Bioinformatics Advance Access published February 19, 2008

Detecting hierarchical structure in molecular characteristics of disease using transitive approximations of directed graphs

Juby Jacob [a], Marcel Jentsch [b], Dennis Kostka [b], Stefan Bentink [a], Rainer Spang [a*]

[a]Computational Diagnostics Group, Institute of Functional Genomics, University of Regensburg, 93053 Regensburg, Germany [b]Max Planck Institute for Molecular Genetics, Ihnestaße 63/73, 14195 Berlin, Germany

Associate Editor: Prof. Thomas Lengauer

ABSTRACT

Motivation: Molecular diagnostics aims at classifying diseases into clinically relevant sub-entities based on molecular characteristics. Typically, the entities are split into subgroups, which might contain several variants yielding a hierarchical model of the disease. Recent years have introduced a plethora of new molecular screening technologies to molecular diagnostics. As a result molecular profiles of patients became complex and the classification task more difficult.

Results: We present a novel tool for detecting hierarchical structure in binary datasets. We aim for identifying molecular characteristics, which are stochastically implying other characteristics. The final hierarchical structure is encoded in a directed transitive graph where nodes represent molecular characteristics and a directed edge from a node A to a node B denotes that almost all cases with characteristic B also display characteristic A. Naturally, these graphs need to be transitive. In the core of our modeling approach lies the problem of calculating good transitive approximations of given directed but not necessarily transitive graphs. By good transitive approximation we understand transitive graphs, which differ from the reference graph in only a small number of edges. It is known that the problem of finding optimal transitive approximation is NP-complete. Here we develop an efficient heuristic for generating good transitive approximations. We evaluate the computational efficiency of the algorithm in simulations, and demonstrate its use in the context of a large genome-wide study on mature aggressive lymphomas.

Availability: The software used in our analysis is freely available from http://pc56269/software/transApproxs.shtml.

Contact: Juby.Jacob@klinik.uni-regensburg.de,Rainer.Spang@klinik.uni-regensburg.de

1 INTRODUCTION

High throughput genomic approaches have entered molecular diagnostics, generating large amounts of molecular data. Recently, clinical studies have been conducted, which screen a fixed set of patients with several complementary high throughput approaches in parallel (Hummel *et al.* (2006)). Among the most frequently used technologies in oncology are transcriptional profiling using gene expression microarrays (Alizadeh *et al.* (2000), van de Vijver *et al.* (2002), Jones *et al.* (2005)), genomic profiling using array CGH or SNP arrays (Solinas-Toldo *et al.* (1997), Pinkel *et al.* (1998), Pollack *et al.* (1999)). These are typically complemented by non-genome-wide molecular approaches like fluorescent in situ hybridization (FISH) and classical immunophenotyping or non-molecular characteristics like the morphology of cancer tissues, and clinical data typically including the age and sex of the patient as well as treatment response or survival. For example, a male cancer patient can carry two SNPs which are associated with his disease. At the same time his tumor displays genetic losses of four different chromosomal locations, and carries one chromosomal translocation. Moreover, the tumor cells display certain morphological properties and immunophenotypically express 3 antigens. Finally a gene expression profile shows a typical pattern for a molecularly defined subtype of the patient's disease. No other patient in the study has the exact same characteristics. Nevertheless, the challenge is to assign this patient to a well defined subgroup in an hierarchical classification scheme of his disease. To do so, intrinsic hierarchical structure in the integrated data sets needs to be made apparent.

Here we describe a novel computational approach, which is aiming at uncovering hierarchical structure in complex data sets. We aim at identifying molecular characteristics, which are stochastically implying other molecular characteristics. We explain this by an example: Assume that 95% of all tumors with the immunohistochemical property **B** also display the expression pattern **A**. Assume also that 91% of all tumors with the chromosomal amplification **C** and 100% of all tumors with the expression pattern **D** display the immunohistochemical property **B**. A toy dataset representing this constellation is shown in Figure 1(a). The patients with property **A** are a noisy superset of those with property **B**, while this set contains the two not necessarily disjoint subsets **C** and **D**, shown in Figure 1(b). Up to a small number of exceptions property **B** implies property **A** and properties **C** and **D** always imply property **B**. We encode this hierarchical structure by the graph shown in Figure 1(c). Our algorithm aims at extracting such hierarchical structure.

We proceed in two steps, which we call graph construction and graph consolidation. The latter is based on the transitive approximation problem and is the main focus of the paper.

