Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Recent advances in gene expression data clustering: a case study with comparative results

George B. Bezerra[1], Geraldo M.A. Cançado[2], Marcelo Menossi[2], Leandro N. de Castro[1] and Fernando J. Von Zuben[1]

[1]Laboratório de Bioinformática e
Computação Bio-Inspirada (LBiC/DCA/FEEC), Caixa Postal 6101
UNICAMP, 13083-852 Campinas, SP, Brasil
[2]Laboratório de Genoma Funcional,
Centro de Biologia Molecular e Engenharia Genética, Caixa Postal 6010
UNICAMP, 13083-970 Campinas, SP, Brasil
Corresponding author: G.B. Bezerra
E-mail: bezerra@dca.fee.unicamp.br

Genet. Mol. Res. 4 (3): 514-524 (2005)
Received May 20, 2005
Accepted July 8, 2005
Published September 30, 2005

ABSTRACT. Several advanced techniques have been proposed for data clustering and many of them have been applied to gene expression data, with partial success. The high dimensionality and the multitude of admissible perspectives for data analysis of gene expression require additional computational resources, such as hierarchical structures and dynamic allocation of resources. We present an immune-inspired hierarchical clustering device, called hierarchical artificial immune network (HaiNet), especially devoted to the analysis of gene expression data. This technique was applied to a newly generated data set, involving maize plants exposed to different aluminum concentrations. The performance of the algorithm was compared with that of a self-organizing map, which is commonly adopted to deal with gene expression data sets. More consistent and informative results were obtained with HaiNet.

Key words: Hierarchical clustering, Gene expression data, Artificial immune systems

INTRODUCTION

The capability to monitor the expression levels of genes on a genomic scale has led to a rapid evolution of molecular biology and functional genomics. By showing DNA transcriptional behavior under particular conditions, expression data give clues about the role of genes in biological processes of interest.

However, the huge amount of data produced by gene expression experiments must be preprocessed in order to reveal potentially useful information. In this preprocessing step, known as cluster analysis, the apparently arbitrary distribution of expression patterns may reveal groups of genes with high degrees of correlation in their expression profiles. As the result of this clustering, the data may expose a more comprehensible structure, which provides meaningful information for intuitive inspections.

Clustering gene expression data are especially challenging computational tasks. The data set is used to present complex characteristics like high dimensionality, low density, and high levels of noise and redundancy, thus preventing traditional computational tools, such as single-linkage hierarchical clustering, from giving satisfactory performance. This fact has motivated the application of several distinct and advanced techniques to the problem, including bio-inspired algorithms. Given that each technique presents its own peculiarities and manipulates the available data set by means of distinct methodologies, it is not possible to determine *a priori* the best choice among these possibilities.

This problem has led to an increasing interaction of people from biological and computational areas. On the one hand, computer scientists and engineers must provide efficient and generic techniques, capable of extracting the most relevant properties of the data set. Furthermore, it is very important that the software tools present a user-friendly interface, supplying intuitive output displays for biologists. On the other hand, biologists have to make it clear which specific properties of the data they are interested in, so that clustering analyses can be better planned and clustering tools can be designed for specific purposes.

We present an advanced clustering tool, named HaiNet (hierarchical artificial immune network) (Bezerra et al., 2004), specifically designed to deal with gene expression data. HaiNet is a powerful algorithm that uses ideas taken from the mammals' immune system, such as a population with a varying size, self-organizing interactions, and affinity maturation. The process somewhat resembles that of antibodies, recognizing invading antigens, with the antigens being associated with the original data set. A network of interconnected antibodies will be obtained with no specific neighborhood until the end of the self-organizing process. After that, a functional neighborhood is built by a minimal spanning tree. Each branch of the network is analyzed and those considered inconsistent are removed, leading to an automatic determination of the number of clusters. Additionally, HaiNet determines a hierarchical relation between clusters, which can be represented as a dendrogram. Larger clusters are initially detected using a small number of antibodies. By increasing the number of antibodies in the network, the representation becomes more and more specific, and refined clusters are then recursively determined.

