Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis: a Distributed Approach

Vitoantonio Bevilacqua, Giuseppe Mastronardi, Filippo Menolascina, Angelo Paradiso and Stefania Tommasi

Abstract—**Microarrays allow biologists to better understand the interactions between diverse pathologic states at the gene level. However, the amount of data generated by these tools becomes problematic. New techniques are then needed in order to extract valuable information about gene activity in sensitive processes like tumor cells proliferation and metastasis activity. Recent tools that analyze microarray expression data have exploited correlation-based approach such as clustering analysis. Here we describe a novel GA/ANN distributed approach for assessing the importance of genes for sample classification based on expression data. Several different approaches have been exploited and a comparison has been given. The developed system was employed in the classification of ER+/- metastasis recurrence of breast cancer tumors and results were validated using a real life database. Further validation has been carried out using Gene Ontology based tools. Results proved the valuable potentialities and robustness of similar systems.**

Index Terms— **Artificial Neural Networks, Breast Cancer Metastasis Recurrence Prediction, Gene Expression Data Analysis, Genetic Algorithms, Microarrays.**

I. Introduction

Introduced for the first time in 1989, microarrays have gained in this time a great fame thanks to their ability to give biologists a quite detailed snapshot of cellular and genomic activity in particular states of the examined organism. Recent advances in microarray technology have allowed studying the expression patterns of thousands of genes in parallel. The principles these devices are based on are really few and simple. Microarrays use hybridisation-based methodology that allows mRNA molecules to bind to their complementary parts (genes). Several probes for each gene are placed on a coated quartz surface (1.28 cm x 1.28 cm); mRNA segments hybridize with probes according to A-T C-G base pairing principle and this allows the monitoring of the expression levels of thousands of genes simultaneously [5]. This enables the measurement of the levels of mRNA molecules inside a cell and, consequently, the proteins being produced. Hence, the role of the genes in a cell at a given moment can be better understood by analyzing their expression levels. In this context, the comparison between gene expression patterns through the measurement of the levels of mRNA in healthy versus unhealthy cells can supply important information about pathological states, as well as information that can lead to earlier diagnosis and more efficient treatment.

The real challenge, then, is to find a set of genes, out of the thousands mapped, which can be used to develop a classifier with the highest accuracy [6]. Similar sets of genes are defined "*gene signatures*" and can be employed in medical common practice in order to provide early diagnosis. Identification of a set of differentially expressed genes could serve to identify disease subtypes that may benefit from distinct clinical approaches to treatment. This was the primary objective of Foekens et al. in [3] and [4]; to predict accurately patient's risk of recurrence is an important aspect of lymph node negative cases treatment planning. Gene signature extracted by Foekens et al. consists of 76 genes (60 for ER+ and 16 for ER- cases). In [3] Foekens et al. have employed statistical methods and supervised/unsupervised clustering techniques in order to extract knowledge from a 286 x 22482 array of gene expression values. Although correlation-based approaches have been widely applied in analyzing the patterns of gene expression [1][2], it is commonly believed they may not fully extract the information from data corrupted by high-dimensional noise. Therefore, these ranking based techniques select the genes which individually provide better classification, but they may not result in meaningful gene combinations for an overall classification task. Hence approaches capable of performing an efficient search in high dimensional spaces, such as evolutionary algorithms (EAs), should prove to be ideal candidates. What is more, while high-throughput technology has significantly accelerated the rate at which biological information is acquired, tools that can successfully mine the resulting large data sets are needed. Some research groups have exploited the potentialities of soft computing techniques applied to bioinformatics and some these works have been carried out in the field of microarray data analysis [7] [8].

With this work we have tried to address the problem of *gene selection* using a distributed Genetic Algorithm that evolves populations of possible solution and uses an Artificial Neural

Vitoantonio Bevilacqua is with the Department of Electronics and Electrical Engineering of the Polytechnic of Bari, Italy, Via E. Orabona, 4 – 70125 Bari, Italy.
Giuseppe Mastronardi is with the Department of Electronics and Electrical Engineering of the Polytechnic of Bari, Italy, Via E. Orabona, 4 – 70125 Bari, Italy.
Filippo Menolascina is with the Department of Electronics and Electrical Engineering of the Polytechnic of Bari, Italy, Via E. Orabona, 4 – 70125 Bari, Italy.
Stefania Tommasi is with the "Istituto Tumori Giovanni Paolo II", Via Samuel F. Hahnemann, 10 – 70126, Bari, Italy.
Angelo Paradiso is with the "Istituto Tumori Giovanni Paolo II", Via Samuel F. Hahnemann, 10 – 70126, Bari, Italy.

Network in order to test the gene signatures' ability to correctly classify cases belonging to the test set. For each of the 286 cases 22482 gene expression levels are measured. Comparing all subsets of genes is an unfeasible approach. It is not possible to examine all the combinations directly, then an efficient method is needed to sample from fewer combinations to find the optimal or near optimal solutions. Although many optimization methods may be in principle appropriate for this task, genetic algorithms provide a general purpose, stochastic search methodology. The techniques described in this paper may well constitute a novel application of similar distributed hybrid systems. A distributed design of the system has been proposed in order to overcome the computational costs, in terms of time, of similar solutions. Therefore, results returned by the GA/ANN based system are then validated using Gene Ontology, a biological validity assessment tool that can show interesting cues of research for biologist and physicians.

This paper firstly gives some details of the problem of classification problem in "post genomic" era. Then a description of the GA and of genetic operators is given. An outlook on the ANN classifier follows. In final paragraphs system results and brief investigations of biological plausibility are exposed.

