CBR System for Leukemia Patients Diagnosis

Juan F. De Paz, Sara Rodríguez, Javier Bajo, Juan M. Corchado

Departamento de Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, España
{fcofds,srg,jbajope,corchado}@usal.es

Abstract. The use of computational methods is fundamental in cancer research. One of the possibilities is the use of Artificial Intelligence techniques. Several of these techniques have been used to analyze expression arrays. However, the new Exon arrays which work with a large amount of data require novel solutions. This paper presents a Case-based reasoning (CBR) system for automatic classification of leukemia patients from Exon array data. The proposed CBR system incorporates novel algorithms for data filtering and classification. The system has been tested and the results obtained are presented in this paper.

Keywords: Case-based Reasoning, ESOINN neural network, leukemia classification.

1. Introduction

During recent years there have been great advances in the field of Biomedicine [1]. The incorporation of computational and artificial intelligence techniques to the field of medicine has yielded remarkable progress in predicting and detecting diseases [1]. One of the areas of medicine which is essential and requires the implementation of techniques that facilitate automatic data processing and extraction of knowledge is genomics. Genomics deals with the study of genes, their documentation, their structure and how they interact [2]. We distinguish different fields of study within the genome. One is transcriptome, which deals with the study of ribonucleic acid (RNA), and can be studied through expression analysis [3]. This technique studies RNA chains thereby identifying the level of expression for each gene studied. It consists of hybridizing a sample for a patient and colouring the cellular material with a special dye. This offers different levels of luminescence that can be analyzed and represented as a data array. Traditionally, methods and tools have been developed to work with expression arrays containing about 50000 data points. The emergence of the Exon arrays [7], holds important potential for biomedicine. However, the Exon arrays require novel tools and methods to work with very large (5500000) amounts of data

This paper presents a CBR system that facilitates the analysis and classification of data from Exon arrays corresponding to patients with leukemia. Leukemia, or blood cancer, is a disease that has a significant potential for cure if detected early [4].

The relationship between the chromosomal alterations and prognosis of leukemia and lymphomas is well established. Recently, conventional array-based expression profiling has demonstrated that chromosomal alterations are associated with distinctive patterns of expression. The four most important types of Leukemia are acute and chronic myelogenous leukemia (AML;CML) and acute and chronic lymphocytic Leukemia(ALL; CLL). About 25000 new cases of both acute and chronic Leukemia appear every year. About 12000 adult cases are diagnosed annually as acute myelogenic Leukemia, 8000 as chronic lymphocytic Leukemia, 500 as chronic myelogenous forms, and about 3500 as acute forms of lymphocytic Leukemia. A study [8] shows that an estimated 19900 new cases of myeloma were diagnosed in USA in 2007.

The system proposed in the context of this work focuses on the detection of carcinogenic patterns in the data from Exon arrays and is constructed from a CBR system that provides a classification technique based on previous experiences. An expression analysis consists basically of three stages: normalization and filtering; clustering and classification; and extraction of knowledge. These stages can be automated and included in a CBR system. The first step is critical to achieve both a good normalization of data and an initial filtering to reduce the dimensionality of the data set with which to work [5]. Since the problem at hand is working with high-dimensional arrays, it is important to have a good pre-processing technique that facilitates automatic decision-making about the variables that will be vital for the classification process [6].

For some time now, we have been working on the identification of techniques to automate the reasoning cycle of several CBR systems applied to complex domains [20] [21] [25]. The objective of this work is to develop a CBR system that allows the identification of patients with various types of cancer. The model aims to improve the cancer classification based on microarray data. The system proposed in this paper presents a new synthesis that brings several artificial intelligence subfields together (filter techniques, clustering and artificial neural networks). The retrieval, reuse, revision and learning stages of the CBR system use these techniques to facilitate the CBR adaptation to the domain of biological discovery with microarray datasets. Specifically, the system presented in this paper uses a model which takes advantage of two novel methods for analyzing Exon array data: a technique for filtering data, and a technique ESOINN [25] for clustering. The first method combines various filtering techniques to dramatically reduce the dimensionality. The second one allows clustering by incorporating both the distribution process of the entire surface of classification, and the separation between groups with low density among them.

