International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:2, 2007

Ranking Genes from DNA Microarray Data of Cervical Cancer by a local Tree Comparison

Frank Emmert-Streib, Matthias Dehmer, Jing Liu, Max Mühlhäuser

***Abstract*— The major objective of this paper is to introduce a new method to select genes from DNA microarray data. As criterion to select genes we suggest to measure the local changes in the correlation graph of each gene and to select those genes whose local changes are largest. More precisely, we calculate the correlation networks from DNA microarray data of cervical cancer whereas each network represents a tissue of a certain tumor stage and each node in the network represents a gene. From these networks we extract one tree for each gene by a local decomposition of the correlation network. The interpretation of a tree is that it represents the n-nearest neighbor genes on the n'th level of a tree, measured by the Dijkstra distance, and, hence, gives the local embedding of a gene within the correlation network. For the obtained trees we measure the pairwise similarity between trees rooted by the same gene from normal to cancerous tissues. This evaluates the modification of the tree topology due to tumor progression. Finally, we rank the obtained similarity values from all tissue comparisons and select the top ranked genes. For these genes the local neighborhood in the correlation networks changes most between normal and cancerous tissues. As a result we find that the top ranked genes are candidates suspected to be involved in tumor growth. This indicates that our method captures essential information from the underlying DNA microarray data of cervical cancer.**

***Keywords*— Graph similarity, generalized trees, graph alignment, DNA microarray data, cervical cancer.**

I. Introduction

COMPARING structured objects such as graphs and trees is a difficult and still outstanding problem. Traditional investigations dealing with distances between graphs are based on isomorphic relations and subgraph isomorphism [6], [11], [15], respectively. An example of such a graph distance is the well-known ZELINKA-distance [18]. The ZELINKA-distance is based on the principle that two graphs are more similar, the bigger the common induced isomorphic subgraph is. ZELINKA was the first who introduced this measure for unlabeled graphs. SOBIK [13], [14] and KADEN [6], [7] generalized this measure for arbitrary graphs, which includes also labeled graphs, of different order and proved that it is a metric.

This paper continues our work started in [4]. There we demonstrated that correlation networks obtained from DNA microarray experiments from cervical cancer of different tumor stages can be classified by a binary graph classifier (BGC) introduced in [4]. These results demonstrated, that the information captured by the DNA microarray experiments is sufficient to differentiate the biologically different tissue stages solely based on the correlation networks extracted form these data. This extends recent finding by GOLUB et al. [5] who demonstrated, that cancer can be classified on a molecular level, however, applying different theoretical methods, which do not involve the description in terms of networks. In this work we will investigate the question: Which genes contribute most to the classification on a network level? For this reason, we introduce a new method for gene ranking. The gene ranking method is based on the comparison of generalized trees, which are locally extracted from the correlation network obtained for each disease stage. More precisely, we determine which local neighborhood of a gene, represented by its corresponding tree, changes most in the correlation network during progression of cancer. For our study we use the data from WONG et al. [17] about cervical cancer. This paper is organized in the following way: In the next section we describe the generalized trees-similarity algorithm (GTSA) to measure the similarity of generalized trees. In section III we present a method to decompose a network locally in generalized trees. We apply these methods in the results section IV to determine the gene ranking for genes from DNA microarray experiments of cervical cancer. The article finishes with a discussion of our obtained results.

Frank Emmert-Streib is with the Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA, e-mail: fes@stowers-institute.org. Matthias Dehmer is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: dehmer@informatik.tu-darmstadt.de. Jing Liu is with the Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, MO 64110, USA, e-mail: jil@stowers-institute.org. Max Mühlhäuser is with the Technische Universität Darmstadt, 64289 Darmstadt, Germany, e-mail: max@informatik.tu-darmstadt.de.

II. Similarity measure of generalized trees

In this section we introduce a similarity measure which operates on a special class of graphs: *unlabeled, hierarchical, and directed graphs*. EMMERT-STREIB et al. [4] called these graphs *generalized trees*, because this graph class generalizes normal trees in the sense that, e.g., connections are allowed that jump over more than one level. In this paper we call the underlying algorithm for measuring the structural similarity of generalized trees the *generalized tree-similarity algorithm* (GTSA) [4]. DEHMER et al. [2] presented an overview of graph similarity measures and the mathematical motivation of the similarity measure in detail. The main idea is based on the derivation of property strings for each generalized tree and then to align the property strings representing the trees by a sequence alignment technique based on *dynamic programming* [1]. From the resulting alignment we obtain a value of the scoring function which is minimized during the alignment process. The similarity of two generalized trees will be expressed by a natural cumulation of local similarity functions which weighs two types of alignments: *out-degree* and *in-degree* alignments.