II. METHODS

A. Data Acquisition and Preprocessing

The dataset used to evaluate performances of the system proposed is publicly available and can be downloaded from the Gene Expression Omnibus (GEO) web site [9]. GEO is a data repository of high-throughput gene expressions and hybridization arrays maintained by the National Center for Biotechnology Information. GEO databases have been used in the recent years by researchers of all over the world in order to give publicity to results of specific researches. Each dataset submitted to GEO receives an ID code (a unique identifier) called "Accession Number".Our focus is on the dataset *GSE2034*, submitted by Tim Jatkoe on the February, 23rd 2005 and provided to us by the I.R.C.C.S. "Mater Dei" of Bari, Italy. This dataset collects the results of a multi-center research carried out by Veridex LLC in collaboration with the Department of Medical Oncology, Erasmus MC-Daniel den Hoed of Rotterdam. The research, which involved 286 patients, aimed at discovering gene signatures able to identify patients at high risk of distant recurrence [3]. The ability to identify patients who have a favorable prognosis could, after independent confirmation, allow clinicians to avoid adjuvant systemic therapy or to choose less aggressive therapeutic options. Obviously this leads to improvement of the quality of treatments and to better living conditions of patients. More details about this research could be retrieved in [3].The dataset was acquired from the web using a function of the new "Bioinformatics Toolbox" included in the seventh release of MATLAB, the development framework chosen for this project. A routine was then developed in order to obtain an array composed by 286 columns, the cases, each of which was defined by 22282 gene expression levels.

```
num=36777;
for i=1:286
   ID(i)=getgeodata(strcat('GSM',int2str(num)));
   data(:,i)=ID(i).Data(:,2);
   num=num+1;
end
```

The *getgeodata* function takes as argument an Accession Number that refers to the single case of a dataset. For this reason iteration is needed to obtain the complete matrix. The structure returned by *getgeodata* function contains 6 fields (substructures).

Scope	'SAMPLE '
Accession	'GSM36777'
Header	<1x1 struct>
ColumnDescriptions	<4x1 cell>
ColumnNames	<4x1 cell>
Data	<22283x4 cell>

Figure 1. Structure returned by getgeodata function

Signal measurements obtained by the microarray scanning are stored in the "Data" array. In the four columns of the "Data" matrix are collected, in order, the gene name, the value observed, the "absolute call" and a p-value that indicates the significance level of the detection call. Gene expression levels are then stored in the second column of the "Data" array.
The data matrix obtained contains quasi-raw data. Gene expression levels, in fact, are characterized by statistical properties that force researchers to apply preprocessing algorithms. In the common practice of microarray data analysis the "normalization" is a key step. Normalization means to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes. Data normalization for microarray experiments is an open field of research and many alternative algorithms have been proposed in order to accomplish this delicate task [10]. The most employed normalization algorithm simply scales the values in each column of microarray by dividing by the mean column intensity. This routine was implemented in the manorm.m M-file in MATLAB. Together with normalization, filtering techniques met more and more often the consensus of researcher thanks to their ability to exclude in early stage of the research less meaningful variables that add computational costs to the data mining of similar datasets [11]. Even in this field many different approaches have been proposed. For the peculiarities of the dataset *GSE2034* an "entropy information" based gene filtering was employed in order to remove genes with low entropy expression values [12]. At the end of these processes a normalised matrix 286 x 20055 was obtained.

B. Gene Selection Using a Genetic Algorithm

In recent years some research groups have focused their attention on the exploitation of GAs' potentialities in

information extraction from biomedical database. In [13] a multi-objective algorithm has been employed in order to build a reliable classification tool small in size and, at the same time, able to produce as accurate a classification as possible. In [14] the problem of gene assessment and sample classification for gene expression data have been addressed using a Genetic Algorithm and a K-Nearest Neighbor classifier. The use of similar hybrid systems has gained a spread consensus in the scientific community in the last years thanks to their ability to generate solutions that inherit strength from each original component.

The proposed approach is based on a hybrid system that uses a GA to select subsets of genes (individuals) and an ANN that classifies cases and returns a metric of the error which is used as a fitness function for the selected subset of genes. Given the high variety of approaches reported in literature we have tried herein to carry out a comparative study of the different systems and their results on the chosen dataset. The use of an EA in bioinformatics can allow researchers to give a coherent solution avoiding the risk of combinatorial explosion brought by statistical exhaustive research of the search space [7][8].

As known GA, are basically inspired by natural evolution and selection. In biological systems, genetic information is stored in chromosomes. Chromosomes are replicated and passed onto the next generation with selection depending on fitness. Genetic information can, however, also be altered through genetic operations such as mutation and crossover. In GAs, each "chromosome" is a set of genes, which constitutes a candidate solution to the problem. In typical implementations a population or subpopulations of "chromosomes" are used. The passage of each "chromosome" to the next generation is determined by its relative fitness, i.e. the closeness of its properties to those desired. Random combinations and/or changes of the transmitted "chromosomes" produce variations in the next generation of "offspring". The Individuals that show higher fitness values (correspondence with desired properties) have greater chances of being selected for transmission. Following these steps and after many generations, optimal or near optimal solutions are obtained. There are four major components of GA:

- Chromosome;
- Fitness;
- Selection;
- Crossover / Mutation.

The modular approach followed in the design phase of the system has allowed to simply evaluating the performances of the other compared classifier systems described below, acting on the fitness function module.

Brief descriptions of particular implementations of the described systems are given below.

1) Chromosome Representation

In this work a 20 genes long chromosome has been used in order to codify a 20 genes long gene signature. A binary codification of the chromosome was selected.

2) Fitness Function

For the calculation of the fitness function the Sum of the Squared Errors (SSE) error returned by an ANN based classifier has been used. This is a metric of the error made by the ANN in the classification task, other evaluation systems could be simply implemented.

3) Selection Criterion

The selection criterion was the elitistic one. Best performing gene signatures from each population are allowed to pass to the other generation producing offspring.