The paper is structured as follows: Section 2 and Section 3 describe the proposed CBR model and how it is adapted to the problem under consideration. Finally, Section 4 presents the results and conclusions obtained after testing the model.

2. CBR System for Classifying Exon Array Data

The CBR developed tool receives data from the analysis of chips and is responsible for classifying of individuals based on evidence and existing data. Case-based

Reasoning is a type of reasoning based on the use of past experiences [9]. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain.

2.1. Retrieve

The retrieve phase starts when a new problem description is received. Contrary to what usually happens in the CBR, our case study is unique in that the number of variables is much greater than the number of cases. This leads to a change in the way the CBR functions so that instead of recovering cases at this stage, important variables are retrieved. In the case study, the number of cases is not the problem, rather the number of variables. For this reason variables are retrieved at this stage and then, depending on the identified variables, the other stages of the CBR are carried out. This phase will be broken down into 5 stages which are described below:

2.1.1. RMA

The RMA (R*obust Multi-array Average*) [10] algorithm is frequently used for pre-processing Affymetrix microarray data. RMA consists of three steps: (i) Background Correction; (ii) Quantile Normalization (the goal of which is to make the distribution of probe intensities the same for arrays); and (iii) Expression Calculation.

2.1.2. Control and Errors

During this phase, all probes used for testing hybridization are eliminated. These probes have no relevance at the time when individuals are classified, as there are no more than a few control points which should contain the same values for all individuals. If they have different values, the case should be discarded. Therefore, the probes control will not be useful in grouping individuals.

On occasion, some of the measures made during hybridization may be erroneous; not so with the control variables. In this case, the erroneous probes that were marked during the implementation of the RMA must be eliminated.

2.1.3. Variability

Once both the control and the erroneous probes have been eliminated, the filtering begins. The first stage is to remove the probes that have low variability. This work is carried out according to the following steps:

1. Calculate the standard deviation for each of the probes j

$$\sigma_{\cdot j} = +\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\overline{\mu}_{\cdot j} - x_{ij}\right)^2} \qquad (1)$$

   Where N is the number of items total, $\overline{\mu}_{\cdot j}$ is the average population for the variable j, x_{ij} is the value of the probe j for the individual i.

2. Standardize the above values

$$z_i = \frac{\sigma_{\cdot j} - \mu}{\sigma} \tag{2}$$

where $\mu = \frac{1}{N}\sum_{j=1}^{N}\sigma_{\cdot j}$ and $\sigma_{\cdot j} = +\sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\overline{\mu}_{\cdot j} - x_{ij}\right)^2}$ where $z_i \equiv N(0,1)$

3. Discard of probes for which the value of z meet the following condition: $z < -1.0$ given that $P(z < -1.0) = 0.1587$. This will effect the removal of about 16% of the probes if the variable follows a normal distribution.

2.1.4. Uniform Distribution

Finally, all remaining variables that follow a uniform distribution are eliminated. The variables that follow a uniform distribution will not allow the separation of individuals. Therefore, the variables that do not follow this distribution will be really useful variables in the classification of the cases. The contrast of assumptions followed is explained below, using the Kolmogorov-Smirnov [18] test as an example. H_0: The data follow a uniform distribution; H_1: The analyzed data do not follow a uniform distribution. Statistical contrast:

$$D = \max\left\{D^+, D^-\right\} \tag{3}$$

where $D^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F_0(x_i)\right\}$ $D^- = \max_{1\le i\le n}\left\{F_0(x_i) - \frac{i-1}{n}\right\}$ with i as the pattern of entry, n the number of items and $F_0(x_i)$ the probability of observing values less than i with H_0 being true. The value of statistical contrast is compared to the next value:

$$D_\alpha = \frac{C_\alpha}{k(n)} \tag{4}$$

in the special case of uniform distribution $k(n) = \sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}$ and a level of significance $\alpha = 0.05$ $C_\alpha = 1.358$.