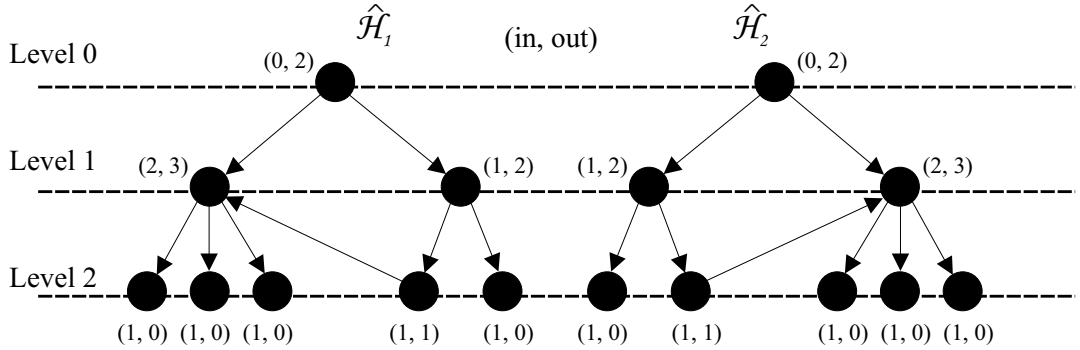

Fig. 1. Shown are two generalized trees $\hat{\mathcal{H}}_1$ and $\hat{\mathcal{H}}_2$ with their property stings. For example, the property string in terms of in-degrees of $\hat{\mathcal{H}}_1$ on level 1 is "2 ∘ 1". Or the out-degrees of $\hat{\mathcal{H}}_2$ on level 2 are "0 ∘ 1 ∘ 0 ∘ 0 ∘ 0".The symbol ∘ denotes usual string concatenation.

Now we are looking for structural characteristics of our generalized trees which are suitable for the definition of a meaningful similarity measure. If we choose degree sequence vectors [2] we see immediately that e.g., simple comparisons of such degree sequences cannot describe the topology of our graphs completely. Since we are examining hierarchical graphs, we take a closer look at the out-degree and in-degree sequences, induced by the vertex sequences $v_{i,1}, v_{i,2}, \ldots, v_{i,\sigma_i}$ and their edge relations (see Figure (1)). If we define the vertex set as

$$\hat{V} := \{v_{0,1}, v_{1,1}, v_{1,2}, \ldots, v_{1,\sigma_1}, v_{2,1}, v_{2,2}, \ldots, v_{2,\sigma_2}, \ldots, v_{h,1}, v_{h,2}, \ldots, v_{h,\sigma_h}\} \quad (1)$$

note that σ_i is maximal in the sense that there is no other vertex sequence such that $v_{i,1}, v_{i,2}, \ldots, v_{i\hat{\sigma}_i}$ with $\hat{\sigma}_i > \sigma_i$. h denote the maximal length of a path from the root $v_{0,1}$ to a leaf. Now, for determining the structural similarity of generalized trees it holds: the more similar the out-degree and in-degree sequences on the levels $i, 0 \leq i \leq h$ are, the more similar is the common structure of the generalized trees, with respect to a *cost function* α. Define $w_k^{\hat{\mathcal{H}}^k} := v_{0,1}^{\hat{\mathcal{H}}^k}, k \in \{1,2\}$, and let $\hat{\mathcal{H}}^1$ be a given graph and $v_{i,j}^{\hat{\mathcal{H}}^1}$, $0 \leq i \leq h_1$, $1 \leq j \leq \sigma_i$ denotes the j-th vertex on the i-th level of $\hat{\mathcal{H}}^1$, analogous to $v_{i,j}^{\hat{\mathcal{H}}^2}$ for $\hat{\mathcal{H}}^2$. As mentioned above, the task of measuring the structural similarity between $\hat{\mathcal{H}}^1$ and $\hat{\mathcal{H}}^2$ is equivalent to determining the optimal alignment of

$$S_1 := v_{0,1}^{\hat{\mathcal{H}}^1} \circ v_{1,1}^{\hat{\mathcal{H}}^1} \circ v_{1,2}^{\hat{\mathcal{H}}^1} \circ \cdots \circ v_{h_1,\sigma_{h_1}}^{\hat{\mathcal{H}}^1}, \quad (2)$$

$$S_2 := v_{0,1}^{\hat{\mathcal{H}}^2} \circ v_{1,1}^{\hat{\mathcal{H}}^2} \circ v_{1,2}^{\hat{\mathcal{H}}^2} \circ \cdots \circ v_{h_2,\sigma_{h_2}}^{\hat{\mathcal{H}}^2}, \quad (3)$$

