Detection of Gene Expressions in Microarrays by Applying Iteratively Elastic Neural Net*

Máx Chacón[1], Marcos Lévano[2], Héctor Allende[3], and Hans Nowak[4]

[1] Universidad de Santiago de Chile; Depto. de Ingeniería Informática,
Avda. Ecuador No 3659 - Casilla 10233; Santiago - Chile
mchacon@diinf.usach.cl
[2] Universidad Católica de Temuco; Escuela de Ingeniería Informática,
Avda. Manuel Montt No 56 - Casilla 15-D; Temuco - Chile
mlevano@uct.cl
[3] Universidad Técnica Federico Santa María; Depto. de Informática
[4] Depto. de Física
Avda. España No 1680 - Casilla 110-V; Valparaíso - Chile
hallende@inf.utfsm.cl, hans.nowak@experimentos.cl

Abstract. DNA analysis by microarrays is a powerful tool that allows replication of the RNA of hundreds of thousands of genes at the same time, generating a large amount of data in multidimensional space that must be analyzed using informatics tools. Various clustering techniques have been applied to analyze the microarrays, but they do not offer a systematic form of analysis. This paper proposes the use of Gorban's Elastic Neural Net in an iterative way to find patterns of expressed genes. The new method proposed (Iterative Elastic Neural Net, IENN) has been evaluated with up-regulated genes of the Escherichia Coli bacterium and is compared with the Self-Organizing Maps (SOM) technique frequently used in this kind of analysis. The results show that the proposed method finds 86.7% of the up-regulated genes, compared to 65.2% of genes found by the SOM. A comparative analysis of Receiver Operating Characteristic (ROC) with SOM shows that the proposed method is 11.5% more effective.

1 Introduction

Modern deoxiribonucleic acid (DNA) microarray technologies [1] have revolutionized research in the field of molecular biology by enabling the study of hundreds of thousands of genes simultaneously in different environments [1].

By using image processing methods it is possible to obtain different levels of expression of thousands of genes simultaneously for each experiment. In this way these techniques generate thousands of data represented in multidimensional space. The process is highly contaminated with noise and subject to measurement errors, finally requiring experimental confirmation. To avoid repeating the whole process experimentally gene by gene, pattern recognition techniques are applied that make it possible to select sets of genes that fulfil given behavior patterns at their gene expression levels.

* This work was supported by projects FONDECYT 1050082, FONDECYT 1040354, FONDECYT 1040365 and MILENIO P02-054-F, Chile.

B. Beliczynski et al. (Eds.): ICANNGA 2007, Part II, LNCS 4432, pp. 355 – 363, 2007.

The most widely used method to determine groupings and select patterns in microarrays is the Self-Organizing Maps (SOM) technique [2], [3], [4]. One of the problems of SOM is the need to have an initial knowledge of the size of the net to project the data, and this depends on the problem that is being studied. On the other hand, since SOM is based on local optimization, it presents great deficiencies by restricting data projections only to its nodes.

One of the recent methods, consensus clustering [5], uses new resampling techniques which should give information about the stability of the found clusters and confidence that they represent real structure. This method is not used in this paper, but will be used and analized in a future contribution.

The Elastic Neural Net (ENN) [6], [7] method generates a controllable net described by elastic forces that are fitted to the data by minimizing an energy functional, without the need of knowing its size a priori. This generates greater flexibility to adapt the net to the data, and like the SOMs it allows a reduction in dimensionality, that improves the visualization of the data, which is very important for bioinformatics applications.

ENNs have been applied to different problems in genetics, such as analysis of base sequence structures (adenine, cytosine, guanine and thymine), where base triplet groupings are discovered [7]; automatic gene identification in the genomes of the mitochondria of different microorganisms [8]. But as far as we can tell, there is no application for finding patterns in microarrays.

This paper proposes the use of IENN to divide clusters iteratively, together with the k-means method and using indices to measure the quality of the clusters, making it possible to select the number of groups formed in each iteration.

To evaluate the results, data from the most widely studied microorganism, the bacterium Escherichia Coli (E.Coli), were used. The levels of gene expression of a set of 7,312 genes were analyzed by means of the microarrays technique. In this set there are 345 up-regulated genes that have been tested experimentally [9] and must be detected with the new method. The results are compared with those of the traditional SOM method.

