A Computational Neural Approach to Support the Discovery of Gene Function and Classes of Cancer

Francisco Azuaje, *Member, IEEE*

***Abstract*—Advances in molecular classification of tumours may play a central role in cancer treatment. Here, a novel approach to genome expression pattern interpretation is described and applied to the recognition of B-cell malignancies as a test set. Using cDNA microarrays data generated by a previous study, a neural network model known as simplified fuzzy ARTMAP is able to identify normal and diffuse large B-cell lymphoma (DLBCL) patients. Furthermore, it discovers the distinction between patients with molecularly distinct forms of DLBCL without previous knowledge of those subtypes.**

***Index Terms*—Bioinformatics, cancer classification, data mining, gene expression analysis, neural networks.**

I. Introduction

THE systematic classification of types of tumours is crucial to achieve advances in cancer treatment and research. 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 [1], [2]. Several limitations about the conventional classification techniques based on morphological features of the tumour have been reported in the literature [1]. Moreover, by analyzing 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].

Thus, 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, and the discovery task refers to the identification of unrecognized subtypes.

In order to achieve a better insight into the problem of cancer classification, systematic approaches based on global gene expression analysis have been proposed [1], [2].

A. Gene Expression Profiling, Cancer Classification, and Related Research

Over the past ten years many scientists have combined efforts in the *Human Genome Project* in order to process and categorise gene sequences, but far fewer researchers have approached the problem of how genes actually contribute to disease using large-scale sequencing and expression data. These massive sources of information extracted from the genome project contain the keys to address fundamental problems relating to the prevention and treatment of diseases, biological evolution mechanisms, and the understanding of particular functional elements in the human organism. The knowledge of the coding sequences of virtually every gene in an organism is an exciting opportunity to develop methods to study the role of a gene in a specific organism or biological function.

One of such methods consists of the monitoring of the level of expression 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 [3].

Generally speaking the expression of a gene provides a measure of "how active" a specific gene is under certain biochemical conditions [4]. This level of expression is related to the relative concentration of messenger RNA $(mRNA)$, which encodes the gene under consideration [3], [4].

The study of gene expression of genes one by one has already provided a wealth of biological insight [5]. Thus, the next challenge has been to analyze the expression of many of the genes in parallel, in order to identify expression patterns at the level of the whole genome of an organism [4], [5].

The generation of quantitative expression patterns of many genes can be achieved by using techniques based on *complementary DNA* $(cDNA)$ *microarrays* [4], [6]. Schena *et al.*, for instance, describe a method for monitoring gene expression, in which differential expression is demonstrated by a simultaneous two-color hybridization scheme [4]. Without going into details, this method is based on the preparation of fluorescent DNA probes from two mRNA sources by using a method known as *reverse transcription* [6]. One set of probe is the "reference probe" and the other is obtained from the tissue where the gene expression needs to be examined (experimental sample). These probes are prepared in the presence of *fluorescein* and *lissamine-labeled nucleotide analogs*, respectively, for instance [4]. The two probes are mixed in equal proportions and allowed to hybridise to a microarray consisting of a series of "known cDNAs" deposited on glass slides. After hybridization, the fluorescence patterns scanned allow to represent a ratio of hybridization of the experimental cDNA probe to the reference probe, that is to say, the relative abundance of the gene in the experimental sample compared with the reference sample. Thus, this method provides a measure of gene expression for a specific sample. The reader is referenced to Schena *et al.* [4] and Eisen and Brown [6] to find detailed information about the process of gene expression monitoring.

Approaches that allow experts to have a systematic understanding of the processes under study are required in order to exploit the full potential of genome-scale experiments [7]. A

Manuscript received June 23, 2000; revised October 31, 2000.

The author is with the Artificial Intelligence Group and Centre for Health Informatics, Trinity College, Dublin 2, Ireland.

Publisher Item Identifier S 0018-9294(01)01573-7.

number of authors have used hierarchical clustering to organize genes into *phylogenetic* trees or dendrograms [2], [8]. 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 visualization task can be directly proportional to the number of elements to be grouped. Recently, a number of expression pattern analysis techniques have been based on machine learning models.

Machine learning techniques such as neural networks are adequate for this type of analysis for their well-known pattern recognition and data organization capabilities [9], [10]. Advanced neural learning algorithms have not only improved the accuracy, reliability and efficiency of many pattern recognition and data mining systems, but they also present several advantages for the implementation of decision support systems in physiological genomics [9], [11]. Tamayo *et al.* have illustrated the value of Kohonen's self-organizing feature maps (SOFMs) [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*. Also based on a SOFM, Golub *et al.* [1] 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 summarized as follows.

- The topology of the network has to be predefined 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 of 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 characterized by high-dimensional representations.