with respect to a *cost function* α. $S_k[i]$ denotes the i-th position of the sequence S_k and it holds $S_1[n] = v_{h_1,\sigma_{h_1}}^{\hat{\mathcal{H}}^1}, S_2[m] = v_{h_2,\sigma_{h_2}}^{\hat{\mathcal{H}}^2}$, $\mathbb{N} \ni n, m \geq 1$, $S_k[1] = w_k^{\hat{\mathcal{H}}_k}$, $k \in \{1,2\}$. The algorithm for finding the optimal alignment of S_1 and S_2 generates a matrix $(\mathcal{M}(i,j))_{ij}, 0 \leq i \leq n, 0 \leq j \leq m$. Hence, its complexity is $O(|\hat{V}_1| \cdot |\hat{V}_2|)$. We express the optimal alignment on the basis of the following algorithm [2]:

$$\begin{aligned}
\mathcal{M}(0,0) &:= 0, \\
\mathcal{M}(i,0) &:= \mathcal{M}(i-1,0) + \alpha(S_1[i], -) \ : \ 1 \leq i \leq n, \\
\mathcal{M}(0,j) &:= \mathcal{M}(0,j-1) + \alpha(-, S_2[j]) \ : \ 1 \leq j \leq m,
\end{aligned}$$

and

$$\mathcal{M}(i,j) := \min \begin{cases} \mathcal{M}(i-1,j) + \alpha(S_1[i], -) \\ \mathcal{M}(i,j-1) + \alpha(-, S_2[j]) \\ \mathcal{M}(i-1,j-1) + \alpha(S_1[i], S_2[j]) \end{cases} \quad (4)$$

for $1 \leq i \leq n$, $1 \leq j \leq m$. Within the GTSA the alignments have both global and local significance. First, the sequence alignments will be implemented in a global sense, to compute the optimal alignment between the sequences S_1 and S_2. For this reason we now express the definition of a distance measure.

Definition 2.1: Let X be a arbitrary set. A positive real valued function $\omega : X \times X \longrightarrow [0,1]$ is called distance measure, if

$$\omega(x,y) = \omega(y,x) \quad \forall\, x, y \in X \quad (5)$$

$$\omega(x,x) = 0 \quad \forall\, x \in X \quad (6)$$

If we set

$$\omega(x,y) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\sigma^2}} \quad (7)$$

we obtain immediately

Lemma 2.1: Let $\omega : \mathbb{R} \times \mathbb{R} \longrightarrow [0,1]$. If we define ω by, $\omega(x,y) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\sigma^2}}$, then ω is a distance measure.

Proof: From the definition of $\omega(x,y)$ we infer $\omega(x,y) \in [0,1]$, $\forall\, x, y \in \mathbb{R}$ and $\omega(x,x) = 1 - 1 = 0$, $\forall\, x \in \mathbb{R}$. Since $(x-y)^2 = (y-x)^2$, $\forall\, x, y \in \mathbb{R}$, the symmetry condition holds. ■

Now, we define

$$\alpha^{out}\left(v_{i_1,j_1}^{\hat{\mathcal{H}}^1}, v_{i_2,j_2}^{\hat{\mathcal{H}}^2}\right) := \omega^{out}\left(\delta_{out}(v_{i_1,j_1}^{\hat{\mathcal{H}}^1}), \delta_{out}(v_{i_2,j_2}^{\hat{\mathcal{H}}^2}), \sigma_{out}^1\right)$$

if $i_1 = i_2$ and

$$\alpha^{out}\left(v_{i_1,j_1}^{\hat{\mathcal{H}}^1}, v_{i_2,j_2}^{\hat{\mathcal{H}}^2}\right) := +\infty$$

else, for $0 \leq i_k \leq h_k$, $1 \leq j_k \leq \sigma_{i_k}$, $k \in \{1,2\}$, where

$$\omega^{out}(x,y,\sigma_{out}^k) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\left(\sigma_{out}^k\right)^2}} \quad (8)$$

with $x, y, \sigma_{out}^k \in \mathbb{R}$ and

$$\alpha^{out}\left(v_{i,j_1}^{\hat{\mathcal{H}}^1}, -\right) := \omega^{out}\left(\delta_{out}(v_{i,j_1}^{\hat{\mathcal{H}}^1}), \xi, \sigma_{out}^2\right), \quad (9)$$

$$\alpha^{out}\left(-, v_{i,j_2}^{\hat{\mathcal{H}}^2}\right) := \omega^{out}\left(\xi, \delta_{out}(v_{i,j_2}^{\hat{\mathcal{H}}^2}), \sigma_{out}^2\right). \quad (10)$$

