Gene Expression Patterns and Cancer Classification: A Self-Adaptive and Incremental Neural Approach

Francisco Azuaje
Artificial Intelligence Group, Department of Computer Science, University of Dublin, Ireland.
Francisco.Azuaje@cs.tcd.ie

Abstract

The automated interpretation of data originating from the human genome may play a crucial role in cancer treatment. Here a new computational approach to the discovery and analysis of gene expression patterns is presented and applied to the recognition of B-cell malignancies as a test set. Using cDNA microarrays data obtained from a previous study, an unsupervised and self-adaptive neural network model known as Growing Cell Structures is able to identify normal and diffuse large B-cell lymphoma (DLBCL) patients. Furthermore, it discovers the distinction between patients with molecularly distinct types of DLBCL without previous knowledge of those subclasses.

Keywords— Bioinformatics, gene expression analysis, cancer classification, data mining, neural networks.

1. Introduction

The completion of the genome of several model organisms represents a fascinating opportunity to explore important normal and abnormal biological phenomena. *DNA microarrays* provide a systematic and comprehensive framework to achieve this goal. This method generates data that reflects the logic and structure of the genetic program. Thus, the automation of the genome expression analysis process should significantly contribute to the understanding of gene function and the development of better cancer diagnosis strategies for instance [1].

The systematic classification of types of tumours is crucial to achieve advances in cancer treatment and research. It has been suggested that the specification of therapies according to tumour types differentiated by *pathogenetic* patterns may maximize the efficacy of the treatment and minimize toxicity on the patients [2], [3]. Several limitations about the conventional classification techniques based on morphological features of the tumour have been reported in the literature [1]. Moreover, by analysing complex patterns defined by molecular markers, it has been demonstrated that there are subtypes of *acute leukaemia*, prostate cancer and *non-Hodgkin's lymphomas* [2].

There are two useful tasks in cancer classification: prediction of classes and discovery of classes. The prediction task consists of the assignment of particular tumour samples to known types of cancer. The discovery task refers to the unsupervised identification of relevant groups of samples and the characterisation of subtypes of cancer. This research aims to implement a discovery task based on a global expression analysis approach.

1.1 Gene expression and cancer classification

One of the methods to study the role of a gene in a specific organism consists of the monitoring of the expression level of a gene. It has been shown that specific patterns of gene expression occur during different biological states such as *embryogenesis*, cell development and during normal physiological responses in tissues and cells [4]. Generally speaking the expression of a gene provides a measure of "how active" a specific gene is under certain biochemical conditions [5]. This level of expression is related to the relative concentration of messenger *RNA* (*mRNA*), which encodes the gene under consideration [3], [4]. The generation of quantitative expression patterns of many genes in parallel can be achieved by using techniques based on *complementary DNA* (*cDNA*) *microarrays* [5], [6]. Schena et al., for instance, describe a method for monitoring gene expression, in which differential expression is demonstrated by a simultaneous two-colour hybridisation scheme [5]. The reader is referenced to Schena et al. [5] and Eisen and

Brown [6] to find detailed information about the process of gene expression monitoring.

A number of authors have applied hierarchical clustering to organise genes into dendrograms based on their expression patterns [2], [7]. Eise et al., for instance, implement a clustering method based on *pairwise average-linkage* cluster analysis [5]. One of the main disadvantages of this type of clustering techniques is that the identification of categories and informational associations is left to the observer. Additionally, the complexity of the cluster visualisation task can be directly proportional to the number of elements to be grouped. Recently, a number of expression analysis techniques have been based on more advanced data mining approaches.

Machine learning techniques such as neural networks are adequate for this type of analysis for their well-known pattern recognition and data organisation capabilities [8], [9]. Advanced neural learning algorithms have not only improved the accuracy, reliability and efficiency of many medical pattern recognition systems, but they also show several advantages for the implementation of decision support systems in physiological genomics [10], [11].