B. Classification of Diffuse Large B-cell Lymphoma (DLBCL) Using Gene Expression Data

A recent effort to understand how genes contribute to disease approaches the discovery of subclasses of DLBCL by using expression analysis [2]. *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 subclasses in DLBCL has not been successful by relying exclusively on morphological features [2]. Alizadeh *et al.* [2] demonstrate that the molecular profile of a tumor 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 class prediction and class discovery techniques presented in this paper are tested on the DLBCL domain. They are based on the analysis of the cDNA microarray data generated by Alizadeh *et al.* in the study referenced above.

C. Aims of this Research

This research aims to implement an automated approach to the prediction and discovery of classes of cancer based on the processing of gene expression data. The proposed technique consists of an artificial neural learning model known as *simplified fuzzy ARTMAP* (SFAM) [14], which addresses some of the weak aspects shown by traditional gene expression analysis methods. This approach may provide an effective, efficient and inexpensive option to support diagnosis tasks and research. The objective of the prediction task is to distinguish normal subjects from those with DLBCL by using a number of genes with known or suspected roles in the development of the disease. The objective of the discovery task is to identify subtypes of cancer from a population of subjects with DLBCL.

The remainder of this article is organized as follows. Section II presents an introduction to the SFAM model. Section III describes the data and methods implemented in this research. Section IV shows the prediction and discovery results obtained from several network architectures. Section V presents a discussion of the results, and their implications to the process of molecular classification of cancer and gene expression analysis. This section also presents possible future work to be developed.

II. Introduction to the SFAM Model

A SFAM is a version of the fuzzy ARTMAP neural network model [14], [15]. SFAM was designed to improve the computational efficiency of the fuzzy ARTMAP model with a minimal loss of learning effectiveness [14]. The "fuzzy" component in the name of this network refers to the fact that its learning process implements fuzzy logic operations in order to achieve a number of key pattern matching and adaptation functions [14]. The essential two-layer network architecture of a SFAM is depicted in Figs. 1–3. During learning, input data are presented to the SFAM, together with their respective teaching stimuli (categories to learn). The raw input data flows to a complement coder, which normalizes and expands the input to twice its original size. This expanded input vector, I, then is located into the *input layer*. Weights (W) from each node of the *output layer* reach down to "sample" the input layer. During the learning phase these weights form the associations between the input patterns and their corresponding category based on a number of adaptation steps. The *category layer* holds the categories or classes that the network has to learn. In the SFAM model, a single-output node can only encode (point to) a single category in the category layer. Moreover, the SFAM does not directly associate inputs at the input layer and the category layer. Such input patterns are firstly self-organized in "prototypical clusters" represented by the nodes in the output layer. Thus, if a cluster in the output

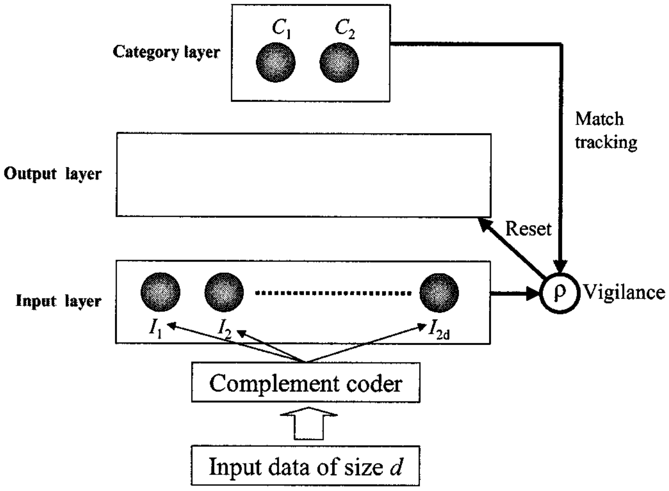


Fig. 1. A SFAM neural network before starting a learning process.

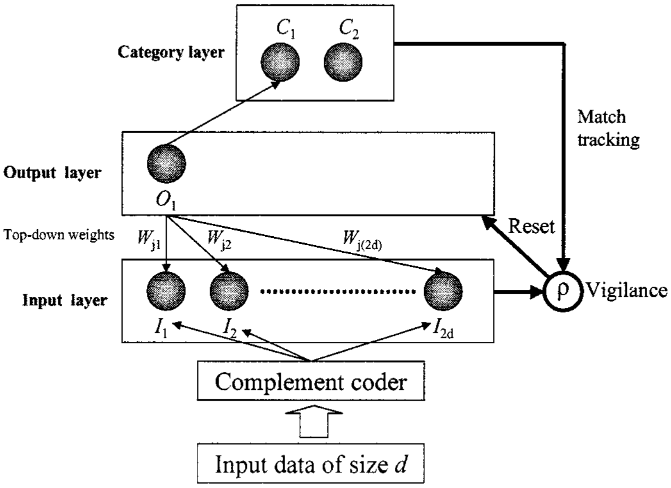


Fig. 2. A SFAM neural network after the first input pattern has been learned.

