A Case-based Reasoning Strategy for Classifying Leukemia Patients

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

Departamento de Informática y Automática, Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, España
{srg, fcofds, 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 [21]. The incorporation of computational and artificial intelligence techniques to the field of medicine has yielded remarkable progress in predicting and detecting diseases [21]. Genomics deals with the study of genes, their documentation, their structure and how they interact [22]. 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 [13]. 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 50.000 data. 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 5.500.000 amounts of data.

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 [16]. 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). 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 [6] for clustering. The first method combines various filtering techniques to dramatically reduce the dimensionality of the data.

The paper is structured as follows: The next section presents the problem that motivates this research, i.e., the classification of leukemia patients from samples obtained through Exon arrays. Section 3 and Section 4 describe the proposed CBR model and how it is adapted to the problem . Finally, Section 5 presents the results and conclusions obtained after testing the model.

2 Computational Methods in Cancer Research

Hematological malignancies like leukemia, have been the ground of chromosomal and genetic analysis for many years. Recently, conventional array-based expression profiling has demonstrated that chromosomal alterations are associated with distinctive patterns of expression. Leukemia is a blood cancer form, originating in a malfunctioning bone marrow that tends to produce abnormal red and white cells at an increased rate [17]. The four most important types of Leukemia are acute and chronic myelogenous leukemia (AML;CML) and acute and chronic lymphocytic Leukemia(ALL; CLL). About 25.000 new cases of both acute and chronic Leukemia appear every year. Most cases appear in adults and persons over 60 years, but acute lymphocytic Leukemia has an increased rate in children. About 12.000 adult cases are diagnosed annually as acute myelogenic

Leukemia, 8.000 as chronic lymphocytic Leukemia, 500 as chronic myelogenous forms, and about 3.500 as acute forms of lymphocytic Leukemia. The rest of the cases are unclassified blood cancer types. A recent study [19], shows that an estimated 19,900 new cases of myeloma were diagnosed in the United States in 2007.

Microarray technology is based on a database of gene fragments called expressed sequence tags (ESTs), which are used to measure target abundance using the scanned intensities of fluorescence from tagged molecules hybridized to ESTs [14]. The high dimensionality of the data provided by each Exon array presents problems of handling and processing. An expression analysis consists basically of two stages: normalization and filtering; clustering and classification. These stages can be automated and included in a CBR system.

3 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. CBR is a type of reasoning based on the use of past experiences [11]. The primary concept when working with CBRs is the concept of case. A case can be defined as a past experience, and is composed of three elements: A problem description, which delineates the initial problem; a solution, which provides the sequence of actions carried out in order to solve the problem; and the final stage, which describes the state achieved once the solution was applied. The way cases are managed is known as the CBR cycle, and consists of four sequential phases: retrieve, reuse, revise and retain.

3.1. RETRIEVE

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. Traditionally, only the similar cases to the current problem are recovered, often because of performance, and then adapted. 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, the other stages of the CBR are carried out. This phase will be broken down into 5 stages:

3.1.1. RMA

The RMA (Robust Multi-array Average) [8] algorithm is frequently used for pre-processing Affymetrix microarray data. RMA consists of three steps: (i) Background Correction; (ii) Quantile Normalization; (iii) Expression Calculation.

3.1.2. Control and Errors

During this phase, all probes used for testing hybridization are eliminated. 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.

3.1.3. 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≡N(0,1)

$$z_i = \frac{\sigma_{\cdot j} - \mu}{\sigma} \qquad (2)$$

3. Discard of probes whose value of $z < -1.0$.

3.1.4. Uniform distributions

Finally, all remaining variables that follow a normal 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 followed is the Kolmogorov-Smirnov [3] test (KS-Test). $\alpha = 0.05$.

3.1.5. 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.

$$r_{x_i y_{\cdot j}} > \alpha \qquad (3)$$

being: $\alpha = 0.95$ $r_{x_i y_{\cdot j}} = \dfrac{\sigma_{x_i x_{\cdot j}}}{\sigma_{x_i}\sigma_{x_j}}$, Where $\sigma_{x_i x_{\cdot j}}$ is the covariance between the probes i and j.