2 Method

2.1 Theorical Foundation

Gorban defines the Elastic Neural Net [6] as a net of nodes or neurons connected by elastic forces (springs), where $Y = \{y^i, i = 1..p\}$ is a collection of nodes, $E = \{E^{(i)}, i = 1..s\}$ is a collection of edges, and $R^{(i)} = \{E^{(i)},E^{(k)}\}$ is the combination of pairs of adjacent edges called ribs denoted by $R = \{R^{(i)}, i=1..r\}$. Each edge $E^{(i)}$ starts at node $E^{(i)}(0)$ and ends at node $E^{(i)}(1)$. The ribs start at node $R^{(i)}(1)$ and end at node $R^{(i)}(2)$, with a central node $R^{(i)}(0)$. The data to be analyzed are $x^j=[x^j_1,...,x^j_M]^T \in R^M$, where M is the dimension of the multidimensional space and $j =1..N$ is the number of data.

The set of data closest to a node is defined as a taxon, $K^i = \{x^j : \| x^j - y^i \| \rightarrow \min\}$. It is clear that there must be as many taxons as nodes. Here $\| x^j - y^i \|$ is the norm of the vector $(x^j - y^i)$, and the Euclidian norm is used. This means that the taxon K^i contains all the vectors of the x^j data whose norms with respect to node y^i are the smallest.

Energy $U^{(Y)}$ between the data and the nodes is defined by (1),

$$U^{(Y)} = \frac{1}{N} \sum_{i=1}^{p} \sum_{x^j \in K^i} \left\| x^j - y^i \right\|^2 , \tag{1}$$

where each node interacts only with the data of its taxon. An elastic energy between the nodes $U^{(E)}$ is added by (2),

$$U^{(E)} = \sum_{i=1}^{s} \lambda_i \left\| E^i(1) - E^i(0) \right\|^2 , \tag{2}$$

where λ_i are the elasticity constants that allow the net's elasticity to be controlled. Additionally, a deformation energy $U^{(R)}$ between pairs of adjacent nodes, is also added by (3),

$$U^{(R)} = \sum_{i=1}^{R} \mu_i \left\| R^i(1) - 2R^i(0) + R^i(2) \right\|^2 , \tag{3}$$

where μ_i are the deformability constants of the net. The same values of λ and μ are chosen for all the λ_i and μ_i. The total energy is now minimized by (4) with respect to the number and position of the y^i nodes for different μ and λ

$$U = U^{(Y)} + U^{(E)} + U^{(R)} . \tag{4}$$

We used the VIDAEXPERT implementation, which can be found in Gorban et al. [6].

In addition to the flexibility offered by the ENNs to fit the net to the data, the projections of the data to the net can be made over the edges and at points within the net's cells, and not only over the nodes as required by the SOMs. This leads to an approximation that has a better fit with the real distribution of the data in a smaller space. This property is very important for applications in bioinformatics, where the specialist has better feedback from the process. The same could be said for image processing where the ENN seems to describe well active contours [10].

2.2 IENN Method

The algorithm used to find groups of genes that have the same behavior patterns consists of four fundamental phases: data preprocessing, ENN application, pattern identification, and finally a stopping criterion and cluster selection based on the level of expression and inspection of the pattern that is being sought.

Phase 1: Preprocessing

The set of *N* data to be analyzed is chosen, $x^j = [x^j_1,\ldots,x^j_M]^T$, j = 1…N, where *M* is the dimension of the multidimensional space. For this application, N corresponds to the 7,312 genes of the E.coli bacterium and M to the 15 different experiments carried out on the genes, and x^j is the gene expression level. The data are normalized in the form $\theta^j = \ln(x^j - \min(x^j) + 1)$ which is used as a standard in bioinformatics [11].