The construction step: In the construction step we screen all pairs of molecular properties for potential noisy subset/superset relations and encode these in a primary graph by drawing directed edges from

[*]to whom correspondence should be addressed

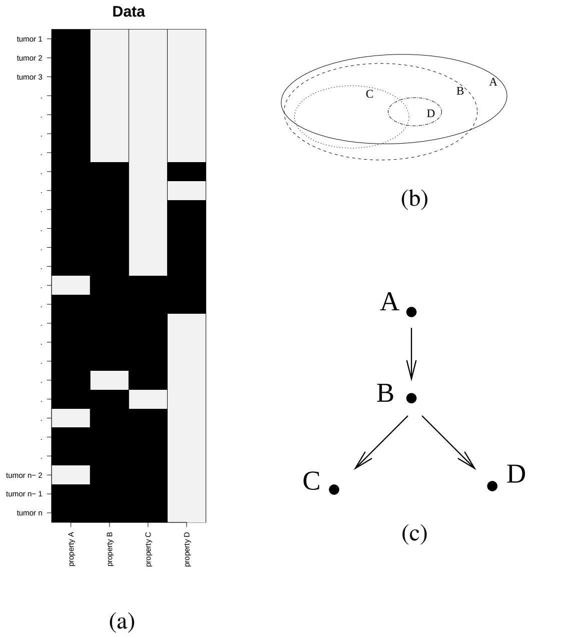

Fig. 1. Subset/superset relations between molecular properties observed in patients. *(a) A toy dataset on molecular properties A,B,C,D observed in patients, represented on the y-axis. Black regions of the data code for the observed property while white regions encode for not observed properties. (b) Subset/superset relations between molecular properties. Tumors with molecular property* **A** *are a (noisy) superset of tumors with molecular property* **B***, which is a (noisy) superset of tumors with properties* **C** *and* **D***. (c) Graphical representation of the subset/superset relations. Directed edges in the graph are drawn from supersets to subsets.*

supersets to subsets. More precisely, for any two molecular features A and B, if α percent or more of patients with feature B also exhibit feature A then we say B is a noisy subset of A and we draw an edge from A to B. We obtain the primary graph by screening all pairs of molecular features for possible subset/superset relation using the above criteria. We note that in case two or more molecular features are identical with respect to the cases they include, then we join them to a single group in the primary graph. The choice of the parameter α is based on practical considerations. Clearly a graph that hardly contains any edge is useless as a hierarchical model and so is an almost fully connected one. We use α to scale the sparseness of the primary graph.

The consistency problem: Note that in the subset relation between the various properties, if property **B** is a subset of **A** (denoted in the graph by the edge $\mathbf{A} \rightarrow \mathbf{B}$) and property **C** is a subset of **B** (denoted in the graph by the edge $\mathbf{B} \rightarrow \mathbf{C}$), then for consistency reasons **C** is a subset of **A** implies also directly. Unfortunately, due to noise in the data the edge $\mathbf{A} \rightarrow \mathbf{C}$ can be missed during graph construction. More generally, logical consistency is requiring that a graph representing a hierarchical structure is transitively closed, while the edgewise construction approach might violate this requirement. For this reason we calculate transitive approximations of the primary graph in the graph consolidation step. This step is based on the transitive approximation problem, which can be formalized as follows: For a given directed graph g find all transitively closed graphs with minimal edit distance to g, where the edit distance between two graphs is the minimal number of edit operations (adding or deleting edges) necessary to transfer the first graph into the second. We refer to all graphs with this optimality property as optimal transitive approximations of g.

For undirected graphs, the transitive approximation problem is as old as Zahn (1964) and has been extensively studied in Delvaux and Horsten (2004), De Clercq and Horsten (2005). Delvaux and Horsten (2004) have shown that the problem is NP-complete. The equivalent problem for directed graphs is less studied. Nevertheless, Natanzon *et al.* (1999) show that the optimal transitive approximation problem is NP-complete also for directed graphs.

Here we develop a heuristic algorithm to calculate *good transitive approximations*, which are graphs with a small but not necessarily optimal edit distance to the original graph. The challenge is to derive a computationally efficient algorithm.

The paper is organized as follows: In the next section we develop a heuristic for calculating good transitive approximations and evaluate it in section 3 using random simulations. In section 4 we give technical details of the graph construction step and demonstrate the use of the two step procedure in the context of a large genomic study on aggressive lymphomas.

2 TRANSITIVE APPROXIMATIONS

2.1 Notations and definitions

Graphs and Sequential Constructions: Throughout we let $\mathcal{G}$, $\mathcal{C}$ denote the set of all directed, respectively transitive directed graphs on n nodes. We let $g = (V, E) \in \mathcal{G}$ denote an arbitrary graph in $\mathcal{G}$ where $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$ denotes its set of nodes and edges. We use the notation $e_k \in g$ to denote that e_k is an edge in g and $e_k \notin g$ to denote e_k is not an edge in g. Next, let $g + e_k$ denote the graph obtained by adding edge e_k to g and $g - e_k$, the graph obtained by removing edge e_k from g. For any random ordering of the edges $e_1, \ldots, e_m$, we obtain a sequence of subgraphs g_k of g, by setting g_0 to be the empty graph (graph with all nodes but no edges) and defining $g_k = g_{k-1} + e_k, 0 < k \leq m$. Note that $g_m = g$. We call the sequence of subgraphs a *sequential construction* $(g_k)_k$ of g.

Distances and Neighborhoods: For two graphs $g_1, g_2 \in \mathcal{G}$ let $d(g_1, g_2)$ denote the *edit distance* between them, i.e. the number of edges that are different in the two graphs. For $g \in \mathcal{G}$, we define the *δ-neighborhood of* g, denoted by $B_\delta(g)$, as the set of all graphs with maximal edit distance δ to g and for $\mathcal{F} \subset \mathcal{G}$, we define $B_\delta(\mathcal{F}) = \bigcup_{f \in \mathcal{F}} B_\delta(f)$.