4) Crossover / Mutation

In the crossover and mutation operators some constraints have been implemented in order to maintain acceptability of the solution (e.g. in order to avoid the repetition of a gene in the same chromosome). Crossover probability was set to 40% and mutation probability was time dependent.

C. Statistical Analysis based Classification

Statistical approaches have been largely explored in the data mining of microarrays derived datasets. Tools like *Analysis of Variance* (ANOVA), *Principal Component Analysis* (PCA) have been employed in order to gain precious knowledge about the examined data. ANOVA has been used in the original work of Foekens [3] in order to extract genes more closely correlated to the metastasis recurrence of cases. PCA, on the other hand, analyze the distribution of variance among original variables and returns new factors that maximize the information content [15]. Although these techniques are very powerful, they show evident limitations in contexts where dataset are characterised by high dimensionality. For these reasons, instead of using similar tools on the entire dataset, a statistical classifier trained on the 20 genes selected by the GA was set up. This was done using as fitness function module that provided for a statistical classifier with a ten-fold cross validation algorithm that returned the result used as fitness value for the gene signature considered.

D. K-NN based Classification

The K-Nearest Neighbor (KNN) estimator is a kind of nonparametric estimator of a function. Given a data set $\{X_i, Y_i\}$ it estimates values of Y for X's other than those in the sample. The process is to choose the k values of X_i nearest the X for which one seeks an estimate, and average their Y values. In the KNN method, the distance is computed between a sample, represented by its pattern vector V_m, and each of the pattern vectors of the training set:

$V_m = (g_{1\,m},\ldots g_{i\,m},\ldots, g_{n\,m})$, where n is the number of genes in the vector (set to 20); $g_{i\,m}$ is the expression level of the *ith* gene in the *mth* sample (m = 1,..,M).

The classification of each sample is accomplished observing the class membership of its K-Nearest Neighbors (calculated in the 20 dimensional space, considering the Euclidean distance). The K value for this classifier has been set to 3. If not all of the K-Nearest Neighbors are of the same class the sample remains unclassified. A GA/KNN classifier for microarray data analysis is described in [14].

E. Support Vector Machine based Classification

Undoubtedly Support Vector Machines (SVMs) are the most employed tool in microarray data analysis. Assumed S = $\{(x_1,y_1),\ldots, (x_n,y_n)\}$, where $x \in X \subset R^m$ and $y_i \in \{0,1,\ldots,c\}$, a training dataset with n samples and N classes. Each x_i is an m-dimensional input vector, and each y_i corresponds to the class associated to x. In the microarrays domain, x_i is the *ith* tissue sample, represented by a set of m genes, and y_i can be different types of cancer, for example. The task of a classification algorithm is to learn a mapping of $x_i \rightarrow y_i$ using data from the set S. SVMs [16] handle this by constructing a hyper plane

$$< w \cdot x > + b = 0 \qquad (0.1)$$

where $w \in R^m$ represents the normal vector associated with the hyper plane and b is the bias that maximise the separation of positive and negative training samples. The margin corresponds to the distance from the separating hyper plane to the closest samples of each class. It is obviously inversely proportional to ||w||. Thus, to have the maximal margin hyper plane one needs to minimize the Euclidean norm of vector w. This task can be translated in function optimization problem and faced with quadratic programming techniques. SVMs have been used in many diagnosis problems since the first studies in this area [1] and provided competitive results.

F. Artificial Neural Network based Classification

Artificial Neural Networks are employed in several different fields. From function approximation, to clustering and classification, these networks have proven to be a powerful tool in complex problems solving. In particular as classifiers both the supervised and unsupervised paradigms have been explored with encouraging results. Feed-Forward, SOM architectures, but even Adaptive Resonance Theory based Networks have been successfully employed in diagnostic tasks, as reported in [17].

Given the characteristics of the problem and the dimensionality of the dataset, an ANN-Feed-Forward (ANN-FF) has been chosen in order to classify cases and to evaluate the ability of single gene signatures to constitute the basis for an accurate classifier. Then, for each fitness function evaluation an ANN-FF is trained on 200 cases each of which is defined by 20 parameters (genes selected by the GA and passed to the fitness function for evaluation) and validated. On the basis of results of research described in [17], and having observed high similarity of dimensionality of the two datasets, a three layer ANN was set up. The selected topology provided for 25 neurons in the first layer, 12 in the hidden layer, and 1 output neuron. Activation functions were: "*tansig*" for the first two layers and "*pure linear*" for the last layer. Stop criterion for the training phase were: 50000 epochs or SSE less than 0.004. Initial learning rate was set to 0.3 and modified by the descent gradient momentum algorithm. These choices provided a solution able reach a good equilibrium between learning and generalization capabilities of the system.

III. Gene Selection

After the preprocessing stage and the system development, the experimentation phase has been carried out. The GA/ANN hybrid system was set up and executed on the *GSE2034* dataset. The GA/ANN as well as the other compared GA/X hybrid systems were executed 100 times and each GA run accounted for 500 generations. Good convergence ability has been reached with described parameters as can be seen in figure 2.