2.1.5. Correlations

At the last stage of the filtering process, correlated variables are eliminated so that only the independent variables remain. To this end, the linear correlation index of Pearson is calculated and the probes meeting the following condition are eliminated.

$$r_{x_i y_{\cdot j}} > \alpha \tag{5}$$

being: $\alpha = 0.95$ $r_{x_i y_{\cdot j}} = \frac{\sigma_{x_i x_{\cdot j}}}{\sigma_{x_i} \sigma_{x_{\cdot j}}}$, $\sigma_{x_i x_j} = \frac{1}{N} \sum_{s=1}^{N} (\bar{\mu}_{\cdot i} - x_{si})(\bar{u}_{\cdot j} - x_{sj})$ Where $\sigma_{x_i x_{\cdot j}}$ is the covariance between probes i and j.

2.2. Reuse

Once filtered and standardized, the probes produce a set of values x_{ij} with i = 1 ... N, j = 1 ... s where N is the total number of cases, s the number of end probes. The next step is to perform the clustering of individuals based on their proximity according to their probes. Since the problem on which this study is based contained no prior classification with which training could take place, a technique of unsupervised classification was used. There is a wide range of possibilities. Some of these techniques are artificial neural networks such as SOM [11] (self-organizing map), GNG [12] (Growing neural Gas) resulting from the union of techniques CHL [13] (competitive Hebbian Learning) and NG [14] (neural gas), GCS [12] (Growing Cell Structure), Growing Grid or the SOINN [15] (self-organizing incremental neuronal network) or methods based on hierarchical clustering [26]. Some of the methods, such as self-organized Kohonen maps, set the number of clusters in the initial phase of training when using the algorithm of the k-means learning method. The self-organized maps have other variants of learning methods that base their behaviour on methods similar to the NG. They create a mesh that is adjusted automatically to a specific area. The greatest disadvantage, however, is that both the number of neurons that are distributed over the surface and the degree of proximity are set beforehand, resulting in the number remaining constant throughout the entire training process. Unlike the self-organizing maps based on meshes, Growing Grid or GCS do not set the number of neurons, or the degree of connectivity, but they do establish the dimensionality of each mesh. This complicates the separation phase between groups once it is distributed evenly across the surface.

After analyzing different techniques and checking the problems they might present so that they might be applied to the problem at hand, we have decided to use a variation of neural network SOINN [15], called ESOINN [16] (Enhanced self-organizing incremental neuronal network). Unlike the SOINN, ESOINN consists of a single layer, so it is not necessary to determine the manner in which the training of the first layer changes to the second. With a single layer, ESOINN is able to incorporate both the distribution process along the surface and the separation between low density groups. It has characteristics of a network CHL in the initial phases, by which it could be understood as a phase of competition, while in a second phase, the network of nodes begins to expand just as with a NG. This process is conducted in an iterative way until it reaches stability. Only the changes in training phase are detailed below:

1. Update the weights of neurons by following a process similar to the SOINN, but introducing a new definition for the learning rate in order to provide greater stability for the model. This learning rate has produced good results in other networks such as SOM [17].

$$\Delta W_{a_1} = n_1(M_{a_1})(\xi - W_{a_1}) \;\; \Delta W_i = n_1(M_{a_i})(\xi - W_{a_i}) \text{ with } i \in N_i \qquad (6)$$

Being $n_1(x) = \frac{1}{\sqrt{x}}$, $n_2(x) = \frac{1}{\sqrt{2+x^2}}$

2. Delete the connections with higher age. The ages are typified and are removed those whose values are in the region of rejection with k>0. α is 0.05.
3. If all input patterns have been passed then a KS-Test [18] is carried out in order to determine if the density distribution for the neurons in each group follows a normal distribution. If so then the learning procedure is finished; otherwise the next pattern is processed. The value of α chosen is 0.05.

