Hybrid Intelligent Data Mining Techniques and Array CGH in Breast Cancer Profiling

Filippo Menolascina[1], Stefania Tommasi[2], Vita Fedele[3], Angelo Paradiso[2], Giuseppe Mastronardi[1], Vitoantonio Bevilacqua[1]

[1] Department of Electronics and Electrical Engineering - Technical University of Bari, Italy
[2] National Cancer Institute of Bari, Italy
[3] Lawrence Berkeley National Laboratory, Berkeley, CA
{bevilacqua, mastrona}@poliba.it
{s.tommasi, a.paradiso}@oncologico.bari.it
filippo.menolascina@gmail.com

Abstract. In this study a cohort of 124 patients has been considered for copy number changes profiling in breast (BC) cancer sub-classes. Array Comparative Genomic Hybridization (aCGH) has been used in order to carry out the gene copy number profiling task. Output of aCGH scanning returns features per patient in the order of some thousands; it is evident that no useful information could immediately be derived from this kind of results unless appropriate data analysis are employed. Powerful techniques are then required in this context, in order to extract biologically plausible information from similar results. Here we propose a hybrid intelligent data analysis technique that combines well-established statistical tools and artificial neural networks (ANN) in order to address the problem of copy number profiling in high-throughput experiments. Meaningful inner features of the datasets have been discovered using statistical preprocessing of the data and intelligent data analysis techniques.

Keywords: Array CGH, Artificial Neural Networks, Gaussian Mixture Model, Hierarchical Analysis, Statistical Analysis, Microarrays, Principal Component Analysis, Self Organizing Maps.

1 Introduction

Chromosomal aberrations have been showed to be frequently involved in human cancers development [1]. Genomic DNA alteration, i.e. loss or amplification of specific genes, in fact, can markedly rise the probability of carcinogenesis in healthy patients. Gene dosage becomes, in this context, a particularly interesting variable to be monitored in order to rise the effectiveness of early diagnosis in human tumors. Different kinds of approaches have been proposed to study such disorders; Fluorescent In Situ Hybridization (FISH) and Representational Difference Analysis (RDA) and Comparative Genomic Hybridization (CGH) [2]. The last is a powerful technique although its usefulness is greatly limited by intrinsic technical limitations that prevent it to become a comprehensive screening tool. However, recent advancements in technologies have allowed researchers to conjugate the strength of

CGH and microarray platforms in Array Comparative Genomic Hybridization (aCGH) [3][4].
Results of aCGH screening are in the form of microarray images (Fig. 1); spot intensities are evaluated as ratios of fluorescent tag concentration and corresponding values are associated to specific probe copy number. Bacterial Artificial Chromosome (BACs) have been commonly used as probes in order to observe copy number changes of regions of the genome that share the same relative copy number on average. The resulting set of values for each patient can be analysed as a profile of genomic segments, as reported in Fig 2.

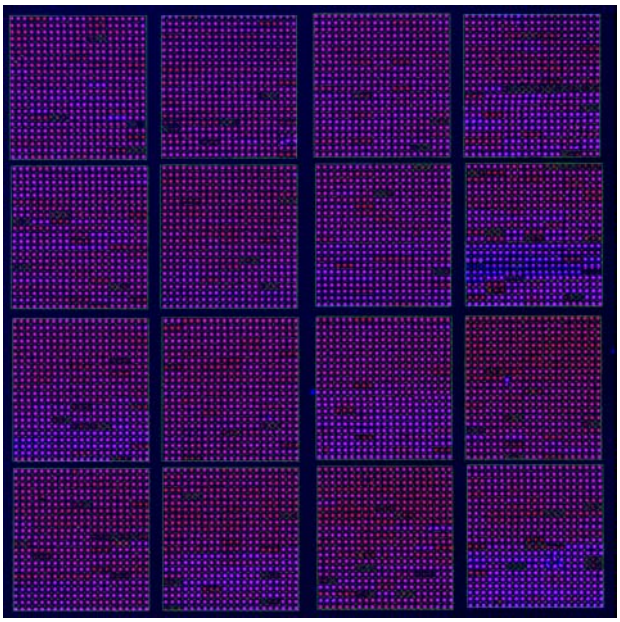

Fig. 1. Each spot in the array corresponds to a single BAC probe. Spot intensity associated to BAC clones is directly dependent on copy number levels of genes included in the clone, i.e. the more the spot is enriched with fluorescent tag, the higher the copy number level of the genes and the more severe the genomic alteration.