We believe that the particular properties of HaiNet are especially suitable for gene expression data clustering. In most clustering techniques, the number of clusters is asked as input data. This is generally a serious drawback because no *a priori* information is available about the data structure, and consequently the user will arbitrarily set this value. HaiNet avoids this problem by performing an automatic determination of the number of clusters. In addition,

the hierarchical relation among clusters produced by the tool may favor the exploration of different degrees of correlation between genes. The specialist can choose which level of detail is more adequate for a given purpose and in a given application.

As an advanced clustering technique devoted to deal with bioinformatics problems, HaiNet is the result of a series of successful applications of immune-inspired algorithms. The first attempt involved solely the artificial immune network (aiNet), with no hierarchical resources (Bezerra and de Castro, 2003). Subsequently, the HaiNet was proposed (Bezerra et al., 2004) and validated by means of an already exhaustively investigated gene expression data set. Here we consider HaiNet as part of a broad research project, with newly generated data sets to be analyzed for the first time, and with alternative clustering approaches to be explored either in a competitive or a collaborative way.

As a case study, a data set was generated based on two strains of maize when exposed to different concentrations of aluminum ions. One strain is tolerant to aluminum, and the other is not. The ultimate purpose of this data analysis is to identify the genes responsible for the tolerance. So, the availability of a flexible and hierarchical clustering device will be an important stage in a complete genomic project. To better assess the performance of HaiNet, we compared the results obtained with those produced by a self-organizing map (SOM) (Kohonen, 1990). More specifically, we adopted as a testbed the approach proposed by Tamayo et al. (1999).

HaiNet

HaiNet is an extension of a clustering technique, called aiNet (de Castro and Von Zuben, 2000, 2001). The kernel of the learning algorithm is the same; however, HaiNet incorporates a hierarchical procedure that increases its potential.

aiNet clustering

The aiNet is a clustering technique well known in the artificial immune system community (de Castro and Von Zuben, 2000, 2001). It combines a preprocessing learning procedure, performed by an immune-based strategy, and a clustering partition step, which is achieved by the use of a minimal spanning tree. The aiNet extracts the most relevant characteristics from the input data by positioning its antibodies in the most representative portions of the data space, thus filtering out noise and redundancy. After that, the minimal spanning tree is built on the antibodies, and the edges of the tree are used as input to a clustering discrimination technique.

The aiNet learning procedure can be explained as follows. A random population of antibodies is initially created. The whole population is then presented to the input patterns, which are directly associated with antigens, and those antibodies that have a high affinity (low Euclidean distance, for example) with the antigens are selected to be cloned. Each antibody is cloned at a rate that is proportional to its affinity with the antigens, and the clones are then mutated at a rate inversely proportional to their affinity value. The aim is to produce antibodies that better recognize the antigens. This process, named clonal expansion, causes a growth in the population of antibodies. It is inspired by one of the most important theories in immunology: the clonal selection principle, originally proposed by Burnet (1959). After the expansion phase, those antibodies presenting an affinity with each other that is higher than a fixed threshold are removed,

thus eliminating the redundancy of the network. Again, this technique is inspired by immune theories, particularly the immune network theory suggested by Niels Jerne (1974). The whole process is then repeated, now with the remaining prototypes as the population of antibodies to be exposed to the antigens. After some iterations, the affinity maturation process leads to the convergence of the network. In our implementation of the aiNet learning algorithm, only the best-matching antibody is selected to be cloned, and the cloning process produces only one individual. The number of iterations used is 10. The computational cost of the aiNet is quadratic with the size of the input instance.

When the learning step is finished, the minimal spanning tree is constructed on the resulting antibodies, thus revealing the structure of the immune network. To extract the information stored in the tree, we use a local criterion proposed by Zahn (1971), which identifies the presence of clusters by evaluating the data distribution in the space. Each edge of the tree is analyzed in relation to its neighbors, and those considered inconsistent, i.e., with a length much greater than its immediate neighbors, are removed, leading to the data partition into clusters, also denoted natural clusters (Everitt, 1993). As a consequence, this criterion tends to preserve the inherent structure of the data set.

