_T, E_T)$ of $G$, $V_T \subseteq V, E_T \subseteq E$, that maximizes the profit $p(T) := \sum_{v \in V_T} p(v) - \sum_{e \in E_T} c(e)$.

Similar to MWCS, every optimal solution $T$ can be reduced to a tree. Now, let $(G, w)$ be an instance of MWCS with positive and negative vertex weights, and let $w' = \min_{v \in V} w_v$ be its smallest node weight. We construct an instance $(G, p, c)$ of PCST by setting the vertex profits to $p(v) = w(v) - w'$ for all $v \in V$ and the edge costs to $c(e) = w'$ for all $e \in E$. Clearly, this is a valid PCST instance since all vertex profits and edge costs are positive. Figure 3 illustrates the transformation.

The following theorem justifies that we can concentrate on the transformed instance in order to solve the MWCS problem.

**Theorem 1.** A prize-collecting Steiner tree $T$ in the transformed instance $(G, p, c)$ is a connected subgraph in $(G, w)$ with $w(T) = p(T) - w'$.

**Proof.** Obviously, $T$ is a connected subgraph. First, observe that its profit is given by

$$p(T) = \sum_{v \in V_T} p(v) - \sum_{e \in E_T} w_e = \sum_{v \in V_T} p(v) + |V_T| w' - |V_T| w' = p(T) - \sum_{v \in V_T} w_v,$$

as required.

Analyzing our lymphoma network we search for genes that are differentially expressed between the ABC and GCB subtype-based fold changes. The $s$-axis shows coefficients of the univariate Cox proportional hazards regression model fitted for each gene separately. A coefficient greater than zero denotes a significant association. Genes scoring positively by our combined scoring function (for a FDR of $0.01$) are colored. This evidences that our score selects genes specifically associated with the different malignancy of the two tumor subtypes.

The following theorem justifies that we can concentrate on the transformed instance in order to solve the MWCS problem.

**Theorem 1.** A prize-collecting Steiner tree $T$ in the transformed instance $(G, p, c)$ is a connected subgraph in $(G, w)$ with $w(T) = p(T) - w'$.

**Proof.** Obviously, $T$ is a connected subgraph. First, observe that its profit is given by

$$p(T) = \sum_{v \in V_T} p(v) - \sum_{e \in E_T} w_e = \sum_{v \in V_T} p(v) + |V_T| w' - |V_T| w' = p(T) - \sum_{v \in V_T} w_v,$$

as required.

Analyzing our lymphoma network we search for genes that are differentially expressed between the ABC and GCB subtype-based fold changes. The $s$-axis shows coefficients of the univariate Cox proportional hazards regression model fitted for each gene separately. A coefficient greater than zero denotes a significant association. Genes scoring positively by our combined scoring function (for a FDR of $0.01$) are colored. This evidences that our score selects genes specifically associated with the different malignancy of the two tumor subtypes.
The advantages over

This mathematical programming-based algorithm is currently the

This simple reduction to the

hard combinatorial optimization problems (Nemhauser and Wolsey,

the implementation of a trade-off between running time and

with a maximal distance to an optimal solution. This allows

the quality of solutions, i.e. each new feasible solution comes

function. (ii) Methods from mathematical programming guarantee

the quality of a model, e.g. the appropriateness of an objective

(i) having provably optimal solutions at hand allows evaluating

from mathematical programming work astonishingly well.

The above algorithm outputs one optimal solution. In practice,

users often like to obtain a list of promising solutions for

manual inspection. Instead of applying straightforward deletion and

re-iteration, we propose a different approach to generate suboptimal

solutions: in our ILP approach, binary variables \( x_{v} \) determine the

presence of nodes in the optimal subgraph \( S \), that is, \( x_{v} = 1 \) if

\( v \in V(S) \) and \( x_{v} = 0 \) otherwise. Now let \( S \) be an optimal subnetwork

as identified by the branch-and-cut algorithm. Adding the Hamming
distance-like inequality

\[
\sum_{v \in V(S)} (1 - x_{v}) \geq a |V(S)|
\]

with \( a \in [0, 1] \) and re-optimizing leads to a best solution differing

in at least \( a |V(S)| \) nodes from \( S \). This procedure can be iterated

\( k \) times. The advantages of this strategy are 2-fold: first, the user

can determine the number \( k \) of suboptimal solutions that should be

reported and, second, he or she may adjust the variety of solutions

via the parameter \( a \).