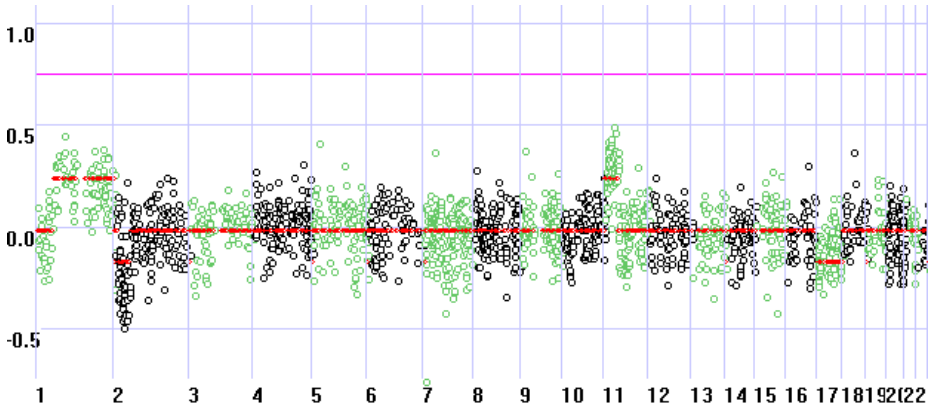

Fig. 2. Whole genomic profile of patient affected by BC. Regions with amplifications and deletions are clearly visible. It is even evident that this kind of approach can easily return a comprehensive snapshot of DNA copy number alterations in a single experiment.

For analysis purposes raw values are transformed applying log2(ratio) transform; this step is meant to give a theoretically 0 median for regions where no alteration occurred. On the other hand segments with positive means represents duplicated regions in the test sample genome and segments with negative means characterize deleted regions of the DNA. It is important to note that although the biological entity (copy number) is intrinsically discrete, the signal under investigation is considered as being continuous; this inconsistency is due to the fact that quantification of copy number levels is based on fluorescence measurement that is of an analogue source.

The obtained profiles constitute quasi raw data; this is the starting point for all the following analysis steps that will guide the researcher to the extraction of useful knowledge about the disease under investigation. After the data acquisition phase a data pre-processing and analysis phase needs to be employed in order to reduce the effects of non biological processes on the data. Background noise and probe saturation are two of the main issues related to high density array analysis. This is a central aspect of aCGH data analysis; accurate signal filtering can, in fact, noticeably reduce the probability of finding inconsistent segments then rising the global accuracy of the system. One of the most used algorithms in this class of methods is often referred to as “Smoothing” [5]. In general statistics based approaches are commonly employed in this phase in order to gain deeper knowledge about some characteristics of the dataset under investigation like noise distribution and bias. In this work we applied Gaussian Mixture Models (GMM) to investigate distributions of different classes underlying the data (patients with amplification, deletion or normal values of copy number).