Transitive Approximations: For $g \in \mathcal{G}$, we let $\delta^*(g)$ denote the *minimum distance* of g to a transitive directed graph. We define *optimal transitive approximations of* g, denoted by $S_0(g)$, to be the set of all transitive directed graphs $c \in \mathcal{C}$ which are at minimal distance to g. Clearly, $S_0(g) = B_{\delta^*}(g) \cap \mathcal{C}$. Usually, there exist more than one optimal transitive approximation to a directed graph. We give an example: Consider the non-transitive graph given in Figure 2(a). Adding the edge AD and deleting the edge DE results in the transitive graph, shown in Figure 2(b). Deleting edges BD and CD also results in a transitive graph, which is shown in Figure 2(c). Both graphs are optimal transitive approximations of the graph in Figure

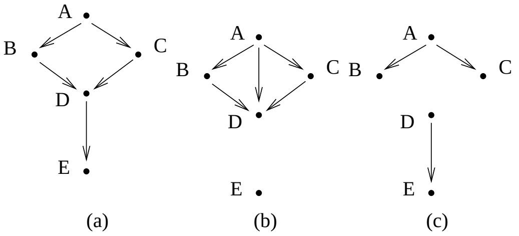

Fig. 2. Multiple optimal transitive approximations. *A non-transitive reference graph (a) with two different optimal transitive approximations (b) and (c).*

2(a), with a minimum distance of 2. Finally, we define *suboptimal transitive approximations* of g denoted by $S_\gamma(g)$ as the set of all transitive graphs that are γ edges short of being at minimal distance to g.

2.2 Preliminaries

LEMMA 1. *Let g be a directed graph with sequential construction $(g_k)_k$. Then for any $h \in \mathcal{G}$, we have*

$$|d(g_{k-1}, h) - d(g_k, h)| = 1.$$

Proof: Let $l = d(g_{k-1}, h)$. We have $g_k = g_{k-1} + e_k$. It is easy to see that if e_k is an edge of h, then $d(g_k, h) = l - 1$ and if e_k is not an edge of h then $d(g_k, h) = l + 1$. In either case it follows that $|d(g_{k-1}, h) - d(g_k, h)| = 1$. □

We next show that when going from g_{k-1} to g_k in a sequential construction, the minimum distance to transitively closed graphs can at most change by 1. Depending on the edge e_k that is added, the distance can either increase or decrease by 1.

LEMMA 2. *Let g be a directed graph with sequential construction $(g_k)_k$ and let $\delta^*(g_k)$ denote the minimum distance of g_k to the set of transitive directed graphs. Then*

$$\delta^*(g_{k-1}) - 1 \leq \delta^*(g_k) \leq \delta^*(g_{k-1}) + 1.$$

Proof: We have $g_k = g_{k-1} + e_k$. Clearly $\delta^*(g_{k-1}) + 1$ is an upper bound for $\delta^*(g_k)$, since by Lemma 1 all optimal transitive approximations of g_k can have at most this distance. On the other hand e_k could be an edge of an optimal transitive approximation $h \in S_0(g_{k-1})$. Then $d(g_{k-1}, h) = \delta^*(g_{k-1})$. Also since e_k is an edge of h, by Lemma 1 we have $d(g_k, h) = \delta^*(g_{k-1}) - 1$. We now show that $\delta^*(g_{k-1}) - 1$ is a lower bound. Suppose that there existed a transitively closed graph $c \in \mathcal{C}$ with distance $d(g_k, c) \leq \delta^*(g_{k-1}) - 2$. Then by using Lemma 1, we have $d(g_{k-1}, c) \leq \delta^*(g_{k-1}) - 1$, a contradiction to the optimality of $\delta^*(g_{k-1})$. □

LEMMA 3. *Let g be a directed graph with sequential construction $(g_k)_k$. Then for all $f \in S_\gamma(g_k)$ $(0 < k \leq m)$, we have*

$$f \in \bigcup_{\gamma-2 \leq \gamma' \leq \gamma+2} S_{\gamma'}(g_{k-1}).$$

Proof: Let $f \in S_\gamma(g_k)$, from the definition of suboptimal transitive approximations it follows that

$$d(g_k, f) = \delta^*(g_k) + \gamma. \tag{1}$$

From Lemma 1, we have $|d(g_{k-1}, f) - d(g_k, f)| = 1$.
That is, $-1 \leq d(g_{k-1}, f) - d(g_k, f) \leq 1$. Using (1) and Lemma 2, we have

$$\begin{aligned} d(g_{k-1}, f) &\leq d(g_k, f) + 1 \\ &\leq \delta^*(g_k) + \gamma + 1 \\ &\leq \delta^*(g_{k-1}) + 1 + \gamma + 1 \\ &\leq \delta^*(g_{k-1}) + \gamma + 2. \end{aligned}$$

Also

$$\begin{aligned} d(g_k, f) - 1 &\leq d(g_{k-1}, f) \\ \delta^*(g_k) + \gamma - 1 &\leq d(g_{k-1}, f) \\ \delta^*(g_{k-1}) + \gamma - 2 &\leq d(g_{k-1}, f). \end{aligned}$$

Hence we have

$$\delta^*(g_{k-1}) + \gamma - 2 \leq d(g_{k-1}, f) \leq \delta^*(g_{k-1}) + \gamma + 2$$

The lemma now follows from the definition of suboptimal transitive approximation. □

The following corollary immediately follows from the previous lemma.