Tamayo et al. have illustrated the value of Kohonen's Self-Organising Feature Maps (SOFM) [12] to interpret gene expression patterns during yeast growth cycle and *haematopoietic* differentiation [13]. They identify predominant gene expression patterns in those biological processes that suggested, for instance, novel hypotheses about haematopoietic differentiation useful for the treatment of *acute promyelocytic leukaemia*. Similarly based on a SOFM, Golub et al. [2] approach the problem of molecular classification of cancer. They propose a procedure that automatically discovers the distinction between *acute myeloid leukaemia and acute lymphoblastic leukaemia* based on the clusters obtained after training the network with a small set of cases. Some of the disadvantages of this approach can be summarise as follows:

- The topology of the network has to be pre-defined by the user, and it is not adapted according to the distribution of the data under consideration;
- The learning process is governed by a number of parameters highly dependent of the user (learning rates and number of learning cycles, for instance). Moreover, the user has to define the time-dependence of a number parameters;
- Training the network with a small number of samples can significantly reduce its prediction capabilities. Similarly, the training phase may become computationally expensive by processing massive data sets characterised by high dimensional representations.

1.2 Application domain and research goals

A recent effort to understand how genes contribute to disease approaches the discovery of sub-classes of diffuse large B-cell lymphoma (DLBCL) by using expression analysis [3]. *B-lymphocytes* are a fundamental component of the body's immune system. DLBCL is a malignancy of mature *B-lymphocytes*, with a high annual incidence in western countries. It has been shown that the discovery of sub-classes in DLBCL has not been successful by relying exclusively on morphological features [3]. Alizadeh et al. [3] demonstrate that the molecular profile of a tumour obtained from cDNA microarrays can indeed be interpreted as a robust and clearer picture of the tumour's biology. Additionally, they demonstrate the existence of two molecularly distinct forms of DLBCL that indicate different stages of B-cell differentiation.

The knowledge discovery system presented in this paper is tested on the DLBCL domain. It is based on the analysis of the cDNA microarray data generated by Alizadeh et al. in the study referenced above. The research reported here aims to implement an automated and unsupervised approach to the identification of expression patterns and classes of cancer. The proposed method is based on an artificial neural learning model known as *Growing Cell Structures* (GCS) [14], which addresses some of the weak aspects shown by traditional gene expression analysis methods. This approach may provide an effective, efficient, inexpensive and user-friendly option to support diagnosis tasks and research [11].

The remainder of this article is organised as follows. Section 2 presents an introduction to the GCS-based clustering model. Section 3 describes the methods implemented and relevant results from the classification and knowledge discovery experiments. Section 4 presents a discussion of the results and possible future work to be developed.

2. The GCS clustering model

GCS neural networks [14] are a variation of the Kohonen's SOFM [15]. GCS offers several advantages for the implementation of gene expression pattern interpretation systems over both non-self-organising neural networks and the Kohonen SOFM

[16], [17]. Some of those advantages are summarised as follows:

a) GCS is a neural network with a self-adaptive topology which is highly independent of the user;
b) The network structure is automatically determined from the domain data and thus reflects the underlying characteristics of the domain in the very topology of the network;
c) In contrast to Kohonen self-organising maps, the GCS self-organising model consists of a small number of constant parameters; there is no need to define time-dependent or decay schedule parameters;
d) In contrast to Kohonen self-organising maps, the adaptation rules of the GCS also enable the processing of data with changing probability distributions [17];
e) Its ability to perform problem-dependent error measures allows the implementation of better adaptive data representations (insertion and deletion of cells) in comparison with static-topology models;
f) Its ability to insert and delete cells allows the accurate estimation of probability densities of input signals;
g) Its ability to interrupt a learning process or to continue a previously interrupted one permits the constructions of incremental and dynamic learning systems;
h) GCS networks have demonstrated their capacity to process both small and high dimensionality data in several application domains in contrast to back-propagation networks for instance [11];
i) Because the final structure of a GCS network depends on the underlying data, it can be used for multiple data analysis and knowledge discovery tasks such as visualisation, clustering, and navigation.