Several algorithms have been proposed for aCGH data analysis purposes. All of them can be grouped in two main classes: the first is mainly focused on “segment finding” [6] while the second is based on the purpose of clustering individual data points into a finite number of groups that will be next classified as normal, amplified or deleted regions of DNA according to some metric related to the means of the data points in each segment [7]. In this work we have experienced the employment of CGH-Miner software [8] for the pre-processing of raw data. CGH-Miner is based on ‘Cluster

along chromosomes' (CLAC) algorithm for the analysis of array CGH data. CLAC builds hierarchical clustering-style trees along each chromosome arm (or chromosome), and then selects the 'interesting' clusters by controlling the False Discovery Rate (FDR) at a certain level. The CLAC algorithm uses a variation of a standard agglomerative clustering algorithm, a bottom-up strategy that generates a binary tree to represent the similarities in the data. Agglomerative clustering algorithms begin with every observation representing a singleton cluster. At each of the n - 1 steps (n = total number of objects) the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level. Under this point of view CLAC could be intended as a segmentation algorithm featuring higher level analysis techniques. The choice of CGH-Miner as pre-processing tool for our data has been influenced by the strong statistical bases it is built on; reproducibility of results is a consequence of the inner characteristics of the algorithm.

The data analysis stage followed. In this phase a formal method is applied to data in order to retrieve information underlying the dataset. Many different approaches have been proposed for this task. Both well known statistical methods [9] and machine learning based systems [10] have been proposed. Researchers paid a great attention on this step for obvious reasons: a good information extraction system can discriminate between a successful work and a failure. Many different alternatives, then, have been proposed. However looking at the recent literature it seems that intelligent systems are gaining a great attention because of their ability to be affective just in the fields where common statistical approaches start to highlight main limitations. Statistical support should never be dropped; formal theoretical backup is necessary in order to maintain confidence with results.

For these reasons hybrid intelligent systems based on common statistical tools like PCA, Hierarchical Analysis and advanced methods like Self Organizing Maps Artificial Neural Networks (SOM-ANN) have then been employed in gene clusters formation. A mesh built on the bi-dimensional map obtained by SOM training highlighted interesting association among genes known to be involved in cancerous processes and previously unstudied genes.

In addition a novel validation approach of the experimental results has been proposed as alternative/complementary to data driven validation: it's the knowledge driven approach [11]. The automated integration of background knowledge is fundamental to support the generation and validation of hypotheses about the function of gene products. One such source of prior knowledge is the Gene Ontology (GO), which is a structured, shared vocabulary that allows the annotation of gene products across different model organisms. The GO comprises three independent hierarchies: molecular function (MF), biological process (BP) and cellular component (CC). Researchers can represent relationships between gene products and annotation terms in these hierarchies. Potentialities of GO in knowledge driven validation of the experimental results is an evident result of its design. In this work we propose a biological interpretation of the discriminating clusters of genes based on GO trees. Results returned by this analysis showed interesting trends in the genomic instability associated to subgroups of BC. Biological validation of obtained results is given through the use of Gene Ontology; this tool allowed to gain deeper insights in the biological mechanisms underlying the disease under investigation and the subclasses

put in evidence. Correlation of previously unconsidered genes with known BC biomarker emerged and pushed further investigation on these genes. This study shows the practical consequences of the use of composite criteria. This point constitutes the main purpose of the discussion (see the Discussion Section).

2 Preliminary Statistical Analysis

In a preliminary stage statistical analyses have been carried out in order rise the resolution of knowledge regarding the dataset under investigation.
First analysis was mainly intended to discover the evidence of classes underlying the dataset. In this stage the choice of optimal cut-off thresholds for class separation has even been investigated. For these reasons a Gaussian Mixture Model (GMM) has been employed. Briefly GMM algorithms try to decompose a given distribution in a number of normal distribution defined by μ e σ^2; in our experiment the Expectation Minimization (EM) algorithm has been used to solve the problem of class identification. The EM algorithm is one way to compute the missing memberships of data points in a distribution model. It is an iterative procedure, where one starts with initial parameters for the model distribution. The estimation process proceeds iteratively in two steps, the Expectation Step and the Maximization Step. The assumption of normal distribution for "loss", "normal" and "amplification" classes has been made in this context. This assumption is coherent with results reported in previous works [12]. A self explaining plot of the classes underlying the DMPC-HFF#1-61H8 clone (which contains the ErbB2/NEU gene, a well known gene in breast cancer disease dynamics) is given in figure 3. Two classes were expected in this case because all of the samples collected in this dataset belonged to diseased patient in which the gene should never found to be deleted; this restricts the classes to "normal" or "amplified" options only. This thesis is sustained by a second experiment

