Advanced Soft-Computing techniques and Clustering Algorithm for Gene Expression Microarray Data Classification.

Olga Valenzuela, Fernando Rojas, Francisco Ortuño, Jose Luis Bernier, M. Jose Saez, Belen San-Roman, Luis Javier Herrera, Alberto Guillen, Ignacio Rojas

Department of Computer Architecture and Technology. University of Granada. 18060 Granada, Spain. olgavc@ugr.es

Abstract. This paper is intended to meet two main objectives. The first is the development of an Parallel Genetic Algorithm using clustering fitness function, for Gene Expression Microarray automatic classification (is called PGA-GEM), which is focused on the processing data contained in genomic microarrays. Within this area, are of special interest the implemented fitness function and the genetic operators within evolutionary algorithm. The fitness function determines the quality of the grouping obtained by statistical analysis of the data, discovering patterns that allow an automated classification and clustering of its. The second objective is to compare the results obtained by other algorithms with PGA-GEM , such as Support Vector Machines (SVM), showing than even it is not possible an superviser methodology, the presented algorithm can obtain similar results, performing an un-superviser training, than other well-known method as SVM

Keywords: Gene Expression Microarray, Parallel Genetic Algorithm, Support Vector Machines

1 Introduction

The conclusion of various genome projects has provided an important information to researchers, but also opened the doors to new questions, such as processes regulating the expression of genes and characterization of genome -level differences between different individuals of the same species and how the more subtle alterations of each of these individual operations predispose to illness. Understanding how genetic variants regulate cell phenotype , tissue and organ research are great challenged of the near future years. There are an estimated 8,000 hereditary diseases, but today only about 200 can detect before birth and genetic tests exist for a few hundred others.

To answer these questions that are beyond genomic studies has developed what is known as "Post- Genomics". Within this category " Post- Genomic " can include comparative genomics , individual genomics, proteomics and transcriptomics among others [3][4]. The development of all this kind of approach is inspired by genomic studies and the possibility of having tools and technologies that enable and support the development of its own mass approaches the " - Omics " approaches. It was in this

environment that microarrays of biological material, microarrays or biochips have been developed as a tool for analysis and mass production of biological information.

With the advent of microarray technology, it has succeeded in controlling the expression levels of thousands of genes during important biological processes. Due to the large volume of data and the large amount of noise embedded in them, interpret the experimental results has been a challenge. To discover hidden patterns in the expression of the genes of a microarray, in our project we have developed a novel clustering algorithm PGA-GEM .

PGA-GEM encodes a cluster on a chromosome where each gene represents a cluster. Based on this structure makes use of a set of play operators that facilitate the exchange of information between chromosomes. The discovery of patterns and reclassification fitness function shows how relevant is the value of a property to determine a particular grouping, instead of considering pairs of local distances also considered clusters are globally organized . PGA-GEM not require the number of clusters is determined a priori. Patterns hidden in each cluster can be explicitly revealed for easy interpretation by the fitness function.

We have performed tests with simulated data and real data from a microarray genomic database of leukemia. Experimental results show that PGA-GEM is very robust in the presence of noise . It is capable of finding near optimal solutions and find the statistical significance of association rules or patterns , for noisy data , indicating the significance of each of the groups of the solution.

Compared with other algorithms as K-Mean or SVM , PGA-GEM is much slower and the time required for a process taking place exceeds the playback time of one iteration of the other algorithms . However, as for the analysis of microarray normal is that the solution to a problem is not immediately required, the time it takes to run PGA-GEM compensates the quality of the results provided. In fact, better performance PGA-GEM is obtained, in term of accuracy, compared with SVM [6][7], and something very important, it can support un-supervised training. However, as for the analysis of microarray normal is that the solution to a problem is not immediately required, the time it takes to run PGA-GEM compensates the quality of the results provided. Also, an important characteristic of PGA-GEM is the intrinsic computational parallelism. In fact, for large volume of data consisting of any microarray can be processed more easily using the cluster computer (parallelism) present in PGA-GEM genetic algorithm.

2 Gene Expression using Microarray

Microarray experiment is an experiment that is performed to determine if certain previous hypotheses are true or false (although they may also lead to new hypotheses) .Like any experiment, it is subject to errors that can come from multiple sources of variability and be of different types. In general these sources of variability usually have two origins : systematic or random .

The systematic variability is one that affects all measurements similar manner. Is mainly due to two aspects: a) The amount of material available; b) The laboratory apparatus used for the experiment.

