A Systems Biology approach for the classification of DNA Microarray Data

Frank Emmert-Streib[1] and Matthias Dehmer[2]

[1] Stowers Institute for Medical Research,
1000 E. 50th Street, Kansas City, MO 64110, USA,
fes@stowers-institute.org
[2] Technische Universität Darmstadt, 64289 Darmstadt, Germany
dehmer@informatik.tu-darmstadt.de

Abstract. In this paper we present a binary graph classifier (BGC) which allows to classify large, unweighted, undirected graphs. The main idea of this classifier is to decompose a graph locally in generalized trees forming the tree set of a graph and to compare the tree sets of graphs by a generalized tree-similarity algorithm (GTSA). We apply our BGC to networks representing co-expressed genes from DNA microarray experiments of cervical cancer and demonstrate, that different tumor stages of the disease can be distinguished on this level of description.

1 Introduction

The structural comparison and classification of graphs and networks is an important and still outstanding problem if the size of these objects becomes large in the number of nodes and edges. Traditional investigations dealing with distances of graphs are based on isomorphic relations and subgraph isomorphism [13, 17], respectively. An example of such a graph distance is the well-known Zelinka-distance [20]. The Zelinka-distance is based on the principle that two graphs are more similar, the bigger the common induced isomorphic subgraph is. In other words, graphs which have a large common induced isomorphic subgraph have a small distance and vice versa. Zelinka was the first to introduce this measure for unlabeled graphs. Sobik [15, 16] and Kaden [8, 9] generalized this measure for arbitrary (also labeled) graphs of different order and proved that it is a metric.

In this work we present an extention of our recently introduced binary graph classifier (BGC) [6] which allows to classify large, unweighted, undirected graphs. This classifier is based on a local decomposition of the graph for each node in generalized trees, which are *unlabeled, hierarchical, and directed graphs* [12]. The main idea of this similarity measure is based on the derivation of property strings for each generalized tree and then to align the property stings representing the trees by a *dynamic programming* [1] technique. From the resulting alignment one obtains a value of the scoring function which is minimized during the alignment process. The similarity of two generalized trees will be expressed by a cumulation

of local similarity functions which weighs two types of alignments: *out-degree* and *in-degree* alignments on a generalized tree level. We call this method the generalized tree-similarity-algorithm (GTSA). The obtained trees, forming the tree set of the graph, are then pairwise compared by the GTSA and the resulting similarity scores determine a characteristic similarity distribution of the graph. Classification in this context is defined as mutual consistency for all pure and mixed tree sets and their resulting similarity distributions in a graph class.

An application for our BGC is provided by the data from DNA microarray experiments. This technique allows the monitoring of the expression level of thousands of genes simultaneously. Based on the obtained data it is possible to reconstruct the underlying regulatory network representing co-expressed genes. With respect to the classification of tumors of different disease stages, one could hypothesis, that the network of co-expressed genes should reflect the molecular modifications due to the disease. This can hold of course only for diseases which classification is possible on a molecular level measured by DNA microarrays. The work by Golub et.al. [7] showed, that this classification is possible for cancer for acute mzeloid leukemia and acute lymphoblastic leukemia. Recent publications by Lapoointe et.al [11] and van't Veer et.al. [18] confirmed this result for prostate and breast cancer. The results of [7, 11, 18] demonstrate only, that the molecular level measured by DNA microarray experiments seems to be appropriate for the distinction of different cancer stages. They do not compare the networks of co-expressed genes itself but select a subgroup of genes according to some criteria and judge based on these feature vectors. See e.g. [3] for a description of the technical details for some commonly used methods. In this work we make an attempt to investigate the question, if the distinction of cancer of different stages can be made on the level of co-expressed gene networks by comparing these networks. Hence, we are utilizing a view from system theory [2] by treating the system under investigation as a whole, in our case as a network.