COROLLARY 1. *With the assumptions of Lemma 3 and for $f \in S_0(g)$, we have*

$$f \in S_0(g_{m-1}) \cup S_1(g_{m-1}) \cup S_2(g_{m-1}).$$

The next lemma shows that suboptimal approximations of g_k of order at most 2 can be constructed from small modifications of suboptimal approximations of order at most 4 of the previous subgraph g_{k-1} in the sequential construction. From Lemma 2 we observe that in going from g_{k-1} to g_k, depending on the edge e_k that is added, the minimum distance can at most change by 1. Consequently, for g_k with minimum distance $\delta^*(g_k)$ we have that $\delta^*(g_k)$ can equal (i) $\delta^*(g_{k-1}) - 1$, (ii) $\delta^*(g_{k-1})$, or (iii) $\delta^*(g_{k-1}) + 1$. We have

LEMMA 4. *Let $g \in \mathcal{G}$ be a directed graph with sequential construction $(g_k)_k$. For each case indicated by the first column of Table 1, optimal and suboptimal transitive approximations of g_k have the properties with respect to g_{k-1} as indicated by the respective columns.*

Proof: If e_k is an edge of an optimal transitive approximation in $S_0(g_{k-1})$, then we have $\delta^*(g_k) = \delta^*(g_{k-1}) - 1$ and row (i) in Table 1 follows from the definitions and previous lemmas. Similarly, if e_k is an edge of a graph in $S_1(g_{k-1})$, the minimum distance to a transitive graph remains the same, yielding row (ii). Finally, in the case when e_k is an edge of a graph in $S_2(g_{k-1})$ we have $\delta^*(g_k) = \delta^*(g_{k-1}) + 1$ and row (iii). □

The previous lemma provides updating rules for suboptimal approximations of g_k using small modifications of suboptimal approximations of g_{k-1}. Updating S_0 requires S_0, S_1 and S_2 of the previous subgraph. Updating S_1 requires S_0, S_1, S_2, S_3 whereas updating S_2

Table 1. **Suboptimal transitive approximations of g_k of distance at most 2, with respect to g_{k-1} as discussed in Lemma 4.**

	If $\delta^*(g_k) =$		$f \in S_0(g_k)$ if and only if	$f' \in S_1(g_k)$ if and only if	$f'' \in S_2(g_k)$ if and only if
(i)	$\delta^*(g_{k-1}) - 1$		$f \in S_0(g_{k-1})$ and $e_k \in f$	$f' \in S_1(g_{k-1})$ and $e_k \in f'$	$f'' \in S_0(g_{k-1})$ and $e_k \notin f''$ or $f'' \in S_2(g_{k-1})$ and $e_k \in f''$
(ii)	$\delta^*(g_{k-1})$		$f \in S_1(g_{k-1})$ and $e_k \in f$	$f' \in S_0(g_{k-1})$ and $e_k \notin f'$ or $f' \in S_2(g_{k-1})$ and $e_k \in f'$	$f'' \in S_1(g_{k-1})$ and $e_k \notin f''$ or $f'' \in S_3(g_{k-1})$ and $e_k \in f''$
(iii)	$\delta^*(g_{k-1}) + 1$		$f \in S_0(g_{k-1})$ and $e_k \notin f$ or $f \in S_2(g_{k-1})$ and $e_k \in f$	$f' \in S_1(g_{k-1})$ and $e_k \notin f'$ or $f' \in S_3(g_{k-1})$ and $e_k \in f'$	$f'' \in S_2(g_{k-1})$ and $e_k \notin f''$ or $f'' \in S_4(g_{k-1})$ and $e_k \in f''$

requires S_0, S_1, S_2, S_3, S_4. Lemma 3 shows that computing S_3 and S_4 require S_5 respectively S_6 of the previous subgraph in the construction. In the following we give an alternative to updating S_3 and S_4.

LEMMA 5. *With the assumptions of Lemma 4, let* $P = B_3(S_0(g_{k-1})) \cup B_2(S_1(g_{k-1})) \cup B_1(S_2(g_{k-1}))$ *and* $Q = B_4(S_0(g_{k-1})) \cup B_3(S_1(g_{k-1})) \cup B_2(S_2(g_{k-1})) \cup B_1(S_3(g_{k-1}))$. *Moreover, let* $\overline{P} = \{h \in P \cap \mathcal{C} \mid d(h, g_{k-1}) = \delta^*(g_{k-1}) + 3\}$ *and* $\overline{Q} = \{h \in Q \cap \mathcal{C} \mid d(h, g_{k-1}) = \delta^*(g_{k-1}) + 4\}$. *Then* $S_3(g_{k-1}) = \overline{P} \cup R$ *with* $R = \{f \in S_4(g_{k-2}) \cup S_5(g_{k-2}) \mid d(f, g_{k-1}) = \delta^*(g_{k-1}) + 3\}$ *and* $S_4(g_{k-1}) = \overline{Q} \cup R'$ *with* $R' = \{f \in S_5(g_{k-2}) \cup S_6(g_{k-2}) \mid d(f, g_{k-1}) = \delta^*(g_{k-1}) + 4\}$.