The random variable is one that can affect different way each component of the experiment. It depends on factors such as: The quality of the material; The efficiency of the laboratory procedures.

To obtain correct results is addressing the effects of variability may be introduced in the experiment. Depending on the type of treatment of this variability is focused in one way or another. In the case of systematic variability , can be estimated from the necessary data by standardization or calibration techniques corrections . To treat random variability, certain models pose error and is often used to control the experimental design and statistical inference to draw conclusions. All these processes should be integrated into a workflow or lifecycle of a microarray experiment .

For the correct development of the experiment is necessary to specify three aspects that influence throughout the life cycle: a) What is the purpose of study.; b) That objective is pursued ; c) What limitations and that kind presents.

Today it is possible to catalog the main applications of microarrays depending on the type of results that are sought in two broad categories:

a) Quantitative tests: the aim is to analyze and quantify the presence of analytes of interest in a complex sample. These analytes can be RNAs (gene expression assays), DNAs (CGH type assays) or proteins. The quantification is performed in such trials is relative, since due to numerous factors is not possible to establish an absolute quantification of the analytes in the sample studied. All quantitative assays are both qualitative time.

b) Qualitative Test**:** the interest is focused on the identification and / or characterization of the analytes of interest present in a sample. In such applications the amount of analyte present in the sample is not significant except for the determination of whether the signal that can be detected is positive or not. As in the case of quantitative assays , a wide variety of analytes that may be studied (RNA, DNA, protein) to be the primary application for this type of arrays designed for resequencing trials and study of mutations or SNPs. Other applications for protein microarrays would interest for testing ELISA assays or other biological molecules which are immobilized

Another criterion that could be used to describe the group and the main applications of microarrays to be related to the area where you will use this technology :

1. Transcriptome or gene expression that are identified and quantified the RNA present in a sample analysis .
2. Genomic analysis in which we can distinguish two groups of applications, applications for resequencing, mutation detection and identification and analysis of type CHG in which quantitative variations are characterized in the number of copies of certain regions of the genome.
3. Protein analysis , studies with protein arrays allow the identification and quantification of the protein of interest present in a sample.
4. Analysis of arrays in which other molecules are immobilized.

Furthermore, it is worth noting the important role that is currently developing the technology of DNA microarrays in different areas of medicine, especially its applications in the area of cancer, where in the past 3 years the growth in the number of scientific publications that suggest a revolutionary role of these techniques in the clinical application against cancer has been exponential .

3 Parallel Genetic Algorithm for Gene Expression Microarray Data Analysis.

In this section, the parallel genetic algorithm for gene expression microarray data is briefly described. To create efficient parallel programs must be able to create, destroy and specify processes and the interaction between them [1],[2]. There are basically three ways to parallelize a program:

1. fine-grained parallelization: program parallelization is done at the instruction level. Each processor takes a portion of each step of the algorithm (selection, crossover and mutation) on the common population.
2. medium-grain parallelization: the programs are parallelized loop level . This parallelization is usually done in an automated way in compilers.
3. coarse-grained parallelization: based on domain decomposition data between processors, each being responsible for performing the calculations on their local data.

Parallel computing has become a critical part in all areas of scientific computing , allowing improved performance simply by using a larger number of processors , memories, and the inclusion of elements of communication that allow processors to work together to solve a particular problem. We can say that genetic algorithms have a structure that is perfectly suited to parallelization. In fact the natural evolution is itself a parallel process and evolves using several individuals. The main methods of parallelization of genetic algorithms consist of dividing the population into several subpopulations. The size and distribution of the population between the different processors will be one of the key factors when parallelizing the algorithm. There are several ways to parallelize a genetic algorithm. The first and more intuitive is global, that is basically the evaluation parallelize maintaining a population of individuals.

Otherwise global parallelization is to perform a sequential execution of different genetic algorithms simultaneously (these two forms of parallelization not change the structure of the algorithm used). The other approaches do change the structure of the algorithm and divide the population into subpopulations that evolve independently and exchange individuals every few generations. If populations are few and large, have coarse-grained parallelization. If the number of populations is large with few individuals in each population have fine-grained parallelization. Finally, there are algorithms that combine properties of the latter two are called mixed. Besides getting shorter execution times , to parallelize a genetic algorithm are modifying the algorithmic behavior, and this allows us to obtain other solutions and experiment with different possibilities of implementation and the different factors that influence it.