This paper is organized in the following way. First, we explain in section (2) the construction of the similarity measure and the generalized tree-similarity-algorithm (GTSA). Then we show in section (3) how we use the GTSA to define a binary graph classifier (BGC) for undirected, unweighted graphs. In section (4) we present results for the classification of networks representing different stages of cervical cancer. For our analysis we use the data from Wong et.al. [19].

2 Comparison of generalized trees

In this section we give a short explanation of the generalized tree-similarity-algorithm (GTSA) we introduced in [4].

In figure 1 we depicted an example of two generalized trees. One can clearly see, that this is a hierarchical, directed graph. We represent each tree $\hat{\mathcal{H}}^k$ as string sequences $S_i^{\hat{\mathcal{H}}^k}$ for each level i:

$$S_0^{\hat{\mathcal{H}}^1} := v_0^{\hat{\mathcal{H}}^1}, \tag{1}$$

$$S_1^{\hat{\mathcal{H}}^1} := v_{1,1}^{\hat{\mathcal{H}}^1} \circ v_{1,2}^{\hat{\mathcal{H}}^1} \circ \cdots \circ v_{i,\sigma_1^{\hat{\mathcal{H}}^1}}^{\hat{\mathcal{H}}^1}, \tag{2}$$

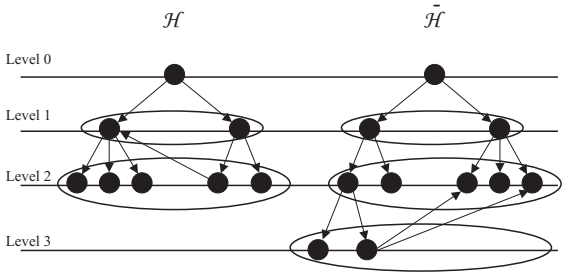

Fig. 1. Two simple examples of generalized trees. Each vertex sequence of the form $v_{i,1}, v_{i,2}, \ldots, v_{i,\sigma_i}$ induces sequences of out- and in-degrees.

$$\vdots \tag{3}$$

$$S_{h_1}^{\hat{\mathcal{H}}^1} := v_{h_1,1}^{\hat{\mathcal{H}}^1} \circ v_{h_1,2}^{\hat{\mathcal{H}}^1} \circ \cdots \circ v_{h_1,\sigma_{h_1}}^{\hat{\mathcal{H}}^1}, \tag{4}$$

and

$$S_0^{\hat{\mathcal{H}}^2} := v_0^{\hat{\mathcal{H}}^2}, \tag{5}$$

$$S_1^{\hat{\mathcal{H}}^2} := v_{1,1}^{\hat{\mathcal{H}}^2} \circ v_{1,2}^{\hat{\mathcal{H}}^2} \circ \cdots \circ v_{1,\sigma_1^{\hat{\mathcal{H}}^2}}^{\hat{\mathcal{H}}^2}, \tag{6}$$

$$\vdots \tag{7}$$

$$S_{h_2}^{\hat{\mathcal{H}}^2} := v_{h_2,1}^{\hat{\mathcal{H}}^2} \circ v_{h_2,2}^{\hat{\mathcal{H}}^2} \circ \cdots \circ v_{h_2,\sigma_{h_2}}^{\hat{\mathcal{H}}^2}, \tag{8}$$

Here, 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 $\hat{\sigma}_i > \sigma_i$. $v_{i,j}^{\hat{\mathcal{H}}^k}$ is the j'th vertex on the i'th level of tree $\hat{\mathcal{H}}^k$ and h_k is the depth of the corresponding tree. The structural similarity between $\hat{\mathcal{H}}^1$ and $\hat{\mathcal{H}}^2$ is then defined via the optimal alignment (alignments of out-degree and in-degree sequences) of these string sequences $S_i^{\hat{\mathcal{H}}^k}$ on each level. That means, the more similar with respect to a *cost function* α the out-degree and in-degree sequences on the levels i are, the more similar is the common structure of the trees. The dynamic programming algorithm with complexity $O(|\hat{V}_1| \cdot |\hat{V}_2|)$ for finding the optimal alignment of the sequences is described in detail in [4]. Here we give only a short discussion of this algorithm. We start by a definition of a distance measure.