Phase 2: Elastic Neural Net (ENN)
The package of Gorban et al. [6], which uses the following procedures, is applied:

(a) The data to be analyzed are loaded.
(b) The two-dimensional net is created according to an initial number of nodes and elastic and deformability constants λ and μ with values between 2 for rigid grids and 0.01 for soft grids.
(c) The net is fitted to the data, minimizing the energy U. For that purpose the initial values of λ and μ are reduced three times (four pairs of parameters are required to be entered by the user). The decrease of λ and μ results in a net that is increasingly deformable and less rigid, thereby simulating annealing, allowing the final configuration of the ENN to correspond to an overall minimum of U or a value very close to it [6].
(d) The data are projected over the net on internal coordinates. In contrast with the SOM, in which piecewise constant projecting of the data is used (i.e., the data are projected on the nearest nodes), in this method piecewise linear projecting is applied, projecting the data on the nearest point of the net [6]. This kind of projection results in a more detailed representation of the data.
(e) Steps (c) and (d) are repeated for different initial values of the nodes, λ and μ, until the best resolution of the patterns found is obtained.

Phase 3: Pattern identification
The data are analyzed by projecting them on internal coordinates for the possible formation of clusters or other patterns such as accumulation of clusters in certain regions of the net. As a typical dependence of the data in a cluster on the dimensions of the multidimensional space, the average of the data for each dimension is calculated (cluster's centroid for the dimension).

For the formation of possible clusters the k-means method is used together with the quality index I [12], which gives information on the best number of clusters. The centroids of each cluster are graphed and analyzed to find possible patterns.

Phase 4: Cluster analysis
Once the best number of clusters is obtained, the centroids' curves are used to detect and extract the possible patterns. In general, the centroid curve of a cluster may present the pattern sought, may be a constant, or may not show a definite trend. Also, the values of the curve can be in a range that is outside the interest of possible patterns (low levels of expression). To decide if the clusters found in a first application of the ENN contain clear patterns, the behavior of the centroids' curves are analyzed. If the centroids' levels are outside the range sought, the cluster is discarded; if the patterns sought are detected, the cluster that contains the genes sought will be obtained (in both cases the division process is stopped), otherwise phases 2 and 3 are repeated with each of the *clusters* and the analysis of phase 4 is carried out again, repeating the process.

2.3 Data Collection

The data correspond to the levels of gene expression of 7,312 genes obtained by the microarray technique of E.Coli [9]. These data are found in the GEO database (Gene Expression Omnibus) of the National Center for Biotechnology Information[1]. The

[1] http://www.ncbi.nlm.nih.gov/projects/GEO/goes

work of Liu et al. [9] provides the 345 up-regulated genes that were tested experimentally. Each gene is described by 15 different experiments (which correspond to the dimensions for the representation of each gene) whose gene expression response is measured [9] on glucose sources. Specifically there are 5 sources of glucose, 2 sources of glycerol, 2 sources of succinate, 2 sources of alanine, 2 sources of acetate, and 2 sources of proline. The definition of up-regulated genes according to [9] is given in relation to their response to the series of sources of glucose considering two factors: that its level of expression is greater than 8.5 on a $\log_2$ scale, and that its level of expression increases at least 3 times from the first to the last experiment on the same scale. For our evaluation we considered a less restrictive definition that includes the genes that have only an increasing activity of the level of expression with the experiments; since the definition given in [9] for up-regulated genes contains very elaborate biological information for which a precise identification of the kind of gene to be detected is required.

The original data have expression level values between zero and hundreds of thousands. Such an extensive scale does not offer an adequate resolution to compare expression levels; therefore a logarithmic normalization is carried out. In this case we preferred to use the natural logarithm [11] instead of the base 2 logarithm used by Liu, because it is a more standard measure. The limiting value for the expression level was calculated using our own algorithm by determining the threshold as the value that best separates the initial clusters (θ_{min}). This expression level allows discarding groups of genes that have an average level lower than this value.

3 Results

First, the net's parameters were calibrated, i.e. the size of the net was set and the series of pairs of elasticity (λ) and deformability (μ) parameters were selected. The strategy chosen consisted in evaluating different net sizes and pairs of parameters λ and μ for the total data set that would allow minimizing the total energy U.