4 RESULTS

4.1 Functional modules in lymphoma network

Using our novel approach we identify the optimal-scoring

subnetwork (Fig. 4) for the combined score using a restrictive

FDR of 0.001. This subnetwork consists of 46 nodes and has a
cumulative score of 70.2. The 37 positive-scoring nodes attain

a weight of 102.9 and the 9 negatively scoring nodes have a

\begin{figure}
\centering
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{a}
\caption{Example of an MWCS instance (a) and its transformed PCST counterpart (b). The minimum weight in (a) is \(-15\). Optimal solutions are marked in red color. The MWCS has weight 36, the optimal PCST has profit \(126 - 75 \times 51 = 36 + 15\).}
\end{subfigure}
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{b}
\end{subfigure}
\caption{Example of an MWCS instance}
\end{figure}
Fig. 4. Optimal subnetwork detected using a score based on the \( P \)-values of a gene-wise two sided \( t \)-test and an univariate Cox regression hazard model. A restrictive threshold equivalent to an FDR of 0.001 was used. The derived subnetwork captures the characteristically differentially expressed-interaction modules associated with the increased malignancy of the ABC subtype. Coloring is according to the fold change where red denotes an over-expression in ABC and green in GCB. Diamond nodes represent negative-scoring genes additionally included in the optimal solution.

score of \(-32.8\). The theoretical upper bound of the solution in a completely connected graph, given by the cumulative score of all positive nodes, is in this case 145.4. For the given network and under these restrictive conditions our algorithm collects 70.8\% of all positive scores. Figure 5 shows the next best solutions with \( \alpha = 0.5 \).

Further we capture interactome modules that have been described to play major biological roles in the GCB and ABC DLBCL subtypes. For example, the proliferation module which is more highly expressed in the ABC DLBCL subtype (Rosenwald et al., 2002) is also evident in our current analysis and includes the genes MYC, CCNE1, CDC2, APEX1, DNTTIP2 and PCNA. Likewise, genes IRF4, TRAF2 and BCL2, which are associated with the potent and oncogenic NF\( \kappa \)B pathway, also clustered together as illustrated in Figure 4. Whereas the two previously described interactome modules were derived from genes/proteins expressed in the malignant cells, our algorithm also identified interactome modules (Fig. 6) derived from non-malignant by-stander cells in the lymphoma specimens. In particular, Fibronectin, SPARC, MMP9, CTSK, ITGA5 and ITGB5 showed tight clustering and represent proteins that are expressed in non-malignant fibroblasts and histiocytes (Rosenwald et al., 2002).

4.2 Validation

To validate the performance of our approach including the scoring function and search algorithm we simulated an artificial module according to Rajagopalan and Agarwal (2005). Based on the topology of our lymphoma network we selected two subnetworks of biologically relevant sizes (46 and 143) as signal components. Following the proposal of Rajagopalan and Agarwal (2005) we set signal \( P \)-values uniformly distributed between 0 and \( 10^{-3} \) and background noise \( P \)-values uniformly distributed between 0 and 1.

Since our approach allows for the fine-tuning of the signal noise decomposition by the FDR we scan a large range of FDRs and evaluate the obtained solutions in terms of recall (true-positive rate) and precision (ratio of true positives to all positively classified). To assess the variability of the solutions we ran 10 repetitions for each single FDR step. The silhouette of the recall/precision curve, (adapted from Sing et al., 2005) for the module of size 46 includes the optimal solutions with a maximum recall and precision of exactly one (Fig. 7, upper plot). In particular we find a large
Thus we obtain three discrete solution spaces visualized as shaded polygons representing their convex hulls in Figure 7. Clearly none of these solutions falls within the region of high precision and recall in the upper right corner. Instead one obtains a set of overly large subnetworks from 865 nodes on average, corresponding to 42.5% of the entire network and 18.8 times the size of the hidden signal component. This is reflected by a poor precision of 0.05. After two recursive iterations the number of false positives was reduced substantially and the resultant subnetworks were considerably smaller ranging from 11 to 36 nodes. However, this solution displayed a large variance especially of the recall ranging from 23 to 71%. A similar behavior was observed for the larger solution displayed a large variance especially of the recall ranging from 23 to 71%.

5 DISCUSSION

In the recent years, the integrated analysis of gene expression data in the context of PPI has received considerable attention (Cabusora et al., 2005; Guo et al., 2007; Ideker et al., 2002; Nacu et al., 2007; Rajagopalan and Agarwal, 2005; Sohler et al., 2004). The main objective of these analyses is the derivation of biologically interesting subnetworks of interpretable size from large scale PPI data. This can be expressed as the problem of finding optimal-scoring subgraphs as stated, for the first time, by Ideker et al. (2002). Here we transform the problem to the well-known PCST problem from Operations Research. Thus, we give an alternative NP-completeness proof and, more importantly, we are able to solve large instances of this problem in reasonable computation time to provable optimality by an ILP approach for the transformed problem. Additionally, this allows us to calculate suboptimal solutions with given Hamming distances to previously found solutions. We present an application of our approach using a large PPI network from HPRD (Mishra et al., 2006; Peri et al., 2003) in combination with the comprehensive and well-established microarray dataset from lymphoma patients (Rosenwald et al., 2002). This dataset also provides valuable information of patients' survival which can be used in a Cox regression hazard model to measure the contribution of each gene to malignancy of the tumor. In contrast to previous studies we do not restrict our analysis to differences in expression between conditions (in our case two tumor subtypes) but also include the P-values of the Cox regression into our analysis to derive functional modules that are specifically associated with the different malignant behavior of the tumor subtypes.