$\xi > 0$ prevents an alignment between two leaves being better evaluated as an alignment between a leaf and a gap ('-'). With

$$\omega^{in}(x, y, \sigma_{in}^k) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\left(\sigma_{in}^k\right)^2}} \tag{11}$$

we define analogously $\alpha^{in}\left(v_{i_1,j_1}^{\hat{\mathcal{H}}^1}, v_{i_2,j_2}^{\hat{\mathcal{H}}^2}\right)$, $\alpha^{in}\left(v_{i,j_1}^{\hat{\mathcal{H}}^1}, -\right)$ and $\alpha^{in}\left(-, v_{i,j_2}^{\hat{\mathcal{H}}^2}\right)$. Second, the alignments will be evaluated on the levels of the generalized trees. For the evaluating of the alignments on each level, we set

$$\text{align}\left(v_{i,j_1}^{\hat{\mathcal{H}}^1}\right) := \begin{cases} v_{i,j_2}^{\hat{\mathcal{H}}^2} & : \quad \text{align}^{-1}\left(v_{i,j_2}^{\hat{\mathcal{H}}^2}\right) = v_{i,j_1}^{\hat{\mathcal{H}}^1} \\ - & : \quad else. \end{cases}$$

This mapping determines for a vertex $v_{i,j_1}^{\hat{\mathcal{H}}^1}$ the vertex $v_{i,j_2}^{\hat{\mathcal{H}}^2}$ during the traceback [2]. Furthermore we state

$$\gamma_{\hat{\mathcal{H}}^k}^{out}(i) := \frac{\sum_{j=1}^{\sigma_i^k} \hat{\alpha}_{out}\left(v_{i,j}^{\hat{\mathcal{H}}^k}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^k}\right)\right)}{\sigma_i^k}, \tag{12}$$

$$\gamma_{\hat{\mathcal{H}}^k}^{in}(i) := \frac{\sum_{j=1}^{\sigma_i^k} \hat{\alpha}_{in}\left(v_{i,j}^{\hat{\mathcal{H}}^k}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^k}\right)\right)}{\sigma_i^k}, \tag{13}$$

$k \in \{1, 2\}$, which are similarity values for out-degree and in-degree alignments. Finally, if we define the functions $\hat{\alpha}_{out}$ and $\hat{\alpha}_{in}$ in the same way as α_{out} and α_{in}, we obtain the normalized and cumulative functions

$$\begin{aligned} \gamma^{out}(i, \hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2) := 1- \\ \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^1} \hat{\alpha}^{out}\left(v_{i,j}^{\hat{\mathcal{H}}^1}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^1}\right)\right) \right\} \\ + \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^2} \hat{\alpha}^{out}\left(v_{i,j}^{\hat{\mathcal{H}}^2}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^2}\right)\right) \right\} \end{aligned} \tag{14}$$

and

$$\begin{aligned} \gamma^{in}(i, \hat{\sigma}_{in}^1, \hat{\sigma}_{in}^2) := 1- \\ \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^1} \hat{\alpha}^{in}\left(v_{i,j}^{\hat{\mathcal{H}}^1}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^1}\right)\right) \right\} \\ + \frac{1}{\sigma_i^1 + \sigma_i^2} \cdot \left\{ \sum_{j=1}^{\sigma_i^2} \hat{\alpha}^{in}\left(v_{i,j}^{\hat{\mathcal{H}}^2}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^2}\right)\right) \right\} \end{aligned} \tag{15}$$

which detect the similarity of an out-degree and in-degree alignment on a level i. $\hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2$ and $\hat{\sigma}_{in}^1, \hat{\sigma}_{in}^2$ are the parameters of $\hat{\alpha}^{out}$ and $\hat{\alpha}^{in}$, respectively. For constructing the final similarity measure d with respect to our trees we need a the definition of a special kind of similarity measures.

Definition 2.2: Let U be a set of units and a mapping $\phi : U \times U \longrightarrow [0, 1]$. We call ϕ a backward similarity measure if it satisfies the conditions

$$\phi(u, v) = \phi(v, u), \forall\, u, v \in U \tag{16}$$

and

$$\phi(u, u) \geq \phi(u, v), \forall\, u, v \in U. \tag{17}$$

Now, we state our key result which has been proven in [2].

Theorem 2.1: Let $\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2$ be two generalized trees with $0 \leq i \leq \rho$, $\rho := \max(h_1, h_2)$.

$$d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2) := \frac{(\rho + 1)}{\sum_{i=0}^{\rho} \gamma^{fin}(i, \hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2, \hat{\sigma}_{in}^1, \hat{\sigma}_{in}^2)} \cdot \prod_{i=0}^{\rho} \gamma^{fin}(i, \hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2, \hat{\sigma}_{in}^1, \hat{\sigma}_{in}^2), \tag{18}$$

is a backward similarity measure, where γ^{fin} is defined as

$$\begin{aligned} \gamma^{fin} &= \gamma^{fin}(i, \hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2, \hat{\sigma}_{in}^1, \hat{\sigma}_{in}^2) \\ &:= \zeta \cdot \gamma^{out} + (1 - \zeta) \cdot \gamma^{in} \end{aligned} \tag{19}$$

with $\zeta \in [0, 1]$.