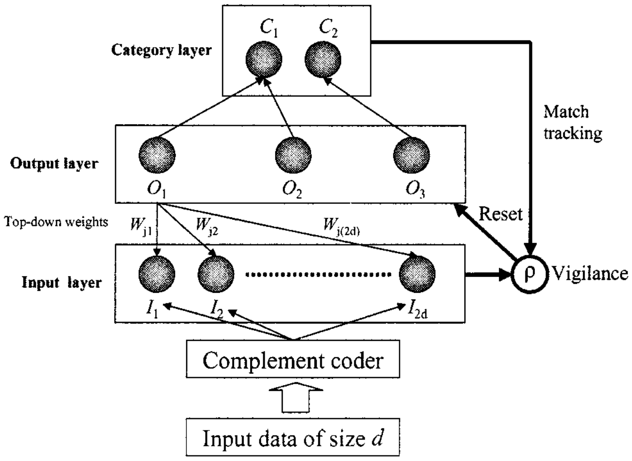


Fig. 3. A SFAM neural network after a number of learning steps.

layer does not match with the teaching category described in the category layer, a re-set signal is generated at the output layer, forcing the input pattern to be re-classified into an appropriate node (cluster) in the output layer. If no such output node exists, a new output node is created to classify the input. In this way, one output node is linked to only one category, but one category may be encoded by several nodes in the output layer. The user-selectable parameter, ρ, or vigilance parameter, determines how close a match is required between an input pattern and output node that encodes a category. The vigilance parameter, ρ, has value between zero and one, and indirectly controls the size of the output layer that will form during the learning phase. Generally, a higher value of ρ provides a better classification performance, although this must be balanced against the potential proliferation of output nodes. The reader is referenced to Kasuba [14] and Downs *et al.* [16] to find detailed information about the learning algorithm of the SFAM and its advantages in comparison to other pattern classification approaches.

As a way of illustration, Figs. 1–3 depict a number of situations during the learning process of a SFAM. Fig. 1 illustrates the initial state of a SFAM before learning to classify two categories, C_1 and C_2. In this case, there are no nodes represented in the output layer until the network has its first opportunity to "learn" a pattern. Once an input pattern is presented, an output node is formed to represent it (Fig. 2). Such an output node is linked to the category label indicated in the category layer. The matching and/or creation of output nodes as well as the adaptation of weights are based on the steps outlined above [14]. After a number of learning steps, the network consists of a number output nodes that encode a number of input patterns. Fig. 3 illustrates this situation without showing the weight connections of nodes O_2 and O_3. The output nodes can also be seen as subclasses of the taught categories C_1 and C_2 (Fig. 3). For instance, nodes O_1 and O_2 encode (cluster) input patterns that belong to category C_1, while O_3 groups patterns that belong to category C_2.

At the end of the learning phase, supervisory stimuli are removed, and the network can be tested by using new input patterns that "recall" previously learned associations.

In comparison to other neural network approaches, SFAM offers several advantages for the development of automated pattern recognition tools in gene expression analysis. Such advantages are summarized as follows.

1) In contrast to Kohonen self-organizing maps and backpropagation networks for instance, SFAM is a neural network based on a self-adaptive topology, which is highly independent of the user.
2) SFAM has demonstrated to improve significantly both the effectiveness and efficiency of several medical applications [16].
3) In contrast to Kohonen self-organizing maps and backpropagation networks for instance, the network structure is determined automatically from the domain data.
4) SFAM consists of only one constant user-dependent parameter.
5) SFAM, its originating models and related versions can operate in both supervised and unsupervised learning modes.
6) Successful learning can be achieved with only one pass through the data set [14], [16].
7) The SFAM does not perform optimization of an objective function, therefore, it is not constrained by the problem

of local minima that can occurs with back-propagation networks [16].

III. Materials and Methods

A. Obtaining the Gene Expression Data

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*. These data were obtained from a recent study published by Alizadeh *et al.* [2], who identified subgroups of DLBCL based on the analysis of the patterns generated by a specialized cDNA microarray technique [2], [6]. The full data and experimental methods are available on the World-Wide Web site of Alizadeh *et al.* (http://llmpp.nih.gov/lymphoma).

B. The SFAM Model

The SFAM, preprocessing and evaluation programs were written in C++ and Java. The SFAM algorithm was implemented based on the algorithmic descriptions presented in [14].

C. Prediction of Classes

The SFAM neural network was implemented to predict two classes of subjects: normal and DLBCL, based on data described above. The prediction performances were calculated by applying the data sampling method known as *round robin* or *leave-one* method [17]. This method can be summarized as follows: if there are n input patterns (cases or samples), the network performs its learning process on $(n-1)$ of them, and then it is tested on the input pattern that has been left out. This process is repeated n times so that every input pattern in the database is used as a testing example. Thus, the prediction performances obtained from the n cases are averaged in order to visualise the general prediction performance of a network. The main advantage of this technique is that it takes into account almost all the informational patterns available to train the network, without affecting the statistical significance of the testing results.