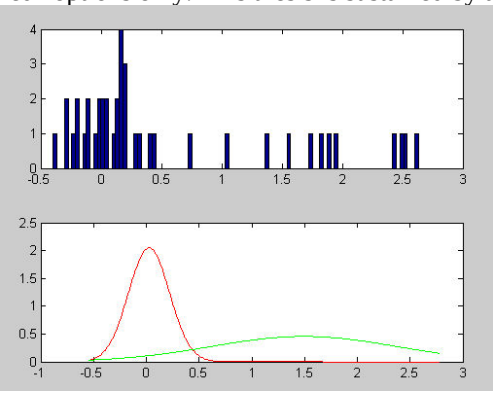

Fig. 3a. In the first plot the histogram of data distribution for ErbB2 gene. Below the GMM of the same dataset. The two Gaussians classes extracted represent the "normal" (red) and "amplification" (green) classes.

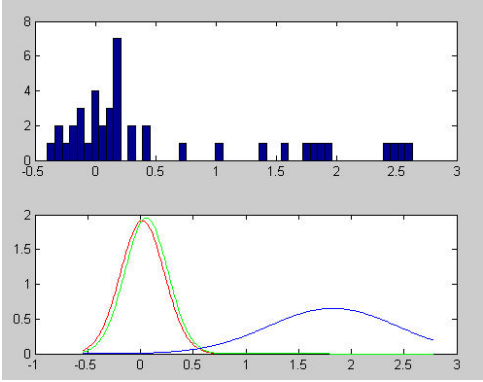

Fig. 3b. The same analysis of figure 4 carried out asking the algorithm to extract three classes instead of 2. It is evident that only two classes are really underlying the global distribution; actually the first and the second distributions represent the same class (corresponding means differ by 0.09 only).

carried out using the same approach; in this case the algorithm was asked to extract three distributions. As it is shown in figure 4, the three classes returned by this further GMM analysis are, actually, 2, since the algorithm has selected two distributions with quite similar means it is highly probable that a single actual class is underlying both the distributions. In this same context it is even possible to observe how the two classes could be optimally separated. Solving the system of equations (in the case of two normal distribution expected):

$$\begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma_1}\ \exp\left(-\dfrac{(x-\mu_1)^2}{2\sigma_1^{\,2}}\right) \\ \dfrac{1}{\sqrt{2\pi}\sigma_2}\ \exp\left(-\dfrac{(x-\mu_2)^2}{2\sigma_2^{\,2}}\right) \end{cases} \qquad (2.1)$$

with μ_1 and σ_1^2 respectively mean and variance of the first normal distribution and μ_2 and σ_2^2 characteristic parameters of the second Gaussian curve. The only finite result of this calculation is a value approximately equal to 0.5 which is even a quite probable cut-off threshold for the dataset considered.

In order to gain a deeper knowledge of genetic associations among genes a hierarchical clustering analysis has been carried out using Mahalanobis distance metric. Mahalanobis distance is a distance measure based on correlations between variables by which different patterns can be identified and analysed. It is a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements. Results of the hierarchical analysis is presented in figure 4.