Hierarchy of networks

The aiNet can be extended to implement a hierarchical approach, which provides a topological structure of the correlation between clusters. This extension is denoted hierarchical aiNet, or simply HaiNet. Large clusters, representing more general differences within the data points, are initially identified. They are then analyzed individually with more constrained parameters, leading to more refined clusters, i.e., clusters with a higher level of similarity among components. This iterative process continues until no more natural clusters are identified.

To achieve this property, HaiNet makes use of a parameter called a suppression threshold (σ_s), which controls the level of refinement of the search. The suppression threshold is a value that determines the maximum similarity that two antibodies may have so that one does not recognize the other. It actuates in the network interaction phase of the aiNet learning procedure. Beyond this value, two antibodies are considered to recognize each other, and one of them must be suppressed to reduce the network redundancy. If σ_s is very high, the prototypes will represent the data set with a high degree of generality, and the number of antibodies in the network will be very small. As a result, the clustering device looks at the data set with gross eyes, and only large-scale divisions can be detected. When σ_s is low, antibodies are able to get closer to each other, leading to a more accurate representation of the data, with a larger number of prototypes. Under these circumstances, the minimal spanning tree can detect the presence of smaller divisions, revealing refined clusters with a higher density and specificity.

The hierarchical procedure consists in recursively running the aiNet, starting with a high suppression threshold. This parameter is then slowly reduced, and the aiNet is run again for each new value. Each value of σ_s corresponds to a novel hierarchical level. The algorithm can be summarized as follows:

1. Parameter definition: define the initial value for the suppression threshold σ_s.
2. aiNet learning: run the aiNet learning algorithm with the given σ_s.
3. Tree branching: each cluster detected by the minimal spanning tree, derived from the

obtained network of antibodies, generates an offspring network in the next level of the tree, i.e., a new branch of the tree. The clusters already detected will indicate the portion of the data set to be attributed to each newly generated branch.

4. Parameters' updating: reduce σ_s, e.g., by geometrically decreasing them by a factor α.
5. Offspring network evaluation: run each offspring network with the corresponding attributed portion of the data set.
6. Tree convergence: if the offspring network does not detect a novel cluster, the process is halted for that branch of the tree, and the tree expansion is completed at that branch. Each branch of the tree represents a cluster and a sequence of branches represents the hierarchy inherent to the data mapped into their corresponding clusters. Otherwise, while a given offspring network (branch) of the tree is still capable of identifying more than one cluster, return to Step 4 and the process continues until no new cluster can be identified in any active branch.

In step 6, the process is halted every time the suppression threshold is low enough so that each antibody of the network represents exactly one point of the data set.

MAIZE DATA

The data set consists of the expression levels of 187 genes of two strains of maize, Cat100-6 (Al-tolerant) and S1787-17 (Al-sensitive). The plants were exposed to different aluminum concentrations (Al ions), in a total of three experimental conditions: control (zero Al), 75Al (75 µM Al) and 283Al (283 µM Al). As both strains were put in the same matrix, the data set to be clustered assumes the form of a 187×6 matrix, in which the first three experiments correspond to Cat100-6 and the other three to S1787-17. Also, each gene was normalized by the mean of its attributes, thus all genes are in the same scale, and the shape of the expression profile becomes more important than the intensity of the values.

Aluminum is the most abundant metal on the earth surface, and most plants are sensitive to it. The toxic effect of aluminum inhibits plant growth, making most kinds of cultivations in soils with high concentrations of this metal impracticable. The main objective of the gene expression analysis is to find the genes involved in the tolerance of Cat100-6. This clustering strategy may elucidate mechanisms involved in the tolerance and in the toxicity of aluminum. This is a project carried out by the LGF (Laboratory of Functional Genome - http://cafe.cbmeg.unicamp.br).

COMPUTATIONAL ANALYSIS AND RESULTS

The self-organizing map

The implementation of the SOM proposed by Tamayo et al. (1999) needs the number of clusters as an input parameter. This is determined by the size of the bi-dimensional grid. Biologists made the choice of the grid size. By visual inspection they selected the grid that yields the most interesting patterns, and at the same time, a relative low deviation. The chosen grid was 5×6, i.e., 30 clusters (Figure 1).