The minimum energy was obtained with a mesh of 28x28 nodes that was used throughout the whole process. Implementation of the ENN [6], [7] requires a set of at least four pairs of λ and μ parameters to carry out the process, because it adapts the mesh's deformation and elasticity in a process similar to simulated annealing that allows approximation to overall minimums. The set of parameters that achieved the lowest energy values had λ with values of {1.0; 0.1; 0.05; 0.01} and μ with values of {2.0; 0.5; 0.1; 0.03}. For the process of minimizing the overall energy U, 1,000 iterations were used. Then the cluster subdivision iteration process was started.

Figure 1 shows the representation of the first division and the expression levels of the centroids for the two clusters selected by the index I (for this first iteration). The expression level value equidistant from the two clusters corresponds to θ_{min}=5.5.

The iteration process generates a tree where each node has branches to a number of subclusters found by the maximum value of the index I. In the particular case of E.Coli, a tree of depth five is generated. The generation of the tree is made together with a pruning by expression level, i.e., only those clusters that present an expression level greater than $\theta_{min} \geq 5.5$ are subdivided.

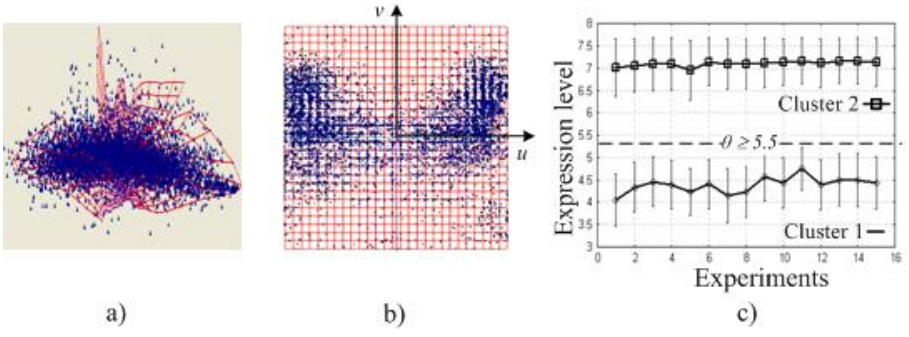

Fig. 1. First iteration of the method. a) Fitted net to original data. b) Projections on internal coordinates. c) Centroids and choice of the minimum expression level.

Finally, to stop the subdivision process of the groups that have an expression level greater than θ_{min}, the behavior of the expression level in the experiments was examined. In this case we only looked for a simple increasing pattern of the expression level in the centroids through the 15 experiments (the strict definition of up-regulated genes given in [9] was not used). Figure 2 shows the tree of subclusters generated by the process applied to the genes of E.Coli.

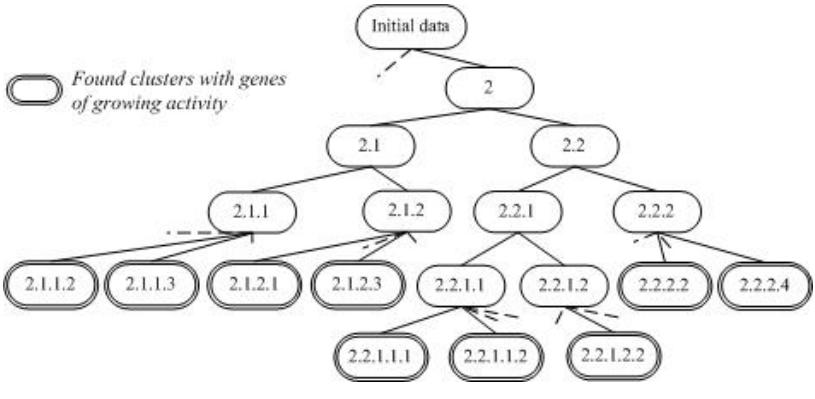

Fig. 2. Prepruned subcluster generation tree. Every leaf shown contains the set of genes with increasing activity where the up-regulated genes to be evaluated are found. The coding of each node shows the sequence of the nodes through which one must go to reach each node from the root.

The results show that the process chooses 1,579 genes, of which 299 correspond to up-regulated genes of the 345 that exist in the total data set, i.e. 86.7% of the total number of up-regulated genes. From the practical standpoint for the biological field, only 19% effectiveness has been achieved because there are 1,280 genes that are not up-regulated, which must be discarded using biological knowledge or by means of individual laboratory tests.