The type of GCS implemented in this work is characterized by a two-dimensional space, where the units (cells) are connected and organised in triangles. The GCS model and programs applied in this research have demonstrated to be successful in a number of medical decision support domains [11]. The initial topology for all the experiments in this work is a single triangle as illustrated in Figure 1a. Each cell in the network is associated with a weight vector, $\boldsymbol{w}$, which is of the same dimension as the input data. At the beginning of a learning process or clustering, the weight vector of each cell is initialised with random values. The basic learning process in a GCS network consists of topology modification and weight vector adaptations [17]. Based on a set of *n training cases*, learning in a GCS is carried out as a sequence of *k input presentation epochs*. Within a single epoch, the network is presented with all training cases, *n*, in a sequential fashion; hence an epoch is characterised by the number, *n,* of training cases. For each training case presentation, the network performs a so-called *learning cycle*, which may result in topology modification and weight vector adaptation that are described in [17]. Figure 1b illustrates a GCS network's topology changes after a number of learning cycles. The cell connections in the diagram reflect the weight vector distances or similarity of corresponding clusters.

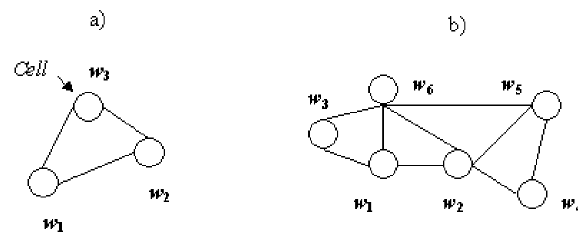

Figure 1. a) Initial topology of a GCS. b) GCS topology and clusters formed after some learning cycles.

GCS modify the overall network structure by *inserting* new cells into those regions that represent large portions of the input data [17]. When higher statistical accuracy is of interest or when the probability density of the input space consist of several separate regions, a better model can be achieved by *removing* cells that do not contribute to the input data representation [17]. The frequency of cell insertion and update is controlled by the parameter λ, which represents the number of learning epochs between two cell insertions. Typically, λ is chosen as a multiple of the number of training cases. Figure 2 illustrates the insertion of a new cell, *s*, between cells *q* and *r* after λ epochs of input presentations, that is, after $\lambda \times n$ learning cycles (*n = number of training cases*). The new cell, *s*, is connected to its neighbour cells, *i*, in such a way that *s* forms a triangle structure with any two of its neighbours. The parameters and expressions that control the insertion process and define a new cell can be found in [17].

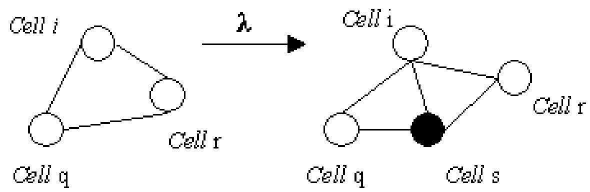

Figure 2. The GCS cell insertion process.

Cells that make little or no contribution to input data representation may be deleted from the GCS topology. A cell's contribution is determined by a number of mathematical expressions and parameters that are described in [17]. Figures 3a and 3b depict a GCS structure before and after deletion of the cell *j*. All the connections to cell *j* are removed and the structure of triangles of the network needs to be preserved. Note, that two clusters of cells have been generated as a consequence of such deletion. One cluster contains cells 1, 6, and 5, the other cluster contains cells 2, 3, and 4.

After learning, each sample (input vector) is categorised into one of the existing cells. Thus, the resulting product is a network of interconnected cells that cluster a number of samples according to their patterns of similarity.