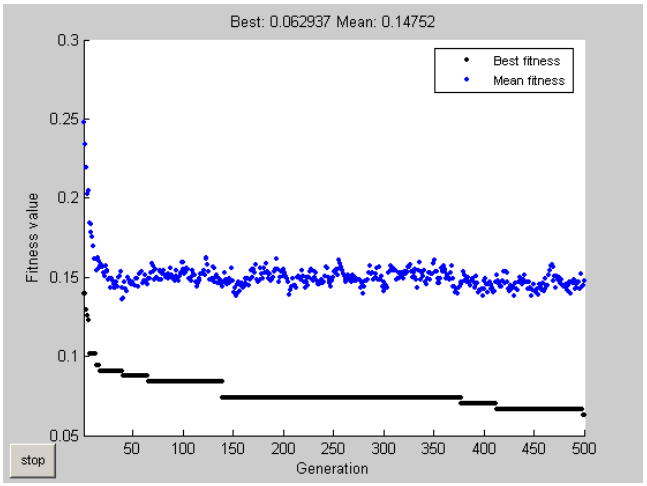

Figure 2. Best fitness, in black, and mean fitness values of generations

From a computational standpoint, it is worth noting that statistical classifiers and K-NN based system are characterized by low CPU times when compared to SVM and ANN hybrid solutions. Even though the GA/ANN system showed a computationally intensive behaviour results returned by this system are characterized by an important factor. The amount of variance in the genes extracted, registered along the 100 runs of the GA, is quite low. This particular aspect distinguishes the GA/ANN approach from the others reported as comparison in this work. As it can be seen in the "Results and Comparisons" section, the GA/ANN returns results coherent with the problem; furthermore it has been observed that this system shows a particular ability in extracting relevant genes with a considerably higher probability than other genes. Simple evaluation of the "standard scores" for each gene selected, calculated as:

$$Z = \frac{E_i - Ex(E_i)}{\sigma} \qquad (1.1)$$

(where E_i is the number of times $gene_i$ was extracted, $Ex(E_i)$, is the expected number of times $gene_i$ was extracted, σ is the square root of the variance) confirmed these observations. This means that this approach shows high robustness and, in general, a good accuracy.

IV. Results and Comparisons

Comparative results of the hybrid systems previously described are provided in this section. After 100 executions of the GA a ranking of selected genes has been compiled. This ranking took in account selection frequencies for each gene extracted and

returned and interesting overview of systems' performances. It is worth noting that all of the solutions analyzed selected sets of genes that demonstrated to be overlapping. In table 1, the rankings of the 20 most selected genes for each hybrid system are reported. As it can be seen, a considerable part of the most selected genes are reported by all of the systems analyzed. However GA/ANN system showed lower variability in results returned; this aspect is peculiar of this particular hybrid system only. Genes extracted by the GA/ANN system remained quite the same over all the 100 runs of the algorithmHowever, as confirmed by quite competitive results returned by the other systems, the choice of using a GA in order to extract relevant genes has a considerably positive impact on the overall system accuracy; evidently the sensitivity of GA employment on system's performances is slightly higher than that of the classifier related choice.

Gene ID	**GA/ANN Rank**	**GA/SVM Rank**	**GA/KNN Rank**	**GA/Stat Rank**
219340_s_at	1	1	1	1
217771_at	2	2	2	2
202418_at	3	3	6	4
206295_at	4	4	4	7
200726_at	5	5	5	3
210314_x_at	6	7	3	6
219588_s_at	7	6	8	8
212567_s_at	8	8	7	10
55081_at	9	10	10	18
218430_s_at	10	9	9	9
217404_s_at	11	11	12	13
205848_at	12	12	13	12
214915_at	13	13	16	-
202687_s_at	14	15	14	-
221241_s_at	15	16	11	11
210593_at	16	14	-	-
204028_s_at	17	17	15	-
201112_s_at	18	19	17	-
209825_s_at	19	18	-	16
209602_s_at	20	-	19	14
209604_s_at	-	20	-	5
201579_at	-	-	18	15
210347_s_at	-	-	-	17
209603_at	-	-	20	19
200827_at	-	-	-	20

Table 1. Genes extracted by the compared systems, rank positions for each solution are reported

However gene signatures returned by similar techniques are useless until an interpretation is formulated regarding their activities, interaction, and possible involvements in critical contexts (e.g. estrogen synthesis).

In this work we provide a two-fold validation of the sets of genes. In current research in the bioinformatics field there are two main trends as for the validation of results returned by any kind of computational approach. The first is referred to as "Data-driven" approach which mainly includes statistical tests or validity indices (e.g. Dunn's index or Silhouette method) applied to the data clustered. The second is referred to as "Knowledge-driven" method [18]. Briefly: the traditional validation method accounts for calculating statistical importance of results, ignoring any knowledge about the context. The "Knowledge-driven" validity assessment technique, on the other hand, is based on a common knowledge-base which is explored in order to find useful information. In the Molecular Biology and Bioinformatics field the "Knowledge-base" model has been translated in a set of "genetic ontology" maintained by a specific consortium: the "Gene Ontology Consortium"[19]. Biological validity assessment of gene signatures obtained is given in the next section. In this section "Statistical-Driven" validation is carried out. The focus is obviously on the gene signature extracted by the GA/ANN, however, given the similarities with other gene subsets, similar analyses could be done for the other results. The first gene, "219340_s_at", was selected 85 times over 100 executions. This gene's product is a transmembrane protein involved in signal transmission between cells. The putative CLN8 protein gene was selected as the most statistically meaningful gene even in the work of Foekens[3]. In the following positions in the ranking there is a set of genes "201112_s_at", "202687_s_at"," 204028_s_at", "205848_at", "206295_at", "210314_x_at" and "221241_s_at" that show frequencies between 62 and 83 and that belong to the same class of genes. In fact all of these genes are involved in the "cell cycle" regulation, included apoptosis that is, the way cells die in an ordered or programmed way. Apoptosis is one of the main types of Programmed Cell Death that allow organism to handle cell proliferation. These mechanisms have been discovered to be strictly correlated to tumor proliferation [20]; in particular down-regulation of apoptosis-related genes can lead to uncontrolled cell proliferation which is a key aspect of cancer. In order to give a comprehensive outlook on all the techniques involved in this comparative study, table 2 collects computational times required by each system. Measurements refer to an Intel P4 EE 2.8 GHz with 2 GB of DDR RAM memory. An average value of all the measurements made was calculated, approximated to next integer number and reported. As shown in table 2, as the complexity of fitness function computation increases, requirements in terms of CPU-time become considerable. An almost one order factor separates the systems reported. It is worth noting that, given the same results, GA/KNN is ten times slower than a simple correlation based solution. The robustness of the GA/ANN approach, instead, makes necessary an intensive computation. For these reasons some kinds of optimizations have been proposed in this paper in order to make this algorithm competitive even from this standpoint.