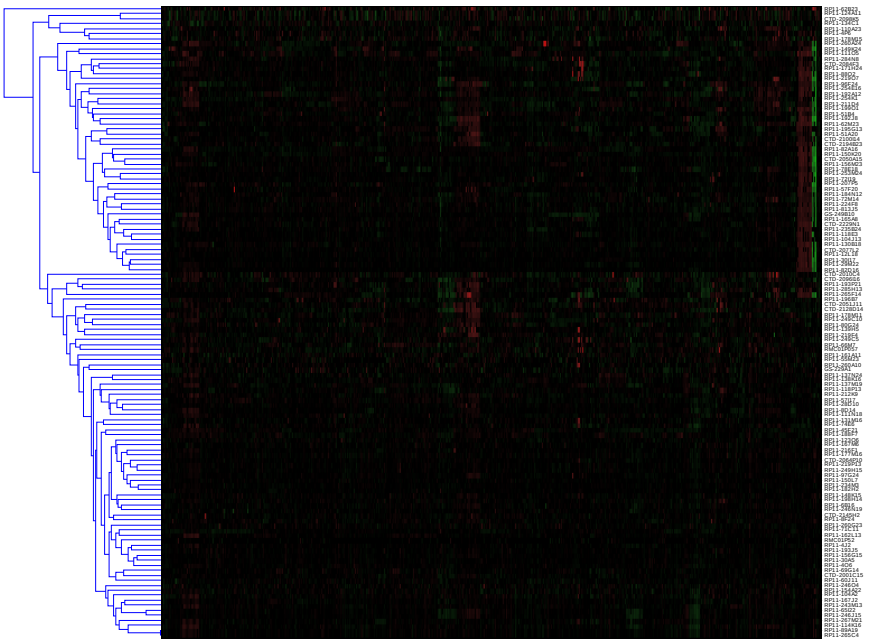

Fig. 4. Dendrogram of the dataset under investigation, cases on the horizontal axis, clones on the vertical one. Cluster of similarly behaving clones are visible in the upper right corner of the heat map.

Closer clones in the heat map in figure 4 tend to show high correlation rate; the evidence of similarly behaving clones pushed for further investigation of correlation among loci and, as a consequence, of DNA regions characterized by genomic instability. For these reasons a correlation analysis has been carried out computing the correlation coefficients of each clone versus all the others as:

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}} \qquad (2.2)$$

with $R(i,j)$ correlation of i and j clones and $C(i,j) = E\left[(i-\mu_i)(j-\mu_j)\right]$.

P-values were computed by transforming the correlation to create a t statistic having $n-2$ degrees of freedom, where n is the number of rows of the matrix. The confidence bounds are based on an asymptotic normal distribution of $\frac{1}{2}\log\left(\frac{1+R}{1-R}\right)$, with an approximate variance equal to $\frac{1}{n-3}$. These bounds are accurate for large samples when the original matrix has a multivariate normal distribution. Results of the correlation analysis are reported in figure 5.