These populations evolve separately to stop at a particular point and the best individuals exchange there between. Technically there are 3 important characteristics that influence the efficiency of a parallel genetic algorithm:

1. The topology that defines the communication between subpopulations.
2. The exchange ratio: number of individuals to be exchanged
3. Migration intervals: frequency with which individuals are exchanged.

Grained algorithms or network communication work from the perspective that there is no hierarchy among subpopulations. In this paper, the population is divided spatially between different processors. The crossing and selection of individuals will be between individuals belonging to the same neighborhood , formed by a set of adjacent individuals according to the abovementioned spatial representation . However, the overlap between neighborhoods is allowed to foster interaction, albeit slight, of all individuals in the population.

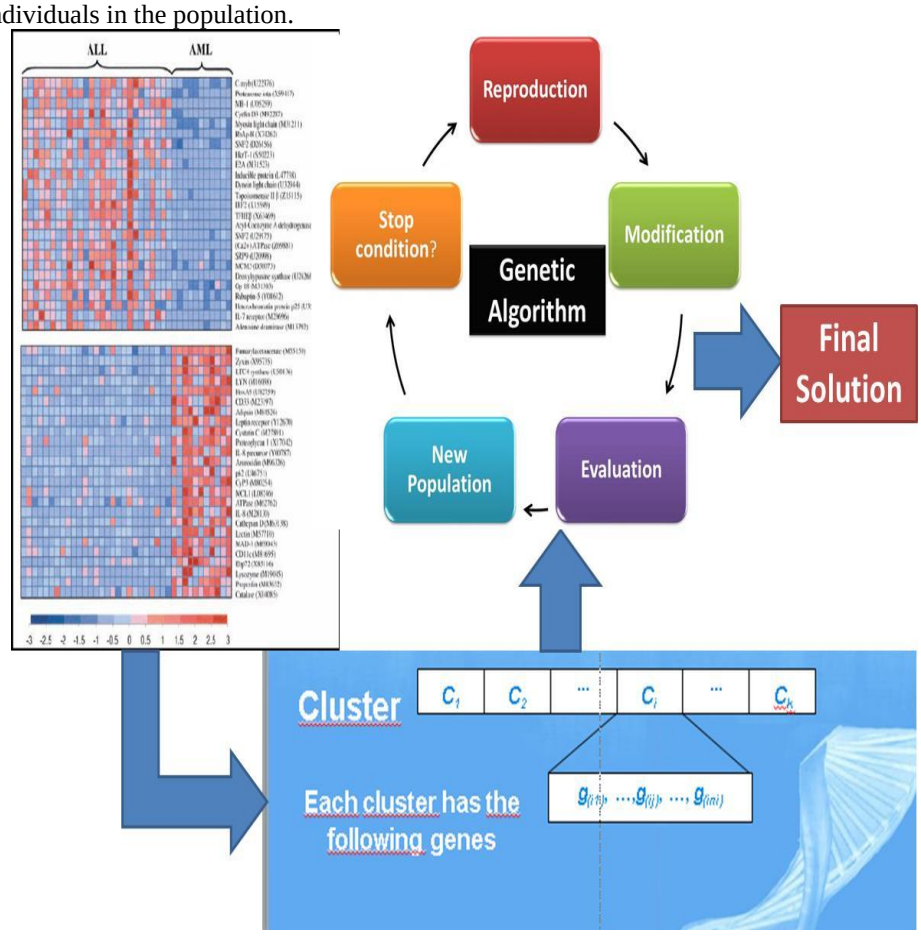

Fig. 1. Flow char of the proposed methodology: Parallel Genetic Algorithm using clustering fitness function, for Gene Expression Microarray automatic classification (PGA-GEM)

PGA-GEM is based on the use of an evolutionary approach (see Figure 1). Compared to others based on both evolutionary and non evolutionary clustering algorithms, PGA-GEM has the following salient features:

1. Encodes a set of clusters on a chromosome so that each gene represents a cluster and each cluster contains labels grouped data records. In the implementation of the algorithm, each of these labels represents a gene identifier of the file containing the DNA microarray.

2.-. There is a set of crossover and mutation operators that facilitate the exchange of information between two chromosomes and allows variations in order to avoid local optima in the solution.

3.- . Use of a fitness function that measures the interest of a particular group of data records encoded in a chromosome is made.