GA/Stat	**GA/KNN**	**GA/SVM**	**GA/ANN**
1385	9910	44500	144200

Table 2. Computational times (in seconds) required by each system

V. The Distributed Approach

As previously described, the computational costs of the proposed approach require some level of optimization in order to reach the necessary level of usability. For these reasons a

distributed version of the system described so far has been designed and developed. The implementation proposed herein takes advantage of the new Distributed Computing Engine included in the latest realease of MATLAB package. Four machines similar to the previously described one have been set up and used in order to create a small cluster. A jobmanager has been set up on the server machine and 2 clients have been started on each of the computer (server included). This approach allowed to reach a double level of optimization: in fact on the one hand the presence of 2 processes could take advantage of the HyperThreading technology implemented in a great portion of the current Intel CPUs and, on the other hand, the distribution of the fitness function evaluation on many client computers reduced significantly the execution time of the proposed algorithm. The solution employed in this experiment provided for a new routine designed to handle the distribution of the workload among the workers and the re-collection of the results at the end of the computation.

The results reach by this approach are shown in table 3.

GA/Stat	GA/KNN	GA/SVM	GA/ANN
460	3506	17921	57702

Table 23 Computational times (in seconds) required by each optimized system

The speed up in terms of time gained is in the order of 60-70%, a quite interesting result indeed.

VI. Biological Validity Assessment

In this section we provide a biological validity assessment of the results obtained in previous steps. Obviously the focus is primarily on the GA/ANN solution, however, given the similarities among the extracted genes subsets, following considerations could be considered applicable even to the other systems.

As we outlined previously, cluster validity assessment may consist of "Data-driven" and "Knowledge-driven" methods, which aim to estimate the optimal cluster partition from a collection of candidate partitions of genes extracted from a microarrays probeset. However, if "Data-driven" methods are mainly based on statistical tools and validity indices,

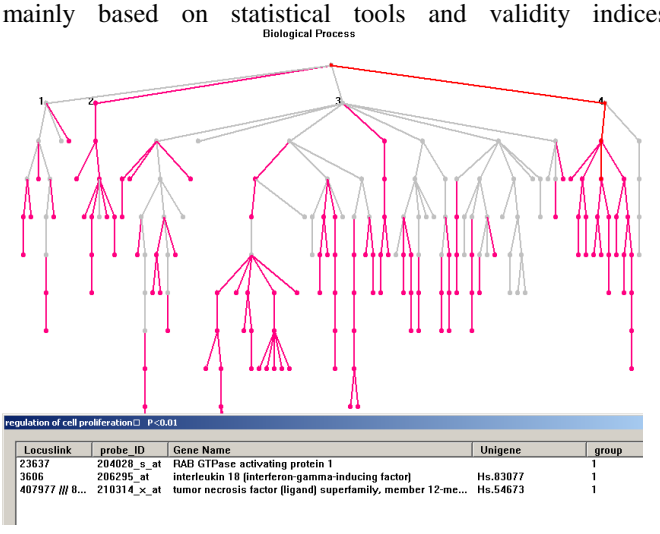

Figure 3. "Biological Process" GO Tree. "Regulation of Cell proliferation" GO Term in evidence

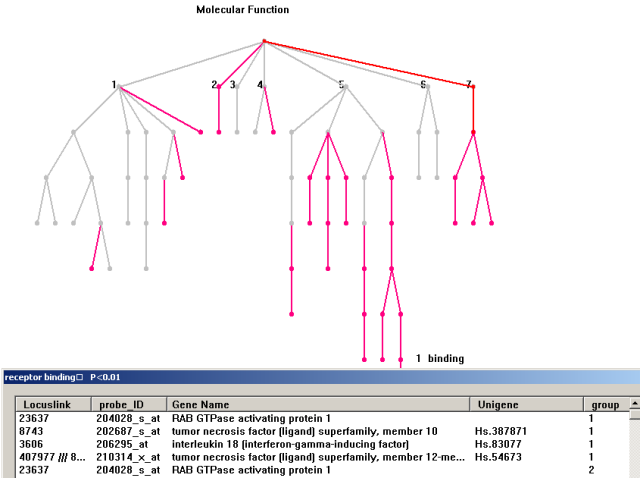

Figure 4. "Molecular Process" GO Tree. "Receptor binding" GO Term in evidence

"Knowledge-driven" methods assess cluster validity based on similarity knowledge extracted from the Gene Ontology. The Gene Ontology (or GO) is composed of three related ontologies covering basic areas of biological research: the molecular function of gene products, their role in multi-step biological processes, and their physical structure as cellular components. GO Consortium provides to constantly control and update databases. The GO defines a shared, structured and controlled vocabulary to annotate molecular attributes across models organisms. As shown in figures 3,4 and 5, each ontology is constructed as a directed acyclic graph. GO is composed by GO terms and each GO term consists of a unique alphanumerical identifier, a common name, and a definition. Terms are classified into only one of the three ontologies. Thanks to this organization of the biological knowledge it is possible to compute similarity of terms emerged in the data analysis step. In fact, given a pair of terms, t1 and t2, a basic method for measuring their similarity consists of calculating the distance between the nodes of the acyclic graph associated with these terms in the ontology (the shorter this distance, the higher the similarity) [18]. In this way sets of genes, extracted by statistical techniques as well as computational intelligence based systems,