The similarity measure $d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2)$ has the following three properties:

$$d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_1) = 1 \tag{20}$$

$$d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2) = d(\hat{\mathcal{H}}_2, \hat{\mathcal{H}}_1) \tag{21}$$

$$0 \leq \quad d(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2) \leq 1 \tag{22}$$

which has been proven in [2].

Finally, we want to mention that the GTSA presented in this section is of course also able to measure the similarity between two (normal) trees, because the class of (normal) trees is a subclass of the graph class of generalized trees. This is important for the following section, because the trees extracted locally from a network are (normal) trees.

III. Decomposing a graph locally in trees

In Section (II) we introduced a method to measure the similarity between a pair of generalized trees. The correlation graphs we are dealing with in the following are unlabeled, unweighted and undirected, hence, we simply call them graphs or networks because no special assumptions on these objects are necessary. Because we can only compare generalized trees and not graphs directly we give here a method which decomposes a graph locally in trees. This decomposition will now be described in detail.

Definition 3.1: A graph G with N nodes can be locally decomposed in a set of trees by the following algorithm: Label all nodes from 1 to N. These labels form the label set $L_S = \{1, \ldots, N\}$. Choose a desired depth of the trees D. Choose an arbitrary label from L_S, e.g., i. The node with this label is the root node of a tree.

1) Calculate the shortest distance from node i to all other nodes in the graph G, e.g. by the algorithm of Dijkstra [3].
2) The nodes with distance k are the nodes in the k'th level of the tree. Select all nodes of the graph up to distance D, including the connections between the nodes. Connections to nodes with distance $> D$ are deleted.
3) Delete the label i from the label set L_S.
4) Repeat this procedure if L_S is not empty by choosing an arbitrary label from L_S, otherwise terminate.

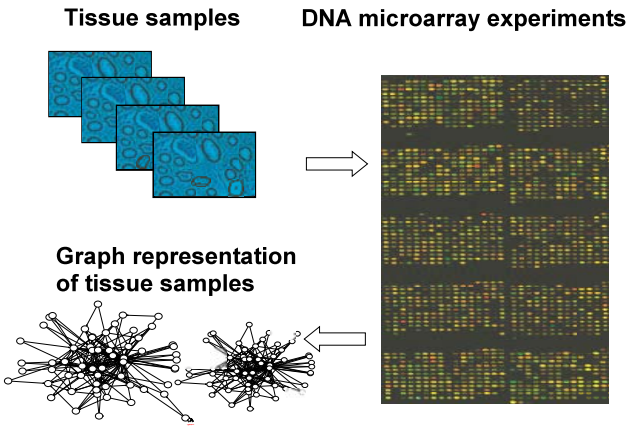

Fig. 2. Schematic representation of the transition from tissue samples of cervical cancer via DNA microarray experiments to the representation as directed, unweighted graph. The graphs in this figure were generated by PAJEK [10].

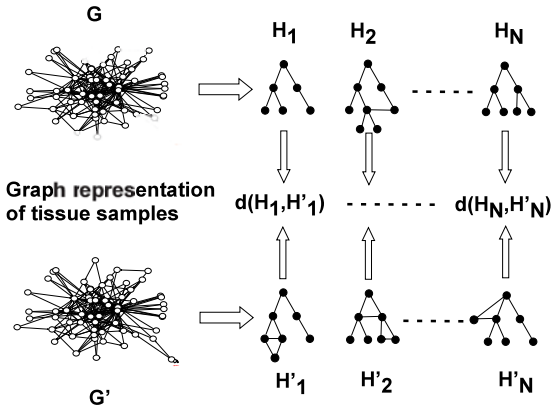

Fig. 3. Comparing two graphs G and G' with N nodes each by comparing locally generalized trees. Only two trees H_i and H'_i resulting from the same node i in the graphs, corresponding to gene i, are compared $d(H_i, H'_i)$. The graphs in this figure were generated by PAJEK [10].

IV. RESULTS

In this section we present results of the application of our method to the DNA microarray data from WONG et al. [17]. They investigated the gene expression levels of different tumor stages of cervical cancer. For a summary of their data set see table I. In general, the higher the integer numbers and the letters of the tumor stages are the more the cancer has grown and spread. The data include also a normal expression profile of cervical tissue indicated in table II as 'normal'. In the following we speak of the network resulting, e.g., form the expression profile of tumor tissue of stage 2A, as the 2A-network, G_{2A}. Similarly, we speak of the 2A-tree set, S_{2A}.

We calculate the correlation networks of the expression data from the DNA microarray experiments by a three step process suggested by Rougemont et al. [12]:

1) Calculate the pairwise correlation coefficient for all gene profiles.
2) Prune the connections if the correlation coefficient is below a threshold Θ_{Co}.
3) Prune the connections to a node i if its clustering coefficient is below a threshold Θ_{Cl}.