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

- $\omega(x,y) = \omega(y,x) \quad \forall\, x, y \in X.$
- $\omega(x,x) = 0 \quad \forall\, x \in X.$

If we set $\omega(x,y) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\sigma^2}}$, we obtain immediately

Lemma 1. *$\omega : \mathbb{R} \times \mathbb{R} \longrightarrow [0,1]$ is defined by $\omega(x,y) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\sigma^2}}$. 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.

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

$$S_0^{\hat{\mathcal{H}}^1}, S_0^{\hat{\mathcal{H}}^1}, \ldots, S_{h_1}^{\hat{\mathcal{H}}^1} \quad \text{and} \quad S_0^{\hat{\mathcal{H}}^2}, S_1^{\hat{\mathcal{H}}^2}, \ldots, S_{h_2}^{\hat{\mathcal{H}}^2}.$$

Now we define

$$\alpha^{out}\left(v_{i_1,j_1}^{\hat{\mathcal{H}}^1}, v_{i_2,j_2}^{\hat{\mathcal{H}}^2}\right) := \begin{cases} \omega^{out}\left(\delta_{out}(v_{i_1,j_1}^{\hat{\mathcal{H}}^1}), \delta_{out}(v_{i_2,j_2}^{\hat{\mathcal{H}}^2}), \sigma_{out}^1\right) & : \quad i_1 = i_2 \\ +\infty & : \quad else\,, \end{cases} \tag{9}$$

$0 \le i_k \le h_k$, $1 \le j_k \le \sigma_{i_k}$, $k \in \{1,2\}$, where $\omega^{out}(x,y,\sigma_{out}^k) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\left(\sigma_{out}^k\right)^2}}$, $x,y,\sigma_{out}^k \in \mathbb{R}$, and

$$\alpha^{out}\left(v_{i,j_1}^{\hat{\mathcal{H}}^1}, -\right) := \omega^{out}\left(\delta_{out}(v_{i,j_1}^{\hat{\mathcal{H}}^1}), \xi, \sigma_{out}^2\right), \tag{10}$$

$$\alpha^{out}\left(-, v_{i,j_2}^{\hat{\mathcal{H}}^2}\right) := \omega^{out}\left(\xi, \delta_{out}(v_{i,j_2}^{\hat{\mathcal{H}}^2}), \sigma_{out}^2\right). \tag{11}$$

$\xi > 0$ prevents an alignment between two leaves being better evaluated as an alignment between a leaf and a gap ('-'). With $\omega^{in}(x,y,\sigma_{in}^k) := 1 - e^{-\frac{1}{2}\frac{(x-y)^2}{\left(\sigma_{in}^k\right)^2}}$ 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 $\alpha^{in}\left(-, v_{i,j_2}^{\hat{\mathcal{H}}^2}\right)$. The alignments will be evaluated on the levels of the generalized trees. Second, for evaluating the alignments on each level (local alignments), we set

$$\text{align}\left(v_{i,j_1}^{\hat{\mathcal{H}}^1}\right) := \begin{cases} v_{i,j_2}^{\hat{\mathcal{H}}^2} & : \quad \text{align}^{-1}\left(v_{i,j_2}^{\hat{\mathcal{H}}^2}\right) = v_{i,j_1}^{\hat{\mathcal{H}}^1} \\ - & : \quad else. \end{cases} \tag{12}$$

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 [4]. On the basis of the functions α^{out}, α^{in} we define analogously $\hat{\alpha}^{out}$, $\hat{\alpha}^{in}$. The parameters of $\hat{\alpha}^{out}$ and $\hat{\alpha}^{in}$ are now denoted by $\hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2$ and $\hat{\sigma}_{out}^1, \hat{\sigma}_{out}^2$, respectively. Furthermore we state

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

$$\gamma_{\hat{\mathcal{H}}^k}^{in}(i) := \frac{\sum_{j=1}^{\sigma_i^k} \hat{\alpha}_{in}\left(v_{i,j}^{\hat{\mathcal{H}}^k}, \text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^k}\right)\right)}{\sigma_i^k}, \tag{14}$$