Proof: The lemma follows directly from the definition of suboptimal transitive approximations. □

2.3 Algorithm

We now present a heuristic algorithm for calculating good transitive approximations of directed graphs. By a good transitive approximation we understand a not necessarily optimal transitive approximation, which from a practical perspective is close enough to the optimum to reveal meaningful hierarchical structure. The main ingredients of the algorithm are Lemma 4 and Lemma 5.

Let $g = (V, E) \in \mathcal{G}$ be an arbitrary directed graph with the set of vertices $V = \{v_1, v_2, \ldots, v_n\}$ and the set of edges $E = \{e_1, e_2, \ldots, e_m\}$ in an arbitrary but fixed order. Let $g_0 \subset g_1 \subset \cdots \subset g_m = g$ be the corresponding sequential construction of g. We will proceed iteratively starting with g_0 by constructing transitive approximations of g_k from transitive approximations of g_l, $l < k$. From Corollary 1 we see that optimal transitive approximations of g_k can be derived from suboptimal approximations of distance at most 2 of earlier subgraphs in the sequential construction by introducing at most 2 modifications. From Lemma 3 approximations of distance at most 2 can be constructed from approximations of distance at most 4 of earlier subgraphs, which again can be constructed from approximations of distance at most 6 of even earlier subgraphs and so on. In this iterative process the number of intermediate approximations increases rapidly requiring inordinate running times for exhaustive enumeration. Therefore we resort to heuristic as a feasible line of attack. The heuristic aspect of the algorithm is to calculate in each step only suboptimal transitive approximations of distances 0,1, and 2 to the optimum and the sets $\overline{P}$ and $\overline{Q}$ ignoring possible contributions from the sets R and R'. The motivation is that for a small number of reference graphs that we examined more closely, we observed that R and R' are small and often empty sets. Nevertheless, ignoring the sets renders our algorithm inexact. Errors can occur and propagate throughout the whole procedure. Simulation experiments will show that the heuristic still produces good transitive approximations for most graphs g.

We start with initializing three sets of graphs called S_0', S_1' and S_2'. During initialization the sets refer to transitive approximations of the empty graph g_0. Since it does not violate transitivity assumptions it is its own unique transitive approximation. Suboptimal transitive approximations of distance 1 and 2 are all transitive graphs with 1, 2 edges respectively. The minimum distance to a transitive graph is 0 and we initialize a variable Δ with this value. At this stage we use Lemma 5 and compute $\overline{P}, \overline{Q}$ which give suboptimal transitive approximations of g_0, of distance 3, respectively 4 stored in variables S_3' and S_4'.

Next we add edge by edge along the chosen sequential construction of g and iteratively update the sets S_0', S_1' and S_2'. Assume that the sets are updated up to g_{k-1}. We now add e_k and use Table 1 and Lemma 5 to update the sets of transitive approximations. We check successively among the current sets S_0', S_1' and S_2' for graphs which contain the edge e_k. If we find a graph in S_0' with edge e_k, we know that this graph has distance $\Delta - 1$ to g_k. We update Δ and assign the graph to the update of S_0'. We then use row(i) of Table 1 to fully update the sets S_0', S_1' and S_2'. Next we compute $\overline{P}$ and $\overline{Q}$ and update S_3' respectively S_4' ignoring possible contributions from R and R'.

If instead we find a graph in S_1' with edge e_k, we know that this graph has distance Δ to g_k. We assign the graph to the update of S_0' and use row (ii) of Table 1 to update the sets S_0', S_1' and S_2'. Sets S_3' and S_4' are again updated using $\overline{P}$ and $\overline{Q}$ respectively.

If however none of the above holds and we find a graph in S_2' with edge e_k then this graph has distance $\Delta + 1$ to g_k. We update Δ and assign the graph to the update of S_0' and use row(iii) of Table 1 to update the sets S_0', S_1' and S_2'. As before we compute $\overline{P}, \overline{Q}$ to update S_3', respectively S_4'. We continue by adding the next edge from the sequential construction and proceed as before. After the last edge is added, the output of our algorithm is the set of good transitive approximations S_0'.

3 EVALUATION ON SIMULATIONS

Running time: To validate the running time performance of the algorithm, we constructed random reference graphs with fixed numbers of nodes N and preset edge probabilities p. For each of the possible N^2 edges a uniformly distributed random number between zero and one was generated. If this number was smaller than p the edge was added to the graph. Thus, the expected number of edges is $p * N^2$. We measured running times on graphs with $10, 15, 20, 25$