Several architectures were implemented to perform supervised learning processes for different values of ρ. Afterwards the systems were tested and evaluated based on their *accuracy, sensitivity* and *specificity*. Moreover, a system performing a voting strategy of the predictions made by a group of selected architectures was implemented. *Accuracy* relates to the full group of *true positives TP* (DLBCL subjects correctly classified), and *true negatives TN* (normal subjects correctly classified), to the total number, N, of tested cases. Thus, accuracy measures the ability of the SFAM to produce correct answers and is defined as follows:

$$Accuracy = (TP + TN)/N. \tag{1}$$

TABLE I
Classification of Normal and DLBCL Subjects Using Five Genes: Prediction Results

SFAM architecture	Prediction accuracy (%)	Prediction sensitivity (%)	Prediction specificity (%)
ρ = 0.1	65	84	17
ρ = 0.3	70	89	22
ρ = 0.5	73	91	28
ρ = 0.7	71	91	22
ρ = 0.95	**76**	**82**	**61**
Voting strategy	71	91	22

Sensitivity relates to the observed frequency of true positives, TP, to the frequency of false negatives, FN (DLBCL subjects classified as normal). *Sensitivity* measures the ability of the model to correctly identify the occurrence of a target subject (DLBCL in this case); it is defined as follows:

$$Sensitivity = TP/(TP + FN). \tag{2}$$

Specificity calculates a ratio based on the number true negatives, TN, and the *frequency of false positives FP* (normal subjects classified as DLBCL). *Specificity* measures the ability of the model to separate the target class, DLBCL, from the normal class; it is defined as follows:

$$Specificity = TN/(TN + FP). \tag{3}$$

D. Discovery of Classes

SFAM networks were trained to categorise two classes of subjects: normal and DLBCL, based on the gene expression data described above. Afterwards, their output layers were examined in order to visualise the subclasses or clusters self-organized during the learning process. Thus, at the end of a learning process a SFAM may recognize subclasses derived from the predefined classes without any additional knowledge. Furthermore, the gene expressions of the samples in each cluster are analyzed in order to discover possible associations between subtypes of DLBCL and the expressions of the genes under consideration. 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*.

IV. Results

A. Prediction of Classes

Table I shows some of the prediction results obtained after performing the experiments explained in Section III-C. The following genes with suspected roles in DLBCL were used as inputs for the SFAM network: BCL-6, BCL-2, CD20, TTG-2, and IRF-4. The different values for ρ are shown in the first column.

The best prediction performance was obtained by using a $\rho =$ 0.95, which was also able to achieve the highest specificity. The best prediction performances for DLBCL subjects were obtained from the networks using $\rho = 0.5$, $\rho = 0.7$ and a voting strategy, but their ability to predict normal cases was drastically reduced. The voting strategy generates a prediction based on the voting of individual predictions made by each network. The prediction performances of classes were not affected by the order of presentation of the input patterns.

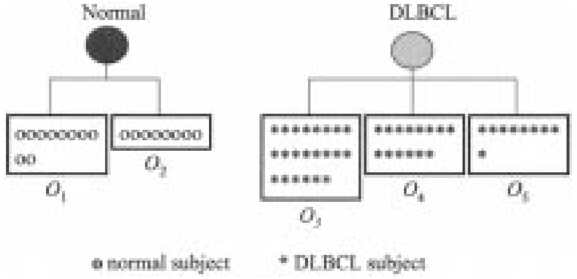


Fig. 4. Discovery of classes: clusters of normal and DLBCL subjects obtained at the end of a learning process with $\rho = 0.3$ and five input features.

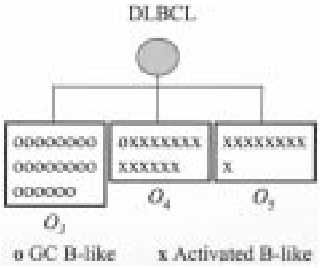


Fig. 5. Discovery of classes: clusters of DLBCL subjects described in terms of recent identified DLBCL subtypes [2] with $\rho = 0.3$ and five input features.

B. Discovery of Classes

Fig. 4 depicts the clusters obtained in the output layer of a SFAM at the end of a learning process with $\rho = 0.3$. The expressions of the genes BCL-6, BCL-2, CD20, TTG-2, and IRF-4 were used as input features for the SFAM network. The normal subjects are grouped into nodes O_1 and O_2, while the DLBLC subjects are categorised into three clusters (nodes O_3, O_4, and O_5). Clusters O_3, O_4, and O_5 categorise 22, 14 and 9 subjects, respectively.