$k \in \{1,2\}$, which are similarity values for out-degree and in-degree alignments. Finally, we obtain the normalized and cumulative functions

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

and

$$\gamma^{in}(i,\hat{\sigma}^1_{in},\hat{\sigma}^2_{in}) := 1 - \tag{16}$$
$$\frac{1}{\sigma_i^1+\sigma_i^2}\cdot\left\{\sum_{j=1}^{\sigma_i^1}\hat{\alpha}^{in}\left(v_{i,j}^{\hat{\mathcal{H}}^1},\text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^1}\right)\right)\right\}$$
$$+\frac{1}{\sigma_i^1+\sigma_i^2}\cdot\left\{\sum_{j=1}^{\sigma_i^2}\hat{\alpha}^{in}\left(v_{i,j}^{\hat{\mathcal{H}}^2},\text{align}\left(v_{i,j}^{\hat{\mathcal{H}}^2}\right)\right)\right\},$$

which detect the similarity of an out-degree and in-degree alignment on a level i. For the construction of the final similarity measure d with respect to our trees we need a the definition of a special kind of similarity measures. For this, we assume a set of units U and a mapping $\phi : U \times U \longrightarrow [0,1]$. We called ϕ a backward similarity measure if it satisfies the conditions

$$\phi(u,v) = \phi(v,u), \forall\, u,v \in U$$

and

$$\phi(u,u) \geq \phi(u,v), \forall\, u,v \in U.$$

We state now the key result that has been proven in [4] for measuring the similarity of generalized trees.

Theorem 1. *Let $\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2$ be two generalized trees, $0 \leq i \leq \rho$, $\rho := \max(h_1,h_2)$.*

$$d(\hat{\mathcal{H}}_1,\hat{\mathcal{H}}_2) := \frac{(\rho+1)}{\sum_{i=0}^{\rho}\gamma^{fin}(i,\sigma_1^{out},\sigma_2^{out},\sigma_1^{in},\sigma_2^{in})}\prod_{i=0}^{\rho}\gamma^{fin}(i,\sigma_1^{out},\sigma_2^{out},\sigma_1^{in},\sigma_2^{in}), \tag{17}$$

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

$$\begin{aligned}\gamma^{fin} &= \gamma^{fin}(i,\sigma_1^{out},\sigma_2^{out},\sigma_1^{in},\sigma_2^{in})\\ &:= \zeta\cdot\gamma^{out} + (1-\zeta)\cdot\gamma^{in}, \quad \zeta\in[0,1].\end{aligned}$$

By construction we have $d(\hat{\mathcal{H}}_1,\hat{\mathcal{H}}_2) \in [0,1]$. As a summary we note that the GTSA measures the similarity of two generalized trees by applying the technique of sequence alignments to out-degree and in-degree sequences (on a level i). These alignments have both global and local significance. On one hand, the sequence alignments will be implemented in a global sense, to compute the optimal alignment between the sequences

$$S_0^{\hat{\mathcal{H}}^1}, S_0^{\hat{\mathcal{H}}^1}, \ldots, S_{h_1}^{\hat{\mathcal{H}}^1} \quad \text{and} \quad S_0^{\hat{\mathcal{H}}^2}, S_1^{\hat{\mathcal{H}}^2}, \ldots, S_{h_2}^{\hat{\mathcal{H}}^2}.$$