Once, the clusters have been made, the new sample is classified. Its association is carried out bearing in mind the similarity of the new case with the recovered variables in the first phase. The similarity measure used is as follows:

$$d(n,m) = \sum_{i=1}^{s} f(x_{ni}, x_{mi}) * w_i \qquad (7)$$

Where s is the total number variables, n and m the cases, w_i the value obtained in the uniform test and f the Minkowski [19] Distance that is given for the following equation.

$$f(x,y) = \sqrt[p]{\sum_i \left| x_i - y_j \right|^p} \;\; \text{con } x_i, y_j \in R^p \qquad (8)$$

This dissimilarity measure weighs those probes that have a less uniform distribution, since these variables don't allow a separation

2.3. Revise/Retain

The revision is carried out by an expert who determines the correction with the group assigned by the system. If the assignation is considered correct, then the retrieve and reuse phases are carried out again so that the system is ready for the next classification. Nevertheless, the system provides an automatic temporal revision considering the retrieved cases. The system calculates the percentage of cases that have already been accurately classified among those retrieved for the current problem. If the percentage of a class is greater than a threshold then the system establishes that the case has been successfully classified. This decision has to be confirmed by the human expert.

3. Case Study

In the case study presented in the framework of this research are available 248 samples are available from analyses performed on patients either through punctures in

marrow or blood samples. The aim of the tests performed is to determine whether the system is able to classify new patients based on the previous cases analyzed and stored.

Figure 1 shows a scheme of the bio-inspired model intended to resolve the problem described in Section 2. The proposed model follows the procedures that are performed in medical centres. As can be seen in Figure 1, a previous phase, external to the model, consists of a set of tests which allow us to obtain data from the chips and are carried out by the laboratory personnel. The chips are hybridized and explored by means of a scanner, obtaining information on the marking of several genes based on the luminescence. At that point, the CBR-based model starts to process the data obtained from the Exon arrays.

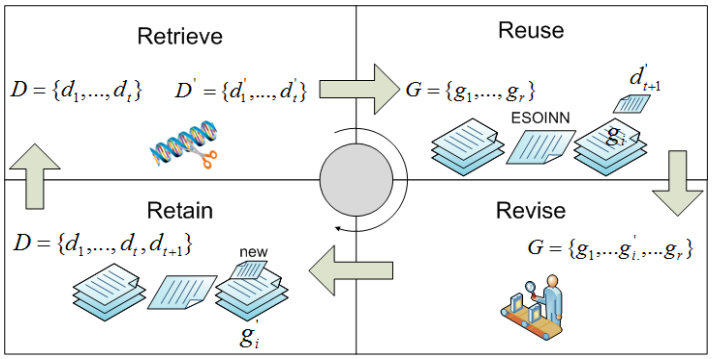

Fig. 1. Proposed CBR model

The retrieve phase receives an array with a patient's data as input information. The retrieve step filters genes but never patients. The set of patients is represented as $D = \{d_1,...,d_t\}$, where $d_i \in R^n$ represents the patient *i* and *n* represents the number of probes taken into consideration. As explained in Section 2.1, during the retrieve phase the data are normalized by the RMA algorithm [10] and the dimensionality is reduced bearing in mind, above all, the variability, distribution and correlation of probes. The result of this phase reduces any information not considered meaningful to perform the classification. The new set of patients is defined through *s* variables $D' = \{d_1',...,d_t'\}\ d_i' \in R^s, s \le n$.

The reuse phase uses the information obtained in the previous step to classify the patient into a leukemia group. The data coming from the retriever phase consists of a group of patients $D' = \{d_1',...,d_t'\}$ con $d_i' \in R^s, s \le n$, each one characterized by a set of meaningful attributes $d_i = (x_{i1},...,x_{is})$, where x_{ij} is the luminescence value of the probe *i* for the patient *j*. In order to create clusters and consequently obtain patterns to classify the new patient, the reuse phase implements a novel neural network based on the ESOINN [16], section 2.2. The network classifies the patients by taking into account their proximity and their density, in such a way that the result provided is a set G where $G = \{g_1,...,g_r\} r < s$. $g_i \subset D$, $g_i \cap g_j = \phi$ with

$i \neq j$ and $i, j < r$. The set G is composed of a group of clusters, each of them containing patients with a similar disease. Once the clusters have been obtained, the system can classify the new patient by assigning him to one of the clusters. The new patient is defined as d'_{t+1} and his membership to a group is determined by a similarity function defined in (7). The result of the reuse phase is a group of clusters $G = \{g_1, ...g'_{i.}, ...g_r\} r < s$ where $g'_i = g_i \cup \{d'_{t+1}\}$.