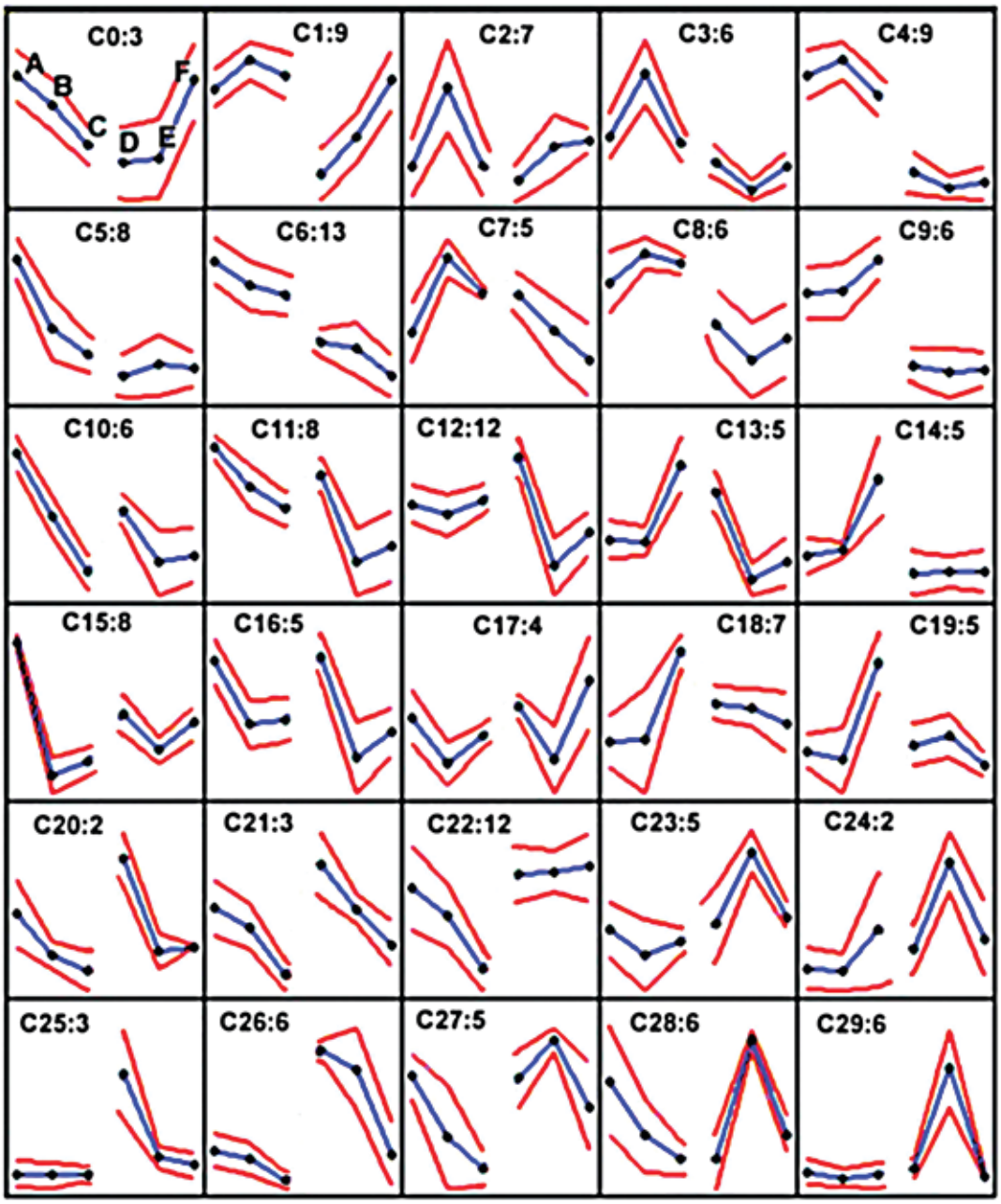

Figure 1. Self-organizing map output clusters, with the number of genes in each one. Letters A to F in cluster C0 indicate the experimental conditions.