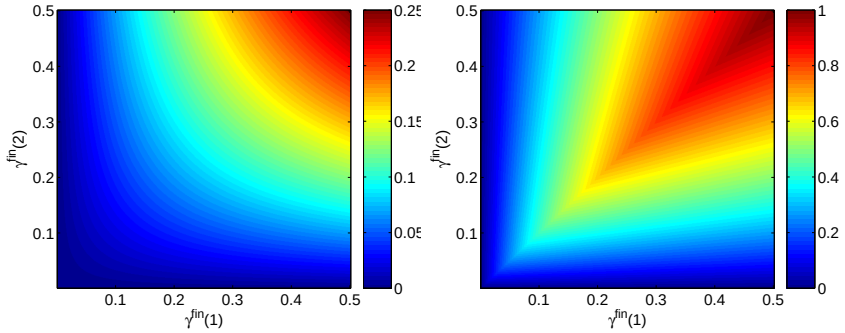

Fig. 2. Similarity measure d_1 (left) and d_{min} (right) as function of $\gamma^{fin}(1)$ and $\gamma^{fin}(2)$ for a fixed $\gamma^{fin}(3) = 0.5$. The values of the similarity measures are color-coded as given in the corresponding color-bars.

On the other hand, the alignments are evaluated on the levels of the generalized trees by the function γ^{fin}. In [6] we used theorem 1 for the classification of artificially generated data and data from microarray experiments. In this paper we present a modified version of theorem 1 which should be more sensitive to structural differences if applied in composition with the binary graph classifier discussed in the next section. We define the similarity measure $d_{\min}$ as follows.

Corollary 1. *Let $\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2$ be two generalized trees, $0 \leq i \leq \rho$, $\rho := \max(h_1, h_2)$.*

$$d_{\min}(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2) := \frac{(\rho+1)}{\sum_{i=0}^{\rho} \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in})} \min_{0 \leq i \leq \rho} \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in}) \tag{18}$$

The measure d_{min} is also a backward similarity measure, with γ^{fin} given in theorem 1.

One can easily show, that $d_{\min}(\hat{\mathcal{H}}_1, \hat{\mathcal{H}}_2) \in [0, 1]$ also holds for this measure. In figure 2 we compared the two similarity measures d (left figure) and $d_{\min}$ (right figure) for the special case $\rho = 2$. The figures show the color-coded similarity measure as function of $\gamma^{fin}(1)$[3] and $\gamma^{fin}(2)$ in the interval $0 \leq \gamma^{fin}(1)$, $\gamma^{fin}(2) \leq \gamma^{fin}(3)$ for a fixed value $\gamma^{fin}(3) = 0.5$. One characteristic difference between both measures is that d_{min} equals one not only for $\gamma^{fin}(1) = \gamma^{fin}(2) = \gamma^{fin}(3) = 1$ but also for $\gamma^{fin}(1) = \gamma^{fin}(2) = \gamma^{fin}(3)$. That means it has a kind of self-normalization if the similarities on all aligned levels are equal. One could now argue, that this property could lead to contradictory results if used as a similarity measure for pairs of generalized trees. This is correct, however, we do not aim to measure the similarity of two generalized trees but the similarity of undirected, unweighted graphs represented as set of generalized trees as explained in detail in the following section. The crucial difference is, that in

[3] Here we used a simplified notation: $\gamma^{fin}(i) = \gamma^{fin}(i, \sigma_1^{out}, \sigma_2^{out}, \sigma_1^{in}, \sigma_2^{in})$ for $i \in \{0, \ldots, \rho\}$

the case of generalized trees an identical similarity on all alignment levels leads inevitably to a wrong result whereas for the classification of graphs this case will have only neglectable effect because of three reasons. First, the decision if two graphs belong to the same class will be given on a statistical level based on sets of generalized trees for each graph and their mutual comparison. Additionally, the number of cases in which all gammas are identical are rare and hence, their effect will be small. Second, from the right figure in 2 one can see, that in the case when all $\gamma^{fin}(i)$ are similar but no longer identical, the value for $d_{\min}$ decreases rapidly. Third, we will use the similarity measure for generalized trees to judge only if two graphs belong to the same class. We make no statement about the similarity between the graphs itself but will introduce a binary classifier.

3 Binary graph classifier

In the preceding sections we introduced a method to measure the similarity between a pair of generalized trees. In this section we extend this method to construct a binary classifier for the classification of graphs. That means we construct a method which allows us to decide if two graphs are similar or not but gives us no information how similar they are. The graphs we deal 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. The basic idea of this extention is a local decomposition of a graph in generalized trees. This decomposition and the construction of the binary classifier will now be described in detail.