An expert from the Cancer Institute is in charge of the revision process. This expert determines if $g'_i = g_i \cup \{d'_{t+1}\}$ can be considered as correct. In the retain phase the system learns from the new experience. If the classification is considered successful, then the patient is added to the memory case $D = \{d_1, ..., d_t, d_{t+1}\}$.

4. Results and Conclusions

This paper has presented a CBR system which allows automatic cancer diagnosis for patients using data from Exon arrays. The model combines techniques for the reduction of the dimensionality of the original data set and a novel method of clustering for classifying patients. The CBR system presented in this work focused on identifying the important variables for each of the variants of blood cancer so that patients can be classified according to these variables.

In the experiments reported in this paper, we worked with a database of bone marrow cases from 248 adult patients with five types of leukaemia, plus a group of 16 samples belonging to healthy persons (no leukemias. The retrieve stage of the proposed CBR system presents a novel technique to reduce the dimensionality of the data. The total number of variables selected in our experiments was reduced to 883, which increased the efficiency of the cluster probe. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Institute of Cancer. The error rates have remained fairly low especially for cases where the number of patients was high. To try to increase the reduction of the dimensionality of the data we applied principal components (PCA) but this reduction of the dimensionality was not appropriate in order to obtain a correct classification of the patients. Figure 2a shows the classification performed for patients from groups CLL. As can be seen in Figure 2a, represented in black, most of the people of the CLL group are together, coinciding with the previous classification given by the experts at the Institute of Cancer. Only a small portion of the individuals departed from the initial classification.

In a similar way we proceeded to evaluate the classification for the rest of the groups. Figure 2b shows the total number of patients from each group and the number of misclassifications. As can be seen in Figure 2b, groups with fewer patients are those with a greater error rate. Once the validity of the method of filtration for selecting the most important variables for classification is verified, the next step in the evaluation was to assess the functioning of the classification process. The system was tested with 15 new patients. The patients were assigned to the expected groups. Only one of the patients was misclassified, being assigned to an erroneous group. The

patient misclassified belongs to the ALL class, while the others individuals belong to the CML and CLL classes.

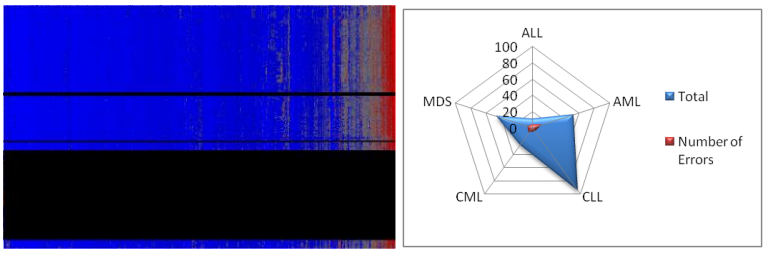

Fig. 2. Classification errors (a) numerical (b) percentage

The final classification was compared with the data obtained using a dendogram [24] and PAM [23] (Partitioning Around Medoids). The proportion of errors in every group was calculated and the Kurskal-Wallis [22] test was applied to determinate if the median of these proportions was equal. The results are shown in table 1.

Table 1. Comparison of methods. * different median and = equal, (-) median of column less than median of row

	CBR	Dendogram	PAM
CBR			
Dendogram	*(-)		
PAM	*(-)	*(-)	

One of the great contributions of the model presented is the ability to work with data from Exon arrays because of its great capacity for the selection of significant variables. The proposed system allows the reduction of the dimensionality based on the filtering of genes with little variability and those that do not allow a separation of individuals due to the distribution of data. It also presents a technique for clustering based on the use of neural networks ESOINN. The results obtained from empirical studies are provided with a tool that allows both the detection of genes and those variables that are most important for the detection of pathology, and the facilitation of a classification and reliable diagnosis, as shown by the results presented in this paper.