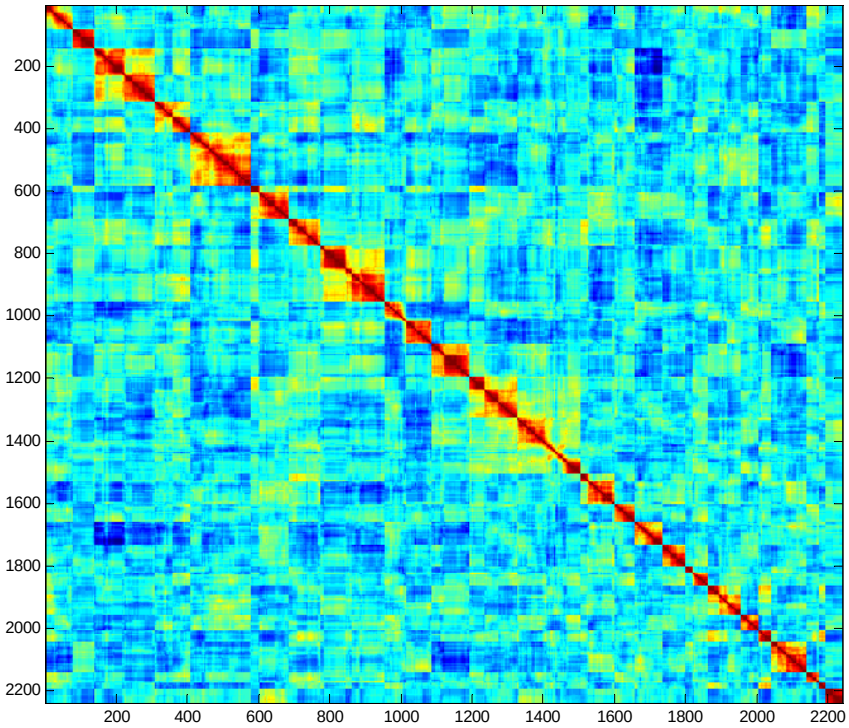

Fig. 5. Heat map showing correlation levels among the filtered clones. Red regions are characterized by highly correlated clones that tend to gain or lose in similar ways gene copies. Regions of genetic instability tend to get smaller along the genome; this is due to the ever reducing dimension of the chromosomes going from the first till the sexual one.

Clones showing absolute low correlation rates and/or low correlation p-values ($p<0.001$) were excluded from further analyses. Sets of clones belonging to the same instability region (sliding window = 10 clones) were analysed using Gene Ontology [13] in order to find functional relations with the disease under investigation. The results of this analysis are covered in "Discussion" section.

Last statistical analyses performed on this dataset are strictly linked with following intelligent analysis methods. Principal Component Analysis (PCA), used in the following step as a data reduction method, has been carried in order to extract preliminary information about clusters of clones and to verify if there was consistency with these and previous results. A plot of the former first three principal components (PC) is shown in figure 6.

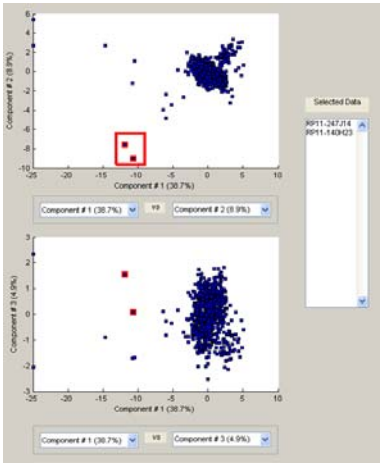

Fig. 6. In first the graph the second versus the first PC plot is shown, while in the second the third versus the first component is plotted. RP11-247J12 e RP11-140H23 clones have been put in evidence in both the graphs.

A finer analysis of the results returned by PCA put in evidence that even in this preliminary phase there are some evident clusters, as it could be observed in figure 7.

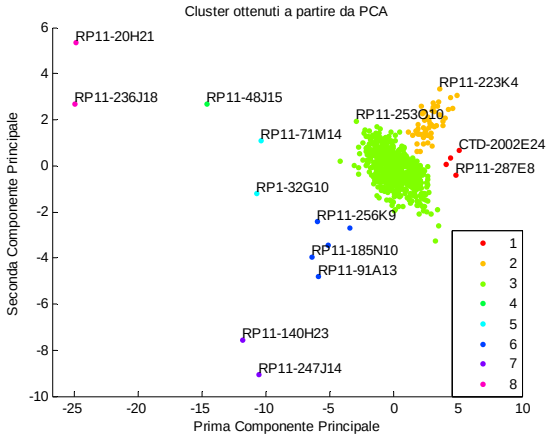

Fig. 7. The plot of the first versus the second PC shows some interesting aspects. Clusters are characterized by different colors. Cluster 2 and 3 are heavily enriched with clones. Clusters 6, 7 and 8 are composed by a small number of clones that show quite similar trends.

K-statistic Analysis carried out on these results showed high comparability and coherence (results not shown) and allowed to pass to the next stage of this research.

3 Intelligent Data Analysis

In this stage the data were analysed in order to find useful biological information about the disease starting from previous statistical results. In particular a Self Organizing Map has been employed in order to verify the associations between clones that were uncovered in the previous step.

A 5x3 Self Organizing Map has been trained on the former first two PC returned by PCA. This approach has been selected in order to compress computational times needed for neural network training at the same time providing the network with the maximum amount of information (supposed to be explained in the very first components returned by PCA).The topology function employed in this experiment is the `gridtop` supplied by MATLAB. The network has been trained for 100 epochs on all of the 124 cases included in the dataset. Using the `plotsom` function to plot a scatter plot of the data we obtained the results shown in figure 8; clusters have been assigned using the nearest node criterion where Euclidean distance have been employed as distance metric.