The next step is to analyze the composition of the DLBCL clusters in order to discover possible relevant biological relationships. A recent study performed by Alizadeh *et al.* [2] demonstrated that the DLBCL subjects analyzed here belong to distinct molecular subtypes of DLBCL. Such a research comprised a systematic study of gene expressions in this type of malignancy. The two subclasses 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 characterized by expression patterns representative of different stages of B-cell differentiation [2].

Fig. 5 describes the DLBCL clusters obtained from Fig. 4 in terms of the molecular subtypes mentioned above. Cluster O_3 comprises all of the GC B-like subjects, except one that was categorised into cluster O_4. Furthermore, O_4 and O_5 represent the clusters encoding the activated B-like subjects.

Based on the resulting DLBCL clusters, the corresponding normalized gene expressions levels of the subjects are compared in order to discover possible associations between them. Tables II and III illustrate the relationship between DLBCL clusters (subtypes of DLBCL) and their corresponding gene expression levels. They compare the GC B-like cluster against the first and second clusters of activated B-like subjects, respectively. The means and standard errors of the normalized expression levels for each cluster are portrayed. Additionally, statistical tests are performed to show the significance of the differences indicated by each subclass of DLBCL (column 4).

TABLE II
SUMMARY OF THE DIFFERENCES BETWEEN GC B-LIKE AND ACTIVATED B-LIKE (O_4) CLUSTERS (MEAN ± SE OF THE NORMALISED EXPRESSION LEVELS). $\rho = 0.3$ AND FIVE INPUT FEATURES

Gene	GC B-like DLBCL cluster (O_3)	Activated B-like DLBCL cluster (O_4)	Significance
CD10	0.56 ± 0.04	0.48 ± 0.06	N.S
BCL-6	0.73 ± 0.03	0.52 ± 0.05	$p < 0.005$
TTG-2	0.67 ± 0.03	0.50 ± 0.06	$p < 0.01$
IRF-4	0.42 ± 0.04	0.59 ± 0.03	$p < 0.002$
BCL-2	0.38 ± 0.04	0.51 ± 0.05	$p < 0.05$

SE: standard error. N.S. no significance.

TABLE III
SUMMARY OF THE DIFFERENCES BETWEEN GC B-LIKE AND ACTIVATED B-LIKE (O_5) CLUSTERS (MEAN ± SE OF THE NORAMLISED EXPRESSION LEVELS). $\rho = 0.3$ AND FIVE INPUT FEATURES

Gene	GC B-like DLBCL cluster (O_3)	Activated B-like DLBCL cluster (O_5)	Significance
CD10	0.56 ± 0.04	0.19 ± 0.04	$p < 0.001$
BCL-6	0.73 ± 0.03	0.49 ± 0.06	$p < 0.002$
TTG-2	0.67 ± 0.03	0.21 ± 0.05	$p < 0.001$
IRF-4	0.42 ± 0.04	0.75 ± 0.03	$p < 0.001$
BCL-2	0.38 ± 0.04	0.68 ± 0.02	$p < 0.001$

SE: standard error.

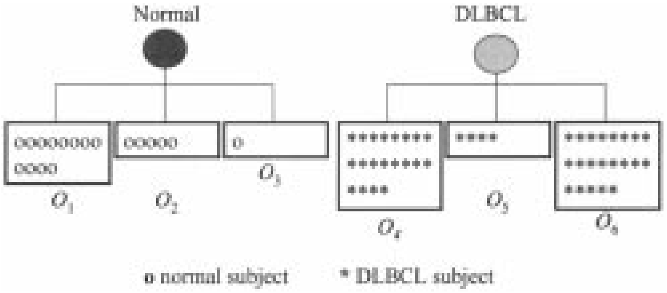


Fig. 6. Discovery of classes: clusters of normal and DLBCL subjects obtained at the end of a learning process with $\rho = 0.5$ and 23 input features.

Fig. 6 depicts the clusters obtained in the output layer of a SFAM at the end of a learning process with $\rho = 0.5$. This time the clustering process uses the expression levels from 23 genes as input features. These genes are shown in the left column of Table IV.

Fig. 7 describes the DLBCL clusters obtained from Fig. 6 in terms of the subtypes GC B-like and activated B-like DLBCL. Clusters O_4 and O_6 are mainly composed by GC B-like and activated B-like subjects, respectively. Tables IV and V compare cluster O_1 (normal) against clusters O_4 and O_6, respectively, in terms of the expressions of 23 genes with suspected roles in the process of DLBCL. Table VI compares the expression levels of clusters O_4 and O_6.