In an effective-algorithmic approach, a well-defined objective function is as important as a good search procedure. Therefore we first combine the set of P-values derived from various experiments by an order statistics-based approach to obtain a P-value of P-values as a scalable measure of overall significance. Then we fit a beta-uniform mixture model on the entire set of raw P-values of all nodes in the interaction network. Thus, we achieve a signal-noise decomposition on which we deduce a scoring function of P-values as a likelihood ratio of the signal and noise component. Thus, we deduce a scalable scoring function with a meaningful interpretation of the adjustment parameter as FDR. The additivity of this logarithmic score allows us to effectively formulate and exactly solve the problem of optimal subgraph identification by an ILP approach. Inspired by the problem of finding local-sequence alignments we strive to identify local maximal-scoring network regions. Given a negative-expectation value of the scoring functions as realized.
by an appropriate choice of the FDR, we achieve an efficient localization of the resulting region of interest. Our approach makes it possible to fine-tune the size of the resultant subnetworks and thereby ‘zoom’ into the maximal-scoring region of interest in the interaction network.

Our order statistics-based approach to aggregate the P-values from different experiments is equivalent to that adopted by Ideker et al. (2002). However, we explicitly allow the user to require a predefined number of P-values to be significant (e.g. the first, second, . . . , nth order statistic) instead of only taking the maximum over all order statistics, which is naturally included in our approach. However, asking for the maximum only can lead to serious problems in cases of highly variable signal content among the P-values of different experiments where the highest signal content will dominate the resulting score. As a case in point, analyzing our lymphoma network, the algorithm of Ideker et al. (2002) only yields solution based on the gene expression P-value. Obviously, the biologically important but statistically weaker signal cannot be detected in combination with a more dominant signal by this approach.

To our knowledge, the presented method allows for the first time the exact decomposition of PPI networks into optimal-scoring subnetworks and suboptimal networks of a given dissimilarity as defined by the Hamming distance of the graphs’ node-incidence vectors. In contrast to previously published methods, our algorithm computes provably optimal solutions without computationally demanding parameter optimization usually necessary in heuristic approaches. Furthermore, heuristic methods do not guarantee to find the optimal solution and are unable to assess the solution quality.

As a representative of these heuristic approaches we chose the algorithm of Ideker et al. (2002) as implemented in Cytoscape (Cline et al., 2007; Shannon et al., 2003) and compared the performance to that of our exact approach. The results clearly demonstrate the shortcomings of the heuristic approaches. Since recursive applications of the algorithm are required, only a limited number of isolated solution sizes can be obtained. None of the solutions comes close to the high-accuracy region showing both, high precision and high recall. The high number of true positives in solutions comes close to the high-accuracy region showing both, high precision and high recall. The high number of true positives in solutions, where true positives neglected in one step will not be contained in any later solution.

Although Ideker et al. (2002) claim that many high-scoring subnetworks highlight biologically interesting regions although not being optimal in the sense of the objective function it must be kept in mind that the solutions provided by their algorithm are quite variable and heavily dependent on the choice of the parameter settings (seed, number of iterations and annealing temperature). More importantly, the scoring system of Ideker et al. (2002) and those related to it lack an explicit signal/noise decomposition and thus provide no estimation about the size of the signal content. This can pose a serious obstacle for these approaches in the case of low-signal content or the even worse scenario of random noise only. Applying batches of P-values randomly drawn from a uniform(0,1) distribution to our network the implementation of Ideker et al. (2002) still reports solutions of 770 nodes (37% of the entire network) on average with scores within the same range as those of containing the signal modules. Subsetting these solutions by reapplying the algorithm still yields networks of sizes between 130 and 210 nodes. Obviously a major drawback of these scoring systems is that due to the lacking estimate of the signal content prior to the search phase no distinction between a true-signal component and a ‘best noise’ aggregate can be made. This problem is solved by the integrated signal-noise decomposition based on a beta-uniform mixture model of our approach. In fact all tests with random P-values as input yielded a fitted model with a mixture parameter \( \lambda \) of 1 corresponding to a signal content of 0. Consistent with that we obtain a parameter \( \mu = 1 \) of the signal beta distribution and consequently a score of zero for all nodes (Equation 3).