Definition 2. *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 *[5].*
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.*

This definition results in a set S_G consisting of N generalized trees of depth D. We apply to this set the GTSA introduced above and obtain a distribution of pairwise tree similarities p_{TS}.

A visualization for the extraction of one tree from a graph is given in figure 3. For didactical reasons, the nodes are regularly arranged on the surface of a

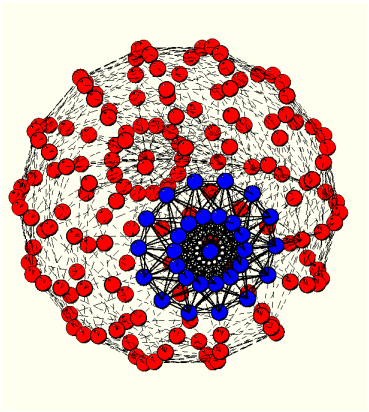

Fig. 3. A spherical graph with regular node arrangement on the surface of a sphere and regular connections between the nodes to the nearest neighbors. Shown is one local tree, resulting from the selection of the nodes up to depth $D = 2$. The root node is in the center of the two surrounding rings of nodes. (Figure was produced by Molscript [10].)

sphere and the nodes are connected to its nearest neighbors. Shown are the nodes which form a tree of depth $D = 2$. The root node is in the center of the two node circles.

Suppose now we have two graphs G_1 and G_2 and we want to decide, if both graphs are similar or not or more precise are both graphs from the same class. As a solution to this problem we suggest not to compare the graphs itself as a whole but to compare local parts which can be compared in an efficient way. This decision is based on the trees similarity distributions $p_{TS}^{G_1}$ and $p_{TS}^{G_2}$ of G_1 and G_2 and the trees similarity distribution p^M which results form the union of the tree sets of the graphs, $S_M = S_{G_1} \cup S_{G_2}$. In the following we call distributions like p^M mixed and distributions like $p_{TS}^{G_1}$ or $p_{TS}^{G_2}$ pure similarity distribution. The binary classifier is based on the following definition.

Definition 3. *Two graphs G_1 and G_2 are similar, iff the three similarity distributions $p_{TS}^{G_1}$, $p_{TS}^{G_2}$ and p^M are similar.*

The idea behind this definition is a consistency check of the mutual compatibility of the contributing tree sets, provided the subset is sufficient large. Definition 3 maps the question of graph similarity to the similarity of distributions. There is more then one possibility to define the similarity between two probability distributions. Here we will use a rather restrictive then relaxed definition.

Definition 4. *Two probability distributions are called similar iff the chi-square (χ^2) test for a significance level α can not reject that both distributions are equal.*

The crucial presumption to apply a χ^2 test is, that the number of samples is large. In our case the sample size is given by $\binom{N}{2}$, with N is the number of nodes

in the graph which corresponds also to the number of trees in the tree set of a given graph. Hence, $\binom{N}{2}$ is the number of different tree pairs one can form from the tree set of a graph. For sufficiently large graphs the χ^2 test will give reliable results. For $N > 150$ the number of different tree pairs is already larger then 10^4. Due to the fact, that we intend to classify large graphs this presumption should be fulfilled for all practical cases. Definition 3 provides now a natural way to define a binary classifier for graphs.

Definition 5 (Binary graph classifier (BGC)). *Two graphs G_1 and G_2 belong to the same class, iff the graphs are similar.*

In the next section we apply this method to networks obtained from microarray experiments for different tumor stages of cervical cancer.

4 Results

The data set we applied our method to is from DNA microarray experiments. We used the data from [19] which investigated the gene expression levels of different tumor stages of cervical cancer. For a summary see table 1. In general, the higher the integer numbers and the letters of the tumor stages the more the cancer has grown and spread. The data include also a normal expression profile of cervical tissue indicated in table 1 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}.