Table IV shows that the expression levels of genes CD22, JAW1, BCL-6, APR, IL-10, c-myc, BLC-2, cyclin D2, CD44, IL-6, TTG-2, and IRF-4 were significantly different in the normal cluster O_1 and the DLBCL cluster O_4. Moreover, there

TABLE IV
SUMMARY OF THE DIFFERENCES BETWEEN NORMAL (O_1) AND DLBCL (O_4) CLUSTERS (MEAN ± SE OF THE NORMALISED EXPRESSION LEVELS). $\rho = 0.5$ AND 23 INPUT FEATURES

Gene	Normal cluster (O_1)	DLBCL cluster (O_4)	Significance
CD21	0.40 ± 0.06	0.44 ± 0.05	N.S
Casein kinase gamma 2	0.50 ± 0.08	0.42 ± 0.03	N.S
CD22	0.57 ± 0.04	0.42 ± 0.04	p < 0.02
WIP/HS	0.62 ± 0.05	0.66 ± 0.04	N.S
JAW1	0.62 ± 0.03	0.48 ± 0.05	p < 0.02
APS	0.54 ± 0.06	0.44 ± 0.05	N.S
PC43	0.50 ± 0.06	0.61 ± 0.03	N.S
BCL-6	0.73 ± 0.04	0.58 ± 0.03	p < 0.005
CD27	0.59 ± 0.05	0.54 ± 0.03	N.S
PKC delta	0.63 ± 0.04	0.59 ± 0.04	N.S
APR	0.31 ± 0.02	0.46 ± 0.05	p < 0.02
IL-10	0.49 ± 0.06	0.64 ± 0.04	p < 0.05
c-myc	0.32 ± 0.05	0.45 ± 0.04	p < 0.05
BCL-2	0.36 ± 0.05	0.50 ± 0.04	p < 0.05
PBEF	0.38 ± 0.03	0.45 ± 0.05	N.S
Cyclin D2	0.29 ± 0.02	0.39 ± 0.04	p < 0.05
CD44	0.25 ± 0.03	0.47 ± 0.04	p < 0.001
IL-6	0.28 ± 0.05	0.48 ± 0.03	p < 0.005
SP100	0.35 ± 0.07	0.44 ± 0.04	N.S
Ld2	0.28 ± 0.04	0.35 ± 0.03	N.S
CD10	0.54 ± 0.05	0.41 ± 0.04	N.S
TTG-2	0.69 ± 0.04	0.53 ± 0.05	p < 0.05
IRF-4	0.41 ± 0.06	0.60 ± 0.03	p < 0.01

SE: standard error. N.S. no significance.

TABLE V
SUMMARY OF THE DIFFERENCES BETWEEN NORMAL (O_1) AND DLBCL (O_6) CLUSTERS (MEAN ± SE OF THE NORMALISED EXPRESSION LEVELS). $\rho = 0.5$ AND 23 INPUT FEATURES

Gene	Normal cluster (O_1)	DLBCL cluster (O_6)	Significance
CD21	0.40 ± 0.06	0.46 ± 0.05	N.S
Casein kinase gamma 2	0.50 ± 0.08	0.40 ± 0.05	N.S
CD22	0.57 ± 0.04	0.54 ± 0.05	N.S
WIP/HS	0.62 ± 0.05	0.65 ± 0.04	N.S
JAW1	0.62 ± 0.03	0.58 ± 0.03	N.S
APS	0.54 ± 0.06	0.34 ± 0.05	p < 0.02
PC43	0.50 ± 0.06	0.43 ± 0.05	N.S
BCL-6	0.73 ± 0.04	0.57 ± 0.05	p < 0.02
CD27	0.59 ± 0.05	0.55 ± 0.06	N.S
PKC delta	0.63 ± 0.04	0.40 ± 0.04	p < 0.001
APR	0.31 ± 0.02	0.43 ± 0.05	p < 0.05
IL-10	0.49 ± 0.06	0.58 ± 0.05	N.S
c-myc	0.32 ± 0.05	0.48 ± 0.05	p < 0.05
BCL-2	0.36 ± 0.05	0.64 ± 0.05	p < 0.001
PBEF	0.38 ± 0.03	0.38 ± 0.05	N.S
Cyclin D2	0.29 ± 0.02	0.40 ± 0.06	N.S
CD44	0.25 ± 0.03	0.37 ± 0.05	p < 0.05
IL-6	0.28 ± 0.05	0.52 ± 0.04	p < 0.001
SP100	0.35 ± 0.07	0.52 ± 0.05	N.S
Ld2	0.28 ± 0.04	0.27 ± 0.05	N.S
CD10	0.54 ± 0.05	0.48 ± 0.07	N.S
TTG-2	0.69 ± 0.04	0.31 ± 0.05	p < 0.001
IRF-4	0.41 ± 0.06	0.57 ± 0.05	p < 0.05

SE: standard error. N.S. no significance.