The hierarchical aiNet (HaiNet)

The results produced by the HaiNet were considerably different from those of SOM.

The minimal spanning tree cutting criterion was capable of automatically detecting 40 clusters, and the content of the clusters are not necessarily consistent with the 30 clusters associated with the SOM strategy. The hierarchical tree that was obtained had six levels (Figure 2). Notice that there are some clusters with attached numbers. These numbers represent SOM clusters that are included into the corresponding HaiNet clusters.

A SOM cluster was considered to be inside of a HaiNet cluster if at least 80% of its genes were included. Only five clusters, C3, C7, C18, C25, and C29, were preserved by HaiNet. In the other cases, HaiNet found inherent subdivisions, most of them already in the initial hierarchical levels of the tree. This is an indication that some of the SOM clusters are not too stable, at least for this configuration. On the other hand, HaiNet found an interesting result concerning topology preservation of the maps. Note that there was no grouping in a same tree node of SOM clusters that were not connected by neighborhood links in the grid. This can be observed in clusters A and B. If, for example, cluster C25 were grouped within node A, this would characterize a topology violation of SOM, because the neuron corresponding to cluster C25 has no neighboring link with any SOM cluster inside A.

Furthermore, some clusters found by HaiNet, which present visible strong characteristic profiles, like G and H, are almost identical to the SOM clusters C29 and C25, respectively, as would be expected. Figure 3 shows these clusters side by side.

As these clusters were more specific in relation to the rest of the data set, they were already identified at the first level of the hierarchy, i.e., they were positioned quite far from the rest of the data set, so that the distance separating them from the other genes was perceivable, using a high value for σ_s.

Determining the number of clusters

The *a priori* definition of the number of clusters is a very risky practice. An incorrect estimate may distort the natural configuration of clusters. In our analysis, whenever SOM was designated to find only 30 clusters, natural clusters could be dissociated and put into two or more clusters in order to fit the restriction. A good example can be seen with clusters C5 and C28. They contain genes that visibly escape from the overall pattern (Figure 4). These genes together form a novel pattern, which was not identified by the SOM. However, HaiNet was capable of extracting this new profile, by removing the discrepant genes from both clusters, and better accommodating them into a novel cluster.

Note that two genes of C5 and one gene of C28 were put together in cluster DC. Also, the remaining genes of C28 were divided into clusters DA and DB, together with other similar genes of the data set. The remaining samples in C5 were also partitioned, but this takes place in another branch of the tree.

The results provided by HaiNet were found to be significantly better than those with SOM. The reason is that more refined clusters reduce the tedious work of biologists in filtering every cluster by visually detecting the genes that escape from the most representative patterns. However, it is fairly valid to say that SOM could have been used with a larger bi-dimensional grid. But, the number of clusters found would still not be automatic. Moreover, one can take advantage of the hierarchical relation produced by HaiNet, by traversing the correlation levels. A higher level, for example, is similar to an SOM executed using a smaller grid, but still with the same benefit of an automatically defined number of clusters.