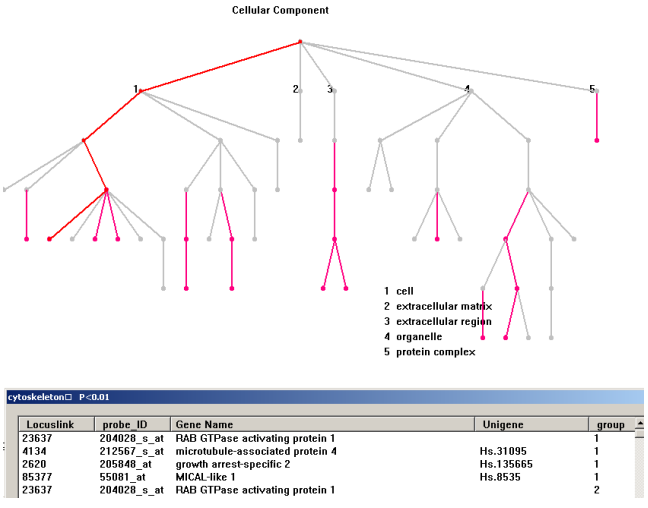

Figure 5. "Cellular Component" GO Tree. "Cytoskeleton" GO Term in evidence

could be analyzed in order to observe if there is any "interaction" or involvement in biological path considered critical for the disease or phenomenon examined. This means that through GO, researchers are able to find quickly aspects of a certain dataset that could reveal to be interesting cues of research.

Several GO based tools for microarrays data analysis results validity assessment have been developed [19]. One of the most interesting is the GO Surfer of the Harvard School of Public Health [21][22]. GoSurfer uses Gene Ontology information in the analysis of gene sets obtained from genome-wide computations or microarray analysis and provides rigorous statistical testing, of the hypothesis. GoSurfer finds all the GO terms that are associated with any genes in the input gene lists, and visualize these GO terms as three hierarchical trees. Each tree corresponds to one of the three general GO categories "biological process, "molecular function," and "cellular component". The Chi-Square test is used to search for the GO terms that are enriched in the annotation of one input lists of genes. Users can click on the GO graph to find the input genes that are associated with a selected GO term [22].

Fed with the list of the 20 most selected genes extracted by the GA/ANN hybrid system, GO Surfer put in evidence quite interesting results. In figure 3 the GO trees of the "biological processes" category are shown. Each node represents an individual GO term and all GO terms at display are associated with at least one out of 20 specific genes. It is remarkable that a great part of these genes belong to GO terms in some way correlated to the regulation of cell life (cellular proliferation, apoptosis and necrosis).

Down-regulation of these genes could results in uncontrolled cell proliferation and then oncogenesis.

In figure 4 GO tree of the molecular function is reported. From this standpoint it is interesting to observe that in one of the most meaningful term ($p < 0.01$), the "receptor binding" one, two "tumor necrosis factor" genes are present. Even in this case irregular activity of these genes can bring to dysfunction in cell cycle regulation. In figure 5, GO tree of the cellular components is shown. In the most interesting branch ($p < 0.01$) "Cytoskeleton" the "205848_at" gene is included. This is a growth arrest specific gene, that is to say a gene that could regulate the growth of cells (furthermore in this process the well known P53 onco-suppressor gene is involved). On the basis of all these considerations and observations it can be argued that the research in the field of Breast Cancer should focus on the activity and regulation of "221241_s_at"," 202687_s_at", "210314_x_at", "205848_at" and "55081_at" genes. As we have previously seen, these genes are not only able to build an accurate classifier, but they are even involved in biological and molecular processes that could be individuated as strictly correlated to typical tumour pattern, just like dysfunction in cell life cycle and receptor activity.

VII. Concluding Remarks and Further Works

The development of gene expression data analysis methods is one of the most important challenges of the post-genomic era. With the advent of microarray technology, the scientific community has assisted at the growth of datasets with a peculiar aspect: high disproportion between the two dimensions. This is a critical challenge for both the data miners and the tools they employed. Datasets, until the 90s, shared a common characteristic: the number of attributes per instance was largely inferior to the number of instances. With the introduction of microarrays this principle was literally reversed. Microarrays, since the first implementations, were able to trace expression levels of thousands of genes at a time but, due to practical and economical constraints, the number of cases per experiment was limited from some tens to few hundreds. Furthermore artifacts and other kinds of noise represent a concrete risk in gene expression levels analysis due to technological limits of the probes. For these reasons the scientific community has focused its research tools that could aid experts in solving these new challenges. Even if in the primary stage, this research has shown interesting potentialities [1][2]. Machine learning techniques, in this context, have demonstrated to be a powerful tool especially for what concerns clustering and combinatorial optimization problems. These are two of the most active research branch in the bioinformatics field. Clustering is commonly used in order to highlight set of genes that show common expressions, or, in general, common trends. Combinatorial optimization, on the other hand, allows exploring the space of combinations of genes that could used to build a system able to discriminate between disease/ disease-free cases. Traditional techniques have been employed so far in order to carry out these tasks. Soft computing techniques, still, are gaining more and more attention due to their abilities and potentialities in the two fields described. In literature can be retrieved reports of researches carried out in the field of computational intelligence applied to bioinformatics [23][24].

In this work we have developed a distributed data analysis system able to extract most relevant genes given dataset. System's performances were tested using a real-world publicly available dataset (*GSE2034*). It is worth noting that the proposed solution shows large applicability: other datasets could be explored with similar results (further experiments were successfully carried out on [26]). Downloaded and preprocessed data have been given in input to a GA that, encoding genes subsets in genotypes, searched for gene signature with high predictive abilities. The fitness of each subset of genes was computed training an ANN-FF and calculating the SSE.