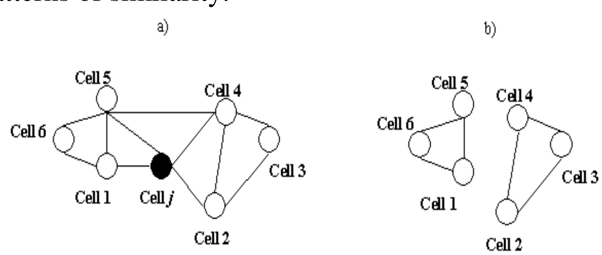

Figure 3. The GCS cell deletion process.

3. Methods and results

The expression levels from a number of genes with suspected roles in processes relevant in DLBCL were used as features for the automatic classification of a number of B-cell samples. The data consisted of 63 cases (45 DLBCL and 18 normal) described by the expression levels of 23 genes such as *CD10*, *BCL-6*, *TTG-2*, *IRF-4* and *BCL-2* [3]. The data was obtained from a recent study published by Alizadeh et al. [3]. The full expression data and experimental methods are available on the WWW site of Alizadeh et al. (http://llmpp.nih.gov/lymphoma).

Several GCS networks were used to cluster the samples based on different learning parameters. Figure 4 depicts one of the clustering topologies obtained after implementing a GCS with 2500 input presentation epochs, inserting a new cell every 500 epochs (λ = 500) and deleting irrelevant cells every 1000 epochs. The reader is referenced to [11] in order to find details about the selection of these and other learning parameter values. The resulting GCS consists of 6 cells or clusters containing the normal and DLBCL samples. In this diagram the cell connections do not reflect the weight vector distances. It shows that each cell corresponds to a representative cluster of the normal and DLBCL classes. For instance, Cells 4 and 6 categorise only normal and DLBCL samples respectively. The majority of the samples recognised by Cells 1 and 5 belong to the class DLBCL. Cell 3 recognises samples belonging only to the category DLBCL.

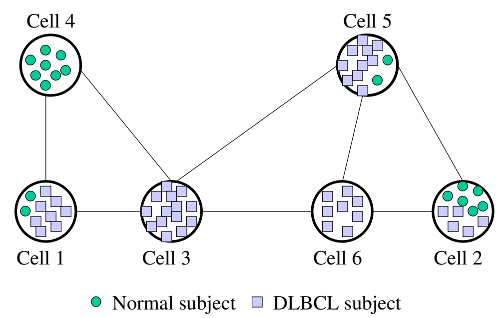

Figure 4. Clustering of normal and DLBCL samples based on a GCS network.

The next step is to analyse the composition of the clusters in order to discover or confirm relevant biological relationships. A recent study performed by Alizadeh et al. [3] demonstrated that the DLBCL subjects analysed here belong to distinct molecular sub-types of DLBCL. The two sub-classes of DLBCL identified by Alizadeh et al. are known as: *germinal centre B-like DLBCL* (*GC B-like DLBCL*) and *activated B-like DLBCL*, which are characterised by expression patterns representative of different stages of B-cell differentiation [3]. Figure 5 describes the DLBCL clusters obtained from Figure 4 in terms of the molecular subtypes mentioned above. Cells 1 and 3 comprise all of the GC B-like subjects. Cells 2, 5 and 6 represent the clusters encoding the activated B-like subjects. Thus, this GCS network consists of two macro-clusters, A and B, which identify the GC B-like and the activated B-like subjects respectively.

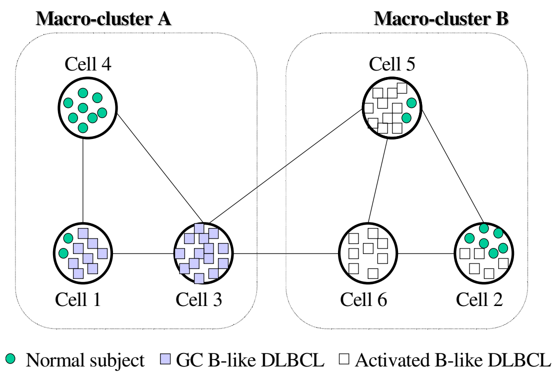

Figure 5. The recognition of DLBCL subtypes.