4.- . Unlike many measures that are based on local distances and do not provide quality results in the presence of noise in the data, the algorithm used a probabilistic measure which takes into account global information contained in certain groups .

5.- The algorithm is able to distinguish between relevant and irrelevant values features during clustering.

6. No one needs to know a priori the number of clusters in the cluster.

During implementation of this algorithm is assumed that a file DNA microarray is a collection of data comprises N genes in M experiments. This allows us to represent the dataset as a set of N records , G = $\{g_{1,} \ldots , g_{i,} \ldots, g_N\}$ with records g_i , $i = 1, \ldots, N$, characterized by M attributes , $E_1, \ldots, E_j, \ldots, E_M$ whose values $e_{i1}, \ldots, e_{ij}, \ldots, e_{iM}$,, where e_{ij} in the domain of E_j attribute represents the value of gene i under experimental condition j.

If a particular chromosome encodes k clusters, $C_{1,} \ldots, C_{i,} \ldots, C_k$ has k genes. Each cluster contains a set of data records. This data set is represented by the identification labels of biological genes that make up the file microarray. For example, assume that the C_i cluster contains n_i genes, $g_{(i1)}, \ldots, g_{(ij)}, \ldots, g_{(in_i)}$ where $g_{(ij)}$ belongs to G = $\{g_{1,} \ldots , g_{i,} \ldots, g_N\}$ are the labels of genes $g_{(i1)}, \ldots, g_{(ij)}, \ldots, g_{(in_i)}$ to cluster *gen* C_i. Therefore, a chromosome that encodes a particular grouping of cluster may be represented as shown in Figure 1. For the initial population is generated randomly each chromosome by PGA-GEM such that the number of clusters k is represented in a chromosome initially randomly generated a valid range for the problem. Each of the genes of G = {g1, … , gi, …, gN} is then assigned, also at random to one of the k clusters.

4 Evaluation of the quality of gene cluster using PGA-GEM

To evaluate the quality of clusters collected during PGA-GEM clustering uses two objective measures: the extent of Davies- Bouldin validity index and F- measure

4.1 Davies- Bouldin validity index

The DBI measure is a function that measures the distance within the cluster and between clusters. These distances are considered good indicators of the quality of a cluster of clusters, thus , a good group should present distances between clusters and relatively small distances within the cluster. In fact many optimizations used for clustering algorithms are mainly developed to maximize the distance between clusters and minimize the distance within the clusters. The DBI measure combines these two distances in a function to measure the average similarity between a cluster and the most similar to it. Assuming that a cluster of cluster k is formed of clusters, we can define the DBI follows:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{d_{\text{int ra}}(i) + d_{\text{int ra}}(j)}{d_{\text{int er}}(i, j)} \right\}$$

$$d_{\text{int ra}}(j) = \left(\sum_{x=1}^{n_j} \left\| g_x - g_{jc} \right\| \right) \Big/ n_j \quad , \quad d_{\text{int er}}(i, j) = \left\| g_{ic} - g_{jc} \right\| \Big/ k \quad ,$$

Where *k* represents the total number of clusters, and d_{inter} d_{intra} represents the distance from the centroid to elements of the same cluster and the distance between centroids of different clusters respectively, and n_j is the number of data records in the cluster *j*.

4.2 F- Measure

The F- Measure is typically used for the evaluation of a clustering process, combining more than two types called precision and recall measures.

- Precision: This measure is defined as the fraction of a cluster that consists of objects of a specific class. The accuracy of a cluster i with respect to class j is represented by precision(i,j) = p_{ij} , where p_{ij} represents the probability of a cluster member *i* to belong to class *j* .
- Recall: This measure is defined as the proportion of objects of a class in a cluster. The recall of a cluster i with respect to class j is represented by recall (i, j) = m_{ij} / m_j, where m_{ij} is the number of objects of class *j* in cluster *i* and m_j is the number of objects of class *j*.

The F- Measure is defined as the rate at which a cluster contains only objects of a particular class and all objects of that class. Thus, the F- Measure of a cluster i with respect to class j is represented by the following expression:

$$F(i, j) = \frac{2 \cdot recall\ (i, j) \cdot precision\ (i, j)}{recall\ (i, j) + precisio\ (i, j)}$$

When the correct classification of data is called the F- measure is useful in the sense that it provides objective information on the degree to which a clustering algorithm is able to recover the original clusters. The F- Measure is measured in the range [0,1] and high values of it indicate good quality of clustering.