A comparative study of performances of other systems has been given in this paper as additional evaluation tool. A distributed approach has been proposed that takes advantage of computational power of a cluster. This paradigm is in line with the main trends in bioinformatics that provide for distributed computational resources employment used to reduce execution times needed to accomplish the experiments. The original design of the system allowed to reduce the number of changes on the code and to reach quite interesting results. Thanks to the high modularity of this particular implementation, moreover, the employment of different clustering methods has been translated in few interventions on the code, underlining the versatility of the proposed approach. As analyzed in previous sections, the results returned by the GA/ANN are quite

competitive: the proposed algorithm demonstrated to be robust and be affected by very low variability of results. This is the most interesting characteristics of the system described. In the various experiments carried out the distributed GA/ANN approach selected small subsets of genes with a high frequency; this can be interpreted as the ability of the proposed system, to focus its research on subsets of features and to avoid local minima entrapment. The validity assessment of returned results has been carried out in two ways. Following the "Data-driven" approach a simple ranking of most frequently selected cases have been reported. The "normal-scores" confirmed the reliability of results. According to the "Knowledge-driven" principles, on the other hand, a Gene Ontology based validity assessment has been carried out. Gene Ontology collects bio-molecular knowledge in databases that are queried in order to discover relationships and interesting aspect of a gene list. As demonstrated in previous sections validity assessment can be considered to be satisfied both from a statistical and from a biological standpoint. Furthermore we observed that some of the most relevant genes extracted were included in the gene signature proposed by Foekens [3].

Further works will be mainly oriented on the optimization of the classifier. Artificial Immune System (AIS) based approaches are being studied. Given the AIS-ANN theory similarity [24], AIS based systems in the limited resources approach [25] could reveal interesting potentialities, optimizing both classification abilities and resource employment.

Further researches has to be carried out but the results in the primary stage are encouraging.

REFERENCES

[1] T.R Golub, D.R. Slonim, P. Tamayo, et al. Molecular Classification of Cancer: Class Discovery and Prediction by Gene Expression Monitoring. *Sciencee,Vol. 286, 15 October 1999*.

[2] Alizadeh AA, Eisen MB, et al.(2000). Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. *Nature* 403, 503-11

[3] Wang Y, Klijn JGM, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet 2005;365:671-9*

[4] John A. Foekens, David Atkins, et al. Multi-center Validation of a Gene Expression Based Prognostic Signature in Lymph Node-Negative Primary Breast Cancer, to appear.

[5] Schena, Mark , Knudsen, Steen. *Guide to Analysis of DNA Microarray Data, 2nd Edition and Microarray Analysis Set*. John Wiley & Sons, 2004. ISBN 0-471-67853-8

[6] Michael Q. Zhang, *Extracting functional information from microarrays: A challenge for functional genomics*. PNAS, vol. 99, no. 20, October 1, 2002, 12509-12511

[7] Anupam Chakraborty and Hitashyam Maka, *Biclustering of Gene Expression Data Using Genetic Algorithm*, CIBCB 2005

[8] Thorhildur Juliusdottir, David Corne, Edward Keedwell and Ajit Narayanan, *Two-Phase EA/k-NN for Feature Selection and Classification in Cancer Microarray Datasets*, CIBCB 2005.

[9] http://www.ncbi.nlm.nih.gov/geo/

[10] B.M. Bolstad 1, R.A Irizarry, M. Åstrand and T.P. Speed, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*, Bioinformatics Vol. 19 no. 2 2003

[11] Virginie M Aris1, Michael J Cody, Jeff Cheng, James J Dermody, Patricia Soteropoulos, Michael Recce, and Peter P Tolias. *Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer*. BMC Bioinformatics 2004, 5:185

[12] S Kohane, I.S., Kho, A.T., Butte, A.J., Microarrays for an Integrative Genomics, MIT Press, 2003.

[13] Kalyanmoy Deb and Raji Reddy. Classification of Two and Multi-Class Cancer Data Reliably Using Multi-objective Evolutionary Algorithms. *KanGAL Report No. 2003006*

[14] Leping Li , et al. *Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/k-nearest Neighbor Method*. Combinatorial Chemistry and High Throughput Screening, (2001), 727—739

[15] Soumya Raychaudhuri , Joshua M. Stuart , and Russ B. Altman, *Principal Components Analysis to Summarize Microarrays Experiments: Application to Sporulation Time Series*.Pac Symp Biocomput. 2000;:455-66.

[16] Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines (And Other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge, UK. 2000.

[17] V.Bevilacqua, G.Mastronardi, F. Menolascina, *Hybrid data analysis methods and Artificial Neural Network desing in breast cancer diagnosis: IDEST experience*, CIMCA 2005.

[18] Nadia Bolshakova, Francisco Azuaje and Pádraig Cunningham,*A knowledge-driven approach to cluster validity, assessment*. Bioinformatics 2005 21(10):2546-2547.

[19] http://www.geneontology.org/

[20] Bonnotte B, Favre N, Moutet M, Fromentin A, Solary E, Martin M, Martin F., Role of tumor cell apoptosis in tumor antigen migration to the draining lymph nodes. Journal of Immunoly 2000 Feb 15;164(4):1995-2000.

[21] Zhong S, Tian L, Li C, Storch FK and Wong WH (2004). *Comparative Analysis of Gene Sets in the Gene Ontology Space under the Multiple Hypothesis Testing Framework*. Proc IEEE Comp Systems

APPENDIX A

219340_s_at	putative transmembrane protein (CLN8) mRNA	**202687_s_at**	Apo-2 ligand mRNA
217771_at	golgi membrane protein GP73	**221241_s_at**	apoptosis regulator BCL-G (BCLG)
202418_at	putative transmembrane protein	**210593_at**	spermidinespermine N1-acetyltransferase mRNA
206295_at	interleukin 18	**204028_s_at**	rab6 GTPase activating protein
200726_at	protein phosphatase 1, catalytic subunit	**201112_s_at**	chromosome segregation 1
210314_x_at	tumor necrosis factor-related death	**209825_s_at**	Similar to uridine monophosphate kinase
219588_s_at	hypothetical protein FLJ20311 (FLJ20311),	**209602_s_at**	GATA binding protein 3
212567_s_at	putative translation initiation factor	**209604_s_at**	GATA-binding protein 3
55081_at	MICAL-like 1 (MICAL-L1), mRNA	**201579_at**	FAT tumor suppressor
218430_s_at	hypothetical protein FLJ12994 (FLJ12994)	**210347_s_at**	C2H2-type zinc-finger protein mRNA
217404_s_at	for alpha-1 type II collagen	**209603_at**	GATA binding protein 3 (GATA3)
205848_at	growth arrest-specific 2 (GAS2)	**200827_at**	procollagen-lysine
214915_at	cDNA FLJ11780 fis		

Bioinformatics Conference. 2004:425-435.