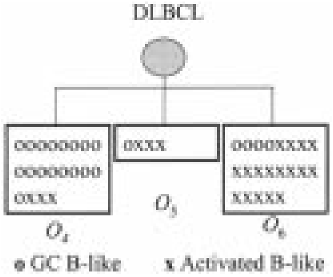


Fig. 7. Discovery of classes: clusters of DLBCL subjects described in terms of recent identified DLBCL subtypes [2], with $\rho = 0.5$ and 23 input features.

TABLE VI
SUMMARY OF THE DIFFERENCES BETWEEN THE DLBCL CLUSTERS O_4 AND O_6 (MEAN ± SE OF THE NORMALISED EXPRESSION LEVELS). $\rho = 0.5$ AND 23 INPUT FEATURES

Gene	DLBCL cluster (O_4)	DLBCL cluster (O_6)	Significance
CD21	0.44 ± 0.05	0.46 ± 0.05	N.S
Casein kinase gamma 2	0.42 ± 0.03	0.40 ± 0.05	N.S
CD22	0.42 ± 0.04	0.54 ± 0.05	N.S
WIP/HS	0.66 ± 0.04	0.65 ± 0.04	N.S
JAW1	0.48 ± 0.05	0.58 ± 0.03	N.S
APS	0.44 ± 0.05	0.34 ± 0.05	N.S
PC43	0.61 ± 0.03	0.43 ± 0.05	p < 0.005
BCL-6	0.58 ± 0.03	0.57 ± 0.05	N.S
CD27	0.54 ± 0.03	0.55 ± 0.06	N.S
PKC delta	0.59 ± 0.04	0.40 ± 0.04	p < 0.002
APR	0.46 ± 0.05	0.43 ± 0.05	N.S
IL-10	0.64 ± 0.04	0.58 ± 0.05	N.S
c-myc	0.45 ± 0.04	0.48 ± 0.05	N.S
BCL-2	0.50 ± 0.04	0.64 ± 0.05	p < 0.05
PBEF	0.45 ± 0.05	0.38 ± 0.05	N.S
Cyclin D2	0.39 ± 0.04	0.40 ± 0.06	N.S
CD44	0.47 ± 0.04	0.37 ± 0.05	N.S
IL-6	0.48 ± 0.03	0.52 ± 0.04	N.S
SP100	0.44 ± 0.04	0.52 ± 0.05	N.S
Ld2	0.35 ± 0.03	0.27 ± 0.05	N.S
CD10	0.41 ± 0.04	0.48 ± 0.07	N.S
TTG-2	0.53 ± 0.05	0.31 ± 0.05	p < 0.01
IRF-4	0.60 ± 0.03	0.57 ± 0.05	N.S

SE: standard error. N.S. no significance.

were significant differences between the normal cluster O1 and the DLBCL cluster O_6, in terms of the expression levels of genes APS, BCL-6, PKC delta, APR, c-myc, BCL-2, CD44, IL-6, TTG-2, and IRF-4 (Table V). Table VI illustrates the significance of the differences between the expression levels of genes PC43, PKC delta, BCL-2 and TTG-2 describing the DLBCL clusters O_4 and O_6.

V. DISCUSSION AND CONCLUSION

These results suggest that a simplified neuro-fuzzy approach can be useful for the prediction and discovery of cancer categories based on gene expression patterns. Despite the small size of the data sets and the disproportion of the number of normal and DLBCL subjects, one of the implemented architectures was able to distinguish between those categories with a considerable degree of accuracy (prediction task). The prediction ability of this approach may be better evaluated by including more subjects from both classes. Similarly, the application of alternative sampling procedures may provide a better visualization of the prediction performance of the model. The implementation of this type of automated classification systems may significantly contribute to the decision support tasks performed in a clinical environment. They may provide diagnostic tools to confirm or clarify unusual cases. The improvements of their prediction capabilities on diverse subclasses, may also collaborate in the maximization of treatment efficacy and the reduction of toxicity to the patients.

This computational model was able to identify DLBCL clusters, which are linked to a number of biological findings recently reported [2]. The SFAM network has shown to identify molecular subtypes of DLBCL without any previous knowledge about their existence. Thus, this class discovery approach not only identified the distinction between normal and DLBCL subjects, but also two subtypes of DLBCL by using the expression levels of five genes. One of the DLBCL categories recognized by the SFAM encodes the GC B-like DLBCL subtype, while the other ones represent subjects belonging to the category of activated B-like DLBCL. The differences between the obtained DLBCL clusters have demonstrated to be significant in terms of the gene expressions under study. Further experiments using the expression levels from 23 genes have also showed the significance of the differences between the obtained clusters. The results portrayed in Tables IV–VI confirm that a number of genes can be used as markers to differentiate between normal and DLBCL subjects, and between subtypes of DLBCL subjects.