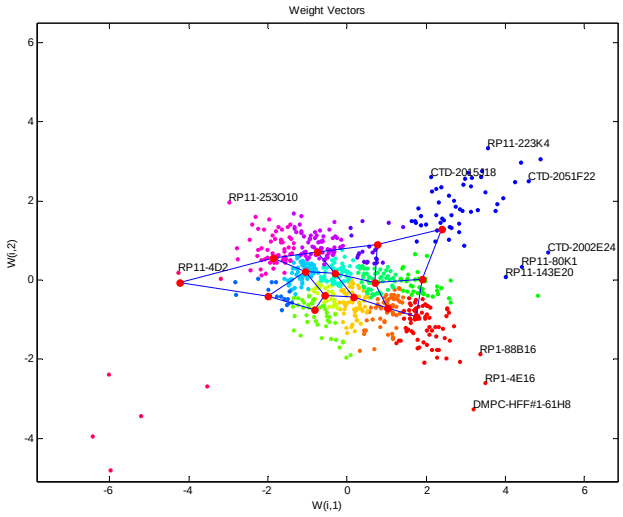

Fig. 8. Scatter plot of the cluster analysis carried out using SOM. The weights of the first input are plotted versus the ones related to the second. The mesh connects all the 15 centroids of the distribution. Clusters are marked with different colors.

Results of this analysis are in many ways strongly overlapping the previous ones: in particular clusters characterized by low cardinalities seemed to be comparable among different methods. Interpretation of obtained results seems to confirm that some underlying and hidden common trends exist among different clones and that this statement is confirmed by different and independent analysis.

We have focused our attention on a single case of interest among the others; it is the case of DMPC-HFF#1-61H8 locus cluster. This clone contains the ErbB2 (or HER 2) gene which is known to be involved in breast cancer development processes. Figure 8 puts in evidence the Euclidean proximity of DMPC-HFF#1-61H8, RP11-88B16 and

RP11-4E16 clones. This aspect suggests a deeper investigation of this and similar trends that could potentially hide answers to unsolved questions about tumorigenesis of breast cancer. This and other results are discussed in detail and validated in the next section.

4 Discussion

Results returned by all these analyses showed interesting trends in the genomic instability associated to subgroups of BC. Several subtypes have been investigated; in particular profiles considering familiarity, estrogen receptor positiveness, progesterone receptor positiveness and age have been extracted from the dataset.
In this section we will focus on the results regarding the general case set, exposing them and trying to assess their validity using Gene Ontology; this tool allowed to gain deeper insights in the biological mechanisms underlying the disease under investigation. Statistical analysis highlighted that major correlated amplifications were more frequent in 1q21-32 (FDR $q<0.01$), 3q21-26 (FDR $q<0.01$), 8q11.1-8q24 (FDR $q<0.01$), 11q13-14 (FDR $q<0.01$), 11q22-25 (FDR $q<0.01$), 13q12-13q14 (FDR $q<0.01$), 13q21-13q22 (FDR $q<0.01$), 13q31-13q34 (FDR $q<0.01$), 16q11-16q13 (FDR $q<0.01$), 16q21-16q24 (FDR $q<0.01$), 17q11 (FDR $q<0.01$), 17q21-17q24 (FDR $q<0.01$), 20p11.2-20p13 (FDR $q<0.01$), 20q13-20qtel (FDR $q<0.01$); more significant correlated deletions were found in 4p11-4p12 (FDR $q<0.01$), 4q21-4q27 (FDR $q<0.01$), 4p13-4p15 (FDR $q<0.01$), 4q21-4q28 (FDR $q<0.01$), 8p23 (FDR $q<0.01$), 11q22-11q24 (FDR $q<0.01$), 13q12 (FDR $q<0.01$), 16q21-16q24 (FDR $q<0.01$). The clones in these loci have been analysed using GO; with this step we tried to discover if interesting functional GO Terms resulted to be enriched with genes. Results of GO analysis are shown in figure 9. Interesting GO Terms were found to be activated by a good amount of genes; the p-value threshold for this analysis was set to 0.01 so that all GO Terms that showed a p-value less than this threshold were considered to be statistically meaningful. Main GO Terms discovered by this analysis were “Development”, “Cell motility”, “Cytoplasm Organization and Biogenesis”, “Morphogenesis” and “DNA Damage Response”. It is quite evident that most of these terms can be involved in tumorigenesis (e.g. “Development” and “DNA Damage Response”) and metastases progression (e.g. “Cell motility”). Genes in clones RP11-88B16 and RP11-4E16 emerged as significant in SOM/PCA analysis are currently under investigation, however they have been located in an hot spot of the 17^{th} chromosome which could suggest an interaction with ErbB2/ HER 2 oncogene.
Results obtained by the employment of hybrid intelligent systems seem to encourage the use of similar approaches in aCGH data analysis. These methods could reveal a great potential in similar data analysis tasks, matching the need of new and powerful approaches to genetic disorders investigation. Interesting aspects of quite huge datasets could be easily investigated through the use hybrid intelligent systems.