5 Results

In this paper we will conduct a study of the behavior of the proposed algorithm PGA-GEM. The execution of the algorithm has been performed with data extracted from the database of leukemia , which contains 72 patients with this disease and 7129 genes. The class type variable distinguishes suffering from leukemia (AML or ALL). The database is divided into two, the part 38 training samples and 34 test part . Because it is not a monitored PGA-GEM , ie has no training phase and qualifying algorithm only been made using one or the other file providing database , namely the test file AMLALL_test.data

The object file analysis presents a large dimensionality , which makes their treatment under the conditions of the proposed algorithm is computationally very expensive in terms of execution time taking into account the resources that are available for test development . Therefore be feasible study on parallel platforms PGA-GEM reduce their computation time. Because the majority of the microarray has high dimensionality has been chosen to undertake the study of the algorithm using simulated data created so as to contain hidden patterns microarray similarity . This will develop a comprehensive study of the programmable features of the algorithm and on the basis of its findings , the study will be applied to real microarray file discussed above. However, the data set that up AMLALL_test.data will apply a dimensionality reduction as unfeasible as explained is run entirely on the platforms we have.

Were made three kinds of tests : executions operators using only guided executions using unguided operators and executions which use a mixture of both types of carriers , fixing the same probability for each. Ranges or limits established in the odds that a gene and a cluster may be selected to participate in crossover or mutation processes have been set to [0.2, 0.8] for all tests . On the other hand , the odds of mutation rates for deletion and reclassification , mixing and division have been established at 0.8 , 0.1 and 0.1 respectively. We divide the maximum and minimum number of cluster have been set at a value equal to 2, the types of mutation and mixed division will have no effect on individuals of the population and that in any case the limits of the number of clusters will be overcome. Therefore, these two types of mutation has been assigned a lower probability .

As selection operator was used by roulette selection operator as operator replacement replacement operator of the worst individuals in the population , as it appears in the original version of the algorithm PGA-GEM . The fitness function used corresponds to the discovery of patterns and reclassification function.

The following table shows the most representative results obtained after performing several runs with the characteristics mentioned above.

Operator	Iter.	Entropy	DBI	F-Measure	Fitness value
Guided	3000	0.31	16.81	0.74	4
		0.31	16.81	0.74	4
		0.31	16.81	0.74	4
No- guided	3000	0.99	23.46	0.49	5
		0.98	22.96	0.5	3
		0.98	22.96	0.5	3
Mixture 50%	3000	0.93	22.08	0.49	5
		0.98	21.62	0.5	3
		0.93	22.08	0.49	5

6 ACKNOWLEDGE

This work has been partially supported by the projects: Spanish Ministry of Science and Innovation SAF2010-20558; Genil Start-Up Projects For Young Researchers PYR-2012-8 and Excellence projects Junta de Andalucia P09-TIC-5476,P07-TIC-02906. Authors want to thank the work performed by Andrea Martínez Trujillo and-Concepción Torres Ceballos.

7 References

[1] S. Zhang, C. Zhang, Q. Yang. Data preparation for data mining. AppliedArtificial Intelligence 17:5-6, 375-381, 2003.

[2] Kim, H. C. and Z. Ghahramani (2006). "Bayesian Gaussian process classification with the EM-EP algorithm." IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12): 1948-1959.

[3] Ortuno, F. M., O. Valenzuela, et al. (2013). "Predicting the accuracy of multiple sequence alignment algorithms by using computational intelligent techniques." Nucleic Acids Research 41(1): e26.

[4] Ortuno, F. M., O. Valenzuela, et al. (2013). "Optimizing multiple sequence alignments using a genetic algorithm based on three objectives: structural information, non-gaps percentage and totally conserved columns." Bioinformatics 29(17): 2112-2121.

[5] Peng, H. C., F. H. Long, et al. (2005). "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy." Ieee Transactions on Pattern Analysis and Machine Intelligence 27(8): 1226-1238.

[6] Urquiza, J. M., I. Rojas, et al. (2012). "Using machine learning techniques and genomic/proteomic information from known databases for defining relevant features for PPI classification." Computers in Biology and Medicine 42(6): 639-650.

[7] Wang, X. D., W. F. Liang, et al. (2006). "Application of adaptive least square support vector machines in nonlinear system identification." WCICA 2006: Sixth World Congress on Intelligent Control and Automation, Vols 1-12, Conference Proceedings: 1897-1900.