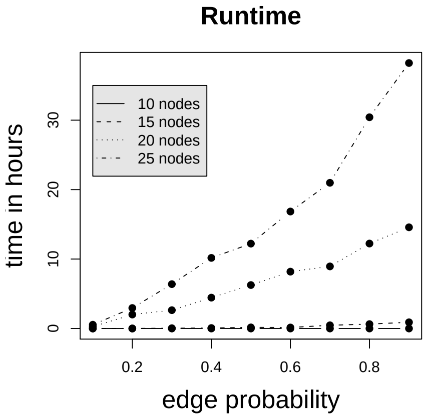

Fig. 3. Running times on random graphs. *The y-axis gives the average running time in hours over 20 runs, the x-axis gives the edge probability encoding the sparseness of the reference graphs.*

nodes and edge probabilities from .1 to .9. For each combination of N and p, 20 random graphs were generated, and the observed running times on an AMD Athlon machine (1800 MHz clock, main memory size: 2GB) were averaged. The results are shown in Figure 3. As expected running times increase both with the number of nodes in the graph and with its density. We obtain practically feasible running times below 20 hours for dense graphs up to 15 nodes. For sparse graphs ($p < 0.2$) even 25 nodes can be computed in less than 3 hours. In our application where we look at the hierarchical disease modelling, sparseness of reference graphs can be controlled by the noise parameter α. Moreover, practical hierarchical models will most likely be sparse. Running times depend on the order of edges in the sequential construction of the graphs. To quantify the range of running times for the same graph with different orderings of edges, we ran the algorithm on the graph shown in Figure 2(a) using 120 random orderings of its edges. Running times varied around an average of 25 seconds with a standard deviation of 6. Orderings producing early subgraphs g_k, which are close to being transitive generally produce short running times. In view of using the algorithm for larger problems ($N > 25$) further research into optimizing the edge ordering appears promising.

Accuracy of approximations: A rigorous validation of the accuracy of transitive approximations is not trivial. Ideally, one would use a large set of reference graphs, for which the complete set of optimal transitive approximations is known. Unfortunately, this validation data is hard to produce due to the NP-completeness of the problem. Instead, we generated random graphs, calculate their transitive closures, modify the closures by a fixed number k of edge deletions or edge insertions, run our algorithm on the modified graphs, and compare the computed distance Δ with k. If Δ is larger than k, we observe evidence for a non-optimal solution. Clearly, $\Delta \leq k$ does not prove optimality of the approximation, but it also does not contradict it, while $\Delta > k$ indicates a non-optimal approximation. Along these lines, we generated 30 graphs of sizes $N \in \{8, 10, 12, 15, 20\}$ and report the results in Table 2. Notably, in no instance we observed evidence for suboptimality of the approximation. On the contrary, in many cases the algorithm found transitive graphs at distances below k.

Number of Nodes	k	$\Delta < k$	$\Delta = k$	$\Delta > k$
8	2	0.15	0.85	0
8	3	0.18	0.82	0
8	4	0.24	0.76	0
8	5	0.24	0.76	0
10	2	0.12	0.88	0
10	3	0.27	0.73	0
10	4	0.24	0.76	0
10	5	0.33	0.67	0
12	2	0.06	0.94	0
12	3	0.18	0.82	0
12	4	0.21	0.79	0
12	5	0.33	0.67	0
15	3	0.35	0.65	0
15	4	0.25	0.75	0
15	5	0.20	0.80	0
15	6	0.31	0.69	0
20	5	0.40	0.60	0
20	6	0.26	0.74	0
20	7	0.29	0.71	0

Table 2. Accuracy of the approximation. *Transitive graphs were modified by performing k edge deletions or insertions and the distance obtained as a result of the algorithm was compared to k. The first column denotes the number of nodes of the transitive graph. The second column denotes the number of edit operations and the following columns give the fraction of graphs where the distance is smaller, equal or larger than the number of modifications.*

4 A HIERARCHY FOR MOLECULARLY DEFINED GROUPS OF LYMPHOMAS

We now show the practical use of the procedure in the context of the classification of mature aggressive lymphomas. We used the data of a large lymphoma study conducted by the research network Molecular Mechanisms of Malignant Lymphoma published in Hummel *et al.* (2006). A total of 220 mature aggressive B-cell lymphomas were analyzed using various molecular approaches including gene expression profiling, immunophenotyping, and molecular cytogenetic analysis. The data is available from www.ncbi.nih.gov/pub/geo through the GEO accession number GSE4475. All molecular characterizations yield binary labels for all lymphomas in the collection.

Gene expression signatures: Gene expression properties were summarized in signatures, which were either present or absent in the profile of an individual lymphoma. Here we used the mBL signature of (Hummel *et al.*, 2006), which characterizes molecular Burkitt lymphoma. Lymphomas not presenting the mBL signature were called non-mBL. Moreover, we used the ABC/GCB signatures (Rosenwald *et al.*, 2002), which defines two subgroups of diffuse large B-cell lymphomas: Lymphomas with expression profiles similar to germinal center B-cells (GCB) and lymphomas with profiles similar to activated B-cells (ABC). Note, that lymphomas with intermediate expression properties between ABC and GCB exist. They were excluded from both groups. Although the GCB/ABC signatures were originally only used for DLBCLs, molecular Burkitt lymphomas also display the GCB signature, and we have included them into the GCB group.

Immunophenotyping: Histological sections of lymphoma tissues were stained with antibodies for the biomarkers CD10, CD5 and Ki-67. The expression of the corresponding proteins was classified by expert pathologists into present and absent. In case of the proliferation marker Ki-67, lymphomas with a score greater than 95% were termed Ki-67 positive.