The networks from the expression data are obtained via a three step process [14]:

1. Calculating 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}.

The size of microarrays used for each experiment in [19] consisted of a total number of 10692 genes. Hence, the networks we have to compare have this number of nodes. Via the local tree decomposition algorithm in definition 2 we obtain tree sets for all graphs consisting of 10692 trees each. From each tree set we choose $N_{RS} = 100000$ tree pairs randomly and calculate their similarity. Additionally, we calculate the distributions for mixed tree sets by choosing

Table 1. Microarray data from [19] for different tumor stages, based on the FIGO (International Federation of Gynecologists and Obstetrics) taging system, of cervical cancer. Each of the 27 (total number of patients) arrays contained 10692 genes.

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

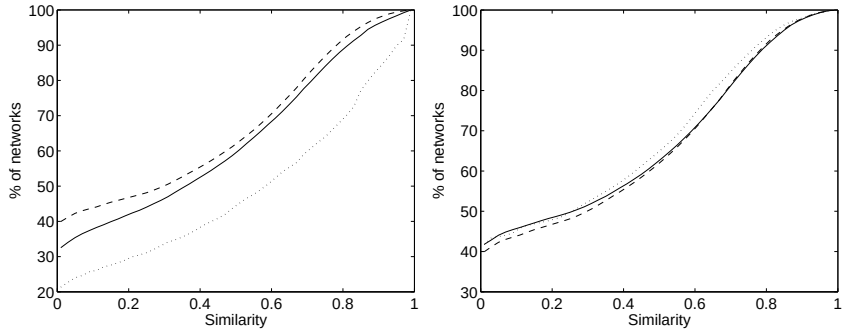

Fig. 4. Resulting similarity distributions obtained by the measure d_{min}. Left: Cumulative similarity distributions for the normal- (dashed line), 1B- (dotted line) and mixed- (full line) tree set. Right: Cumulative similarity distributions for the normal- (dashed line), 2A- (dotted line) and mixed- (full line) tree set.

randomly N_{RS} different pairs of trees from $(S_{G_N}, S_{G_{1B}})$ and $(S_{G_N}, S_{G_{2A}})$. The number of random samples N_{RS} was chosen empirically. We tested several different sizes and found to estimate the underying one-dimensional probability distribution p_{TS} of the tree similarities a value of 100000 samples gives robuts results. The cumulative similarity distributions $P_{TS} = \int p_{TS}(x)dx$ for all comparisons are shown in figure 4. The mixed cumulative similarity distribution is always shown as full line. Visual inspection of the curves in figure 4 reveals, that the 'normal' and 1B-networks (left figure) can be clearly distinguished while the 'normal' and '2A' network (right figure) are more similar but also not equal. To obtain a statistical answer to the question, based on definition 4, we calculated for all distribution pairs the χ^2 value. For a significance level of $\alpha = 0.01$ we had to reject the hypothesis, that two distributions are equal in all cases. Hence, our method was able to judge, that all graphs are mutually different. Even the comparison between normal- and 2A-networks could be classified correctly. This demonstrates that the obtained networks from DNA microarray experiments, representing tissues from different cancer stages of cervical cancer, can be distinguished with our BGC.

5 Conclusion

In this paper we extended our binary graph classifier (BGC), introduced in [6], which allows the classification of large undirected, unweighted graphs. This classifier is based essentially on the generalized tree-similarity-algorithm (GTSA) [4] which provides a similarity measure between tree pairs by an alignment of property strings representing the trees. The application of the GTSA to undirected, unweighted graphs was enabled by a local decomposition of the graph in generalized trees resulting in a tree set on which the GTSA could be applied. The obtained similarity values for all tree pairs of the tree set determine the sim-

ilarity distribution of the graph. Hence, the problem of graph classification was mapped to compare one-dimensional probability distributions. For this comparison we applied the well known chi-square test from statistics to the distributions. We demonstrated by the application of our BGC, with the similarity measure d_{min} for generalized trees, that the classification of co-expressed networks from DNA microarray experiments representing tissues from different tumor stages of cervical cancer [19], is possible. This is non-trivial because we know, in fact, that the networks represent tissues from different tumor stages but we do not know if the information buried in the networks is sufficient to distinguish them. Hence, it is highly probably that essential modifications in the gene expression regulation which are caused by the disease can be captured by DNA microarray experiments and in the corresponding networks. Our approach is in contrast to most existing studies dealing with the classification of microarray data, because we compared and classified networks, representing the relation between co-expressed genes, instead of selecting subsets of genes as feature vectors.