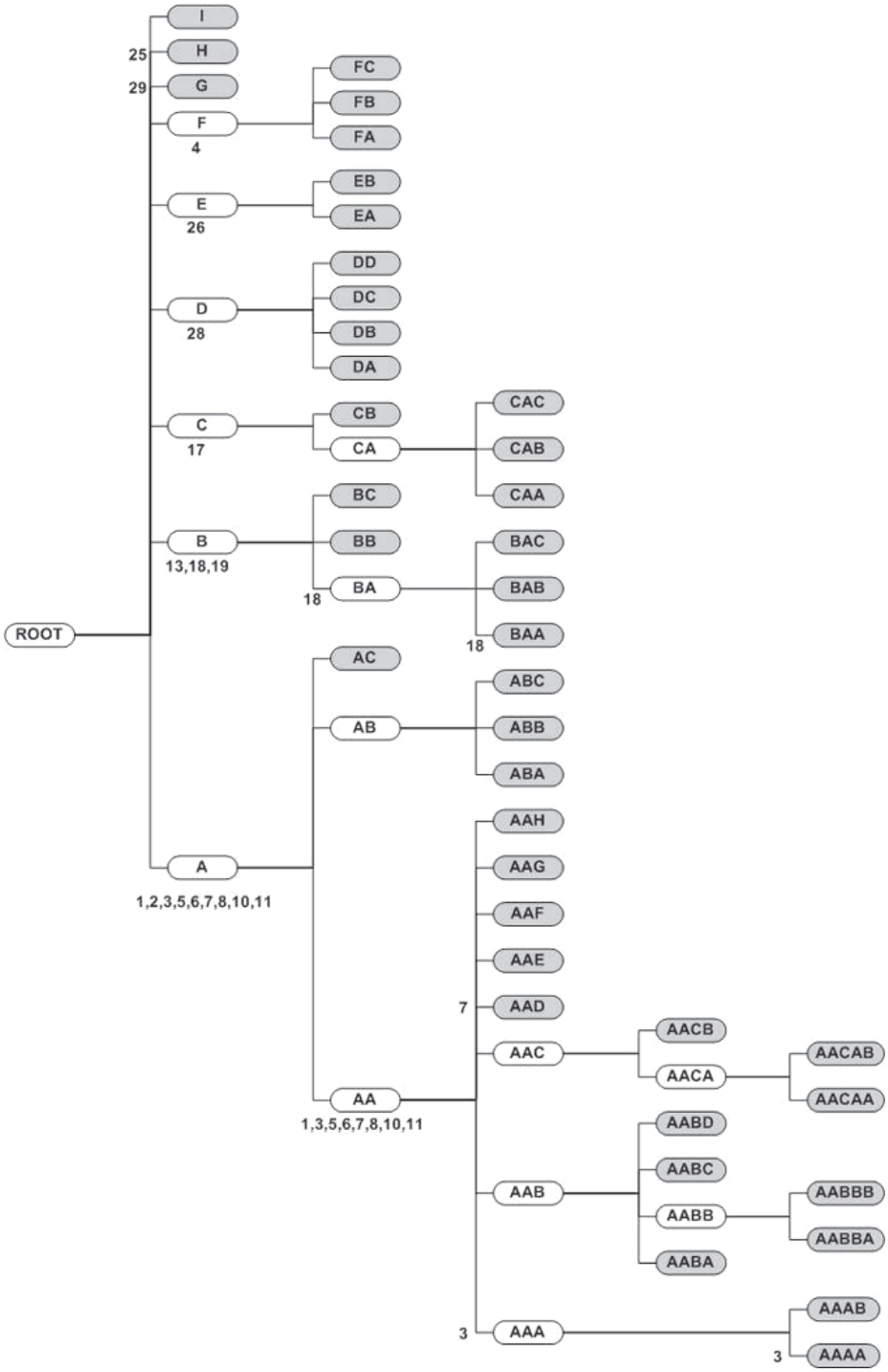

Figure 2. Hierarchical tree with six levels obtained by HaiNet. Numbers next to nodes indicate the self-organizing map clusters that are included.

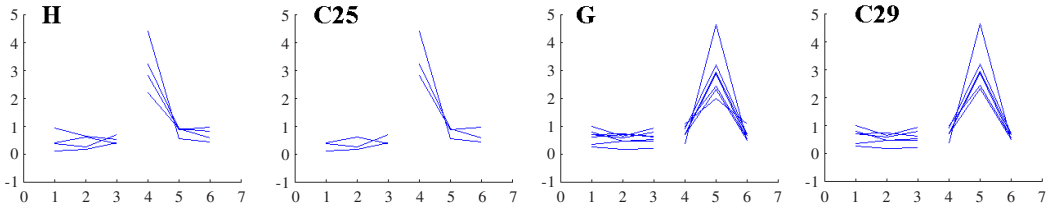

Figure 3. Clusters with the most evident characteristic patterns were identified by both techniques.