3.2. 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. 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 [10] (Self-Organizing Map), GNG [5] (Growing Neural Gas) resulting from the union of techniques CHL [15] (competitive Hebbian Learning) and NG [5] (Neural Gas), GCS [5] (Growing Cell Structure), Growing Grid or the SOINN [20] (self-organizing incremental neuronal network). Some of the methods, such as self-organized maps, set the number of clusters in the initial phase of training when using the algorithm of the k-means. 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 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, thus complicating, the adaptation of the mesh. 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.

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, called ESOINN [6] (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. The operation and training of the network presents many similarities with those used in GCS. Nevertheless, it more closely resembles a merger between a CHL and a NG. The training phase and the various algorithms applied at every stage are detailed below:

1. Create an empty set of nodes A
2. Create an empty set of interconnections between nodes $C \subset AxA$
3. Insert two nodes in the A and assign random values to weights W.
4. Select a pattern p $\xi_p = (x_{p1},...,x_{ps})$ of the data set I with dimension R^s where x_{pj} represents the intensity of luminescence in probe j of individual i.
5. Search nodes: a_1 node closest to the pattern of entry and a_2 the second node closest to the pattern of entry, with $a_i \in A$, using Euclidean distance as the measure distance.

$$a_1 = \arg\min_{a \in A} \left\| \xi_p - W_a \right\|, \quad a_2 = \arg\min_{a \in A \setminus \{a_1\}} \left\| \xi_p - W_a \right\| \tag{4}$$

Up to this point the same steps are taken without changing any of the CHL algorithm techniques. Next the part concerning the ESOINN network begins, and modifications are made in order to automatically adjust various parameters.

6. If the distance $a_1 > T_i$ or $a_2 > T_i$ add a new node in that position and continue with the step

$$T_i = \begin{cases} \max_{j \in N_i} \left\| W_i - W_j \right\| & N_i \neq \phi \\ \max_{j \in N_i} \left\| W_i - W_j \right\| & N_i = \phi \end{cases} \tag{5}$$

Being N_i the set of neighbouring nodes of i

7. Increase the age of nodes connected with a_1

$$e_i(t+1) = e_i(t) + 1 \tag{6}$$

8. If necessary, establish a new connection or delete it between a_1 and a_2 according to the constraints:
   - If either of the nodes is not associated with any subclass or both are in the same, then create a new connection between them and assign an age value of zero.
   - If both nodes are in different subclass A, B, then to calculate the greater density for every subclass A_{max} B_{max}. If some of the next conditions is true, both subclass are joint and a new connections is created, otherwise if there was a connection it is deleted.

$$\min(a_1) > \alpha_A A_{\max} \text{ or } \min(a_1) > \alpha_B B_{\max} \tag{7}$$

Being

$$\alpha_a = \begin{cases} 0.0 & 2.0 * mean_A \geq A_{\max} \\ 0.5 & 3.0 mean_A \geq A_{\max} > 2.0 * mean_A \\ 1.0 & A_{\max} > 3.0 * mean_A \end{cases}$$

$$mean_A = \frac{1}{N_A} \sum_{i \in A} h_i$$

This ensures that connections with nodes with a higher age be subsequently deleted.

9. Update the density of a_1 depending on the distance to the neighbour nodes

$$h_i = \frac{1}{N} \sum_{j=1}^{n} \sum_{k=1}^{\lambda} p_{ik} \tag{8}$$

Where N is the number of $\sum_{k=1}^{\lambda} p_{ik} \neq 0$, n the number of periods, λ the number of pattern for period $\lambda = 50 + \sqrt{\#D}$

$$p_{ik} = \begin{cases} \dfrac{1}{\left(1 + \dfrac{1}{m}\sum_{j=1}^{m} \|W_i - W_j\|\right)^2} & i \text{ winner} \\ 0 & eoc \end{cases} \quad (9)$$

Where m is the number of neighbours to node *i*.