In general, we think that our binary graph classifier (BGC) is a considerable improvement to existing graph classification approaches because we presented not only a theoretical framework suitable for small graphs but also for large graphs of sizes which are relevant for practical applications. In the future we will continue refining our method with special attention to networks from DNA microarray experiments.

Acknowledgment

We would like to thank Jie Chen, Galina V. Glazko, Arcady Mushegian and Chris Seidel for fruitful discussions. Additionally, we would like to thank the anonymous reviewers for useful comments.

References

1. R. Bellman, *Dynamic Programming*. Princeton University Press, 1957
2. L. von Bertalanffy, *General System Theory: Foundation, Development, Application* New York, George Braziller, 1968
3. S. B. Cho and H. H. Won, *Machine Learning in DNA Microarray Analysis for Cancer Classification* Proceedings of the First Asia-Pacific Bioinformatics Conference, Australien Computer Society Inc. 2003, Vol. 19, 1899–198
4. M. Dehmer. and R. Gleim and A. Mehler, *A new method of measuring similarity for a special class of directed graphs*. Tatra Mountains Mathematical Publications, Slovakia, submitted for publication, August 2004
5. E. W. Dijkstra, *A note on two problems in connection with graphs*. Numerische Math., Vol. 1, 1959, 269–271
6. F. Emmert-Streib., M. Dehmer, J. Kilian: *Classification of large Graphs by a local Tree decomposition*, accepted to appear in: Proceedings of DMIN'05, International Conference on Data Mining, in conjuction with the 2005 World Congress in Applied Computing, Las Vegas, USA, 2005

7. T. R. Golub et.al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science, Vol. 286, 1999, 531–537
8. F. Kaden, *Graphmetriken und Distanzgraphen*. ZKI-Informationen, Akad. Wiss. DDR, Vol. 2 (82), 1982, 1–63
9. 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
10. 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
11. J. Lapointe et.al., *Gene expression profiling identifies clinically relevant subtypes of prostate cancer* Proc. of the Natl. Acad. of Sci. USA, Vol. 101, 2004, 811-816
12. A. Mehler et.al., *Towards logical hypertext structure. A graph-theoretic perspective*, Proc. of I2CS'04, Guadalajara/Mexico, Lecture Notes in Computer Science, Berlin-New York: Springer, 2004
13. R. C. Read and D. G. Corneil, *The graph isomorphism disease*. Journal of Graph Theory, Vol. 1, 1977, 339–363
14. J. Rougemont and P. Hingamp, *DNA microarray data and contextual analysis of correlation graphs*. BMC Bioinformatics, Vol. 4, 2003, 4–15
15. F. Sobik, *Graphmetriken und Klassifikation strukturierter Objekte*. ZKI-Informationen, Akad. Wiss. DDR, Vol. 2 (82), 1982, 63–122
16. F. Sobik, *Graphmetriken und Charakterisierung von Graphklassen*. 27. Internat. Wiss. Koll., TH-Ilmenau, Vol. 2 (82), 1982, 63–122
17. J. R. Ullman, *An algorithm for subgraph isomorphism*. J. ACM, Vol. 23 (1), 1976, 31–42
18. L. J. van't Veer et.al., *Gene expression profiling predicts clinical outcome of breast cancer* Nature, Vol. 415, 2002, 530–537
19. 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
20. B. Zelinka, *On a certain distance between isomorphism classes of graphs*. Časopis pro pěst. Mathematiky, Vol. 100, 1975, 371–373