Table 1. Summary of the differences between GC B-like (Cell 3) and activated B-like (Cell 6) clusters (Mean ± SE of the normalised expression levels).

Gene	GC B-like DLBCL (14 samples)	Activated B-like DLBCL (8 samples)	Sig.
CD21	0.51± 0.05	0.28±0.04	$p < 0.005$
CK-G2	0.41± 0.05	0.41± 0.06	N.S
CD22	0.44± 0.03	0.23± 0.05	$p < 0.002$
WIP/HS	0.71± 0.04	0.64± 0.04	N.S
JAW1	0.68± 0.03	0.31± 0.04	$p < 0.001$
APS	0.56± 0.06	0.37± 0.06	$p < 0.05$
PC43	0.55± 0.05	0.58± 0.04	N.S
BCL-6	0.71± 0.04	0.53± 0.03	$p < 0.005$
CD27	0.53± 0.05	0.58± 0.03	N.S
PKC-D	0.61± 0.04	0.65± 0.04	N.S
APR	0.34± 0.03	0.36± 0.04	N.S
IL-10	0.54± 0.05	0.71± 0.06	$p < 0.05$
c-myc	0.36± 0.03	0.43± 0.02	N.S
BCL-2	0.39± 0.04	0.39± 0.03	N.S
PBEF	0.35± 0.02	0.58± 0.03	$p < 0.005$
C- D2	0.27± 0.02	0.43± 0.04	$p < 0.002$
CD44	0.36± 0.03	0.51.± 0.05	$p < 0.01$
IL-6	0.44± 0.05	0.38± 0.02	N.S
SP100	0.36± 0.05	0.41± 0.04	N.S
Ld2	0.33± 0.03	0.48± 0.03	$p < 0.005$
CD10	0.51± 0.04	0.47± 0.05	N.S
TTG-2	0.70± 0.03	0.54± 0.06	$p < 0.05$
IRF-4	0.52± 0.02	0.60± 0.04	N.S

SE: standard error. N.S: no significance. Sig.: Significance. CK-G2: Casein kinase gamma 2.

The clusters that include samples from both normal and DLBCL categories also illustrate other interesting aspects about this expression pattern recognition process. For instance, all of the samples categorised by Cell 1 belong to the GC B-like DLBCL category, except two samples (tonsil GC B cells and tonsil GC centroblast) from the category normal. However, all of the samples encoded by Cell 1 correspond to the same B-cell differentiation stage [3]. Moreover, almost all of the genes that define GC B-like DLBCL were highly expressed in those normal GC B cells [3].

Based on the resulting normal and DLBCL clusters, the corresponding normalised gene expressions levels of the subjects were compared in order to discover possible associations between them. For example, Table 1 illustrates the relationship between two DLBCL clusters (sub-types of DLBCL) and their corresponding gene expression levels. They compare a GC B-like cluster (Cell 3) against a cluster of activated B-like subjects (Cell 6). The means and standard errors of the normalised expression levels for each cluster are portrayed. Additionally, statistical tests are performed to show the significance of the differences indicated by each sub-class of DLBCL (column 4). The significance of the differences between the expression levels of the obtained clusters was established by means of the well-known *two-tailed t-test*.

4. Discussion and conclusions

These results suggest that a GCS approach can be useful for the unsupervised classification and discovery of classes of cancer based on gene expression patterns. This automated approach was able to identify DLBCL clusters, which are linked to a number of biological findings recently reported [3]. The GCS-based model has shown to identify molecular sub-types of DLBCL without any previous knowledge about their existence. Thus, this class discovery approach not only has identified the distinction between normal and DLBCL subjects, but also it has recognised two subtypes of DLBCL by using the expression levels of 23 genes. The differences between the obtained DLBCL clusters have demonstrated to be significant in terms of the gene expressions under study. The results portrayed in Table 1 confirm that a number of genes can be used as markers to differentiate between normal and DLBCL subjects, and between subtypes of DLBCL subjects [3].