Figure 2 shows schematically the overall idea of our approach. We obtain from the DNA microarray data for the tissue samples, representing one tissue type, e.g., tissue of stage 2A, one graph by applying the three step process by Rougemont et al. [12]. That means, the four different tissue types given in table I are transformed to four different unweighted, undirected graphs. Hence, we represent a tissue of a tumor stage as a graph. These four graphs form the starting point of our theoretical analysis.

TABLE I
MICROARRAY DATA FROM [17] FOR DIFFERENT TUMOR STAGES, BASED ON THE FIGO (INTERNATIONAL FEDERATION OF GYNECOLOGISTS AND OBSTETRICS) TAGING SYSTEM, OF CERVICAL CANCER. EACH OF THE 32 (TOTAL NUMBER OF PATIENTS) ARRAYS CONTAINED 10692 GENES.

FIGO stage	Number of patients
normal	8
1B	11
2A	8
2B	5

The size of the DNA microarrays used for each experiment in [17] consisted in a total number of 10692 genes. Hence, the networks have this number of nodes. Via the local tree decomposition algorithm in definition 3.1 we obtain tree sets for all graphs consisting of 10692 trees each. We calculate for all pairs of normal and cancerous tissue the pairwise tree similarity with the generalized trees-similarity algorithm (GTSA) explained in section II. More precisely, we calculate the similarity between tree i from the S_{normal} tree set with tree i from, e.g., tree set $S_{1\text{B}}$. That means, we do only compare the trees originating from the same root node in the correlation graph that corresponds to the same gene. Figure 3 shows our approach schematically. Due to the fact, that all graphs have the same number of nodes, corresponding to the number of genes in the DNA microarray experiments, we can ask the question - how much did the graph change? Here the change referes always to the graph representing normal cervical tissue which serves as reference. More precisely, we can ask - how much did, e.g., the 2A graph change compared with the normal graph? We suggest to answer this question locally, based on the similarity of generalized trees. The application of the local tree decomposition algorithm from the previous section gives us N trees for each graph, because this corresponds to the number of nodes in the graph, and, hence, the comparison of two graphs results in N local similarity values $d(\hat{H}_i, \hat{H'}_i)$ for $i \in \{1, \ldots, N\}$. The obtained similarity values are then rank-ordered in decreasing order of similarity values. Hereof, we calculate the overall rank-order resulting from the linear ranking of all three possible tissue pairs between normal and cancerous tissues. This ranking provides averaged information about the genes which changed most from normal to cancerous tissues.

In Fig. 4 we show in a semi-logarithmic plot the rank-ordered similarity values that result from a comparison bet-

TABLE II
GENES WHICH HAVE BEEN FOUND AMONG THE TOP 100 ($< 1\%$ OF ALL GENES) GENE RANKING LIST. THE GENE ID CORRESPONDS TO THE ENUMERATION OF GENES OF THE DATA [17] PROVIDED AT THE NCBI HOMEPAGE.

Gene ID	Accession no.[1]	Gene name
3640	AA434373	E74-like factor 3 (ets domain transcription factor, epithelial-specific)
3082	AA709143	transcription termination factor, RNA polymerase I
778	R19406	ESTs (Weakly similar to A47582 B-cell growth factor precursor [H.sapiens])
320	T51538	sortilin-related receptor, L(DLR class) A repeats-containing
1020	N20335	clathrin, light polypeptide (Lcb)
2978	AA916327	protective protein for beta-galactosidase (galactosialidosis)
6523	AA195002	myosin 5C
1923	H56944	splicing factor, arginine/serine-rich 11
958	T65211	SFRS protein kinase 2
65	N95249	v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog
1503	AI365523	synovial sarcoma, translocated to X chromosome 298
2710	N54456	ALEX3 protein
194	AA287323	xeroderma pigmentosum, complementation group C
5381	AA256502	proprotein convertase subtilisin/kexin type 5
131	AA455955	proprotein convertase subtilisin/kexin type 7
1576	AI309770	ubiquitin-activating enzyme E1C (homologous to yeast UBA3)
1392	AA521339	chimerin (chimaerin) 2 181

[3] Accession numbers in the GenBank database.

ween normal and 1B tissue (full line) and between normal and 2B tissue (dashed line). One can clearly see, that around position 1000 there is a drop in the similarity values indicating that these tree pairs below this position are quite unsimilar. In Fig. 5 we show a more detailed semi-logarithmic plot by presenting only the first 1000 rank-ordered genes. From these curves it is plausible to select only the top 100 ranked genes. This demonstrated, that a small number of genes change its local environment in the correlation networks more than all others. A more statistical argument can be obtained by calculating the average rank-values for a completely random selection of genes. This result is shown in Fig. 6. In this figure one can see, that less than 2000 genes have values which are less than the average rank-value of 5000. This confirms our visual estimate given above. This result confirms approaches trying to find some marker genes that indicate the onset or progression of a disease [9], [16] in contrast to monitor a large number of genes as indicator.