[22] Zhong S, Storch F, Lipan O, Kao MJ, Weitz C, Wong WH (2004). *GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space*. Applied Bioinformatics. 3(4):261-4

[23] Ke Tang, Ponnuthurai Nagaratnam Suganthan and Xin Yao, *Feature Selection for Microarray Data Using Least Squares SVM and Particle Swarm Optimization*, CIBCB 2005.

[24] Dasgupta, D., (1997), *Artificial Neural Networks and Artificial Immune Systems: Similarities and Differences*, Proc. of the IEEE SMC, 1, pp. 873-878.

[25] Andrew Watkins, Jon Timmis, and Lois Boggess, Artificial Immune Recognition System (AIRS): An Immune Inspired Supervised Machine Learning Algorithm. Genetic Programming and Evolvable Machines, 5(1), March 2004.

[26] Erich Huang, Skye H Cheng et al. *Gene Expression Predictors of Breast Cancer Outcome*, Lancet (2003) 361, 1590-1596.

Vitoantonio Bevilacqua was born in Bari (Italy) in 1969 and obtained both the Bachelor Degree in Electronic Engineering and the Ph.D. in Electrical Engineering from Polytechnic of Bari in 1996 and 2000 respectively. He is currently an Assistant Professor in Computing Systems at the Department of Electrical and Electronic Engineering of Polytechnic of Bari where he teaches C/C++ Programming and Expert Systems. Since 1996 he has been working and investigating in the field of image processing, neural network, evolutionary algorithm, and hybrid expert systems. The main applications of his research consist of industrial informatics, real world applications and recently in medicine and bioinformatics. In 2000 he was involved as Visiting Researcher in an EC funded TMR (Trans-Mobility of Reserchers) network (ERB FMRX-CT97-0127) called CAMERA (CAd Modeling Environment from Range Images) and worked in the field of geometric feature extraction and 3D objects reconstruction. He has published more than 25 papers in refereed journals, books, international conferences proceedings and chaired two session in Speech Recognition and Bioinformatics in two different international conferences.

Giuseppe Mastronardi was born in Bari (Italy) in 1949 and received the doctorate degree in Computer Science in 1976 at the University of Bari discussing a work developped in Torino (Italy), at the CSELT Laboratory, in the environment of Sirio satellite project. Since 1977 he joined as Assistant Professor, since 1982 as Senior Researcher, since 1992 as Associate Professor and since 2003 as Full Professor of Information Technology with the Electrical and Electronic Department of the Politechnic of Bari, where he is teaching Information Security, Medical Informatics and Informatic Systems for the Automation. His interests included computer vision, data security, biometric systems, signal and image processing in manufacturing, medical and environment fields. He organized several scientific international meetings of informatics. Since 1982 he was responsable of several ministerial projects and he collaborates with some local and multinational industries in the filed of quality control by means of computer vision (ALSTOM, ALTANET, BOSCH, DIAMEC, MASMEC, MERMEC, NERGAL e SIEMENS). He was referee of scientific papers for international meetings and issues, and he was evaluator of research pojects. He was scientific coordinator, for the Politechnic of Bari, of PRAI-Puglia project on the innovations by means of biotechnologies. He is a member of Scientific Counsil of Centro Laser and is responsable of Information and Multimediality Systems at the Politechnic of Bari. He is Vice-President of Information Engineering courses and is a member of Technical Committee of Agenda 21 of Trani city. He published more than 90 scientific papers plus two films on "Parallel Computing" (for CNR) and "Personal Identification" (for CSM) and two multimedial didactic supports on "Computer Architectures" and "Data Security". He is a member of AEI, AICA, ISMM (President of the Italian Branch), New York Academy of Science and SIMAI.

Filippo Menolascina was born in Bari (Italy) in 1984 and received his Laurea degree in Computer Engineering from the Polytechnic of Bari in 2006. His main research topics are focused on the development of Intelligent Systems in the biomedical and bioinformatic fields. He is currently supporting the Experimental Clinical Oncology Laboratory activity of IRCSS of Bari in a project funded by NCI. He has published three articles in IEEE international conference proceedings.

Stefania Tommasi was born in Bari (Italy) in 1963 and obtained the Bachelor degree and the Ph.D. in biology from the University of Bari. She is currently head researcher of the Clinical Esperimental Oncology Laboratori at the NCI of Bari. Her main research interests are focused on genetic alterations and expression of genes involved in cancer diseases with particular attention to the cell cycle related genes and growth factor receptors in breast cancer, genes involved in familiar breast cancer, and mitochondrial genome in human neoplasia. She is author of more than 50 papers published on national and international journals and of many posters and talks presented in national and international meetings.

Angelo Paradiso was born in Bernalda (Italy) in 1954 and obtained both the Laurea and Specialisation degrees in Medicine, Oncology and Applied Farmacology from the University of Bari in 1980, 1985 and 1988 respectively. He is author of more the 280 publications in the field of clinical and experimental oncology. He is currently Scientific Director of the Cancer Institute “Giovanni Paolo II” of Bari.