Nevertheless, heuristic approaches may be able to deliver biologically relevant solutions as claimed by Ideker et al. (2002) if the proportion of signal is high enough. Especially in case of low-signal content the biological relevancy of the obtained solution may be questionable, and even after recursive application of the algorithm the quality of the obtained subnetwork is hardly assessable due to the high variability. Inherently, all published heuristic methods based on the approach of Ideker et al. (2002) share one of the discussed drawbacks either in terms of search algorithm or scoring function. Therefore it is highly desirable to attain truly optimal solutions with an explicit estimate of the signal content and control of the FDR as provided by the presented approach.

We emphasize that, despite the underlying computational complexity, our algorithm runs very fast on biologically relevant instance sizes; our software tool heinz computes provably optimal results usually in a few minutes; profiling our implementation we measured a median runtime of 182 seconds for test runs on 1000 score-permuted graphs. Most importantly we demonstrate that our approach discovers biologically meaningful modules in a lymphoma interaction network which include and extend the results reported by Rosenwald et al. (2002) which have been described to be of importance in the pathogenesis of the GCB and ABC DLBCL subtypes. In general, the integration of survival and expression data into the analysis of PPI networks exhibits perturbed interaction modules associated with the malignancy of the tumor and can yield new insights into tumor biology on a cellular level.

In the future, we plan to generalize our method to an even broader application setting. As a first step, we propose to integrate edge weights, which could, for example, be derived from correlation of gene expression as used by Guo et al. (2007) or from P-values of interaction predictions with STRING (von Mering et al., 2007). Furthermore, we intend to provide an interface to non-commercial optimization libraries and to integrate our algorithm into the Bioconductor environment (Gentleman et al., 2004).

ACKNOWLEDGEMENTS

G.W.K. thanks A. Bley for helpful discussions and I. Ljubić and A. Moser for support with the dhea code. MTD thanks the IZKF (MD/PhD program) and the SFB688 (TPA2).

Conflict of Interest: none declared.
REFERENCES

Aitkin,H.T. and Schwobowski,B. (2006) Graph-based methods for analysing networks in cell biology. Brief. Bioinform., 7, 243–255.

Anderson,S.P. and Gill,R. (1982) Cox’s regression model for counting processes: a large sample study. Ann. Stat., 10, 1100–1120.

Blenk,S. et al. (2007) Germline central B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): analysis of molecular predictors, signatures, cell cycle state and patient survival. Cancer Inform., 3, 409–430.

Byrd,R.H. et al. (1995) A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput., 16, 1190–1208.

Cabusora,L. et al. (2005) Differential network expression during drug and stress response. Bioinformatics, 21, 2989–2995.

Carey,V.J. et al. (2005) Network structures and algorithms in Bioconductor. Bioinformatics, 21, 135–136.

Cline,M.S. et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc., 2, 2386–2382.

Fisher,R.A. 1948. Combining independent tests of significance. Am. Stat.

Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. Journal of Computational Biology, 11(3–4), 379–409.

Gentleman,R.C. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol., 5, R80.

Guo,Z. et al. (2007) Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network. Bioinformatics, 23, 2121–2128.

Ida,T. et al. (2002) Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics, 18(Suppl. 1), i231–i240.

Kelley,B.P. et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. USA, 100, 11394–11399.

Lindsay,W. (1993) Statistical Theory. Chapman & Hall, New York.

Ljubić,I. (2006) An algorithmic framework for the exact solution of the prize-collecting steiner tree problem. Math. Program. Ser. B, 105, 427–449.

Mitra,G. et al. (2006) Human protein reference database—2006 update. Nucleic Acids Res., 34(Database issue), D411–D414.

Nacu,S. et al. (2007) Gene expression network analysis and applications to immunology. Bioinformatics, 23, 850–858.

Neuwald,A. and Wootley,J.L. (1998) Integer and Combinatorial Optimization. John Wiley & Sons, New York, USA.

Peri,S. et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res., 13, 2363–2371.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. Bioinformatics, 19, 1256–1264.

Pounds,S. and Morris,S.W. (2003) Integration of biological networks and gene expression data. Genome Res., 13, 2498–2504.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. Nat. Biotechnol., 24, 427–433.

Sing,T. et al. (2005) ROCR: visualizing classifier performance in R. Bioinformatics, 21, 288–290.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol., 3, Article 3.

Sohler,F. et al. (2004) New methods for joint analysis of biological networks and expression data. Bioinformatics, 20, 1517–1521.

Souvannasam,B. et al. (2007) Current progress in network research: toward reference networks for key model organisms. Brief. Bioinform., 8, 318–332.

Therneau,T. et al. (1990) A package for survival analysis in R. R Foundation for Statistical Computing, Vienna, Austria.

von Mering,C. et al. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res., 35(Database issue): D350–D355.

Wasserman,L.A. (2005) All of Statistics: A concise course in statistical inference. 2nd edn, Springer, New York, USA.