10. Increase the number of times winner of the neuron

$$M_{a_1}(t+1) = M_{a_1}(t) + 1 \quad (10)$$

11. 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 [4].

$$\Delta W_{a_1} = n_1(M_{a_1})(\xi - W_{a_1})$$
$$\Delta W_i = n_2(M_{a_i})(\xi - W_{a_i}) \text{ with } i \in N_i \quad (11)$$

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

12. Delete the connections with higher age. The ages are typified z≡N(0,1) and are removed those whose values are in the region of rejection with k>0. α =0.05.
13. If the number of iterations is a multiple of λ
    1. Update subclass to which each neuron belongs bearing in mind the highest local density of each neuron with its neighbours.
    2. Delete nodes that meet any of the following conditions
        1. $\forall a \in A / \#N = 2 \Rightarrow h_a < c_1 \sum_{j=1}^{N_A} h_j / \#N$

           N is the set of neighbour nodes of a, c1=0.001
        2. $\forall a \in A / \#N = 1 \Rightarrow h_a < c_2 \sum_{j=1}^{N_A} h_j / \#N$

           N is the set of neighbour nodes of a, c_2=1.0.
14. The clustering of elements is carried out bearing in mind the connections among the neurons.
15. If all input patterns have been passed then a KS-Test [3] 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. α =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::

$$d(n,m) = \sum_{i=1}^{s} f(x_{ni}, x_{mi}) * w_i \quad (12)$$

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 [7] Distance that is given for:

$$f(x,y) = \sqrt[p]{\sum_i |x_i - y_j|^p} \text{ con } x_i, y_j \in R^p \quad (13)$$

This dissimilarity measure weighs those probes that have a less uniform distribution, since these variables don't allow a separation. In order to validate the selected distance, the Kruskal-Wallis [1] test was carried out. It was verified whether the proportion of errors in the classification of each one of the individuals were the same for each group, bearing in mind the different measure. The results are shown in table 1 α =0.05.

Table 1. Comparing errors. * Different median and = equal, (-) median of column less than median of row.

	Minkowski	Eclidean	Max Absolute
Minkowski			
Eclidean	=		
Max Absolute	*(-)	*(-)	

3.3. 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.

4 Case Study

In the case study presented in the framework of this research 248 samples are available from analyses performed on patients either through punctures in marrow or blood samples, which have been hybridized and analyzed through Exon arrays manufactured by Affymetrix. 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 3. 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.

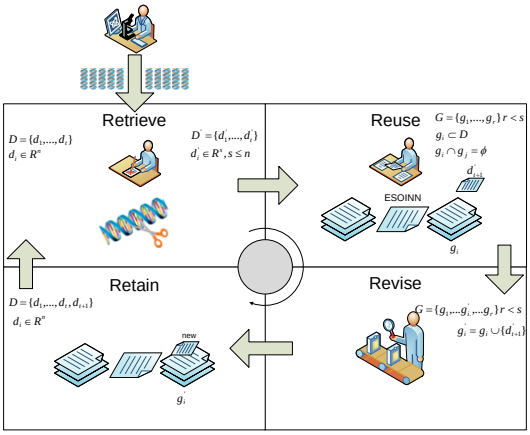

Figure 1. Proposed CBR model

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 3.1. during the retrieve phase the data are normalized by the RMA algorithm [8] and the dimensionality is reduced bearing in mind, above all, the variability, distribution and correlation of probes. 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 patients are first grouped into clusters. 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 [6] one. The structure of this neural network has been described in detail in Section 3.2. 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 (12). 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 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 $D = \{d_1,...,d_t,d_{t+1}\}$.

5 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.

In the study of leukemia on the basis of data from Exon arrays, the process of filtering data acquires special importance. 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). Each case (microarray experiment) stores 289.961 ESTs corresponding to the expression level of thousands of genes. The data consisted of 5.500.000 scanned intensities. 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. In addition, the selected variables resulted in a classification similar to that already achieved by experts from the laboratory of the Institute of Cancer of Salamanca. The error rates have remained fairly low especially for cases where the number of patients was high. Figure 2 shows an Intensity Plot with the classification performed for patients from groups CLL and ALL. 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, it is represented by the black lines that appear separated. Figure 2b shows the classification obtained for the LAL patients. We can see that, although the ranking is not bad, the proportion of individuals misclassified is higher. Groups that have fewer individuals are those with a higher classification error.