References

1. Shortliffe, E.H., Cimino, J.J.: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Springer, (2006).
2. Tsoka, S., Ouzounis C.: Recent developments and future directions in computational genomics. FEBS Letters, Vol. 480 (1), (2000) 42-48
3. Lander E.S. et al.: Initial sequencing and analysis of the human genome. Nature (2001) 409, 860 - 921
4. Rubnitz, J.E., Hijiya, N., Zhou, Y., Hancock, M.L., Rivera ,G.K., Pui C.: Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia. Pediatric Blood & Cancer, Vol. 44 (2), (2005). 138-141

5. Armstrong, N.J., van de, Wiel M.A.: Microarray data analysis: From hypotheses to conclusions using gene expression data. Cellular Oncology, Vol. 26 (5-6), (2004) 279-290.
6. Quackenbush, J.: Computational analysis of microarray data. Nature Review Genetics, Vol. 2(6), (2001) 418-427
7. Affymetrix, GeneChip Human Exon 1.0 ST Array, http://www.affymetrix.com/products/arrays/specific/Exon.affx
8. SEER (Surveillance Epidemiology and End Results), U.S. National Cancer Institute http://seer.cancer.gov/. (2007)
9. Kolodner, J.: Case-Based Reasoning. Morgan Kaufmann (1993).
10. Irizarry, R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. Speed,T.P.: Exploration, Normalization, and Summaries of High density Oligonucleotide Array Probe Level Data. Biostatistics, Vol. 4, (2003) 249-264.
11. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, (1982) 59-69.
12. Fritzke, B.: A growing neural gas network learns topologies. Cambridge MA: G. Tesauro, D. S. Touretzky, and T. K. Leen, , Advances in Neural Information Processing Systems 7, (1995) 625-632.
13. Martinetz, T.: Competitive Hebbian learning rule forms perfectly topology preserving maps. Amsterdam : Springer,. ICANN'93: International Conference on Artificial Neural Networks, (1993). 427-434.
14. Martinetz, T., Schulten, K.: A neural-gas network learns topologies. T Kohonen, et al. Amsterdam: Artificial Neural Networks, (1991). 397-402.
15. Shen, F.: An algorithm for incremental unsupervised learning and topology representation. Tokyo: Ph.D. thesis. Tokyo Institute of Technology, (2006).
16. Furao, S, Ogura, T, Hasegawa, O.: An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks, Vol. 20, (2007), 893-903.
17. Corchado, J. M., Bajo, J., De Paz Y., De Paz J. F. Integrating Case Planning and RPTW Neuronal Networks to Construct an Intelligent Environment for Health Care. International Journal of Computational Intelligence in Bioinformatics and System Biology, In Press.
18. Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition, Vol. 34, (2001) 1625-1637.
19. Gariepy, R., Pepe, W.D.: On the Level sets of a Distance Function in a Minkowski Space. Proceedings of the American Mathematical Society, Vol. 31 (1), (1972), 255-259
20. Corchado, J.M., Corchado, E.S., Aiken, J., Fyfe, C., Fdez-Riverola, F., Glez-Bedia, M.: Maximum Likelihood Hebbian Learning Based Retrieval Method for CBR Systems. In Proceedings of the 5th International Conference on Case-Based Reasoning, (2003) 107-121.
21. Riverola, F., Díaz F., Corchado, J. M.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence,. Vol. 22 (3-4), (2006), 254-268.
22. Kruskal, W, Wallis, W.: Use of ranks in one-criterion variance analysis. Journal of American Statistics Association (1952)
23. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. (1990).
24. Saitou, N., Nie, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Vol. 4, (1987) 406-425
25. Arshadi, N., Jurisica, I.: Data mining for case-based reasoning in high-dimensional biological domains. Knowledge and Data Engineering, IEEE Transactions. Vol. 17 (8), (2005) 1127-1137
26. Sultan M., Wigle D., Cumbaa C.A., Maziarz M., Glasgow J., Tsao M., Jurisica I. Binary tree-structured vector quantization approach to clustering and visualizing microarray data. ISMB (2002) 111-119