Cytogenetics: Chromosomal translocations were determined using interphase fluorescence in situ hybridization (FISH). The group IG-MYC comprises lymphomas with a translocation of the MYC locus involving fusion of MYC to the immunoglobulins IGH, IGK or IGL, "atyp.myc" includes lymphomas with a breakpoint in the MYC locus, but without fusion of the gene with one of the IG genes, while bcl6Br denote lymphomas with breakpoints in the BCL6 locus and IGH-BCL2 denotes fusion of BCL2 to IGH. For every cytogenetic and immunophenotypic label we included two sets of lymphomas into the analysis, one with the lymphomas, which are positive for the label and one with the lymphomas, which are negative for it. The original dataset contains more molecular features of lymphomas than we included in the model. The omitted features were either present in almost all or almost none of the lymphomas and hence did not contribute to a hierarchical model. Also for simplicity of calculations, we excluded all lymphomas where at least one of the 17 modelled features was not assessed, leaving us with 176 remaining lymphomas.

From this data we constructed a graph based on noisy subset/superset relations, which we then corrected for transitivity by using our approximation algorithm. Graph construction was done with several values for the noise parameter α. Small α returned dense graphs with almost 90% edges, whereas large values yielded graphs with hardly any edge. $\alpha = 0.9$ allowing for 10% noise in the subset relations yielded a primary graph with 10% edge density, which is well in the range of practically useful hierarchical models. With this graph as a reference, we ran the transitive approximation algorithm resulting in six good transitive approximations with an edit distance of four to the reference graph. In order to choose one of them, we scored them by averaging the actual accuracy of all edges in them. By accuracy of an edge we understand the percentage of lymphomas in the daughter node that are also contained in the parent node. Clearly this accuracy is bound by α from below. The hierarchical model of lymphomas resulting from the highest scoring approximation is shown in Figure 4. Some caution in interpreting the model is needed. If an entity has two daughter entities this only means that the daughter entities are essentially included in the parent. It does not mean that the two daughter entities are disjoint, nor does it mean that the two daughter nodes fully cover the lymphomas in the parent node. The model reflects known properties of mature aggressive lymphomas. For example, molecular Burkitt lymphomas (mBL) carry the GCB signature, are CD10 positive, and do not carry a IGH-BCL2 fusion nor a breakpoint in the locus of the BCL6 gene, in line with findings of Hummel *et al.* (2006). As described by Rosenwald *et al.* (2002), lymphomas with an BCL2 break are also GCB. Moreover ABC, are CD10 negative non-mBLs with no abberations of the MYC locus. In addition to confirming published results, we can also easily identify contradictions of the present study data to other published data. For example Hans *et al.* (2004) claim that CD10 expression implies the GCB type of DLBCLs. Our model does not show a directed edge between these features and in fact the modelled study data is not supporting this claim at all. Furthermore, the WHO proposes the presence of an IG-MYC fusion and a Ki-67 score larger than 95% as defining criteria for Burkitt lymphoma (Jaffe *et al.* (2001)). Neither IG-MYC nor Ki-67 positivity implies Burkitt lymphoma. The disagreement of Ki-67 may be due to limited reproducibility of the staining of this marker (de Jong *et al.* (2007)). Interestingly, low expression of Ki-67 is a feature of ABC non-mBL cases.

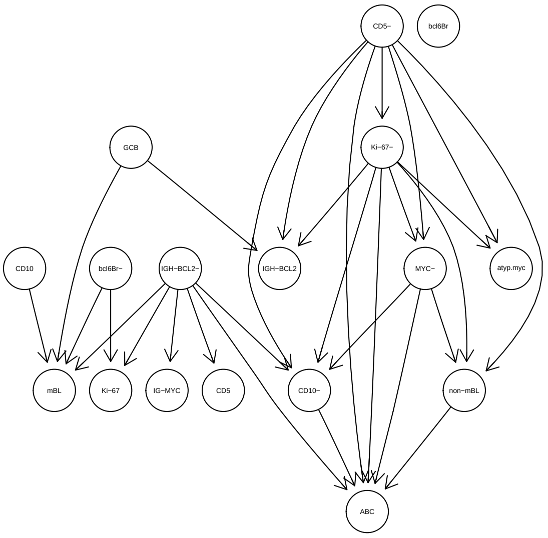

Fig. 4. A hierarchical model of molecular subgroups of mature aggressive B-cell lymphoma.

In order to evaluate the influence of the choice of α on the final transitive graph we ran the transitive approximation algorithm on primary graphs obtained by varying alpha from 0.85 to 0.95 in steps of .01. For each alpha we then chose one transitive approximation by a scoring using the actual accuracy of edges, as described before. A comparison of the transitive approximation with $\alpha = 0.9$ to the transitive approximations with $\alpha = 0.85, \ldots, 0.95$ showed that on an average 85% of the edges in the various transitive approximations are shared by the transitive approximation with $\alpha = 0.9$.

5 DISCUSSION

With the availability of high dimensional data in both biological and medical research, there is an increasing need to develop new and efficient strategies to analyze such data. In genomic clinical studies, where the same patients are characterized by various molecular readouts large and complex datasets are generated. Often these datasets contain hierarchical structure on molecular characteristics. In the present paper, we have introduced a novel tool to uncover hierarchical structure in complex datasets. We aim for identifying molecular characteristics, which imply other molecular characteristics. The final hierarchical structure is summarized in a directed transitive graph.