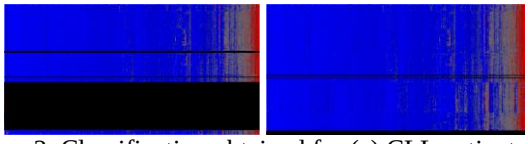
Figure 2. Classification obtained for (a) CLL patients and (b) ALL patients.

In a similar way we proceeded to evaluate the classification for the rest of the groups. Figure shows the total number of patients from each group and the number of misclassifications. As can be seen in Figure 3, groups with fewer patients are those with a greater error rate.

Figure 3b, shows the percentage of error in each group. 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.

Figure 3. Classification errors (a) numerical (b) percentage

The final classification was compared with the data obtained using a dendogram [18] and PAM [9] (Partitioning Around Medoids). The proportion of errors in every group was calculated and the Kurskal-Wallis [12] test was applied to determinate if the median of these proportions was equal. The results are shown in table 2.

Table 2. Comparation of methods. * different median and = equal, (-) median of column less than median of row

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

As demonstrated, 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 allow the detection of pathology, and the facilitation of a classification and reliable diagnosis, as shown by the results presented in this paper

Acknowledge

This development has been partially supported by the Spanish Ministry of Science project TIN2006-14630-C03-03. Special thanks to the Salamanca Institute of Cancer for the information and technology provided.

References

[1] Aaronson, J., Juergen, H, Overton, G.: Knowledge Discovery in GENBANK. In Proceedings of the First International Conference on Intelligent Systems for Molecular Biology, (1993) 3–11.

[2] Affymetrix, GeneChip Human Exon 1.0 ST Array, www.affymetrix.com/products/arrays/specific/Exon.affx

[3] Brunelli, R.: Histogram Analysis for Image Retrieval. Pattern Recognition, Vol. 34, (2001) 1625-1637.

[4] 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.

[5] 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.

[6] Furao, S, Ogura, T, Hasegawa, O.: An enhanced self-organizing incremental neural network for online unsupervised learning. Neural Networks, Vol. 20, (2007). 893-903.

[7] 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, No. 1 (1972), 255-259

[8] 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, (2003) 4, 249-264.

[9] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. (1990).

[10] Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics, (1982) 59-69.

[11] Kolodner J.: Case-Based Reasoning. Morgan Kaufmann (1993).

[12] Kruskal W, Wallis W.: Use of ranks in one-criterion variance analysis. Journal of American Statistics Association (1952)

[13] Lander E.S. et al.: Initial sequencing and analysis of the human genome. Nature (2001). 409, 860 - 921

[14] Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. H.: High density synthetic oligonucleotide arrays. Nature Genetics, (1999) 21:20–24.

[15] Martinetz, T.: Competitive Hebbian learning rule forms perfectly topology preserving maps. Amsterdam : Springer,. ICANN'93: International Conference on Artificial Neural Networks. (1993). 427-434.

[16] Riverola F., Díaz F., Corchado J. M.: Gene-CBR: a case-based reasoning tool for cancer diagnosis using microarray datasets. Computational Intelligence. ISSN 0824-7935. Vol. 22 , num. 3-4, 254-268.(2006)

[17] 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

[18] Saitou, N., Nie, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. (1987) Evol. 4: 406-425

[19] SEER (Surveillance Epidemiology and End Results), U.S. National Cancer Institute http://seer.cancer.gov/. (2007)

[20] Shen, F.: An algorithm for incremental unsupervised learning and topology representation. Tokyo: Ph.D. thesis. Tokyo Institute of Technology, (2006).

[21] Shortliffe, E.H., Cimino, J.J.: Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer. (2006).

[22] Tsoka S., Ouzounis C.: Recent developments and future directions in computational genomics. FEBS Letters, Vol. 480 (1), (2000) 42-48