An alternative method for comparing these results is to use SOMs with the same data and conditions of the application with IENN. For this purpose the methodology proposed by Tamayo et al. [4] was followed, which suggests using SOMs in a single iteration, where the initial SOM mesh is fitted in such a way that at each node the patterns that present an increasing activity are identified. In this case the process shows that with a mesh of size 5x6 (30 nodes) it was possible to obtain patterns of increasing activity on the nodes of the SOM. The selected clusters are obtained directly from the patterns with increasing activity. With the SOMs 1,653 increasing activity genes were selected, 225 of which were up-regulated genes, and therefore in this case 65.2% of the 345 up-regulated genes were detected, and a practical efficiency of 13.6% was achieved, because 1,428 genes that do not correspond to up-regulated genes must be discarded.

Since in this application to the genes of E.Coli we can count on the 345 up-regulated genes [9] identified in the laboratory, it is possible to carry out an evaluation considering both methods (IENN and SOM) as classifiers. Moreover, if the expression level θ is considered as a classification parameter, it is possible to make an analysis by means of Receiver Operating Characteristic (ROC), varying the expression level θ over an interval of [4.4 - 8.9]. Figure 3 shows the ROC curves for IENN and SOM.

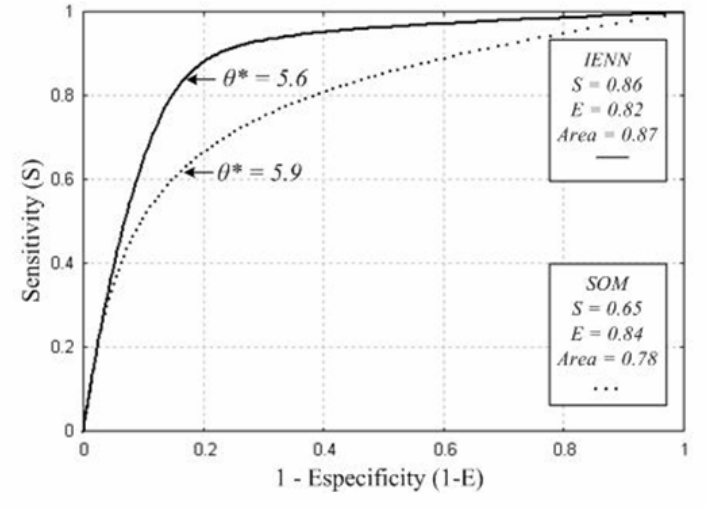

Fig. 3. ROC curves for IENN and SOM

The optimum classification value for IENN is achieved at θ^*=5.6. At this point a sensitivity of 86% and a specificity of 82% were reached, covering an area of 0.87 under the ROC curve. When the same data, normalization values and expression level ranges were considered for SOM, an optimum classification value of θ^*=5.9 is obtained, achieving a sensitivity of 65%, a specificity of 84%, and an area under the ROC curve of 0.78.

4 Discussion and Conclusion

When the results of the proposed method (which uses ENN) are compared with those of the traditional SOM method, it is seen that the IENN method detects 74 up-regulated genes more than the SOM, which correspond to 21.5% of those genes. For practical purposes it must be considered that these genes are not recoverable in the case of the SOM because they are mixed up with the group of 5,659 undetected genes. On the other hand, the efficiency of the method that uses the ENN is better, because it requires discarding 1,280 genes that are not expressed, compared to the 1,428 that must be discarded with the SOM. Since the final objective of the experiment with E.Coli consists in detecting the up-regulated genes, it is possible to consider the IENN and SOM methods as classifiers and carry out an analysis of the merit of the classification by means of an ROC curve.

When considering an overall analysis of the classifier using the expression level θ as a parameter, it is important to consider the area under the ROC curve. In this case the area for the proposed method is 0.87, compared to 0.78 for the SOM, which represents an 11.5% improvement. In relation to the sensitivity at the optimum decision level, the proposed method is 21% more sensitive than the SOM.