Some genes that have been found among the top 100 ranked

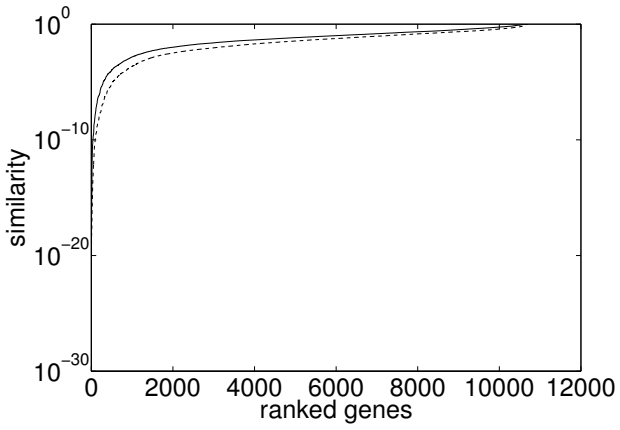

Fig. 4. Rank-ordered similarity values of tree pairs between normal and 1B tissue (full line) and between normal and 2B tissue (dashed line) for all 10692 genes.

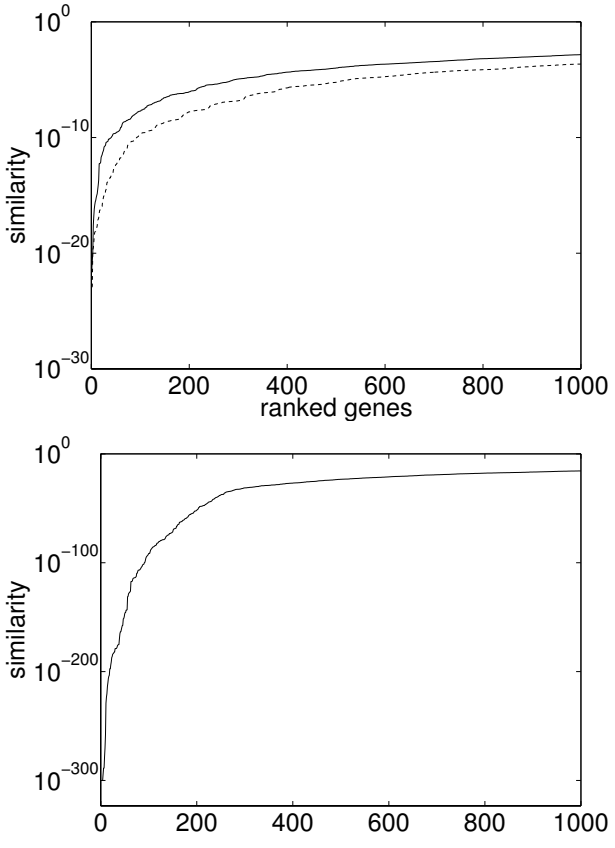

Fig. 5. Rank-ordered similarity values of tree pairs for the first 1000 genes. Upper figure: Between normal and 1B tissue (full line) and between normal and 2B tissue (dashed line). These curves are a magnification of Fig. 4. Lower figure: Between normal and 2A tissue.

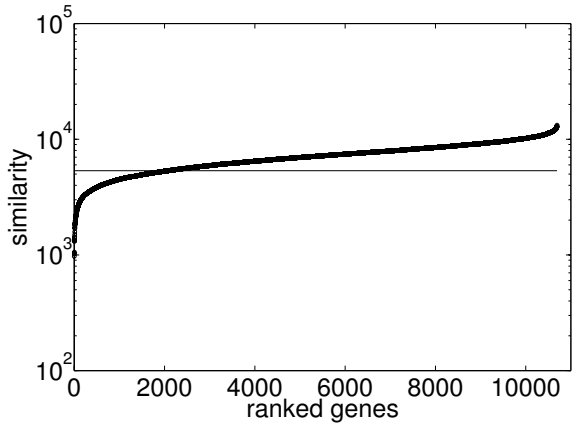

Fig. 6. Averaged rank-ordered similarity values for the three graph comparisons (bold line). The horizontal line corresponds to the average rank-value obtained by a random selection of N genes.

genes are shown in table II. These genes correspond to the trees, which differ most from normal to cancerous tissues. We found by our gene ranking method genes, which are involved in transcription (3640,3082), growth factors (778), cell signaling (1392), endocytosis (320,1020, 2978,6523), post translation regulation (1923,958) and cancer related genes (65, 1503,2710,194) to name only some given in table II[2]. All these genes are expected to be involved in tumor growth. Moreover, our list of genes is qualitatively comparable to the gene list compiled by Wong et al. [17]. This indicates, that our method is appropriate to select relevant genes involved in the progression of cervical cancer.