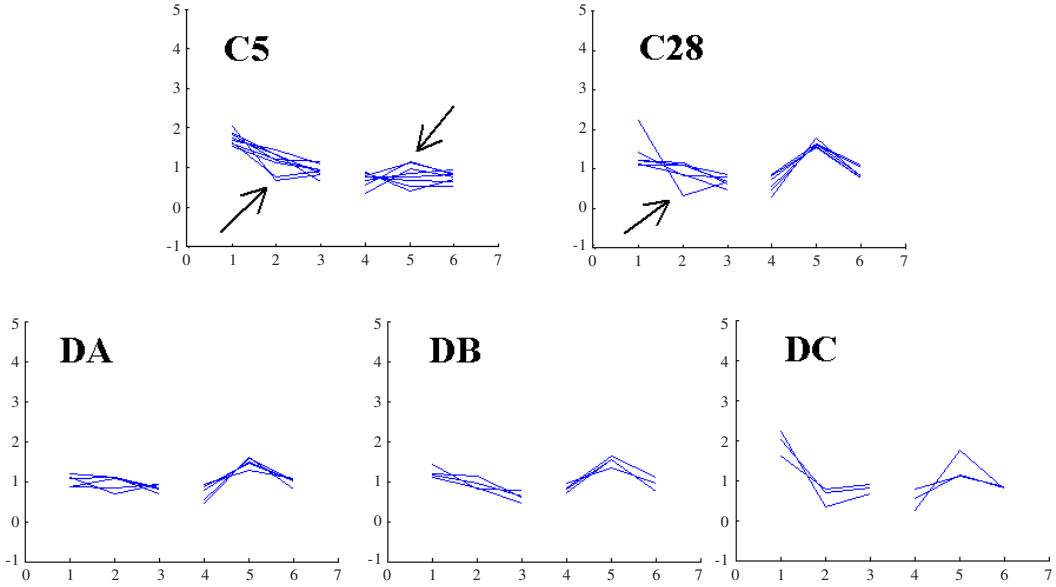

Figure 4. HaiNet put the discrepant genes in clusters C5 and C28 into a novel cluster, DC.

Exploring the hierarchical topology

In the hierarchical procedure explained in the "Hierarchy of networks" section, large-scale divisions are first detected, followed by more and more refined evaluations. In order to illustrate this property, Figure 5 gives prominence to one branch of the tree, concerning the expansion of cluster B. The intention is to demonstrate the fine-tuning process by detailing two descending hierarchical levels. A similar treatment could be adopted for the remaining branches of the tree.

Notice that cluster B presents a very noisy pattern. However, its genes are considerably distinct from the rest of the data set for this level of the hierarchy, being grouped to form an isolated cluster. As we descend to the second level of the branch, the most visible patterns within B are extracted, leading to clusters BA, BB and BC. Nevertheless, BA is still not good enough. By fine-tuning the HaiNet parameters, subdivisions in cluster BA are found, and the more refined clusters, BAA, BAB and BAC are thus identified.