Closely associated with constructing hierarchical models of disease is the problem of transitive approximations of directed graphs. Formally stated, given a reference graph, the problem is to find transitive approximations of the reference graph, which are transitive graphs with minimal edit distance to the reference graph. The problem is known to be NP-complete for both directed and undirected graphs. Here we have developed an efficient heuristics for directed graphs and to our knowledge it is the first of its kind. Note that in principal our algorithm can be used to approximate a reference graph g by graphs from an arbitrary set $\mathcal{G}$. We did not use that $\mathcal{G} = \mathcal{C}$ in our formal arguments. Of course our validations of the efficiency and accuracy of the algorithm refer only to the transitive approximation problem. We obtain transitive approximation for sparse graphs with up to 25 nodes in less than three hours, with good accuracy. Finally we used the algorithm to retrieve a hierarchical model of mature aggressive B-cell lymphomas, which reproduces well known results, suggest new relationships between lymphoma groups and uncovers discrepancies between different studies.

Besides the hierarchical modelling of disease the transitive approximation algorithm might also be useful in reconstructing signalling pathways using nested effects models (see Markowetz *et al.* (2007)). Here the graph of up- and down-stream relations in molecular signalling networks is estimated, from phenotypic screens of RNAi experiments. In a modular approach, this is done by deciding for each pair (or triple) of signalling molecules, whether they are up- or down-stream from each other. Combining these estimated relations for all analyzed signalling molecules yields a reference graph for the signalling network. Although consistency requires that up/down-stream relations need to be transitive, noise in the data leads to non-transitive reference graphs, such that the transitive approximation algorithm can be used to consolidate the signalling networks.

The hierarchical modelling approach can be interpreted as a generalization of the clustering problem. Clustering the binary characterization vectors of lymphomas detects sets of molecular features that are present in essentially the same lymphomas, suggesting equivalence of the features. Going beyond equivalence, we detect with our approach features that tend to imply each other. While clustering uncovers a noisy equivalence relation, our algorithm uncovers a noisy order relation.

Here we have focused on the discrete transitive approximation problem. The graph construction part of the algorithm is relatively simple. In a more elaborate modelling approach it could be replaced by a full Bayesian model, which specifies posterior probabilities for each potential edge. This would lead to a generalization of the transitive approximation problem, since then one needs to find transitively closed graphs with optimal posterior probability across all edges.

We believe that models of hierarchical structure will be a valuable complementation of the current repertoire of data mining tools. Due to the general nature of the underlying concept, we expect it to be useful in many fields of application.

ACKNOWLEDGEMENT

We thank Florian Markowetz for his intellectual input. We also thank all former and present members of the Computational Diagnostics Group for inspiring discussions. JJ and RS were supported by the *Bayerisches Genomforschungsnetz (BayGene)*. DK was supported by the Deutsche Krebshilfe in the context of the Molecular Mechanisms in Malignant Lymphomas Network Project.

REFERENCES

Alizadeh, A., Eisen, M., and Davis, R. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–11.

De Clercq, R. and Horsten, L. (2005). Closer. *Synthese*, **146**, 371–93.

de Jong, D., Rosenwald, A., and Chhanabhai, M. *et al.* (2007). Immunohistochemical prognostic markers in diffuse large B-cell lymphoma: validation of tissue microarray as a prerequisite for broad clinical applications–a study from the Lunenburg Lymphoma Biomarker Consortium. *J Clin Oncol*, **25**(7), 805–12.

Delvaux, S. and Horsten, L. (2004). On best transitive approximations to simple graphs. *Acta Informatica*, **40**(19), 637–55.

Hans, C., Weisenburger, D., and Greiner, T. *et al.* (2004). Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, **103**(1), 275–82.

Hummel, M., Bentink, S., and Berger, H. *et al.* (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N Engl J Med*, **354**(23), 2419–30.

Jaffe, E., Harris, N., Stein, H., and Vardiman, J., editors (2001). *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues*. IARC Press, Lyon, France.

Jones, J., Otu, H., and Spentzos, D. *et al.* (2005). Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res*, **11**(16), 5730–9.

Markowetz, F., Kostka, D., and Troyanskaya, O. *et al.* (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, **23**(13), i305–12.

Natanzon, A., Shamir, R., and Sharan, R. (1999). Complexity classification of some edge modification problems. In *Workshop on Graph-Theoretic Concepts in Computer Science*, pages 65–77.

Pinkel, D., Segraves, R., and Sudar, D. *et al.*. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20**(2), 207–11.

Pollack, J., Perou, C., and Alizadeh, A. *et al.* (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, **23**(1), 41–6.

Rosenwald, A., Wright, G., and Chan, W. *et al.* (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med*, **346**(25), 1937–47.

Solinas-Toldo, S., Lampel, S., and Stilgenbauer, S. *et al.* (1997). Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**(4), 399–407.

van de Vijver, M., He, Y., and van't Veer, L. *et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, **347**(25), 1999–2009.

Zahn, C. (1964). Approximation symmetric relations by equivalence relations. *Journal of the Soceity for Industrial and Applied Mathematics*, **12**(4), 840–47.