V. Conclusions

We introduced in this paper a method for gene ranking from DNA microarray data and presented first results for expression data of Wong et al. [17] of cervical cancer. Our method is based on a correlation graph that can be calculated from the data representing tissue of a tumor stage. From these networks we extract one tree for each gene by a local decomposition from the correlation network. For the obtained trees we measure the pairwise similarity between trees rooted by the same gene from normal to cancerous tissues. This evaluates the modification of the tree topology due to progression of the tumor. Finally, we rank and average the obtained similarity values from all tissue comparisons and select the top ranked genes. For these genes the local neighborhood in the correlation networks changes most between normal and cancerous tissues. As a result we found genes that are suspected to be involved in tumor growth, e.g., genes involved in transcription, growth factors, cell signaling, endocytosis, post translation regulation and cancer related genes. These are promising results that indicate that our method captures essential information from the underlying DNA microarray data of tissues from different tumor stages.

[2]The number in brackets gives the gene ID in the first column in the results table II.

In future work we will continue demonstrating that our approach to select genes from DNA microarray data based on a similarity ranking of a tree comparisons is capable to uncover biological information by applying our method to various data sets from DNA microarray data of cancer experiments. We think that non-monogenetic diseases as cancer are more likely to be understood by the application of a method which is based on a systems view as ours, because such methods can detect changes in activity patterns of interconnected genes rather than just alterations in a the activity of a single gene. We hope that our work can contribute to unravel the molecular mechanisms of cancer in the long run and by this provide a base for a better treatment of this disease.

Acknowledgments

We would like to thank Galina V. Glazko and Chris Seidel for fruitful discussions and Mike Coleman and Daniel Thomasset for computer support.

References

[1] R. Bellman, *Dynamic Programming*. Princeton University Press, 1957.
[2] M. Dehmer, *Strukturelle Analyse web-basierter Dokumente*, Ph.D Thesis, Department of Computer Science, Technische Universität Darmstadt, 2005.
[3] E. W. Dijkstra, *A note on two problems in connection with graphs*. Numerische Math., Vol. 1, 1959, 269–271.
[4] F. Emmert-Streib., M. Dehmer, J. Kilian: *Classification of large Graphs by a local Tree decomposition*, Proceedings of the 2005 International Conference on Data Mining (DMIN'05), Editors: H.R. Arabnia, A. Scime (2005) 200–207.
[5] T. R. Golub et.al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science, Vol. 286, 1999, 531–537.
[6] F. Kaden, *Graphmetriken und Distanzgraphen*. ZKI-Informationen, Akad. Wiss. DDR, Vol. 2 (82), 1982, 1–63.
[7] F. Kaden, *Graph metrics and distance-graphs*. In: Graphs and other Combinatorial Topics, ed. M. Fiedler, Teubner Texte zur Math., Leipzig, Vol. 59, 1983, 145–158.
[8] P. J. Kraulis, *Molscript: A Program to Produce Both detailed and schematic plots of protein structures*. Journal of Applied Crystallography, Vol. 24, 1991, 946–950.
[9] K.Mori et al., *Highly specific marker genes for detecting minimal gastric cancer cells in cytology negative peritoneal washings*, Biochem. Biophys. Res. Commun. 23;313(4):931-937 (2004).
[10] V. Batagelj and A. Mrvar, *Pajek - Program for Large Network Analysis*, Connections 21:47-57 (1998).
[11] R. C. Read and D. G. Corneil, *The graph isomorphism disease*. Journal of Graph Theory, Vol. 1, 1977, 339–363.
[12] J. Rougemont and P. Hingamp, *DNA microarray data and contextual analysis of correlation graphs*. BMC Bioinformatics, Vol. 4, 2003, 4–15.
[13] F. Sobik, *Graphmetriken und Klassifikation strukturierter Objekte*. ZKI-Informationen, Akad. Wiss. DDR, Vol. 2 (82), 1982, 63–122.
[14] F. Sobik, *Graphmetriken und Charakterisierung von Graphklassen*. 27. Internat. Wiss. Koll., TH-Ilmenau, Vol. 2 (82), 1982, 63–122.
[15] J. R. Ullman, *An algorithm for subgraph isomorphism*. J. ACM, Vol. 23 (1), 1976, 31–42.
[16] Y. Wang et al., *Gene expression profiles and molecular markers to predict recurrence of Dukes'B colon cancer*, J. Clin. Oncol. 1;22(9):1564-1571 (2004).
[17] Y. F. Wong et.al. *Expression Genomics of Cervical Cancer: Molecular Classification and Prediction of Radiotherapy Response by DNA Microarray*. Clinical Cancer Research, Vol. 9, 2003, 5486–5492.
[18] B. Zelinka, *On a certain distance between isomorphism classes of graphs*. Časopis pro pěst. Mathematiky, Vol. 100, 1975, 371–373.