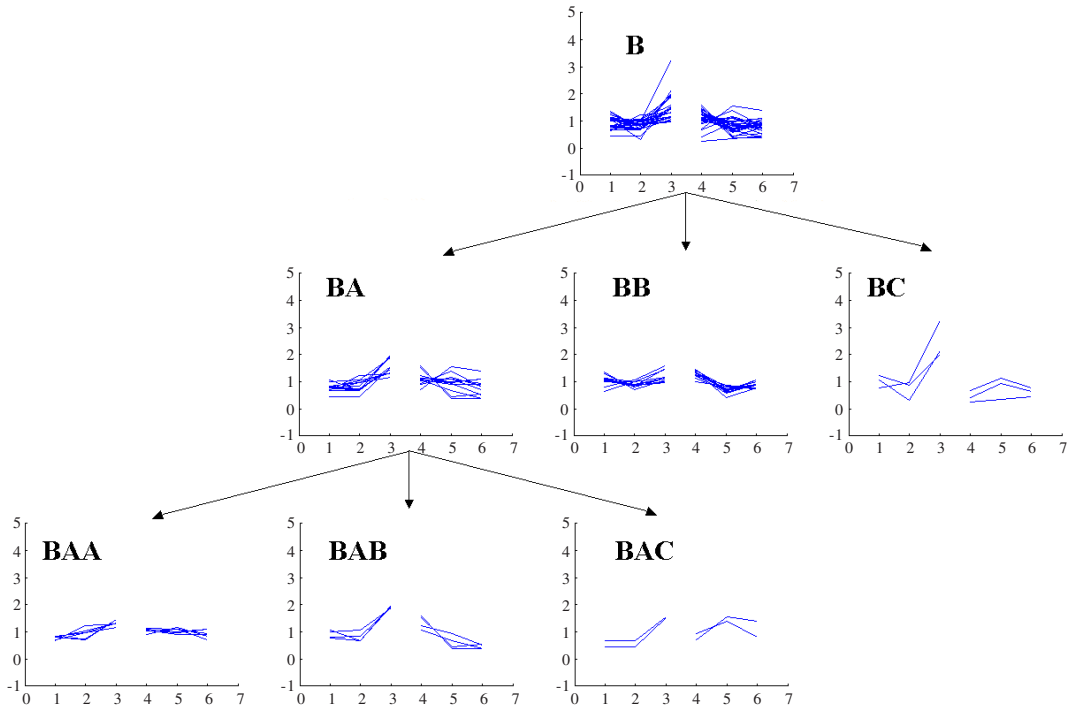

Figure 5. Detailed branch of the hierarchical tree concerning the expansion of cluster B.

CONCLUSION

We have presented a flexible and hierarchical clustering device for gene expression data analysis. HaiNet possesses desirable properties, including an automatic definition of the number of clusters and a hierarchical topology. Comparing HaiNet results with those produced with SOM, more consistent and informative outputs can be obtained with the former. This is a relevant aspect in the context of gene expression data analysis, mainly because the SOM approach is a well-disseminated clustering device for biologists.

As a demonstration of the practical applicability of HaiNet, we applied the algorithm to a data set concerning maize plants exposed to different aluminum ion concentrations. The results obtained will be further explored by biologists toward better understanding the genetic tolerance mechanisms of plants to aluminum.

ACKNOWLEDGMENTS

The authors thank CNPq and FAPESP for financial support.

REFERENCES

Bezerra, G.B. and **de Castro, L.N.** (2003). Bioinformatics data analysis using an artificial immune network. *Lecture Notes in Computer Science - Proceedings of the Second International Conference on Artificial Immune Systems*, Edinburgh, UK, *2787*: 22-33.

Bezerra, G.B., de Castro, L.N. and **Von Zuben, F.J.** (2004). A hierarchical immune network applied to gene expression data. *Lecture Notes in Computer Science - Proceedings of the Third International Conference on Artificial Immune Systems*, Springer, Catania, Italy, *3239*: 14-27.

Burnet, F.M. (1959). *The Clonal Selection Theory of Acquired Immunity*. Cambridge University Press, Cambridge, UK.

de Castro, L.N. and **Von Zuben, F.J.** (2000). An Evolutionary Immune Network for Data Clustering. *Proceedings of the Brazilian Symposium on Neural Networks*, Las Vegas, NV, USA, pp. 84-89.

de Castro, L.N. and **Von Zuben, F.J.** (2001). aiNet: An artificial Immune Network for Data Analysis. In: *Data Mining: A Heuristic Approach* (Abbass, H.A., Saker, R.A. and Newton, C.S., eds.). Idea Group Publishing, University of New South Wales, Australia.

Everitt, B. (1993). *Cluster Analysis*. Heinemann Educational Books, London, UK.

Jerne, N.K. (1974). Towards a network theory of the immune system. *Ann. Immunol. 125C*: 373-389.

Kohonen, T. (1990). The self-organizing map. *Proc. of the IEEE 78*: 1464-1480.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrvsky, E., Lander, E.S. and **Golub, T.R.** (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA 96*: 2907-2912.

Zahn, C.T. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Trans. on Computers C-20*: 68-86.