This automated discovery technique may support the crucial task of identifying clinical cases and subtypes of cancer. The analysis of the obtained clusters confirms the suspected roles of the selected genes in processes relevant to DLBCL.

Future work comprises the implementation of clustering and discovery procedures involving expression patterns of higher dimensionality (additional cases and genes for instance). These automated methods may represent not only a promising tool to explore genetic mechanisms in the development of a type of cancer, but also to support the search for key processes (genes) that differentiate multiple classes of cancer [3]. Other aspects that deserve further investigations include the automatic discovery of attribute relevance or weight, i.e. the importance of a gene for the prediction of a category, and the implementation of advanced gene selection procedures. Furthermore, the GCS-based model is currently being compared against other advanced gene expression analysis approaches such as the SFAM approach [10].

The techniques outlined in this paper may yield significant benefit in the improvement of decision support systems for the analysis and interpretation of genome expression patterns.

References

[1] F. Azuaje, Interpretation of genome expression patterns: computational challenges and opportunities, to be published by *IEEE Engineering in Medicine and Biology, November 2000*.

[2] T.R. Golub, D.K Slonim, P. Tamayo, C. Huard, M. Gassenbeck, J.P. Mesirov, H. Coller, M.L Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science,* 286, 531-537, 1999.

[3] A.A. Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature,* 403, 503-511, 2000.

[4] P.J. Russel, *Fundamentals of genetics*, Second Edition, San Francisco, Addison Wesly Longman Inc., 2000.

[5] M. Schena, D. Shalon, R.W. Davis and P.O, Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270, 467-471, 1995.

[6] M.B Eisen and P.O. Brown, DNA arrays for analysis of gene expression, *Methods Enzymol.*, 303, 179-205, 1999.

[7] W.M. Fitch and E. Margoliash, Construction of phylogenetic trees, *Science,* 155, 279-284, 1967.

[8] B. Ripley, *Pattern Recognition and Neural Networks,* Cambridge, England, Cambridge University Press, 1996.

[9] F. Azuaje, W. Dubitzky, P. Lopes, N. Black, K. Adamson, X. Wu, and J. White, Predicting Coronary Disease risk Based on Short-Term RR Intervals Measurements: A Neural Network Approach, *Artificial Intelligence in Medicine*, 15, 275-298, 1999.

[10] F. Azuaje, Making Genome Expression Data Meaningful: Prediction and discovery of classes of cancer through a connectionist learning approach, to be published in the *Proceedings of the IEEE Symposium on Bioinformatics and Biomedical Enineering* (BIBE-2000).

[11] F. Azuaje, W. Dubitzky, N. Black, and K. Adamson, Discovering Relevance Knowledge in Data: A Growing Cell Structure Approach, *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 30, 448-460, 2000.

[12] T. Kohonen, *Self-Organizing Maps*, Heidelberg, Springer, 1995.

[13] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmistrovsky, E. Lander and T.R. Golub, Interpreting patterns of gene expression with self-organising maps: methods and applications to hematopoietic differentiation, *The Proceedings of the National Academy of Sciences of U.S.A,* 96, 2907-2912, 1999.

[14] B. Fritzke, "A Self-Organizing Network that Can Follow Non-Stationary Distributions", *in Proceedings of International Conference on Artificial Networks (ICANN-97),* 1997, pp. 613-618.

[15] T. Kohonen, *Self-Organizing Maps*. Heidelberg, Germany: Springer, 1995.

[16] B. Fritzke, "Growing Self-Organizing Networks-Why?", *in ESANN'96: European Symposium on Artificial Neural Networks*, 1996, pp. 61-72.

[17] B. Fritzke, "Growing Cell Structures-A Self-Organizing Network for Unsupervised an Supervised Learning", *Neural Networks*, vol. 7, pp. 1441-1460, 1994.