The numerical advantages derived from the application of the proposed method for the detection of the up-regulated genes of E.Coli are clear, but there are other aspects that must be analyzed with the purpose of projecting these results to the search of genes expressed in microarrays. The IENNs present several advantages that allow reinforcing the proposed method of iteration divisions. On the one hand, the IENNs have a greater capacity for adapting the net to the data because they have a set of parameters that control the deformation and elasticity properties. By carrying out the minimization of the overall energy in stages (evaluating different combinations of parameters λ and μ), a process similar to annealing is induced, making it possible to approach the overall minimum and not be trapped in local minimums. The same minimization methods allow the automatic selection of parameters that are fundamental for the later development of the process, such as the minimum expression level θ_{min} and the size of the net.

The other important advantage of the ENNs refers to their representation capacity, because the use of piecewise linear projecting makes it possible to increase the resolution of the data projected on the space having the lowest dimensions (internal coordinates). In the case of the microarray analysis this better representation becomes more important, since a common way of working in the field of microbiology and genetics is based on the direct observation of the data. On the other hand, the SOMs only allow a projection on the nodes when using *piecewise constant projecting* or the alternative U-matrix projections [2], [3], [4], which approximate only sets of data to the plane but do not represent directly each data.

A valid point that should be analyzed when comparing SOMs with IENNs is to consider the argument that an iteration process of divisions with SOMs can improve the results of the method. But the iteration process presented is based on the automatic selection of parameters (particularly the size of the net and the minimum expression level) for its later development, which is achieved by a global optimization method like ENN. The SOM does not allow the expression level to be determined automatically, and that information must come from the biological knowledge of the

expression levels of particular genes. The alternatives of using the minimum error of vector quantization of SOM as an alternative the minimum energy of ENN did not produce satisfactory results.

The results of the application to the discovery of up-regulated genes of E.Coli show a clear advantage of the proposal over the traditional use of the SOM method.

We chose to carry out a comparison with well established methods that are used frequently in the field of bioinformatics, but it is also necessary to evaluate other more recent alternatives such as flexible SOMs [13].

References

1. Molla M, Waddell M, Page D and Shavlik J. Using machine learning to design and interpret gene-expression microarrays. Artificial Intelligence Magazine 25 (2004) 23-44.
2. Kohonen T. Self-organizing maps, Berlin: Springer-Verlag (2001).
3. Hautaniemi S, Yli-Harja O, Astola J. Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps, Machine Learning **52** (2003) 45-66.
4. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E and Golub T. Interpreting patterns of expression with self-organizing maps: Methods and application to hematopoietic differentiation. Genetics **96** (1999) 2907-12.
5. Monti S, Tamayo P, Mesirov and Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Springer Netherlands, Vol. **52**, (2003) 91-118.
6. Gorban A, and Zinovyev A. Method of elastic maps and its applications in data visualization and data modeling. International Journal of Computing Anticipatory Systems, CHAOS. **12** (2001) 353-69.
7. Gorban A, Zinovyev A, Wunsch D. Application of the method of elastic maps. In analysis of genetic texts. Proc. International Joint Conference on Neural Networks (IJCNN), Portland, Oregon (2003) July 20-24.
8. Zinovyev AY, Gorban A, and Popova T, Self-organizing approach for automated gene identification, Open Sys and Information Dyn **10** (2003) 321-33.
9. Liu M, Durfee T, Cabrera T, Zhao K, Jin D, and Blattner F Global transcriptional programs reveal a carbon source foraging strategy by *E. Coli*. J Biol Chem **280** (2005) 15921-7.
10. Gorban A and Zinovyev A. Elastic principal graphs and manifolds and their practical applications, Computing **75** (2005) Springer-Verlag 359-379.
11. Quackenbush J, Microarrays data normalization and transformation, Nature Reviews Genetics **2** (2001) 418-27.
12. Maulik U, Bandyodpadhyay S Performance evaluation of some clustering algorithms and validity indices, IEEE PAMI **24** (2002) 1650-4.
13. Salas R, Allende H, Moreno S and Saavedra C Flexible architecture of self-organizing maps for changing environments, CIARP 2005, LNCS **3773**, (2005) 642-53.