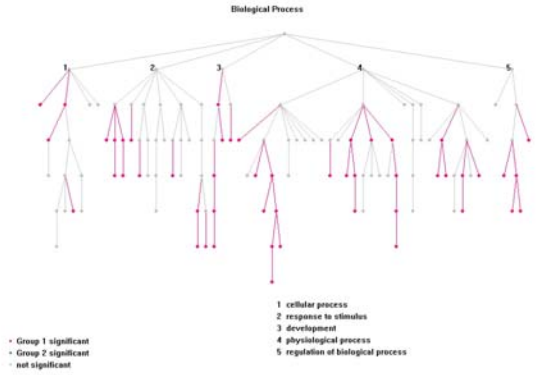

Fig. 9. Biological Process tree returned by GO. GO Terms are represented as graph nodes and statistically relevant nodes ($p < 0.01$) are highlighted in res.

References

1. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO., Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12963-8.
2. Beheshti B, Park P, Braude I, Squire J: Molecular Cytogenetics: Protocols and Applications Humana Press; 2002.
3. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P: Matrix-based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances. Genes, Chromosomes and Cancer 1997, 20:399-407.
4. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C,Kuo W, Chen C, Zhai Y, Dairkee S, Ljung B, Gray J: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics 1998, 20:207-211.
5. Eilers P, Menezes R: Quantile smoothing of array CGH data. Bioinformatics 2004
6. Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M, Kallioniemi A: CGH-plotter: MATLAB toolbox for CGH-data analysis. Bioinformatics 2003, 13:1714-1715.
7. Jong K, Marchiori E, van der Vaart A, Ylstra B, Weiss M, Meijer G: Applications of Evolutionary Computing: EvoWorkshops 2003: Proceedings, Springer-Verlag Heidelberg, chap. chromosomal breakpoint detection in human cancer 2003, 2611:54-65.
8. Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R: A method for calling gains and losses in array CGH data. Biostatistics 2005, 6(1):45-58.
9. Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ, A statistical approach for array CGH data analysis. BMC Bioinformatics, Vol. 6 (2005)
10. Aliferis CF, Hardin D, Massion PP., Machine learning models for lung cancer classification using array comparative genomic hybridization. Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. Proc AMIA Symp. 2002;:7-11.
11. Nadia Bolshakova, Francisco Azuaje and Pádraig Cunningham, A knowledge-driven approach to cluster validity assessment. Bioinformatics 2005 21(10):2546-2547;
12. Albertson DG., Profiling breast cancer by array CGH. Breast Cancer Res Treat. 2003 Apr;78(3):289-98.
13. http://www.geneontology.org