This automated discovery technique may support the crucial research task of identifying subtypes of cancer. It may provide a better understanding of the biological functions of specific genes in the development of disease. The analysis of the obtained clusters confirms the suspected roles of the selected genes in processes relevant to DLBCL. These results suggest (Tables II and III), for instance, that the expression levels of genes CD10, BCL-6 and TTG-2 are higher in GC B-like than in activated B-like DLBCL. Similarly, the expression levels exhibited by IRF-4 and BCL-2 were significantly higher in the clusters encoding activated B-like DLBCL subjects. The genes CD10 and BCL-6 are well-established germinal centre markers [18]. Additionally, BCL-6 has been demonstrated to be the most frequently translocated gene in DLBCL [2]. Similarly, TTG-2, IRF-4, and BCL-2 have been shown to be translocated genes in lymphoid malignancies [19].

Future work should comprise the implementation of prediction and discovery procedures involving expression patterns of higher dimensionality (thousands of 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 [7]. Therefore, 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 SFAM-based model should be compared against other advanced machine learning approaches [11] and alternative quality evaluation procedures will be considered. The results and software systems of this and future research will be publicly available on the *Internet*.

The techniques implemented in this research may yield significant benefit in the improvement of decision support systems in cancer classification, and provide a better insight into the process of genome-wide expression analysis.

REFERENCES

[1] T. T. 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*, pp. 531–537, 1999.

[2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Bird, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[3] P. J. Russel, *Fundamentals of Genetics*, 2nd ed. San Francisco: Addison Wesly Longman Inc., 2000.

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

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *The Proceedings of the National Academy of Sciences of U.S.A.*, vol. 95, pp. 14 863–14 868, 1998.

[6] M. B. Eisen and P. O. Brown, "DNA arrays for analysis of gene expression," *Meth. Enzymol.*, vol. 303, pp. 179–205, 1999.

[7] F. Azuaje, "Interpretation of genome expression patterns: Computational challenges and opportunities," *IEEE Eng. Med. Biol. Mag.*, p. 119, Nov. 2000.

[8] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, pp. 279–284, 1967.

[9] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[10] 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 meausurements: A neural network approach," *Artif. Intell. Med.*, vol. 15, pp. 275–298, 1999.

[11] F. Azuaje, W. Dubitzky, N. Black, and K. Adamson, "Discovering relevance knowledge in data: A growing cell structure approach," *IEEE Trans. Syst., Man, Cybern. B*, vol. 30, pp. 448–460, June 2000.

[12] T. Kohonen, *Self-Organizing Maps*. Heidelberg, Germany: 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-organizing maps: Methods and applications to hematopoietic differentiation," *Proc. National Academy of Sciences of U.S.A.*, vol. 96, pp. 2907–2912, 1999.

[14] T. Kasuba, "Simplified fuzzy ARTMAP," *IEEE AI Expert*, pp. 19–25, Nov. 1993.

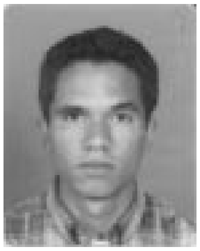
[15] G. A. Carpenter, S. Grossberg, J. H. Markuzon, J. H. Reynolds, and D. B. Rosen, "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," *IEEE Trans. Neural Networks*, vol. 3, pp. 698–712, Sept. 1992.

[16] J. Downs, R. Harrison, R. Kennedy, and S. Cross, "Application of the fuzzy ARTMAP neural network model to medical pattern classification tasks," *Artif. Intell. Med.*, vol. 8, pp. 403–428, 1996.

[17] F. Tourassi and C. Floyd, "The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis," *Med. Decision Making*, vol. 17, pp. 186–192, 1997.

[18] D. Weir, *Handbook of Experimental Immunology*, D. Weir, Ed. Oxford, U.K.: Blackwell Scientific, 1996.

[19] H. W. Mittrucker, T. Maysuyama, A. Grossman, T. M. Kündig, J. Potter, A. Shahinian, A. Wakeham, B. Patterson, P. S. Ohashi, and T. W. Mak, "Requirements for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function," *Science*, vol. 275, pp. 540–543, 1997.



Francisco Azuaje (M'96) received the B.Sc. degree in electronic engineering in from Simon Bolivar University, Caracas, Venezuela, in 1995. He performed graduate studies on Policy and Management of Technological Innovation (Central University of Venezuela, Caracas, Venezuela, in 1996). He received the Ph.D. degree in computational intelligence from the University of Ulster at Jordonstown, Newtownabbey, U.K. in 2000.

He is currently a Lecturer at the Department of Computer Science, Trinity College, Dublin, Ireland, where he develops research in the areas at the intersection of computer science and life sciences. One of his fundamental topics of research is to investigate computational techniques to solve challenging problems in biology and medicine (for example, genomic data mining and linking genotype to phenotype), as well as the application of biological knowledge to design new computational methods and architectures. He has published several papers in conference proceedings, journals and books related to computational intelligence applications in biology and medicine.

Dr. Azuaje is a Guest Editor of IEEE *Engineering in Medicine and Biology*.